MACHINE LEARNING
Earth

Part of the MACHINE LEARNING Series

PAPER • **OPEN ACCESS**

Interpretable neural networks to predict momentum fluxes of orographic gravity waves

To cite this article: Elias Haslauer *et al* 2026 *Mach. Learn.: Earth* **2** 025007

MACHINE LEARNING
Earth

PAPER

RECEIVED
18 April 2026

REVISED
9 June 2026

ACCEPTED FOR PUBLICATION
16 June 2026

PUBLISHED
23 July 2026

Interpretable neural networks to predict momentum fluxes of orographic gravity waves

Elias Haslauer[1,*], Mierk Schwabe[1], Andreas Dörnbrack[1], Edwin P Gerber[2], Markus Rapp[1,3], Nedjeljka Žagar[4] and Veronika Eyring[1,5]

[1] Deutsches Zentrum für Luft- und Raumfahrt e.V., Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany
[2] Courant Institute, New York University, New York, NY, United States of America
[3] Meteorologisches Institut, Ludwig-Maximilians-Universität München, Munich, Germany
[4] Meteorological Institute, University of Hamburg, Hamburg, Germany
[5] Institute of Environmental Physics (IUP), University of Bremen, Bremen, Germany
[*] Author to whom any correspondence should be addressed.

E-mail: elias.haslauer@dlr.de

Keywords: climate models, gravity waves, parameterisations, subgrid processes, explainable AI

Abstract

State-of-the-art Earth system models (ESMs) cannot explicitly resolve many small-scale atmospheric processes such as atmospheric gravity waves, and thus must represent, or parameterise, their effects on the resolved state. Machine learning (ML) has the potential to improve these parameterisations. In our study, we train neural networks (NNs) on ERA5 reanalysis data to predict momentum fluxes of orographic gravity waves as a function of the state variables at the resolution of a coarse ESM. Employing a full year of data, we extract inertia-gravity waves using the software MODES, which applies linear theory for wave filtering, and train ML models on data coarse-grained to the ESM's target resolution. We consider four different cases: the full spectrum of inertia-gravity waves resolved in ERA5, or just the part of the spectrum that is subgrid-scale in the target ESM, both over all land or just over mountainous terrain. Our NNs successfully predict momentum fluxes, with a global coefficient of determination, R^2, ranging from 0.72 to 0.56, depending on the case, when evaluated offline with data from another year. An analysis of our models using SHAP values, an explainable AI technique, suggests that the networks learned physically meaningful relationships. In addition, we give a comparison with the physics-based parameterisation scheme by Lott and Miller. This work forms the basis for the development of operational ML-based parameterisations to improve the representation of gravity waves and their effects in climate models.

1. Introduction

Earth system models (ESMs) are key to a better understanding of the Earth system, projecting future climate change, and deriving suitable strategies for mitigation and adaptation in response to global warming. While models have significantly improved over the last decades regarding the simulation of the mean climate and its variability for many large-scale indicators of climate change—mainly due to increasing model resolutions and a better representation of physical processes—they still face systematic errors and large uncertainties (Vicente-Serrano *et al* 2022, Vautard *et al* 2023, Beverley *et al* 2024, Eyring *et al* 2024). A main reason for these uncertainties is the difficulty to represent small-scale processes like clouds, convection, and microphysics, processes that cannot be captured explicitly by state-of-the-art models, which typically run at grid resolutions of around 50–100 km. The statistical effects of these subgrid processes are represented by so-called parameterisations, which are traditionally based on physical process understanding and empirical relationships, but often come along with severe simplifications necessary to make them computationally efficient (Kim *et al* 2003, Stensrud 2007, Achatz *et al* 2024, Eyring *et al* 2024).

Gravity waves are atmospheric disturbances in which buoyancy acts on air parcels displaced from hydrostatic equilibrium. They are called inertia-gravity waves if the waves are of large scale, such that the Coriolis force has substantial additional influence. These waves constitute an important class of subgrid-scale processes (Fritts and Alexander 2003). Gravity waves have traditionally been differentiated depending on their sources: orographic gravity waves arising from wind flowing over mountains, vs non-orographic gravity waves, induced by, e.g. convection, jets, and fronts. As gravity waves propagate through the atmosphere, they transport momentum, and influence the atmospheric circulation when they finally break and deposit their momentum. Since gravity waves occur at any scale, with horizontal wavelengths ranging from hundreds of metres to several thousands of kilometres, a significant part of their spectrum cannot be captured at the resolution of today's climate models. Even though the evolution of computational capabilities allows the integration of global storm resolving models with resolutions on the order of kilometres, gravity wave parameterisations are still needed since even these highly resolved models do not capture the smallest scales of gravity waves, and coarse resolution models are needed, e.g. to run large model ensembles (Polichtchouk *et al* 2023, Achatz *et al* 2024).

Initially, gravity wave parameterisations were introduced to correct for a westerly bias in models (Palmer *et al* 1986), improving their numerical stability and prediction skill. However, even today's operational gravity wave parameterisations (e.g. Hines 1997, Lott and Miller 1997, Lott 1999) operate with oversimplifying assumptions, the severest being neglecting horizontal propagation (Alexander *et al* 2010, Stephan *et al* 2019, Eichinger *et al* 2023, Achatz *et al* 2024). The consequences of these shortfalls include, e.g. an unrealistic or absent quasi-biennial oscillation in many models (Richter *et al* 2022), missing gravity wave drag (GWD) over the Southern Ocean (McLandress *et al* 2012), and deficiencies in the simulation of the wintertime polar vortex and sudden stratospheric warmings (McLandress *et al* 2013). These issues can partly be addressed by tuning. Also, it is an open question whether the models will generalise in a changing climate (Achatz *et al* 2024). Hence, improvements in accuracy and physical adequacy are highly desirable.

In the last years, much research has been conducted to replace conventional parameterisation schemes in climate models with machine learning (ML) approaches (e.g. Gentine *et al* 2018, Rasp *et al* 2018, Yuval and O'Gorman 2020, Grundner *et al* 2022, Yuval and O'Gorman 2023, Heuer *et al* 2024, Sarauer *et al* 2025, for an overview see Burgh-Day and Leeuwenburg 2023). ML techniques enable the description of complex non-linear relationships based on data and have revolutionised many fields of science in the last decade. Neural networks (NNs) are trained on large amounts of 'known' data, allowing predictions also for unseen data. For a general review of ML, deep learning techniques, and terminology see Alzubaidi *et al* (2021).

In the context of ML-based parameterisations, there are two main approaches: first, emulations, which mimic conventional schemes with the goal of reducing the computational requirements. Second, training ML models with high-resolution simulations or observations that resolve the process of interest. In regard to the parameterisation of gravity waves, Chantry *et al* (2021), Espinosa *et al* (2022), Hardiman *et al* (2023), Connelly and Gerber (2024), Sun *et al* (2024) follow the first strategy, emulating existing schemes. While these ML-based parameterisations might imitate their reference models adequately and improve computational performance, they inherit the limitations of traditional gravity wave parameterisations discussed above.

Related to the second approach, NNs have been applied using reanalysis data to predict gravity wave momentum fluxes (GWMFs) locally over Japan (Matsuoka *et al* 2020), as well as non-orographic gravity waves over sea (Amiramjadi *et al* 2023). Dong *et al* (2023) train an ML model on high-resolution data to speed up a physics-based model. Gupta *et al* (2024c) aim for a global approach, also including the representation of non-local gravity wave effects by taking into account data of neighbouring grid cells, as well of the whole globe. This is further pursued in Gupta *et al* (2025), where the authors predict GWMFs by fine-tuning a foundation model pretrained for weather and climate applications.

In this work, we train NNs on GWMFs from atmospheric reanalysis data, and successfully predict fluxes associated with mesoscale inertia-gravity waves based on the coarse state variables. As a first step, we focus primarily on orographic gravity waves, since their subgrid sources can be identified more easily. To this end, the ERA5 global reanalysis dataset (Hersbach *et al* 2020) is used as high-resolution training data. Since ERA5's resolution captures only part of the gravity wave spectrum, this is only a first step towards using higher resolution training data in the future.

Gravity waves are extracted using the spherical linear theory for the decomposition of three-dimensional circulation (Kasahara and Puri 1981) with the software MODES (Žagar *et al* 2015). We investigate both the case of the full spectrum of gravity waves (IG, i.e. all inertia-gravity waves), as well as a limited part of the spectrum which is 'subgrid' (SG), i.e. unresolved at a resolution as it would be used in a climate model. After the computation of the respective momentum fluxes, the data are coarse-grained to this lower resolution, in which we train a modified U-Net architecture (Ronneberger *et al* 2015) on atmospheric columns. The results are physically interpretable when analysing SHAP values (SHapley Addiditve exPlanations, Lundberg and Lee 2017), an explainable AI (XAI) technique, and yield encouraging results when comparing to the conventional gravity wave parameterisation scheme by Lott and Miller (Lott and Miller 1997, Lott 1999), which is used operationally in many weather and climate models.

The remainder of this paper is structured as follows: section 2 describes the dataset used in this study, our method of extracting gravity waves and calculating corresponding momentum fluxes, the architecture of the NNs, and the different experiments of this study. We also give a short overview of SHAP values. In section 3, we present offline (i.e. the parameterisation is not coupled to a climate model) results of the NNs, the analysis of SHAP values, and a comparison of our networks with the conventional Lott and Miller scheme. We briefly discuss our results in section 4. Section 5 gives an outlook and sketches the road ahead, in particular extending the approach to non-orographic gravity waves and adapting the NNs for use in the ICON XPP model (Müller *et al* 2025).

2. Data and methods

2.1. Dataset and general approach

This study uses the ERA5 global reanalysis dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF, Hersbach *et al* 2020). ERA5 is produced with the ECMWF Integrated Forecasting System (IFS) Cy41r2 on 137 hybrid σ-pressure levels with a model top at 0.01 hPa and $\sim$31 km horizontal resolution (TL639). We use the dynamical variables on 37 vertical pressure levels up to 1 hPa, with a horizontal resolution of $0.25^\circ \times 0.25^\circ$ (i.e. a grid spacing of $\sim$28 km at the equator), and the time resolution of 1 h.

The full year 2024 is used for training of the NNs. Three additional datasets were prepared for validation and testing. Days 1, 11, and 21 of each month of the year 2022 (denoted 2022a) serve for hyperparameter validation and testing. Days 6, 16, and 26 of each month of 2022 (denoted 2022b), and days 1, 11, and 21 of the year 2020 (denoted 2020) provide two additional independent test sets. A large amount of training data distributed over an entire year is needed due to the seasonal variability of gravity wave generation and propagation. The skill of the NNs in predicting momentum fluxes on the test sets declines substantially when trained with less data.

We are interested in the prediction of GWMFs of both the full spectrum of gravity waves, as well as of subgrid-scale GWMFs, as a function of the coarse state variables. Therefore, our strategy is as follows: The original ERA5 data, given on a 720×1440 latitude-longitude grid, are considered as high-resolution ground truth for the purpose of this study. Approximately eight grid points are needed to resolve a gravity wave adequately, which determines the effective resolution of the grid (Sun *et al* 2023). In our case, waves with horizontal wavelengths of $\sim$200 km and more are fully resolved (Gupta *et al* 2024a). This means that we are dealing with mesoscale inertia-gravity waves and are missing a substantial fraction of the gravity wave spectrum in this initial work.

This limitation regarding ERA5 has been reported in several studies comparing the structure and magnitudes of GWMFs resolved in ERA5 with satellite data and high-resolution IFS simulations. Amiramjadi *et al* (2020) reported that GWMFs in ERA5 are weaker by a factor of up to 3 compared to those derived from GRACILE satellite data. Consistently, Gupta *et al* (2024a) compared resolved momentum fluxes in ERA5 with those in a 1.4 km IFS simulation and found the fluxes in ERA5 to be weaker by a factor of 2–2.5. They pointed out that ERA5 resolves only 40% of the fluxes associated with the full gravity wave spectrum, with small-scale waves contributing substantially to the overall flux. In line with these results, Lear *et al* (2024), Yoshida *et al* (2024) reported smaller fluxes in ERA5 than those observed in satellite, radar, and higher-resolution IFS data. It is unclear what fraction of the missing flux is attributable solely to horizontal resolution. Polichtchouk *et al* (2023) reported that increasing the horizontal resolution from $\mathcal{O}(10\,\text{km})$ to $\mathcal{O}(1\,\text{km})$ in the IFS model roughly doubles the zonal-mean resolved fluxes in the stratosphere, suggesting that the missing fluxes in ERA5 may be associated with its limited horizontal resolution, i.e. they are unresolved. The vertical resolution is another source of uncertainty. Even though the 137 levels of the ECMWF IFS model resolve more features than a typical climate model, the model does not fully capture observed variability in the vertical structure of gravity waves (e.g. Dörnbrack *et al* 2018) or the strongly sheared background flow through which waves propagate (Žagar *et al* 2025).

Despite its substantial improvements over previous reanalyses (e.g. Simmons 2022), ERA5 is known to still contain biases in middle-atmospheric temperature and winds. It exhibits a cold bias in the lower stratosphere and a warm bias near the stratopause (Hersbach *et al* 2020). Wright and Hindley (2018) reported a decreasing correlation between temperatures in ERA5 and satellite data with increasing height in the stratosphere. For winds, Hersbach *et al* (2020) describe an occasionally overly strong tropical westerly jet and, rarely, excessively high wind speeds near the surface. More generally, the ECMWF analyses systematically underestimate the vertical shear of the horizontal wind when compared with Aeolus satellite wind profiles (Žagar *et al* 2025). Any circulation biases, associated either with a lack of wind observations or with model deficiencies, are part of our training data. The resulting uncertainties are difficult to quantify, but we believe they are smaller than other uncertainties related to the gravity wave filtering, estimation of vertical velocity, and model resolution.

We train the NNs at the coarser target resolution of 64×128 grid points ($\sim$300 km at the equator). At that resolution, only gravity waves of scale $\sim$2000 km and larger are fully resolved. Thus, it is possible to regard the 'small-scale' part of the gravity wave spectrum (wavelengths between $\sim$200 km and $\sim$2000 km) which is resolved in the high-resolution data, but not at the coarse resolution, as the subgrid part. This part needs to be parameterised in a climate model running at the coarse resolution. Such an approach will allow replacing the conventional gravity wave parameterisation in a coarse model in future work. Note that all statements made here about resolution depend on the latitude when using a lat-lon grid, and the values given above are estimated at the equator. Regarding the vertical resolution, we keep all 37 pressure levels.

2.2. Extraction of gravity waves using normal mode functions

A crucial decision is how to identify and separate the gravity waves from the global dynamical fields, and the filtering for specific wavelengths. Sun *et al* (2023) give insight into different methods for the extraction of gravity waves. Following a different approach, we use the software MODES (Žagar *et al* 2015) which is based on the theory of normal-mode function expansion (Hough 1898, Longuet-Higgins 1968, Kasahara and Puri 1981).

Normal-mode functions constitute a spectral basis in which the global wind and mass fields can be expressed simultaneously, and permit the decomposition into two basic types of motion, inertia-gravity waves and Rossby waves, with different horizontal length scales and vertical structures. The calculation of this basis starts from a linearised system of primitive equations, which is used to obtain differential equations describing the vertical and horizontal structure of the atmosphere. The basis functions are eigensolutions of these differential equations. For a detailed mathematical derivation, we refer to Kasahara and Puri (1981) and Žagar *et al* (2015).

Once the horizontal and vertical basis functions are obtained for a given background stratification and vertical grid, the global circulation can be projected onto them. The expansion completeness enables the filtering of the desired wave type (here, only gravity waves) and wavelengths in physical space to reconstruct the velocity and temperature perturbations of waves of interest. The filtering does not involve wave frequencies. This process is conducted at every time step independently using the precomputed basis functions. MODES gives a precise linear separation of inertia-gravity waves from the dynamical background, since the dispersion relations of the different wave types are inherent to the underlying theory of normal-mode functions. Although this is not of interest here, MODES provides both wind and temperature perturbations of gravity waves, since it is a multivariate decomposition. For examples of gravity-wave wind and temperature filtering, see Žagar *et al* (2017).

We apply MODES at the full ERA5 resolution. To obtain the full spectrum of IG waves, all wavelengths of the inertia-gravity wave type are projected back to physical space; for the subgrid-scale part (SG) at our selected coarse resolution, we filter for all inertia-gravity waves with zonal wavenumber (i.e. the number of wavelengths fitting into a circle of latitude) greater than 16. This corresponds to wavelengths smaller than $\sim$2500 km at the equator, which fits well with the typically used cut-off scale in many gravity wave studies (e.g. Polichtchouk *et al* 2023). This wavelength decreases as moving towards the poles, e.g. it is $\sim$1200 km at 60° latitude. We consider these waves to be unresolved by the coarse grid. The full set of parameters used for running MODES can be found in appendix A.1.

The applied version of MODES operates on terrain-following σ-levels. We define σ-levels using our pre-defined 37 pressure levels, and interpolate the input ERA5 data from pressure to σ-levels. After applying the backward projection to filter gravity waves, results are interpolated from σ-levels back to pressure levels. The vertical interpolations are carried out using Climate Data Operators (CDO, Schulzweida 2023) with the function *intlevelx3d*. This is different from a typical MODES setup which operates on the hybrid σ-pressure model levels of the ECMWF IFS model and interpolates them to σ-levels. The additional interpolation steps involved in our case are not considered detrimental to the study

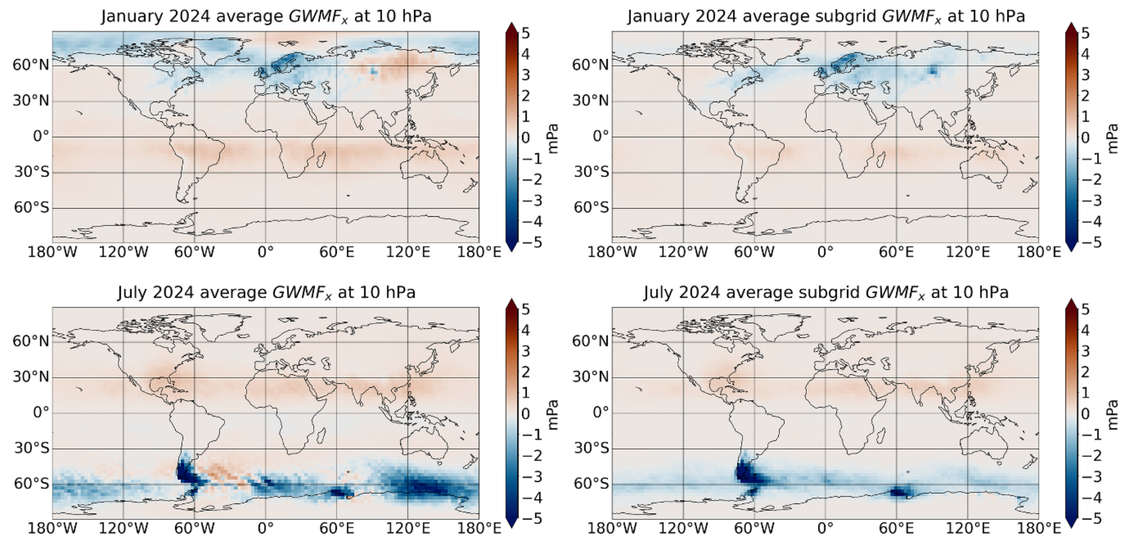

Figure 1. Monthly averages of zonal GWMFs in ERA5 for the full spectrum of gravity waves (IG, left) and the subgrid part (SG, right) at 10 hPa, for January (top) and July (bottom) 2024.

focusing on the middle atmosphere. The new version of MODES operating in the pressure vertical coordinate (Žagar *et al* 2023) will facilitate future work.

2.3. Calculation and evaluation of GWMFs

MODES yields global fields of horizontal wind perturbations (u', v') corresponding to the inertia-gravity waves and (in the subgrid-scale case) the filtered, unresolved wavelengths. Since vertical velocities ω in ERA5 are given in Pa s^{-1}, we calculate zonal and meridional momentum fluxes using the formulas

$$\mathrm{MF}_x = -g^{-1}\overline{u'\omega}$$
$$\mathrm{MF}_y = -g^{-1}\overline{v'\omega}.$$

Here, $g = 9.81\ \mathrm{m\,s^{-2}}$ is the gravitational acceleration of Earth and $\overline{\cdot}$ denotes averaging, which we implicitly perform by coarse-graining the data horizontally to a 64×128 grid with first order conservative remapping, using the CDO function *remapcon*. The version of MODES used for this study does not provide the pressure velocity perturbations ω'. Therefore, we use ω directly from ERA5, assuming that the true average is approximately zero (that is, $\omega' \gg \overline{\omega}$), and that in the areas occupied by gravity waves, ω stems mainly from small-scale motions associated with the waves and not from the large-scale balanced flow. This simplification follows Procházková *et al* (2025).

Monthly averages of zonal momentum fluxes for January and July at 10 hPa are shown in figure 1 for the full spectrum and the subgrid-scale part, respectively. Monthly averages of the months January, April, July, and October 2024 can be found in the appendix (figures A1 and A2). For further discussion of the dataset itself, additional figures and videos, as well as a detailed evaluation of the GWMFs in ERA5 calculated for this study, see Haslauer *et al* (2026).

An evaluation of GWMFs with observational data is difficult. Since we consider global data, only estimations of fluxes based on temperature measurements with satellites provide the necessary spatial coverage. Valuable references are Ern *et al* (2011, 2018), Geller *et al* (2013), and Hindley *et al* (2020), considering data collected by the HIRDLS, SABER, and AIRS satellites. However, quantitative comparison remains delicate, since data is from different years, and satellite measurements cover heights of ∼30 km and above only. Here, we only give a very short summary and refer to Haslauer *et al* (2026) for more details.

Considering the full spectrum of inertia-gravity waves, the fluxes found with MODES in ERA5 on 1 hPa and 10 hPa are roughly of the same order of magnitude as those obtained from HIRDLS and SABER satellite measurements (Ern *et al* 2011, 2018) for January and July, but are weaker on average. We attribute this to missing smaller scale waves in ERA5. This finding is consistent with prior analyses by, e.g. Lear *et al* (2024), Yoshida *et al* (2024). There is, however, a good qualitative agreement. For example, we observe relatively strong negative GWMFs between 50°S and 75°S in the months June, July, and August, which stem primarily from the wave activity associated with strong winds flowing across the high ridge of the Andes mountains. Such strong negative fluxes are typical for this season and region, and are also found in the literature, for climatologies (Hindley *et al* 2020) as well as for single events (Plougonven *et al* 2008). Also, there is good qualitative agreement with the analysis of GWMFs conducted by Gupta *et al* (2024b).

Regarding the subgrid part of the spectrum, we found no appropriate data available to compare to. As expected, both mean and maxima of momentum fluxes associated with small-scale gravity waves are clearly smaller in magnitude than those of the full spectrum. However, the general patterns of the fluxes are similar to those of the full spectrum.

Taking a closer look at figure 1, the large-scale structure of fluxes at 10 hPa in the Southern Hemisphere deserves some attention, related to the linear decomposition and the use of the total ω field. Linear wave decomposition by MODES separates 3D circulation between geostrophically balanced Rossby waves on the sphere and the remaining signal projecting on IG modes. Linear, spherical Rossby waves have a small divergence associated with the β-term, proportional to $v_g \beta / f$, where v_g is the meridional geostrophic wind, f is the Coriolis parameter and β is its meridional gradient. These isallobaric motions, i.e. most of ageostrophic dynamics related to the baroclinic Rossby wave dynamics in midlatitudes, project on the IG modes. This means that large-scale midlatitude IG modes are a mixture of ageostrophic circulation and inertia-gravity (or gravity) waves. The zonal wavenumber 3 structure of the flux at 10 hPa in midlatitudes thus largely reflects ageostrohic dynamics associated with vertically-propagating Rossby waves in the winter hemisphere (see discussion in Žagar *et al* 2023). After filtering out large scales, this part of the signal is greatly weakened.

2.4. Architecture of the NNs

As a first step towards developing physically consistent parameterisations for both orographic and non-orographic gravity waves, this study focuses primarily on orographic gravity waves. More precisely, we consider GWMFs only over land and use input features that describe the unresolved topography. NNs are trained on full atmospheric columns to incorporate the vertical propagation of gravity waves. The target variables of the networks are vectors of zonal and meridional momentum fluxes on 37 levels $\{\mathrm{MF}_x, \mathrm{MF}_y\}$, summing up to 74 variables in total. As feature variables, we use columns of three-dimensional wind and temperature $\{u, v, \omega, T\}$, and five scalar variables describing the orography within a grid cell, namely mean geopotential at the ground level z, standard deviation μ, anisotropy γ, angle σ (i.e. orientation of the terrain relative to an eastward axis), and slope θ of subgrid-scale orography. This amounts to 153 feature variables in total, all of which are obtained by coarse-graining ERA5 data.

Since we are interested in orographic gravity waves and do not consider horizontal propagation in the current study, we only use columns over land, i.e. all columns where the land-sea-mask of ERA5 is non-zero. Further, we exclude all columns north of 70°N and south of 70°S in the training data, to omit over-representation of these areas, where grid points lie increasingly dense. Starting from $64 \times 128 = 8192$ columns per time step, this reduction leaves us with 2794 columns per time step, which we use as samples for training the NNs.

Harnessing the full year 2024 for training, this gives us 24 542 496 training samples. As gravity waves are strongly related to specific atmospheric conditions (such as strong winds flowing over mountains) which often endure for several hours or even days, picking training and test data randomly from the same year seems unsuitable. Consecutive time steps are not independent in the setup with a time resolution of 1 h. We therefore use data from separate years for validation and testing (2022a) and independent testing (2022b, 2020). Each of these test sets consists of 2 414 016 samples.

All variables are normalised independently for each level by subtracting the mean and dividing by the standard deviation. While normalisation is a common step in many ML applications, we found it crucial for this task, since the magnitudes of momentum fluxes depend on the model level due to the decrease of density ρ with height. Without normalisation, upper levels would matter less.

We ran experiments with fully connected, convolutional, U-Net, and LSTM architectures. A slightly modified U-Net operating on atmospheric columns turned out to be the best performing type of NN for predicting orographic gravity waves. The U-Net (Ronneberger *et al* 2015), originally developed in the context of biomedical image segmentation and widely applied for image processing tasks, consists of a contracting path (encoder) of convolutional layers, and a corresponding upsampling path (decoder), which is able to include information from the contracting path by so-called skip connections. This type of model has already proven to perform well also in the field of atmospheric science (e.g. Heuer *et al* 2024, Gupta *et al* 2025). Here, we adopt the principle to our problem, using one-dimensional convolutions, which are applied along the columns of winds and temperature. In the so-called bottleneck layer between contracting and up-sampling path, we inject the five scalar variables describing the orography,

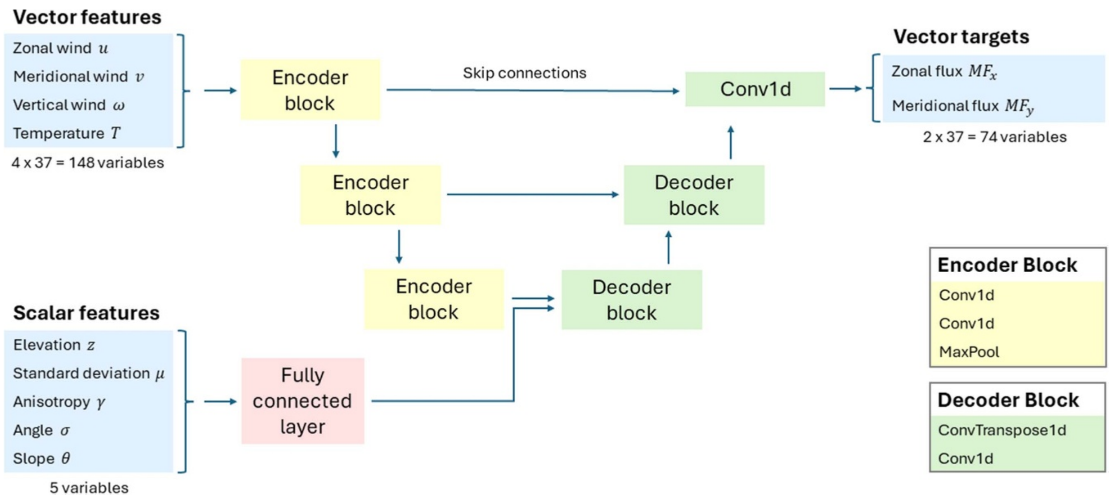

Figure 2. Modified U-Net architecture used in this study. First, vector features are passed through three encoder blocks. In the bottleneck, they are combined with the scalar features, which have passed a fully connected layer before. Then, the data are passed through two decoder blocks and a final 1D-convolution.

having passed them through one fully connected layer before. Optimising for the coefficient of determination, R^2, a setup with three convolutional blocks (including the bottleneck) and respective up-sampling blocks, with a total number of 2.6 million trainable parameters proves best for all our experiments. Figure 2 is an illustration of this architecture.

We implement the networks in Python using the ML framework PyTorch and train for 40 epochs with a batch size of 1024, using the *Adam* optimiser, mean squared error as loss function, a learning rate of 10^{-4}, and *LeakyReLU* as activation function. For regularisation, we apply a weight decay parameter of 10^{-5}.

We also tried adding an attention mechanism (Vaswani *et al* 2017) to the U-Net, which did not lead to significant improvements. In addition to adding the weight decay parameter, we tried to reduce overfitting by varying the complexity of the NNs, early stopping, and by means of L^1 and L^2 penalties, batch normalisation, and drop-out layers. None of these efforts lead to significant improvements of the observed overfitting (see section 3). For the experiments regarding model architecture, hyperparameters, and regularisation, we only used the training set and test set 2022a.

2.5. Experiments

In total, we present four experiments: We consider (1) the full spectrum of gravity waves (IG) and (2) the subgrid-scale part (SG), using the same coarse input variables. Further, for both cases, we investigate (a) training the networks over all land, and (b) training only over mountainous terrain, precisely only using the 50% of the columns with a standard deviation of the subgrid orography greater than the median standard deviation over land. We keep the architecture of the NNs fixed, but train four different models, independently, for the four cases. All NNs are applied to predict the GWMFs in columns over their respective training regions (all land or mountainous terrain) as a function of the coarse state variables.

2.6. Interpretability of the NNs

To ensure physical consistency and to strengthen the trust in ML models, it is important to understand how they work (Heuer *et al* 2024). This becomes increasingly difficult with more complex models. A common technique to analyse the predictions of a NN in relation to its inputs are SHAP values (Lundberg and Lee 2017), based on the game theoretic approach of Shapley (1953). SHAP values quantify the contribution of an input feature to the model's predictions. In brief, a positive SHAP value assigned to a positive input signifies that this input pushes the prediction to a higher value. Similarly, a negative SHAP value associated with a positive input implies that the respective input decreases the predicted value, and the reverse for negative inputs. The absolute value of a SHAP value is thus a measure of the impact of a feature. If it is high, the respective variable influences the model's prediction heavily, if it is close to zero, its contribution is of minor relevance.

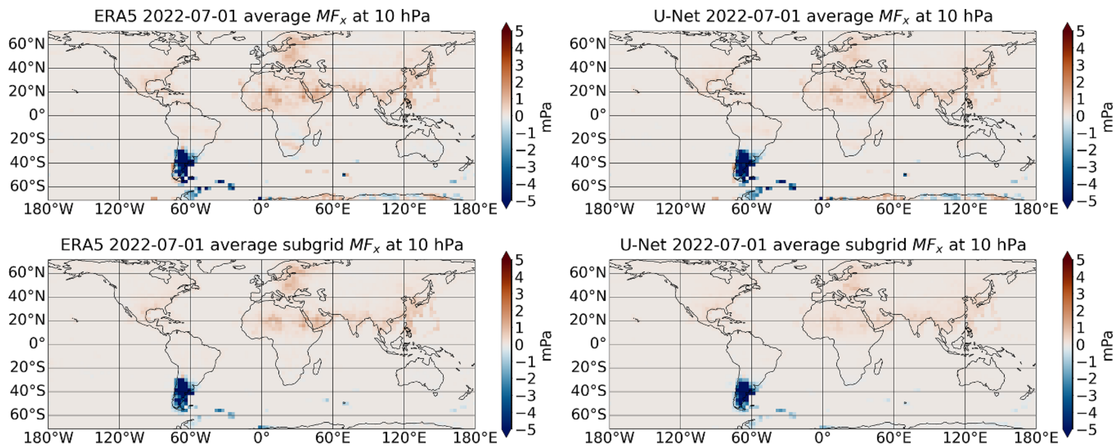

Figure 3. GWMFs in ERA5 (ground truth, left) and predictions of the U-Nets trained and applied over all land (right), for the full spectrum of gravity waves (IG, top) and the subgrid part (SG, bottom). The maps show fluxes averaged over 21 July 2022 at 10 hPa. Points outside the training/test regions are set to zero. Non-zero points over oceans are due to small islands, implying a non-zero land-sea-mask.

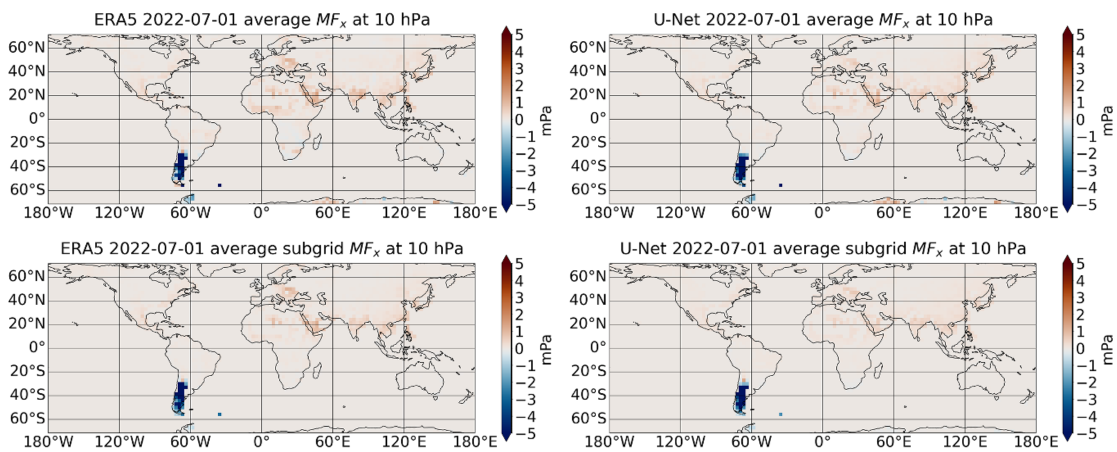

Figure 4. Like figure 3, but for the U-Nets trained and applied over mountainous terrain only.

3. Results

3.1. General performance

Figures 3 and 4 show zonal GWMFs in ERA5 and predictions of the U-Nets trained and applied over all land, and only over mountainous terrain, respectively, at 10 hPa, both for IG and SG, averaged over 21 July 2022 (24 time steps). The U-Nets are able to predict momentum fluxes of gravity waves very well, both for the full spectrum of gravity waves and for the small-scale waves. Comparing ground truth and the predictions of the NNs, we find a good agreement in structure and magnitudes globally. In particular, the NNs capture the regions of strong negative momentum fluxes, such as over the southern tip of South America around 50°S, 70°W, the belt of positive fluxes in the subtropics at 20°N of the Northern Hemisphere, and the positive and negative fluxes over islands in the Southern Atlantic (60°S, 30°W) and the Indian Ocean (45°S, 40–70°E).

Table 1 quantifies the skill of the U-Net in each of the four experiments in terms of the coefficient of determination, R^2, for the four datasets: the training set and test sets 2022a, 2022b, and 2020. First we observe the consistency of the NNs across the three test datasets. All three exhibit the same skill, with nearly identical R^2 values differing maximally by ± 0.02. This assures us that using 2022a for both validation and testing did not bias the results, and indicates that, at least globally, inferences based on different meteorological conditions remain consistent. We also emphasise the robustness of the skill across different years.

Second, we observe the decline in skill from the training to test sets, which could be an indication that the NNs overfit the training data. We tried to reduce this employing various common techniques,

Table 1. Overview of the skill of various setups of NNs on training set and different test sets. For every experiment, we state the values of the coefficient of determination, R^2, calculated over the respective training regions. We applied the NNs either over all land or only over mountainous terrain (standard deviation of the subgrid orography greater than its median over land), either for IG, or only for SG. Test sets 2022a and 2020 consist of days 1, 11, and 21 of each month in the year 2022, respectively 2020. Test set 2022b consists of days 6, 16, and 26 of each month in the year 2022.

	R^2 training set	R^2 test set 2022a	R^2 test set 2022b	R^2 test set 2020
IG all land model	0.82	0.69	0.69	0.70
IG mountains only model	0.86	0.72	0.72	0.72
SG all land model	0.70	0.56	0.55	0.55
SG mountains only model	0.78	0.63	0.61	0.62

Table 2. Skill of the four NNs applied over different test regions on the test set 2022a. For every experiment, we state the R^2 values calculated on the test regions of all land, mountains only (standard deviation of the subgrid orography greater than its median over land), high mountains (mountains only and top 20% elevation, i.e. average elevation within the grid cell higher than 1078 m), mountains and wind (mountains only and top 20% near-surface wind speed over mountains, i.e. absolute wind speed at the lowermost level higher than 4.25 m s^{-1}), high mountains and wind, and flat land (standard deviation of the subgrid orography lower than its median over land).

Test over ...	All land	Mountains	High mts	Mts & wind	High mts & wind	Flat land
IG all land model	0.69	0.71	0.75	0.73	0.77	0.64
IG mountains only model	0.64	0.72	0.76	0.74	0.78	0.49
SG all land model	0.56	0.62	0.62	0.66	0.69	0.37
SG mountains only model	0.52	0.63	0.64	0.67	0.70	0.20

including complexity reduction, early stopping, L^1 and L^2 penalties, batch normalisation, and drop-out layers (see section 2.4), all of which had only little impact. As described, we only kept a weight decay parameter. Since reducing the complexity (i.e. the number of trainable parameters) of the NNs leads to worse results for both training and test set, we believe the poorer skill could reflect low frequency variability of gravity wave processes, or a class imbalance, or sample size effects in the test set. Substantially more data may be needed to minimise the gap between training and test set, a conclusion which is supported by tests with varying amounts of training data.

Focussing on the performance of the NNs in the four different experiments, we find that the skill is better for the full spectrum of gravity waves (IG) than the subgrid scale (SG) component. We attribute this mainly to the fact that some of these waves are partially resolved on the coarse grid, such that there is information on these waves in the input data. Also, the GWMFs of the full spectrum have larger values and more pronounced structures, which seem to be easier to capture by the ML models.

Training and testing on regions with strong orography only (standard deviation greater than median) improves R^2 values for both IG and SG experiments, but much more for the SG setup. Although this restriction reduces the size of the training set by a factor of 2, the remaining columns seem to be more consistent in terms of the physical mechanisms related to orographic gravity waves, which the NNs are supposed to describe. In particular, we presume that the data contain less contamination from gravity waves of non-orographic sources, and potentially less noise from other atmospheric processes, leading to the observed improvements.

This intuition can be verified by considering the skill of the schemes trained over all land, but scored only over regions of strong orography. As seen in the second column of table 2, for both IG and SG, the all land schemes perform almost equivalently to the schemes trained only over strong orography. The poorer overall skill is due to the difficulty of capturing gravity wave effects over flat land. This is shown in the sixth column of table 2. While we focus on the mountains only region and scheme for the remainder of the study, the results are equivalent if we use the all land scheme.

To get a better understanding of the behaviour of the NNs depending on orography and atmospheric conditions we evaluated it on further subsets of the mountains region using test set 2022a. Columns 3–5 in table 2 show the skill of the different networks over high mountains (top 20% elevation, i.e. higher than 1078 m), mountains and near-surface wind (top 20% absolute wind speed, i.e. higher than 4.25 m s^{-1} wind speed at the lowermost level), and both combined. While all changes in R^2 values are moderate (maximal difference 0.07), the restrictions to the different subsets improve the results. In the IG case, in particular, the restriction to higher mountains seems to be effective, while in the SG case, information regarding near-surface wind speeds helps the model. In all cases, the skill on the subset with both conditions combined is best.

Figure 5 shows the performance of the NNs as a function of atmospheric pressure (left) as well as in different regions (right) for the SG case, trained and applied over mountainous orography (with test set

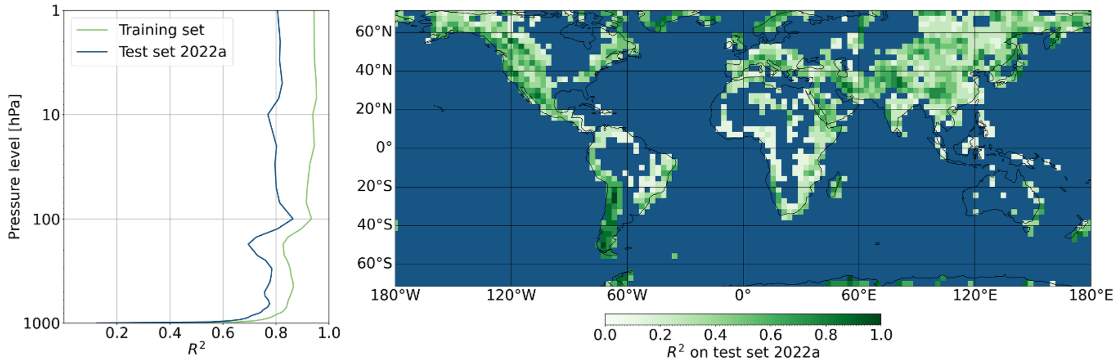

Figure 5. R^2 values of the U-Net trained and applied over mountainous terrain for the SG case. The left plot shows R^2 values of training and test set (2022a) for all grid cells and time steps on various model levels, the right plot the R^2 values of the test set for all levels depending on the region. Grid cells outside the training/test regions are shown in dark blue; 1.5% of the 'active' grid cells have negative R^2 values.

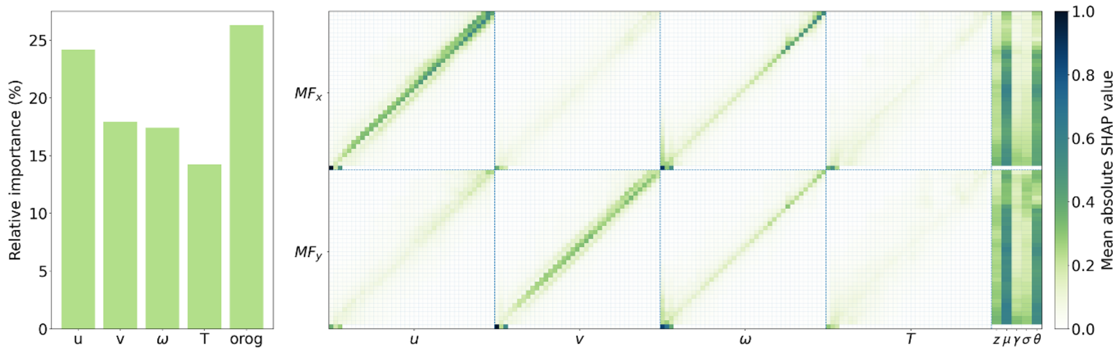

Figure 6. Relative importance of variables u, v, ω, T, and orographic variables z, μ, γ, σ, θ, ('orog'), summed over all levels (left) and mean absolute SHAP values for all levels separately (right), for the SG case with the U-Net trained over mountainous terrain. In the right plot, each square depicts the relation of one of the two target variable classes MF_x, MF_y and one of the four feature variable classes u, v, ω, T, for all combinations of model levels; in each square, the height z of the respective model level is decreasing from top to bottom and increasing from left to right. The boxes on the very right show SHAP values of the orographic variables z, μ, γ, σ, θ (columns from left to right). All values are normalised to 1 by dividing by the maximal value.

2022a). We find that the NNs perform similarly well at all levels, except for a minor drop of R^2 around 200 hPa. This is near the tropopause, and we expect the fluxes to change significantly at this level. We also observe decreasing performance at lower heights below 500 hPa. Generally, the quality of the predictions is better and more stable in the stratosphere (above ∼100 hPa). This could be related to the fact that in the troposphere, convective fluxes could contaminate the data, compared to a higher importance of the local wind related to wave breaking in the stratosphere. Considering the horizontal distribution of R^2 values, the quality of the predictions is spread quite heterogeneously over the globe. The skill of the NNs is best over mountainous regions, such as the Andes, the Rocky Mountains, Greenland, and Scandinavia.

For the IG case, the NNs show quite similar behaviour in terms of spatial distribution, but perform better on average (appendix, figure A3, see also table 1). Also, for the full spectrum, we observe a smaller decrease in performance in the lower atmosphere, where the R^2 of the test set basically stays around 0.8 all the way down to the ground.

3.2. SHAP analysis

A key question for the trust in our NNs is whether the predictions are physically meaningful. To test this, we analyse the networks using SHAP values (see section 2.6). Figure 6 shows absolute SHAP values for the U-Net trained with GWMFs of small-scale waves over all land, but the results are similar for all experiments.

First, we observe that the orographic variables and zonal winds u constitute the most important feature classes for prediction of GWMFs, followed by meridional and vertical winds, v and ω, and temperature T. Note that in the bar plot of figure 6, u, v, ω, and T relate to full columns of the respective

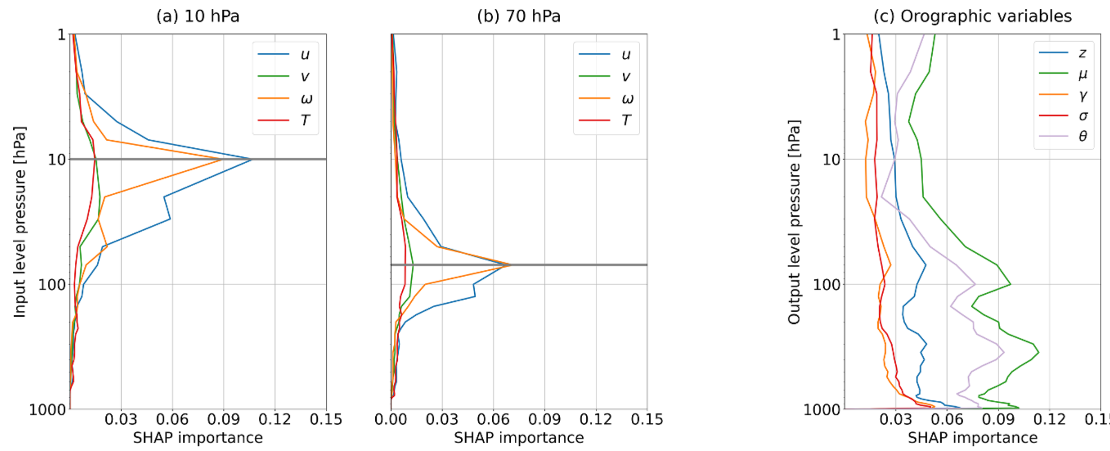

Figure 7. Absolute SHAP values for the prediction of zonal GWMFs in the SG case with the U-Net trained over mountainous terrain, averaged over all locations and times. The left plots show the SHAP importance of the input variables u, v, T, and ω at different pressures (y-axis) for the prediction of zonal GWMFs at (a) 10 hPa and (b) 70 hPa (indicated by the horizontal lines). (c) shows the SHAP importance of the orographic variables z, μ, γ, σ, and θ for the prediction of zonal GWMFs at different pressures (y-axis).

feature (37 variables each), while for orographic variables, 'orog' relates to five variables only. Thus, orographic features, especially standard deviation μ and slope θ, are most important when comparing single inputs. While in the SG case considered here, the class of orographic features is slightly more important than the zonal winds u, winds contribute strongest in the IG case, with percentages still being quite similar (appendix figure A4).

Taking a closer look at the relations between feature and target variables, the pronounced diagonals in the right panel of figure 6 clearly show that, picking a specific level of zonal momentum fluxes MF_x, u and ω on this level, as well as μ and θ, are features of dominant importance. In turn, for meridional momentum fluxes MF_y, this holds for v and ω. In both cases, neighbouring levels of u and v, especially the levels just below the considered level, also contribute significantly, as can be seen in the secondary diagonals, an effect intensifying in the upper levels.

This can be seen in greater detail in figures 7(a) and (b), which show the importance of the variables u, v, ω, and T on different levels for the prediction of zonal GWMFs at (a) 10 hPa and (b) 70 hPa. Similar results can be found in Connelly and Gerber (2024), and are consistent with our physical understanding of the mechanisms of gravity wave breaking and the related deposition of momentum. As gravity waves are propagating primarily upward, we expect physical relationships below and at the level of momentum deposition. The background wind at each respective level plays a decisive factor for critical level filtering. The levels above a given output level are also assigned higher SHAP values, suggesting that the vertical shear plays an important role in momentum deposition. Non-zero SHAP values well above the prediction level could be related to secondary wave generation, but could also indicate that the U-Net has learned to use correlated, but not causal relationships.

The SHAP values of the five scalar orographic variables are shown in the rectangles on the very right of figure 6 and, in more detail, in figure 7(c). Especially standard deviation σ and slope θ show significant contributions throughout the atmosphere. However, we found that the correlation between σ and θ is 0.98. The SHAP values of all features are weakening in higher regions. The corresponding plots for merdional GWMFs, and the IG case can be found in the appendix (figures A5–A7).

3.3. Comparison with Lott & Miller 1997 parameterisation scheme

The main application of the NNs developed in this study is the parameterisation of subgrid-scale gravity waves in climate models. For this purpose, the relevant physical quantity is the GWD, i.e. the force per mass that gravity waves exert on the atmosphere. The GWD is the pressure gradient of the momentum flux, multiplied by the gravitational acceleration g:

$$\mathrm{GWD}_x = g\frac{\partial \mathrm{MF}_x}{\partial p}$$

$$\mathrm{GWD}_y = g\frac{\partial \mathrm{MF}_y}{\partial p}$$

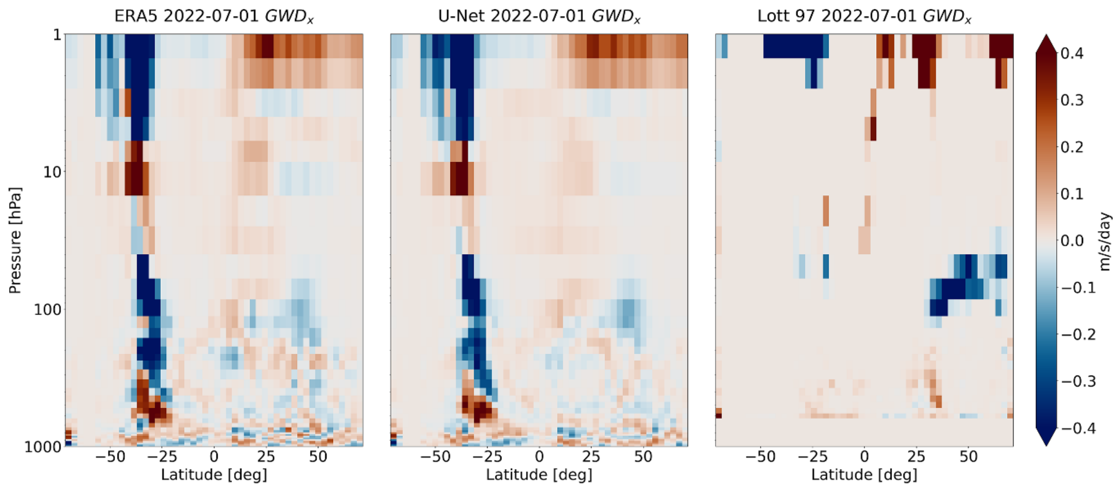

Figure 8. Zonal means of zonal gravity wave drag (GWD_x) due to small-scale (subgrid) gravity waves in ERA5 (left), the prediction of the U-Net trained for the SG case over mountains only (middle), and the modified conventional Lott and Miller scheme (right). Data are averaged over 1 July 2022 (24 time steps).

We investigate how our scheme compares to the GWD parameterisation developed by Lott and Miller (Lott and Miller 1997, Lott 1999), which is adopted in many models, including the ICON model (Zängl *et al* 2015, Giorgetta *et al* 2018). For our purposes, we recreated the FORTRAN implementation in Python, introducing slight simplifications and building an interface enabling the processing of ERA5 data. A more detailed description of the modifications can be found in appendix A.2. The original scheme is designed to parameterise tendencies due to three processes associated with unresolved topography: the gravity wave stress as a function of height over the full atmospheric column, low-level drag associated with blocked flow, and mountain lift. As we focus only on the tendencies due to GWD, we ignore the second and third part of the scheme for the purposes of this comparison, setting the respective tuning parameters to zero. However, even when switching these components on, they had almost no influence. Since the original scheme is tuned for the use in ICON and not for ERA5, we also tuned the parameterisation to get the best, i.e. most similar to ERA5, results achievable in terms of GWD. For details, see appendix A.2.

Figure 8 shows zonal means of zonal GWD due to small-scale waves averaged over 1 July 2022, for the ground truth ERA5 data, the prediction of our U-Net, trained on the SG case over mountains, and the conventional parameterisation. In accordance with the results above, the U-Net (middle) captures the structures of the ERA5 data (left) very well, also matching in terms of magnitudes, with only some slight local differences (e.g. weaker positive fluxes at ∼3 hPa, −40°, and at ∼100 hPa, −30°, missing negative fluxes at ∼100 hPa, 20°).

Many of the large-scale structures visible in ERA5 also appear in the conventional parameterisation's output (right). First, there is strong positive GWD in the uppermost levels of the atmosphere between the equator and the north pole in ERA5 and the Lott and Miller parameterisation. Second, we observe strong negative GWD in the NN predictions as well as in Lott and Miller's scheme at the highest levels between the south pole and the equator, as well as an attached, slightly tilted band of negative fluxes ranging from the top of the atmosphere down to around 100 hPa. This band is interrupted by a region of positive drag (ERA5 and U-Net at 10 hPa), while the drag simply vanishes in the conventional scheme.

In the lower atmosphere, below 100 hPa, there are more heterogeneous structures with positive and negative areas of GWD in both schemes. Lott and Miller's parameterisation captures some of the bigger patterns, such as the negative fluxes at around 100 hPa poleward from 25° latitude and positive fluxes at $\sim -75°$ and $\sim -25°$ latitude above the ground, but it is not able to reproduce the details found in ERA5.

In summary, the predictions of our U-Net are in strong agreement with ERA5 and match it clearly better than the conventional parameterisation, which is also reflected in the respective R^2 values (0.76 for the U-Net, −0.48 for the conventional scheme). This is somewhat to be expected, as the U-Net was trained to capture all of ERA5's 'subgrid' waves, while Lott and Miller target only orographic waves, and also finer scales not resolved in ERA5. The poor R^2 of −0.48 for the conventional scheme reflects the lack of smaller scale detail. The scheme was designed to capture the gross features in the distribution of the GWD, which are more similar. Figures A8–A10 showing the same comparison for three other dates exhibit a similar behaviour. This is encouraging, even if we point out that the comparison with the conventional scheme is made on a qualitative basis. A more thorough quantitative analysis will be possible as soon as the parameterisation is coupled to a climate model.

In terms of runtime, the U-Net and our Python version of the conventional parameterisation are comparable. Estimating GWD for an entire day (24 time steps) globally in the example case shown here took 142.3 s for the U-Net and 210.2 s for the conventional parameterisation. The computation time required for the other days shown in the appendix is similar.

4. Discussion

Before concluding, we summarise some of the strengths and limitations of this study and comment on how our approach compares to the study by Gupta *et al* (2025).

First, using ERA5, we have a large and well-known dataset, with one-hourly global output over one year. However, at the resolution of 0.25°, corresponding to ∼30 km at the equator, a large part of the gravity wave spectrum, i.e. waves with wavelengths smaller than ∼200 km, is still not resolved. In particular, many waves which need to be parameterised in climate models operating at resolutions below 100 km are not explicitly resolved, as discussed in more detail in section 2.1. Therefore, our study shows that in principle, gravity wave effects can be estimated by a data-driven model, but this approach still has to be extended to datasets based on higher-resolution integrations. Also, the highest level in our setup is at 1 hPa, which corresponds to ∼50 km, while state-of-the art climate models usually work with a higher model top, and gravity waves strongly impact the dynamics of the mesosphere above.

Second, while filtering with MODES is the physically most precise approach, since it respects the dynamical properties of gravity waves, the software does not differentiate GWs with respect to their sources. This work focuses on orographic gravity waves, while being aware that despite considering only regions over land (and particularly mountainous terrain) there is contamination by the effects of non-orographic waves. Also, using full ω for calculation of the fluxes possibly incurs some noise from other sources.

Finally, we want to point out the differences of our study and the work of Gupta *et al* (2025), which also employs ERA5 data to train NNs for GWMFs. Differently from our approach, the authors aim to include horizontal propagation of gravity waves, using a U-Net with attention mechanism operating on the full 3D model state and predicting full 3D gravity wave fields. While this eliminates one of the severest shortcomings of gravity wave parameterisations, coupling such a non-local scheme to a climate model is challenging. The column approach taken in our work makes it possible to use significantly smaller networks (2.6 million trainable parameters, compared to 38 million in Gupta *et al* (2025)), and would allow integration in current climate model parameterisation frameworks. In addition, using winds, temperature, and topographic parameters, we are able to give a physical interpretation of the predictions of our NNs in terms of SHAP values, as shown in section 3.2.

5. Conclusions

This study demonstrates that NNs can predict the force exerted by gravity waves over topographic regions. Employing the full year 2024 of the ERA5 reanalysis, momentum fluxes of gravity waves were calculated in the original ERA5 resolution, and then coarsened for the training of various ML models designed to predict the flux based on the coarsened model state alone. Four different NNs estimate momentum fluxes for either the full or the subgrid-scale gravity wave spectrum, over all land or over mountainous terrain only. A U-Net acting on atmospheric columns, with scalar features injected in the bottleneck layer, proved to perform best for this task in an offline setting.

Evaluating the networks on selected dates from the year 2022 and 2020, we found that the full spectrum of gravity waves was easier to predict compared to only the subgrid-scale waves (R^2 values 0.69 and 0.56 for the full and subgrid spectra, respectively, on the test set 2022a over all land). We attribute this difference in performance mainly to the fact that a part of the full spectrum of gravity waves is still resolved in the coarse resolution, and that the subgrid-scale part has more fine-scale structure. Training and testing on only regions with mountainous terrain, or nearly equivalently, scoring the all land models only over mountainous regions, improves the performance (R^2 values 0.72 and 0.63 on the test set 2022a), especially in the subgrid-scale case. These results suggest the GWMFs over flatter land are associated with other processes (e.g. convection, fronts), which are not well represented in the input features to our models, targeted to orographic gravity waves. Overall, we found exceptional robustness when testing the NNs on three different test sets, including one from another year. An issue still to be addressed is that our networks appear to overfit the training data. This might be improved by use of a larger and more diverse dataset, and will be investigated in future work.

Analysis of the NNs with XAI methods reveals that the contribution of input features is in accordance with our understanding of the physical mechanisms of orographic gravity waves. In particular, variables describing the orography as well as zonal (meridional) wind, and vertical wind are the dominant features for predicting zonal (meridional) momentum fluxes. Considering GWMFs on a given level, winds at the same and at neighbouring levels are most important, consistent with the role of dissipation by critical layers in our theoretical understanding of gravity wave dynamics.

Finally, a comparison with the conventional, physics-based parameterisation by Lott and Miller (1997) and Lott (1999) in terms of the GWD shows a good qualitative agreement. Key structures of the drag are captured in both approaches, which further supports the credibility and physical consistency of the NNs. However, the NNs capture the variability of the high-resolution data substantially better.

In a next step, we will extend our NNs to the case of non-orographic gravity waves. By including additional feature variables related to sources like convection, jets, or fronts, and training also over oceans, we aim for a scheme addressing the various wave types all in one. Also, we are investigating techniques of transfer learning (Gholizade *et al* 2025) to adapt our parameterisation to the ICON XPP model (Müller *et al* 2025) and run coupled online simulations. Training from a dataset based on a higher resolution model will help to overcome the limitations of ERA5 regarding the representation of GWMFs. A question still to be answered is how to tackle horizontal propagation of gravity waves, which is not covered by the current one-column approaches and poses challenges in coupling the ML-based parameterisation to the climate model. Nonetheless, this work demonstrates the potential of data-driven approaches to significantly improve the representation of gravity waves in models.

This study shows the capabilities of ML to accurately estimate gravity wave effects in weather and climate models and thereby improve their predictions and projections. Even as the resolution of models increases in the future, parameterisations will still be relevant for capturing the smallest scales and running large ensembles. To this end, for physically complex, non-local, transient, and multi-scale phenomena like gravity waves, NNs can play a key role. In addition, precise representations will help to improve our physical understanding, especially regarding the effects of gravity waves in the complex Earth system and their behaviour in a changing climate.

Acknowledgments

The authors thank the entire matrix group *Middle Atmosphere* at the Institute of Atmospheric Physics, Deutsches Zentrum für Luft- und Raumfahrt e.V., Oberpfaffenhofen, for fruitful discussions, constructive ideas, and their support. Also, we thank Claudia C Stephan, Leibniz Institute of Atmospheric Physics at the University of Rostock, Ostseebad Kühlungsborn, for interesting conversations, her input and assistance, as well as Corwin J Wright, Centre for Space, Atmospheric and Oceanic Science, University of Bath, and François Lott, Laboratoire de Météorologie Dynamique, Sorbonne Université Paris, for valuable comments.

Data availability statement

The code of this work will be published under https://github.com/EyringMLClimateGroup/haslauer26MLE_NNs_GravityWaves_ERA5/. The software MODES (Žagar *et al* 2015) can be obtained upon request at https://modes.cen.uni-hamburg.de/.

The data that support the findings of this study are openly available at the following URL/DOI: https://cds.climate.copernicus.eu/. In detail, we used the datasets *ERA5 hourly data on pressure levels from 1940 to present* and *ERA5 hourly data on single levels from 1940 to present* (Hersbach *et al* 2020, 2023a, 2023b).

Author contributions

Conceptualisation: E H; M S; A D; M R; V E. Methodology: E H; M S; A D; E G; M R; N Z; V E. Data curation: E H. Data visualisation: E H. Writing original draft: E H; M S; V E. All authors approved the final submitted draft.

Funding

Funding for this study was provided by the European Research Council (ERC) Synergy Grant 'Understanding and Modelling the Earth System with Machine Learning (USMILE)' under the Horizon 2020 research and innovation programme (Grant agreement No. 855 187). The authors gratefully acknowledge the Earth System Modelling Project (ESM) for funding this work by providing computing time on the ESM partition of the supercomputer JUWELS (Jülich Supercomputing Centre, 2021) at the Jülich Supercomputing Centre (JSC). M S acknowledges support from the DLR Quantum Computing Initiative and the Federal Ministry of Research, Technology and Space. E G acknowledges support from the US NSF through award OAC-2004 572, and Schmidt Sciences, as part of the Virtual Earth System Research Institute (VESRI). V E was additionally supported by the Deutsche Forschungsgemeinschaft (German Research Foundation) through the Gottfried Wilhelm Leibniz Prize awarded to V E (Reference No. EY 22/2-1).

Appendix A

A.1. Settings for MODES

For applying the software MODES (Žagar *et al* 2015), the following parameters were used:

First, describing up to which vertical and horizontal mode normal mode function decomposition is performed, we chose:

- Number of vertical modes: `num_vmode=30`
- Zonal wavenumbers: `num_zw=400`
- Number of Hough modes for `EIG`, `WIG`, and `ROT`: `maxl=230`

Second, for the projection back to physical space, we kept all vertical modes, all `EIG` and `WIG` modes (corresponding to inertia-gravity waves) and excluded all `ROT` modes. For the IG case, we kept all `kmode` when projecting back to physical space. For the SG, we chose `kmode_s=0` and `kmode_e=16`.

A.2. Python implementation of Lott & Miller 1997 GWD scheme

Our version of the parameterisation scheme by Lott and Miller (1997) and Lott (1999) is based on the FORTRAN code used in ICON (Zängl *et al* 2015, Giorgetta *et al* 2018), namely the module `mo_ssodrag`. We translated the code into Python as accurately as possible, with some changes to render it for the ERA5 case and to tune it to get the best results possible:

- All loops and the corresponding commands for parallelisation over horizontal grid points were removed. Parallelisation is implemented in ICON to reduce computation time, but is not required for our tests.
- The parts of the scheme not touched by our parameterisation, namely blocked low-level flow drag and mountain lift, were disabled. This was done by setting the tuning parameters `gkwake` and `gklift` to zero. However, switching these parts on with comparable tuning parameters led to only very slight differences.
- The tuning parameter `gkdrag`, which controls the strength of the GWD part of the parameterisation was set to `0.00001`, the parameter `nktopg` to 36.
- We built an interface preparing all relevant variables from ERA5, and passing them to the parameterisation, which yields zonal and meridional drag (in the terminology of ICON, `pdu_sso` and `pdv_sso`).
- In the sense of 'tuning' the parameterisation to the ERA5 data, indices corresponding to atmospheric levels in some loops were changed from `1` to `0`, such that GWD occures also in the uppermost level.

A.3. Additional figures

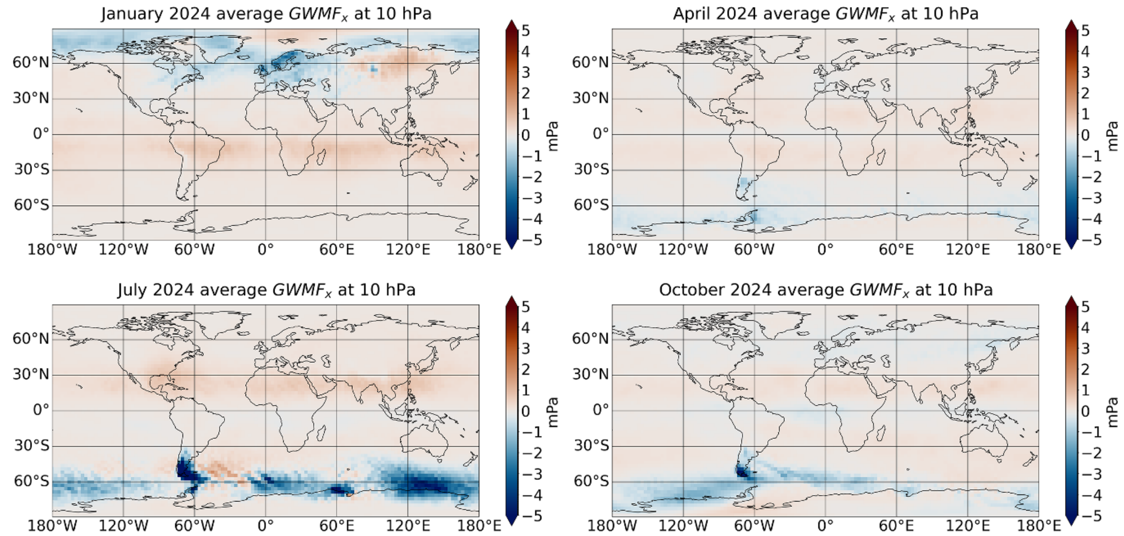

Figure A1. Monthly averages of zonal GWMFs in ERA5 of the full spectrum of gravity waves for January, April, July, and October 2024.

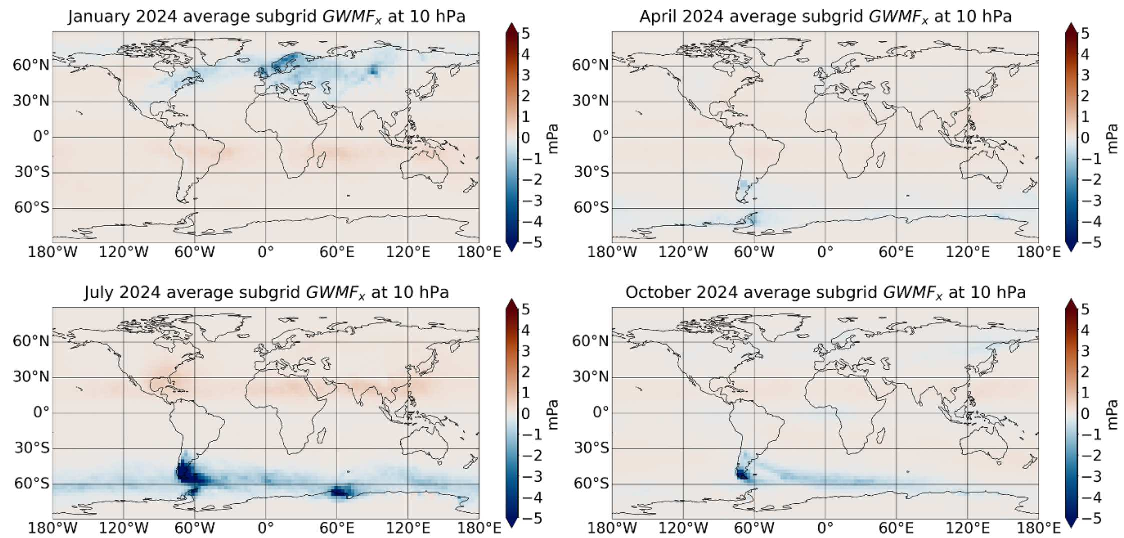

Figure A2. Like figure A1, but for the subgrid part of the gravity wave spectrum.

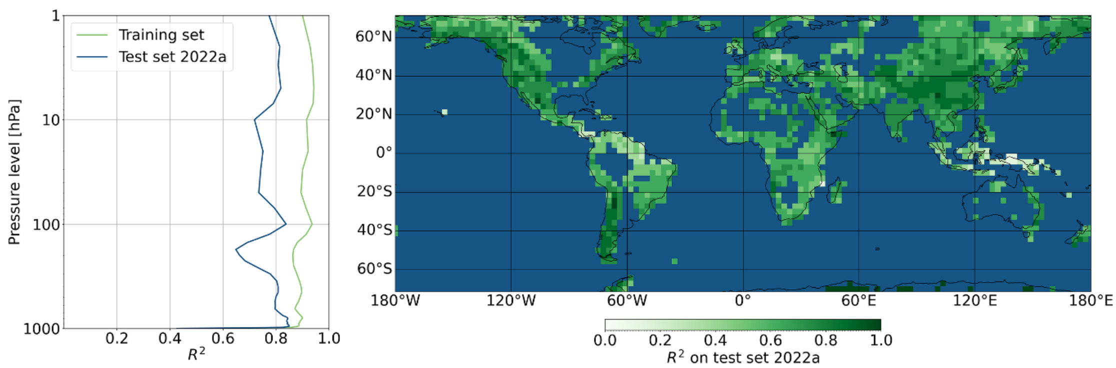

Figure A3. R^2 values of the U-Net trained and applied over mountainous terrain for the IG case. The left plot shows R^2 values of training and test set (2022a) for all grid cells and time steps on various model levels, the right plot the R^2 values of the test set for all levels depending on the region. Grid cells outside the training/test regions are shown in dark blue; none of the 'active' grid cells have R^2 values below 0.

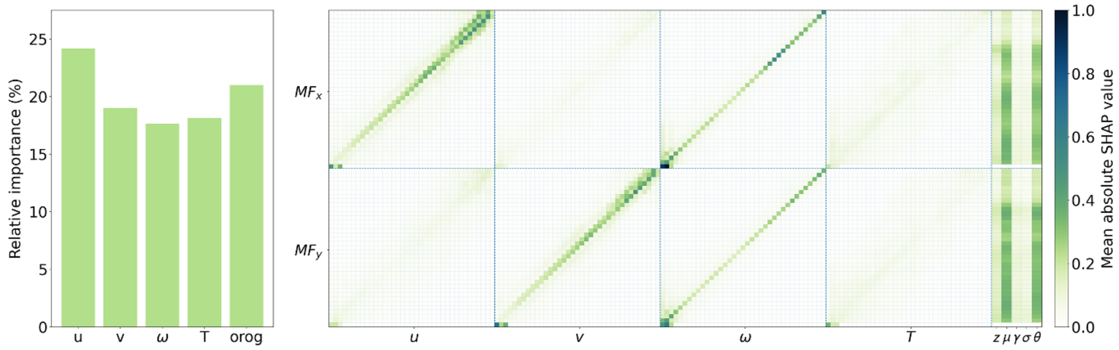

Figure A4. Relative importance of variables u, v, ω, T, and orographic variables z, μ, γ, σ, θ, ('orog'), summed over all levels (left) and mean absolute SHAP values for all levels separately (right), for the IG case with the U-Net trained over mountainous terrain. In the right plot, each square depicts the relation of one of the two target variable classes MF_x, MF_y and one of the four feature variable classes u, v, ω, T, for all combinations of model levels; in each square, the height z of the respective model level is decreasing from top to bottom and increasing from left to right. The boxes on the very right show SHAP values of the orographic variables z, μ, γ, σ, θ (columns from left to right). All values are normalised to 1 by dividing by the maximal value.

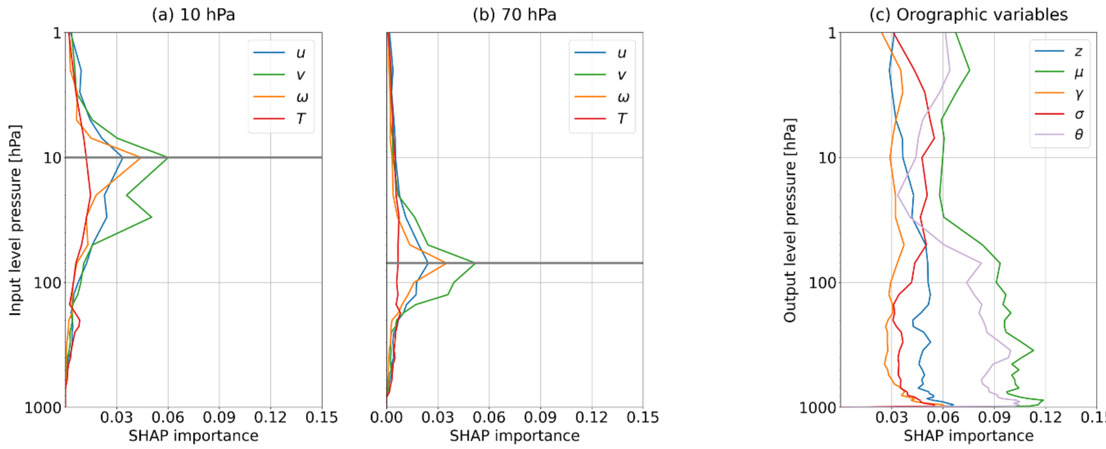

Figure A5. Absolute SHAP values for the prediction of meridional GWMFs in the SG case with the U-Net trained over mountainous terrain, averaged averaged over all locations and times. The left plots show the SHAP importance of the input variables u, v, T, and ω at different pressures (y-axis) for the prediction of zonal GWMFs at (a) 10 hPa and (b) 70 hPa (indicated by the horizontal lines). (c) Shows the SHAP importance of the orographic variables z, μ, γ, σ, and θ for the prediction of zonal GWMFs at different pressures (y-axis).

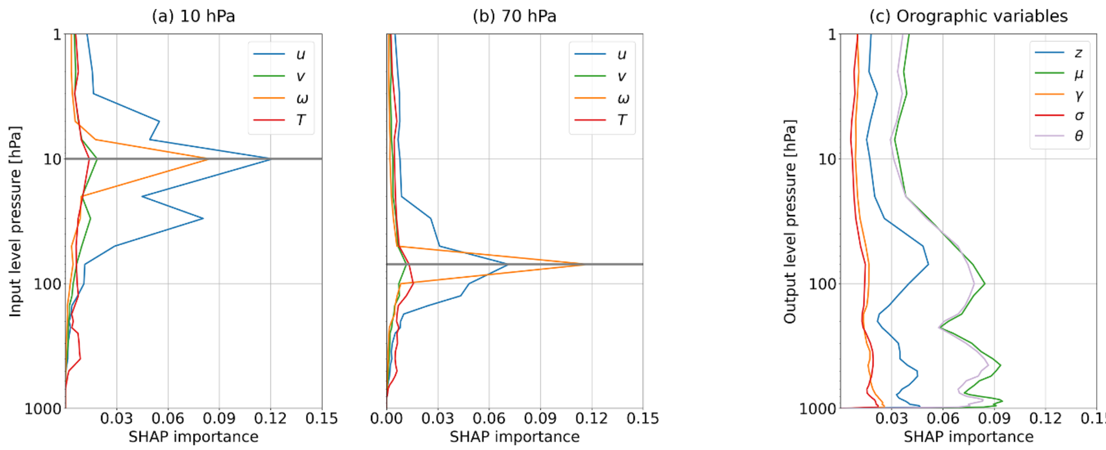

Figure A6. Like figure A5, but for zonal GWMFs in the IG case.

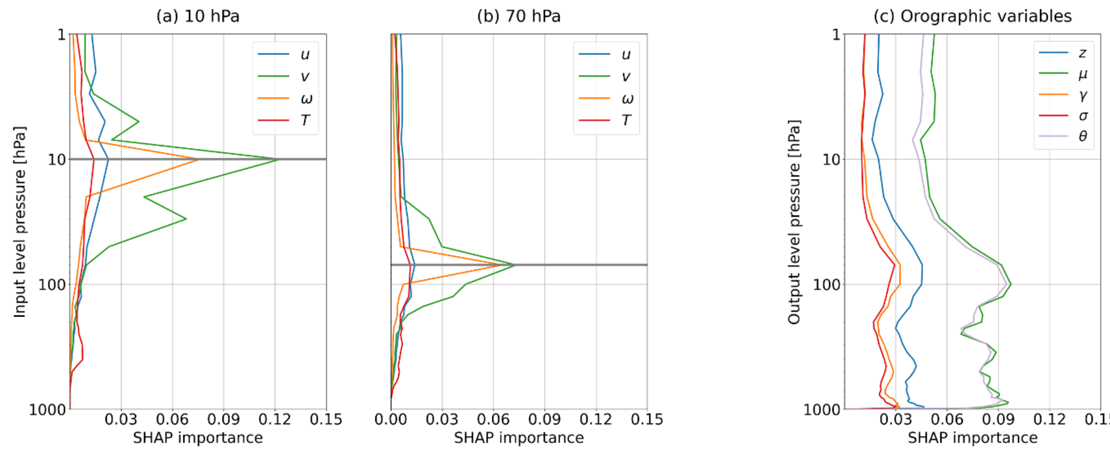

Figure A7. Like figure A5, but for meridional GWMFs in the IG case.

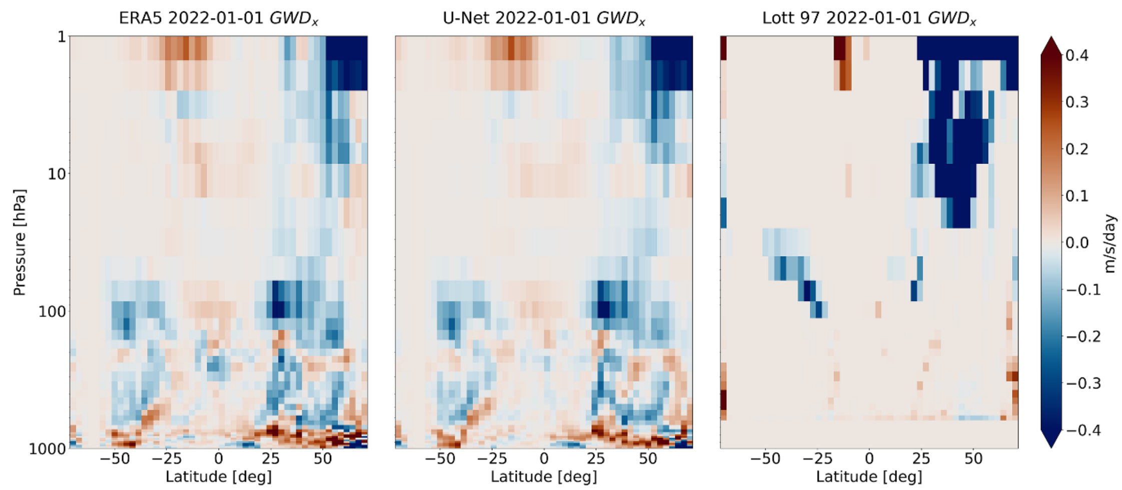

Figure A8. Zonal means of zonal gravity wave drag (GWD_x) due to small-scale (subgrid) gravity waves in ERA5 (left), the prediction of the U-Net trained for the SG case over mountains only (middle), and the modified conventional Lott and Miller scheme (right). Data are averaged over 1 January 2022 (24 time steps). R^2 values are 0.68 (U-Net) and −32.43 (conventional scheme).

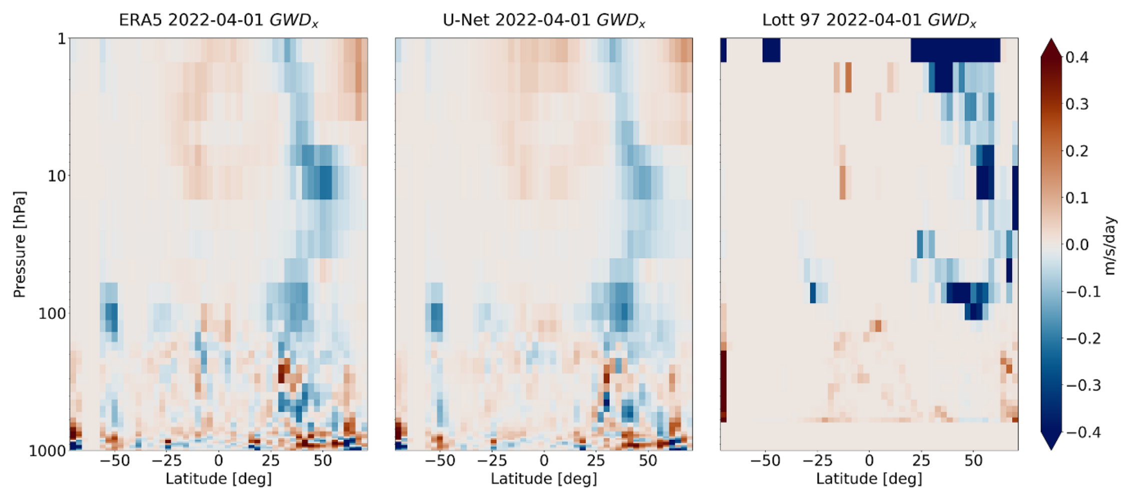

Figure A9. Like figure A8, but for 1 April 2022. R^2 values are 0.68 (U-Net) and −111.49 (conventional scheme).

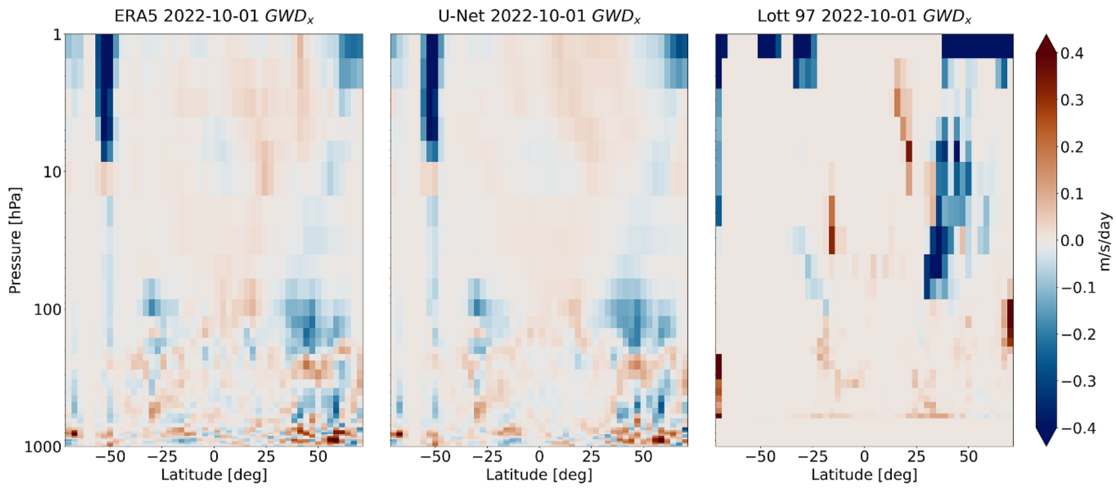

Figure A10. Like figure A8, but for 1 October 2022. R^2 values are 0.71 (U-Net) and −14.87 (conventional scheme).

ORCID iDs

Elias Haslauer 0009-0004-1265-3209
Mierk Schwabe 0000-0001-6565-5890
Andreas Dörnbrack 0000-0003-0936-0216
Edwin P Gerber 0000-0002-6010-6638
Markus Rapp 0000-0003-1508-5900
Nedjeljka Žagar 0000-0002-7256-5073
Veronika Eyring 0000-0002-6887-4885

References

Achatz U *et al* 2024 Atmospheric gravity waves: processes and parameterization *J. Atmos. Sci.* **81** 237–62
Alexander M J *et al* 2010 Recent developments in gravity-wave effects in climate models and the global distribution of gravity-wave momentum flux from observations and models *Q. J. R. Meteorol. Soc.* **136** 1103–24
Alzubaidi L, Zhang J, Humaidi A J, Al-Dujaili A, Duan Y, Al-Shamma O, SantamarÍa J, Fadhel M A, Al-Amidie M and Farhan L 2021 Review of deep learning: concepts, CNN architectures, challenges, applications, future directions *J. Big Data* **8** 53
Amiramjadi M, Mohebalhojeh A R, Mirzaei M, Zülicke C and Plougonven R 2020 The spatiotemporal variability of nonorographic gravity wave energy and relation to its source functions *Mon. Weather Rev.* **148** 4837–57
Amiramjadi M, Plougonven R, Mohebalhojeh A R and Mirzaei M 2023 Using machine learning to estimate nonorographic gravity wave characteristics at source levels *J. Atmos. Sci.* **80** 419–40
Beverley J D, Newman M and Hoell A 2024 Climate model trend errors are evident in seasonal forecasts at short leads *npj Clim. Atmos. Sci.* **7** 285
Burgh-Day C O d and Leeuwenburg T 2023 Machine learning for numerical weather and climate modelling: a review *Geosci. Model Dev.* **16** 6433–77
Chantry M, Hatfield S, Dueben P, Polichtchouk I and Palmer T 2021 Machine learning emulation of gravity wave drag in numerical weather forecasting *J. Adv. Model. Earth Syst.* **13** e2021MS002477
Connelly D S and Gerber E P 2024 Regression forest approaches to gravity wave parameterization for climate projection *J. Adv. Model. Earth Syst.* **16** e2023MS004184
Dong W, Fritts D C, Liu A Z, Lund T S, Liu H-L and Snively J 2023 Accelerating atmospheric gravity wave simulations using machine learning: Kelvin-Helmholtz instability and mountain wave sources driving gravity wave breaking and secondary gravity wave generation *Geophys. Res. Lett.* **50** e2023GL104668
Dörnbrack A, Gisinger S, Kaifler N, Portele T C, Bramberger M, Rapp M, Gerding M, Faber J, Žagar N and Jelić D 2018 Gravity waves excited during a minor sudden stratospheric warming *Atmos. Chem. Phys.* **18** 12915–31
Eichinger R, Rhode S, Garny H, Preusse P, Pisoft P, Kuchař A, Jöckel P, Kerkweg A and Kern B 2023 Emulating lateral gravity wave propagation in a global chemistry-climate model (EMAC v2.55.2) through horizontal flux redistribution *Geosci. Model Dev.* **16** 5561–83
Ern M, Preusse P, Gille J C, Hepplewhite C L, Mlynczak M G, Russell J M and Riese M 2011 Implications for atmospheric dynamics derived from global observations of gravity wave momentum flux in stratosphere and mesosphere *J. Geophys. Res.* **116** D19107
Ern M, Trinh Q T, Preusse P, Gille J C, Mlynczak M G, Russell III J M and Riese M 2018 GRACILE: A comprehensive climatology of atmospheric gravity wave parameters based on satellite limb soundings *Earth Syst. Sci. Data* **10** 857–92
Espinosa Z I, Sheshadri A, Cain G R, Gerber E P and DallaSanta K J 2022 Machine learning gravity wave parameterization generalizes to capture the QBO and response to increased CO_2 *Geophys. Res. Lett.* **49** e2022GL098174
Eyring V *et al* 2024 Pushing the frontiers in climate modelling and analysis with machine learning *Nat. Clim. Change* **14** 916–28
Fritts D C and Alexander M J 2003 Gravity wave dynamics and effects in the middle atmosphere *Rev. Geophys.* **41** 1003
Geller M A *et al* 2013 A comparison between gravity wave momentum fluxes in observations and climate models *J. Clim.* **26** 6383–405

Gentine P, Pritchard M, Rasp S, Reinaudi G and Yacalis G 2018 Could machine learning break the convection parameterization deadlock? *Geophys. Res. Lett.* **45** 5742–51

Gholizade M, Soltanizadeh H, Rahmanimanesh M and Sana S S 2025 A review of recent advances and strategies in transfer learning *Int. J. Syst. Assur. Eng. Manag.* **16** 1123–62

Giorgetta M A *et al* 2018 ICON-A, the atmosphere component of the ICON Earth system model: I. Model description *J. Adv. Model. Earth Syst.* **10** 1613–37

Grundner A, Beucler T, Gentine P, Iglesias-Suarez F, Giorgetta M A and Eyring V 2022 Deep learning based cloud cover parameterization for ICON *J. Adv. Model. Earth Syst.* **14** e2021MS002959

Gupta A, Reichert R, Dörnbrack A, Garny H, Eichinger R, Polichtchouk I, Kaifler B and Birner T 2024a Estimates of southern hemispheric gravity wave momentum fluxes across observations, reanalyses and kilometer-scale numerical weather prediction model *J. Atmos. Sci.* **81** 583–604

Gupta A, Sheshadri A and Anantharaj V 2024b Gravity wave momentum fluxes from 1 km global ECMWF integrated forecast system *Sci. Data* **11** 903

Gupta A, Sheshadri A, Roy S and Anantharaj V 2025 Offline performance of a nonlocal deep learning parameterization for climate model representation of atmospheric gravity waves *J. Adv. Model. Earth Syst.* **17** e2025MS004977

Gupta A, Sheshadri A, Roy S, Gaur V, Maskey M, and Ramachandran R 2024c Machine learning global simulation of nonlocal gravity wave propagation (arXiv:2406.14775)

Hardiman S C, Scaife A A, van Niekerk A, Prudden R, Owen A, Adams S V, Dunstan T, Dunstone N J and Madge S 2023 Machine learning for nonorographic gravity waves in a climate model *Artif. Intell. Earth Syst.* **2** e220081

Haslauer E, Schwabe M, Chen J, Dörnbrack A, Žagar N, and Eyring V 2026 Gravity wave momentum fluxes in ERA5 - A MODES analysis of the year 2024 In preparation

Hersbach H *et al* 2020 The ERA5 global reanalysis *Q. J. R. Meteorol. Soc.* **146** 1999–2049

Hersbach H *et al* 2023a 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 *et al* 2023b ERA5 hourly data on single levels from 1940 to present *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)* (https://doi.org/10.24381/cds.adbb2d47)

Heuer H, Schwabe M, Gentine P, Giorgetta M A and Eyring V 2024 Interpretable multiscale machine learning-based parameterizations of convection for ICON *J. Adv. Model. Earth Syst.* **16** e2024MS004398

Hindley N P, Wright C J, Hoffmann L, Moffat-Griffin T and Mitchell N J 2020 An 18-Year climatology of directional stratospheric gravity wave momentum flux from 3-D satellite observations *Geophys. Res. Lett.* **47** e2020GL089557

Hines C O 1997 Doppler-spread parameterization of gravity-wave momentum deposition in the middle atmosphere. Part 2: broad and quasi monochromatic spectra, and implementation *J. Atmos. Sol.-Terr. Phys.* **59** 387–400

Hough S S 1898 On the application of harmonic analysis to the dynamical theory of the tides. Part II. On the general integration of Laplace's dynamical equations *Proc. R. Soc.* **62** 209–10

Kasahara A and Puri K 1981 Spectral representation of three-dimensional global data by expansion in normal mode functions *Mon. Weather Rev.* **109** 37–51

Kim Y, Eckermann S D and Chun H 2003 An overview of the past, present and future of gravity-wave drag parametrization for numerical climate and weather prediction models *Atmos. Ocean* **41** 65–98

Lear E J, Wright C J, Hindley N P, Polichtchouk I and Hoffmann L 2024 Comparing gravity waves in a kilometer-scale run of the IFS to AIRS satellite observations and ERA5 *J. Geophys. Res.* **129** e2023JD040097

Longuet-Higgins M S 1968 The eigenfunctions of Laplace's tidal equations over a sphere *Phil. Trans. R. Soc. A* **262** 511–607

Lott F 1999 Alleviation of stationary biases in a GCM through a mountain drag parameterization scheme and a simple representation of mountain lift forces *Mon. Weather Rev.* **127** 788–801

Lott F and Miller M J 1997 A new subgrid-scale orographic drag parametrization: its formulation and testing *Q. J. R. Meteorol. Soc.* **123** 101–27

Lundberg S M and Lee S-I 2017 A unified approach to interpreting model predictions *Proc. 31st Int. Conf. on Neural Information Processing Systems, NIPS'* vol 17 (Curran Associates Inc) pp 4768–77 (arXiv:1705.07874v2)

Matsuoka D, Watanabe S, Sato K, Kawazoe S, Yu W and Easterbrook S 2020 Application of deep learning to estimate atmospheric gravity wave parameters in reanalysis data sets *Geophys. Res. Lett.* **47** e2020GL089436

McLandress C, Scinocca J F, Shepherd T G, Reader M C and Manney G L 2013 Dynamical control of the mesosphere by orographic and nonorographic gravity wave drag during the extended northern winters of 2006 and 2009 *J. Atmos. Sci.* **70** 2152–69

McLandress C, Shepherd T G, Polavarapu S and 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? *J. Atmos. Sci.* **69** 802–18

Müller W A *et al* 2025 The ICON-based Earth system model for climate predictions and projections (ICON XPP v1.0) *Geosci. Model Dev.* **18** 9385–415

Palmer T N, Shutts G J and Swinbank R 1986 Alleviation of a systematic westerly bias in general circulation and numerical weather prediction models through an orographic gravity wave drag parametrization *Q. J. R. Meteorol. Soc.* **112** 1001–39

Plougonven R, Hertzog A and Teitelbaum H 2008 Observations and simulations of a large-amplitude mountain wave breaking over the Antarctic peninsula *J. Geophys. Res.* **113** D16113

Polichtchouk I, van Niekerk A and Wedi N 2023 Resolved gravity waves in the extratropical stratosphere: effect of horizontal resolution increase from $O(10)$ to $O(1)$ km *J. Atmos. Sci.* **80** 473–86

Procházková Z, Zajíček R and Šácha P 2025 Climatology, long-term variability and trend of resolved gravity wave drag in the stratosphere revealed by ERA5 *Weather Clim. Dyn.* **6** 927–47

Rasp S, Pritchard M S and Gentine P 2018 Deep learning to represent subgrid processes in climate models *Proc. Natl Acad. Sci.* **115** 9684–9

Richter J H *et al* 2022 Response of the quasi-biennial oscillation to a warming climate in global climate models *Q. J. R. Meteorol. Soc.* **148** 1490–518

Ronneberger O, Fischer P and Brox T 2015 U-Net: Convolutional networks for biomedical image segmentation *Medical Image Computing and Computer-Assisted Intervention - Miccai 2015* vol 9351, ed N Navab, J Hornegger, W M Wells and A F Frangi (Springer) pp 234–41

Sarauer E, Schwabe M, Weiss P, Lauer A, Stier P and Eyring V 2025 A physics-informed machine learning parameterization for cloud microphysics in ICON *Environ. Data Sci.* **4** e40

Schulzweida U 2023 CDO User guide (available at: https://doi.org/10.5281/zenodo.10020800)

Shapley L S 1953 A value for n-person games *Contributions to the Theory of Games (AM-28), Volume II*, ed H W Kuhn and A W Tucker (Princeton University Press) pp 307–18

Simmons A 2022 Trends in the tropospheric general circulation from 1979 to 2022 *Weather Clim. Dynam.* **3** 777–809

Stensrud D J 2007 *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models* 1st edn (Cambridge University Press)

Stephan C C, Strube C, Klocke D, Ern M, Hoffmann L, Preusse P and Schmidt H 2019 Intercomparison of gravity waves in global convection-permitting models *J. Atmos. Sci.* **76** 2739–59

Sun Y Q, Hassanzadeh P, Alexander M J and Kruse C G 2023 Quantifying 3D gravity wave drag in a library of tropical convection-permitting simulations for data-driven parameterizations *J. Adv. Model. Earth Syst.* **15** e2022MS003585

Sun Y Q, Pahlavan H A, Chattopadhyay A, Hassanzadeh P, Lubis S W, Alexander M J, Gerber E P, Sheshadri A and Guan Y 2024 Data imbalance, uncertainty quantification and transfer learning in data-driven parameterizations: lessons from the emulation of gravity wave momentum transport in WACCM *J. Adv. Model. Earth Syst.* **16** e2023MS004145

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I 2017 Attention is all you need *Proc. 31st Int. Conf. on Neural Information Processing Systems, NIPS'* vol 17 (Curran Associates Inc) pp 6000–10 (arXiv:1706.03762v7)

Vautard R *et al* 2023 Heat extremes in Western Europe increasing faster than simulated due to atmospheric circulation trends *Nat. Commun.* **14** 6803

Vicente-Serrano S M *et al* 2022 Do CMIP models capture long-term observed annual precipitation trends? *Clim. Dyn.* **58** 2825–42

Wright C J and Hindley N P 2018 How well do stratospheric reanalyses reproduce high-resolution satellite temperature measurements? *Atmos. Chem. Phys.* **18** 13703–31

Yoshida L, Tomikawa Y, Ejiri M K, Tsutsumi M, Kohma M and Sato K 2024 Large-amplitude inertia gravity waves over Syowa station: Comparison of PANSY radar and ERA5 reanalysis data *J. Geophys. Res.* **129** e2023JD040490

Yuval J and O'Gorman P A 2020 Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions *Nat. Commun.* **11** 3295

Yuval J and O'Gorman P A 2023 Neural-network parameterization of subgrid momentum transport in the atmosphere *J. Adv. Model. Earth Syst.* **15** e2023MS003606

Žagar N, Jelić D, Blaauw M and Bechtold P 2017 Energy spectra and inertia-gravity waves in global analyses *J. Atmos. Sci.* **74** 2447–66

Žagar N, Kasahara A, Terasaki K, Tribbia J and Tanaka H 2015 Normal-mode function representation of global 3-D data sets: open-access software for the atmospheric research community *Geosci. Model Dev.* **8** 1169–95

Žagar N, Neduhal V, Vasylkevych S, Zaplotnik Ž and Tanaka H L 2023 Decomposition of vertical velocity and its zonal wavenumber kinetic energy spectra in the hydrostatic atmosphere *J. Atmos. Sci.* **80** 2747–67

Žagar N, Pilch Kedzierski R, De Chiara G, Healy S, Rennie M and Sielmann F 2025 ESA's Aeolus mission reveals uncertainties in tropical wind and wave-driven circulations *Geophys. Res. Lett.* **52** e2025GL114832

Zängl G, Reinert D, Rípodas P and Baldauf M 2015 The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: description of the non-hydrostatic dynamical core *Q. J. R. Meteorol. Soc.* **141** 563–79