Fast, accurate parameterization of atmospheric gravity waves, accounting for transient dynamics

David S. Connelly, Edwin P. Gerber, Ulrich Achatz, Georg Sebastian Voelker

Key points

- Transient dynamics affect gravity wave momentum transport, but are neglected in most parameterizations due to computational constraints
- The transient scheme MS-GWaM is configured to be highly accurate with an order of magnitude less runtime and memory than previous versions
- Low-cost configurations lie in a "sweet spot" balancing sufficient physical and spectral resolution against computational constraints

Abstract

Gravity waves with length scales from 1 km to 100 km, too small to be properly resolved in atmospheric models, transport momentum from the surface to the stratosphere and beyond. These waves propagate through the atmosphere on time scales of hours to days. Atmospheric model parameterizations, however, typically assume that gravity waves propagate instantaneously with respect to the model time step, and so neglect both the effects of unsteady large-scale winds on wave trajectories, as well as any transient, non-dissipative interactions between waves and the large-scale flow. This simplification is taken as a cost-saving measure, as capturing transient propagation has, to date, required an order of magnitude or more computational resources. Here, we systematically explore the tradeoff between computational cost and accuracy in the transient parameterization MS-GWaM, which resolves transient dynamics using ray tracing. Working first with a stand-alone version of the scheme, we find that careful calibration of key resolution parameters allows MS-GWaM to be run at low cost while still achieving accuracy comparable to much more computationally demanding configurations. The key is to optimally coarsen the spectral and vertical resolution together to make full use of the rays allowed by computational constraints. We then validate these low-cost configurations in an atmospheric model, where the best-performing one reproduces the key features of stratospheric variability observed in an expensive control integration.

Plain language summary

Gravity waves are oscillations in the atmosphere driven by buoyancy and the pull of Earth's gravity. Common sources include wind blowing over mountains and disturbances associated with convective storms. But while they are typically generated at the surface or in the lower atmosphere, gravity waves can propagate long vertical distances into the stratosphere and affect the velocity and direction of the winds there. Atmospheric models usually represent this propagation as occurring instantaneously, but in reality it takes hours or days, and this transience can significantly alter where and by how much gravity waves accelerate the wind. Transience has historically been too computationally expensive to include in numerical simulation of the atmosphere, but in this paper we find efficient configurations of a model of transient propagation. It proves important to balance the length and time scales of the modeled waves against the available computational power. Using a simplified model, we systematically identify the optimal configuration parameters under significant computational constraints. We then verify that this low-cost configuration performs well when included in a full atmospheric model.

1 Introduction

Atmospheric inertia-gravity waves arise as the superposition of buoyancy and Coriolis oscillations. In the troposphere, they are generated by a diverse set of sources (convection, flow over topography, and frontal activity, among others) and typically propagate upwards into the stratosphere and mesosphere, where they play a first-order role in the momentum budget of the middle atmosphere (Achatz, 2022). However, a substantial portion of the gravity wave spectrum has spatial scales on the order of a few tens of kilometers (Fritts & Alexander, 2003), below what can reliably be resolved by climate models. As a result, models must employ gravity wave parameterizations (GWPs) to capture the momentum transport associated with gravity waves. These parameterizations remain a key locus of model uncertainty (Richter et al., 2020).

All operational GWPs sacrifice physical fidelity to achieve computational efficiency. Most assume that gravity waves propagate instantaneously (or at least within one model time step) and only in the vertical. Although these assumptions reduce the computational cost of the GWP, they are at odds with physical reality. Gravity wave group velocities are on the order of $0.1\,\mathrm{m\,s^{-1}}$ to $1\,\mathrm{m\,s^{-1}}$, such that they take anywhere between several hours and a few days — much longer than a model time step — to travel tens of kilometers into the middle atmosphere. Gravity waves also propagate laterally in the atmosphere. For example, relative to observations, climate models consistently underestimate gravity wave activity in the Drake Passage, where powerful waves generated over the Andes move southward (Kruse et al., 2022). In the tropics, too, Kim et al. (2024) found that equatorward-propagating extratropical waves were crucial to modeling the descent rate and latitudinal structure of the easterly phase of the Quasi-Biennial Oscillation (QBO).

Ray tracing is a framework for parameterizing gravity waves that avoids both these troubling assumptions. Wave packets are modeled as Lagrangian tracers that evolve according to ordinary differential equations (ODEs). Their state persists across time steps, such that wave transience is represented, and they can be allowed to travel laterally between grid cells. Ray tracing schemes have been validated against wave-resolving simulations (Jochum et al., 2025; Muraschko et al., 2015) and used to study atmospheric phenomena such as the propagation of waves generated by convection (Song & Chun, 2008) and the refraction of gravity waves into the polar night jet (Amemiya & Sato, 2016). But the theoretical advantages of ray tracing come at considerable practical cost: for example, experiments with the Multiscale Gravity Wave Model (MS-GWaM), a ray tracing scheme, found it to be 50 times more demanding than an operational GWP, leading to a roughly fourfold increase in total integration time even without lateral propagation (Bölöni et al., 2021). Such computational demands are a significant obstacle to operational use of these parameterizations.

The tradeoff between the computational costs of ray tracing GWPs and their accuracy, however, has

not been well-studied. To the extent that a Lagrangian scheme has a "numerical resolution," it is largely determined by two source scale parameters δz and δc_{p}, which set the extent in physical and spectral space, respectively, of each wave packet. Reducing these parameters allows the ray tracer to respond to smaller-scale changes in the mean state and capture finer spectral features. But doing so also increases the rate at which new wave packets are added to the system, and therefore necessitates increasing $N_{\max}$, the number of packets allowed to propagate at once. Since the number of ODEs the ray tracer must solve, and therefore its computational cost, is set by $N_{\max}$, there is a fundamental tension between accuracy and efficiency in the choice of these three parameters. Exploring this tradeoff is the key theme of this work.

In particular, Bölöni et al. (2021) coupled MS-GWaM to the upper-atmosphere configuration of the Icosahedral Non-hydrostatic (ICON) model (Borchert et al., 2019; Zängl et al., 2014) with $N_{\max} = 2500$ and found the parameterization to be roughly converged but computationally onerous. Further studies have used values of $N_{\max}$ of similar or larger orders of magnitude (Kim et al., 2024; Voelker et al., 2024). Here, we develop a cheap-to-run stand-alone version of MS-GWaM and use it to investigate whether configurations with $N_{\max} = 250$ can be made sufficiently accurate for operational use. We then verify, to the extent possible, the accuracy of these configurations in an atmospheric model. Although we retain the assumption of purely vertical propagation, our results are predicated on dramatically reducing the number of Lagrangian tracers needed by the GWP. As such, we feel that our approach is readily extensible to the case of three-dimensional propagation.

2 Models and data

This work is centered around experiments with an implementation of MS-GWaM in the Model of an idealized Moist Atmosphere (MiMA), both of which are outlined in this section. We then characterize the particular wave source and the MiMA integrations from which we draw test scenarios. Finally, we describe the single-column ray tracer we use to conduct calibration and sensitivity studies.

2.1 Ray tracing and MS-GWaM

Here we sketch the MS-GWaM parameterization and provide a brief account of the theory that underpins it. For a more detailed presentation of this scheme, we refer the reader to Bölöni et al. (2021).

MS-GWaM is a *transient* parameterization, in that it models gravity wave packets propagating on the same time step as the parent model. Trajectories are affected by changes in the large-scale wind $\bar{\boldsymbol{u}}$, and $\bar{\boldsymbol{u}}$ in turn may be driven by non-dissipative accelerations: momentum flux gradients associated not with

wave breaking, but rather with waves changing amplitude as they propagate through a region. These accelerations, as well as the time dependence of the eventual breaking levels of each wave, are ignored by instantaneous schemes. A transient approach also allows for the simulation of reflection of wave packets.

MS-GWaM captures these transient effects by representing wave packets as Lagrangian objects moving according to the *ray tracing* equations. Letting $\mathrm{d}/\mathrm{d}t$ denote the derivative along a wave packet trajectory, the ray tracing equations in two dimensions are

$$\begin{aligned}
\frac{\mathrm{d}z}{\mathrm{d}t} &= \frac{\partial \omega}{\partial m} \\
&= -m \frac{\hat{\omega}^2 - f^2}{\hat{\omega}\left(|\boldsymbol{k}|^2 + m^2\right)} && \text{(1a)} \\
&\equiv c_\mathrm{g} \\
\frac{\mathrm{d}m}{\mathrm{d}t} &= -\frac{\partial}{\partial z}\left[\boldsymbol{k}\cdot\bar{\boldsymbol{u}} + \hat{\omega}\right] \\
&= -\boldsymbol{k}\cdot\frac{\partial \bar{\boldsymbol{u}}}{\partial z} - \frac{NK^2}{\hat{\omega}\left(|\boldsymbol{k}|^2 + m^2\right)}\frac{\partial N}{\partial z} && \text{(1b)} \\
&\equiv \dot{m} \\
\frac{\mathrm{d}\mathcal{A}}{\mathrm{d}t} &= -\mathcal{A}\frac{\partial c_\mathrm{g}}{\partial z} + \mathcal{S} && \text{(1c)}
\end{aligned}$$

where N is the buoyancy frequency; $\boldsymbol{k} \equiv (k, \ell)$ is the horizontal wave vector; m is the vertical wavenumbers; c_g is the vertical group velocity; ω is the extrinsic frequency; and $\hat{\omega} \equiv \omega - k\bar{u}$ is the intrinsic (Doppler-shifted) frequency satisfying the gravity wave dispersion relation

$$\hat{\omega}^2 = \frac{|\boldsymbol{k}|^2 N^2 + m^2 f^2}{|\boldsymbol{k}|^2 + m^2} \tag{2}$$

The wave action density $\mathcal{A} \equiv E/\hat{\omega}$, where E is the energy density associated with the wave packet, is related to the momentum flux by the identity $\overline{\boldsymbol{u}'w'} = \boldsymbol{k}\mathcal{A}c_\mathrm{g}$, and $\mathcal{S}$ represents any non-conservative sources or sinks of wave action. Ray tracing parameterizations launch a spectrum of wave packets, and step the position, wavenumber, and action of each one forward in time according to the set of differential equations (1).

(Strictly speaking, $k\mathcal{A}c_\mathrm{g}$ is the *pseudo*momentum flux, but we will identify it with the momentum flux to ease readability and to agree with much of the GWP literature. See Wei et al., 2019 for further discussion.)

Spectral discretization In practice, equation (1c) can prove troublesome: the derivative $\partial_z c_\mathrm{g}$ can only be computed by referencing neighboring wave packets, and so may be ill-defined if there are multiple packets with different spectral properties at similar altitudes (Muraschko et al., 2015). To overcome this difficulty, MS-GWaM adopts an approach first applied to atmospheric gravity waves by Hertzog et al. (2002). The

gravity wave field is modeled as a superposition of non-interacting wave fields $(m_\beta, \mathcal{A}_\beta)$, indexed by β and satisfying (1) individually. Then, defining the *spectral* wave action density

$$\mathcal{N}(z, m, t) \equiv \sum_\beta \mathcal{A}_\beta(z, t)\delta\left(m - m_\beta\right) \tag{3}$$

where δ is the Dirac delta function, one can show

$$\frac{\partial \mathcal{N}}{\partial t} + c_\mathrm{g}\frac{\partial \mathcal{N}}{\partial z} + \dot{m}\frac{\partial \mathcal{N}}{\partial m} = \mathcal{S} \tag{4}$$

Equation (4) shows that absent explicit sources and sinks, $\mathcal{N}$ is exactly conserved along wave packet trajectories, provided we consider those trajectories in a joint z-m phase space. (The proof of (4) is rather formal and well-presented in both Hertzog et al. (2002) and Muraschko et al. (2015), so we omit it here.) MS-GWaM therefore evolves its Lagrangian wave packets according to (1a), (1b), and the spectral action equation $\mathrm{d}\mathcal{N}/\mathrm{d}t = 0$, such that the trajectories are now interpreted as in the full z-m phase space.

Integration of (3) with respect to m yields

$$\mathcal{A}(z, t) \equiv \sum_\beta \mathcal{A}_\beta(z, t) = \int \mathcal{N}(z, m, t)\,\mathrm{d}m$$

such that the driving of the large-scale wind by parameterized gravity waves is

$$\left(\frac{\partial \bar{\boldsymbol{u}}}{\partial t}\right)_\mathrm{gw} = -\frac{1}{\rho}\frac{\partial}{\partial z}\int \boldsymbol{k}\mathcal{N}c_\mathrm{g}\,\mathrm{d}m \tag{5}$$

Ray volumes The divergence of the velocity field in phase space is

$$\frac{\partial}{\partial z}c_\mathrm{g} + \frac{\partial}{\partial m}\dot{m} = \frac{\partial}{\partial z}\frac{\partial \omega}{\partial m} - \frac{\partial}{\partial m}\frac{\partial \omega}{\partial z} = 0 \tag{6}$$

such that the induced flows are volume-preserving. As a result, rather than representing wave packets as point particles, MS-GWaM discretizes its equation set using Lagrangian *ray volumes* with finite extent in z-m space (Figure 1).

At each time step, the group velocity is calculated at the top and bottom of each ray volume and the vertical position and extent evolve accordingly. (The group velocity c_g depends on N through its dependence on $\hat{\omega}$, and so may vary in the vertical.) The wavenumber tendency $\dot{m}$ is evaluated at the center of the ray volume. Finally, the updated spectral extent is chosen to satisfy $\Delta z_\mathrm{old}\Delta m_\mathrm{old} = \Delta z_\mathrm{new}\Delta m_\mathrm{new}$, so that the numerics respect (6) exactly.

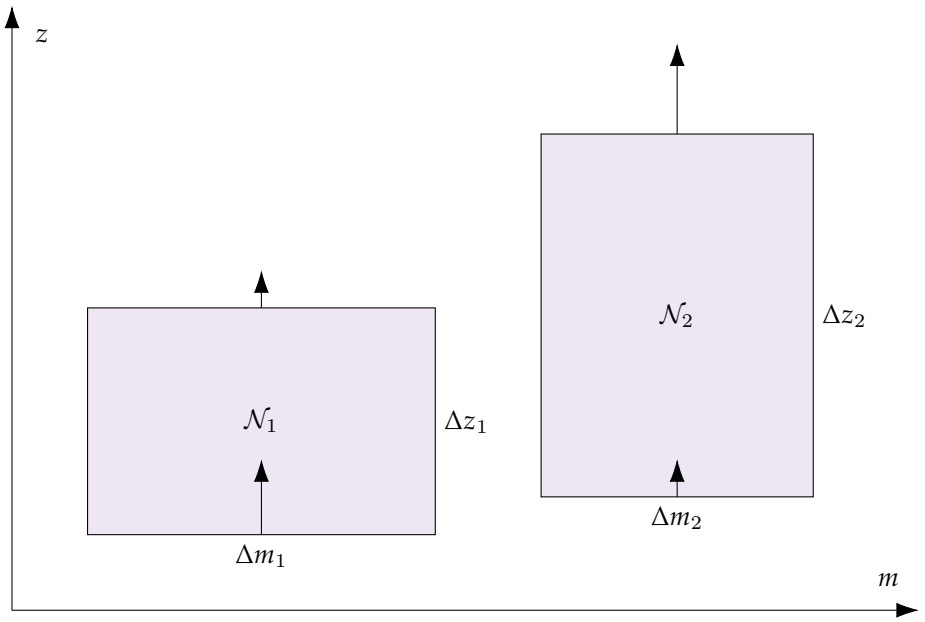

Figure 1 Schematic of ray volumes propagating in z-m phase space. Arrows indicate group velocities as calculated at the top and bottom of each volume.

The gravity wave momentum fluxes at a point z_i on the model vertical grid are calculated via the projection operation

$$\overline{\boldsymbol{u}'w'}\Big|_{z_i} = \sum_j \alpha_{ij} \left(\boldsymbol{k}\mathcal{N}c_{\mathrm{g}}\Delta m\right)_j \tag{7}$$

where j ranges over active ray volumes and $\alpha_{ij} \in [0, 1]$ is the fraction of the grid cell containing z_i intersected by the j^{th} ray volume. Equation (7) is a discrete approximation of the integral in (5), so that after projection, the acceleration of the large-scale wind by parameterized gravity waves can be approximated by finite differences.

Bottom boundary condition MS-GWaM allows the user to specify an arbitrary source spectrum. In an instantaneous GWP, each wave packet would be assumed to undergo its entire life cycle immediately, and so the whole spectrum can be launched each time step. But ray volumes in MS-GWaM move with finite velocity and persist between steps, and so more care must be taken at the source.

Following Bölöni et al. (2021), we impose a constant momentum flux as a function of wavenumber; a similar approach, however, would be necessary if the desired momentum flux varied in time. We fix a source intrinsic frequency $\hat{\omega}$ and a phase speed region of interest, discretized into channels of width δc_{p}. (We find it more convenient to parameterize the source in terms of the phase speed than the wavenumber m, but one can be recovered from the other, given $\hat{\omega}$.)

A *ghost layer* below the prescribed source level is added to ensure control of the flux at the source

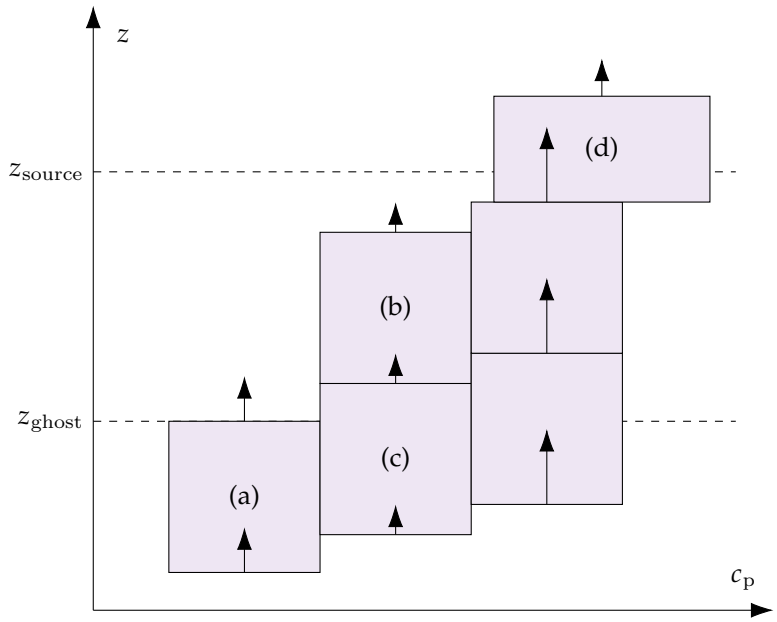

Figure 2 Three phase speed channels in the source. Ray volume (a) has just been instantiated, with its upper edge at z_{ghost}. The bottom edge of volume (b) has cleared z_{ghost}, so a new ray volume (c) is instantiated just below it with the same phase speed. Meanwhile, volume (d) has entered the domain, and so it evolves according to the full equation set: its wavenumber has shifted, and it experiences different group velocities at its upper and lower edges. Its aspect ratio has changed (exaggerated here for illustrative purposes) but its volume has not.

(Figure 2). New ray volumes are instantiated with spectral width δc_{p}, vertical extent δz, and upper edges at z_{ghost}. When the bottom boundary of one volume has cleared z_{ghost}, another is instantiated just below it with the same phase speed. In this way, there is always exactly one ray volume passing through the source level in each phase speed channel. Until its center has cleared the source level and entered the domain interior, each ray volume is prevented from refracting (that is, m is held constant) and has its vertical velocity calculated uniquely at its center, so that its aspect ratio does not change.

Sinks of wave action The spectral action density $\mathcal{N}$ associated with each ray volume is conserved, except for the effects of three sinks. First, if a static instability criterion is met, assuming constructive interference between all wave packets intersecting a given vertical layer, the spectral densities of all ray volumes at that level are scaled down as necessary to prevent convective instability. Second, at each time step, $\mathcal{N}$ is slightly diminished by a damping term with viscosity inversely proportional to the large-scale density ρ. In our implementation, this damping is replaced in the uppermost layers by a sponge that prevents gravity wave momentum from escaping the model domain.

Unlike the previous two sinks, the third is numerical rather than physical. For practical purposes, one must choose a value N_{max} fixing the maximum number of active ray volumes permitted in a given column at once. If there are already N_{max} ray volumes propagating and a new one must be instantiated to enforce

the source boundary condition, an existing ray volume must be *pruned* to make room. (Following Bölöni et al. (2021), we prune the ray volume with the lowest energy density $E = \mathcal{N}\hat{\omega}\Delta m$.) This pruning acts as an additional sink on the gravity wave action, one that is entirely an artiact of the numerical scheme. Ideally, only ray volumes that had already mostly broken or dissipated would be pruned. However, if $N_{\max}$ is too small, pruning can constitute a major source of numerical error.

2.2 MiMA

We implement MS-GWaM in MiMA, an intermediate-complexity atmospheric circulation model. We chose MiMA because it captures key atmospheric processes associated with gravity waves (midlatitude jets, polar vortices, the QBO) but sufficiently computationally efficient to permit robust statistical testing. We run MiMA at T42 spectral truncation (triangular truncation of total wavenumber 42), which corresponds to a spatial resolution of roughly 2.8 degrees.

The model includes interactive moisture with a simple Betts-Miller convection parameterization (Betts, 1986; Betts & Miller, 1986) as well as radiative transfer with the Rapid Radiative Transfer Model (Iacono et al., 2000; Mlawer et al., 1997). A key simplification is to neglect the impact of clouds on radiative transfer. MiMA is integrated coupled with a purely dynamic, or slab, ocean component. See Jucker and Gerber (2017) for a more detailed description of MiMA, and Garfinkel et al. (2020) for an account of the topography, land-sea contrast, and specified ocean heat transport used to produce an Earth-like climate. As detailed in Connelly and Gerber (2024), MiMA produces a large-scale circulation comparable to those of comprehensive climate models, allowing us to test GWPs in a realistic setting.

2.3 Wave source

The wave source is constrained to include only pure zonally- or meridionally-propagating wave packets; that is, each packet has $k \neq 0$ or $\ell \neq 0$, but not both. The horizontal phase speed is therefore either purely zonal or purely meridional. In the following discussion we focus on zonally-propagating waves, but the same source parameterization is used, *mutatis mutandis*, for waves with $k = 0$ and $\ell \neq 0$.

Following Alexander and Dunkerton (1999) and Garfinkel et al. (2022), we specify the launch spectrum by prescribing the source momentum flux that is Gaussian as a function of c_{p}. For zonally-propagating waves, we have

$$F(c_{\mathrm{p}}) = B \,\mathrm{sign}\left(c_{\mathrm{p}} - \bar{u}\right) \exp\left\{-\frac{1}{2}\left(\frac{c_{\mathrm{p}} - c_0}{\sigma}\right)^2\right\} \tag{8}$$

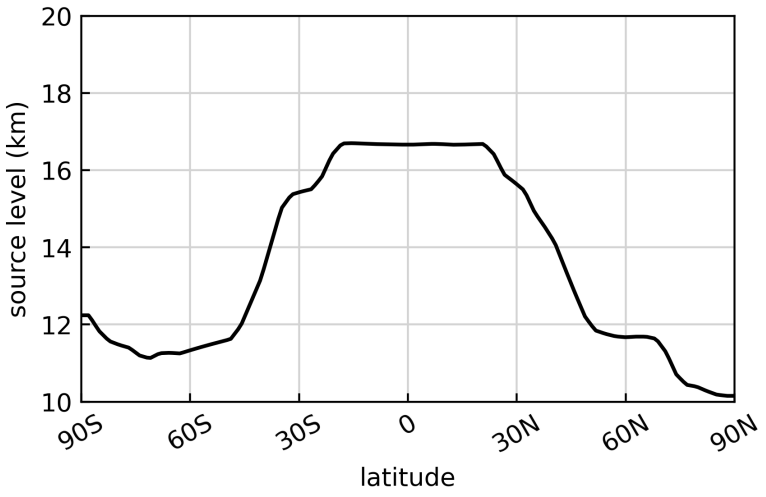

Figure 3 Ray volume source altitude as a function of latitude.

where $\sigma = 35\,\mathrm{m\,s^{-1}}$ and B is chosen so that

$$\sum_{c_\mathrm{p}} F(c_\mathrm{p}) = \begin{cases} 3\,\mathrm{mPa} & |\vartheta| < 20^\circ \\ 8\,\mathrm{mPa} & |\vartheta| > 30^\circ \end{cases} \tag{9}$$

with linear interpolation for intermediate latitudes. These values were chosen so that the polar night jet speeds and QBO period were realistic. We take

$$c_0 = \begin{cases} \bar{u} & |\vartheta| < 25^\circ \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where ϑ is the latitude. Thus waves with greatest flux move with the mean flow in the tropics, where gravity wave sources are predominantly convective, and have phase speed zero in the extratropics, where the sources are largely orographic.

Fully determining the spectral properties of each wave packet requires fixing one more degree of freedom. Note that the dispersion relation (2) constrains the gravity wave intrinsic period to fall between that of buoyancy oscillations (on the order of minutes) and that of inertial oscillations (one day). We therefore take the source intrinsic frequency to be

$$\hat{\omega} = \frac{2\pi}{10\,\mathrm{h}} \approx 1.75 \times 10^{-4}\,\mathrm{s^{-1}}$$

Then the horizontal wavenumber is given by $k = \hat{\omega}/\left(c_\mathrm{p} - \bar{u}\right)$, and the vertical wavenumber is recovered

from (2) as

$$m = -\sqrt{\frac{k^2\left(N^2 - \hat{\omega}^2\right)}{\hat{\omega}^2 - f^2}}$$

The negative root is taken to ensure upward propagation of ray volumes. These choices lead to horizontal wavelengths between 36 km and 1800 km, and vertical wavelengths between roughly 350 m and 18 km (depending on the latitude and the stratification).

The source altitude is set at the climatological minimum of zonal-mean temperature in the final year of a spinup integration, and ranges from roughly 10 km at the poles to just under 17 km at the equator (Figure 3). This shape is intended to approximate the meridional variation of thermodynamic tropopause height, and because such a launch level distribution has been well-studied in MiMA and is known to produce reasonable stratospheric variability (Connelly & Gerber, 2024; Garfinkel et al., 2020).

In all simulations in this work, the discretization of phase speed space covers the interval within $50\,\mathrm{m\,s^{-1}}$ of the center of the flux spectrum c_0. In the MiMA control integration, we subdivide that region into 20 ray volumes per component, which gives $\delta c_\mathrm{p} = 5\,\mathrm{m\,s^{-1}}$. We then take $\delta z = 1500\,\mathrm{m}$ and $N_\mathrm{max} = 2500$. These three parameter choices are quite similar to those used in ICON by Bölöni et al. (2021), though they have been adjusted slightly to achieve minimal pruning and a reasonable climate.

2.4 Control integration and test scenarios

We integrate MiMA with MS-GWaM for 25 years, starting from the same spun-up state used in Connelly and Gerber (2024). Figure 4 shows seasonal climatologies of large-scale wind and gravity wave momentum flux from this integration. The average speed of the polar night jet is $30.4 \pm 3.3\,\mathrm{m\,s^{-1}}$ in boreal winter and $56.3 \pm 1.7\,\mathrm{m\,s^{-1}}$ in austral winter. These values are similar to those from previous integrations using the Alexander and Dunkerton (1999) parameterization (Connelly & Gerber, 2024; Garfinkel et al., 2020). Figure 14a shows the QBO, a climate feature particularly sensitive to the GWP, from this control integration. The QBO period is 29.0 ± 2.5 months, in good agreement with the observed period of around 28 months and notably not phase-locked to the annual cycle.

We also evaluate the frequency of sudden stratospheric warmings (SSWs). This integration produced 7.7 ± 2.3 SSWs per decade; while this value exceeds the observed value of roughly 6 per decade, the fairly wide uncertainty of our estimate (due in part to the relatively short integration time) suggests that the frequency with which this configuration exhibits SSWs is plausible.

Using the control integration, we generate atmospheric column scenarios to use as calibration test cases for the ray tracer. We extract several month-long time series from the final year of the MiMA integration (Table 1). For each scenario, we save the large-scale variables needed by the ray tracer at 4-minute intervals:

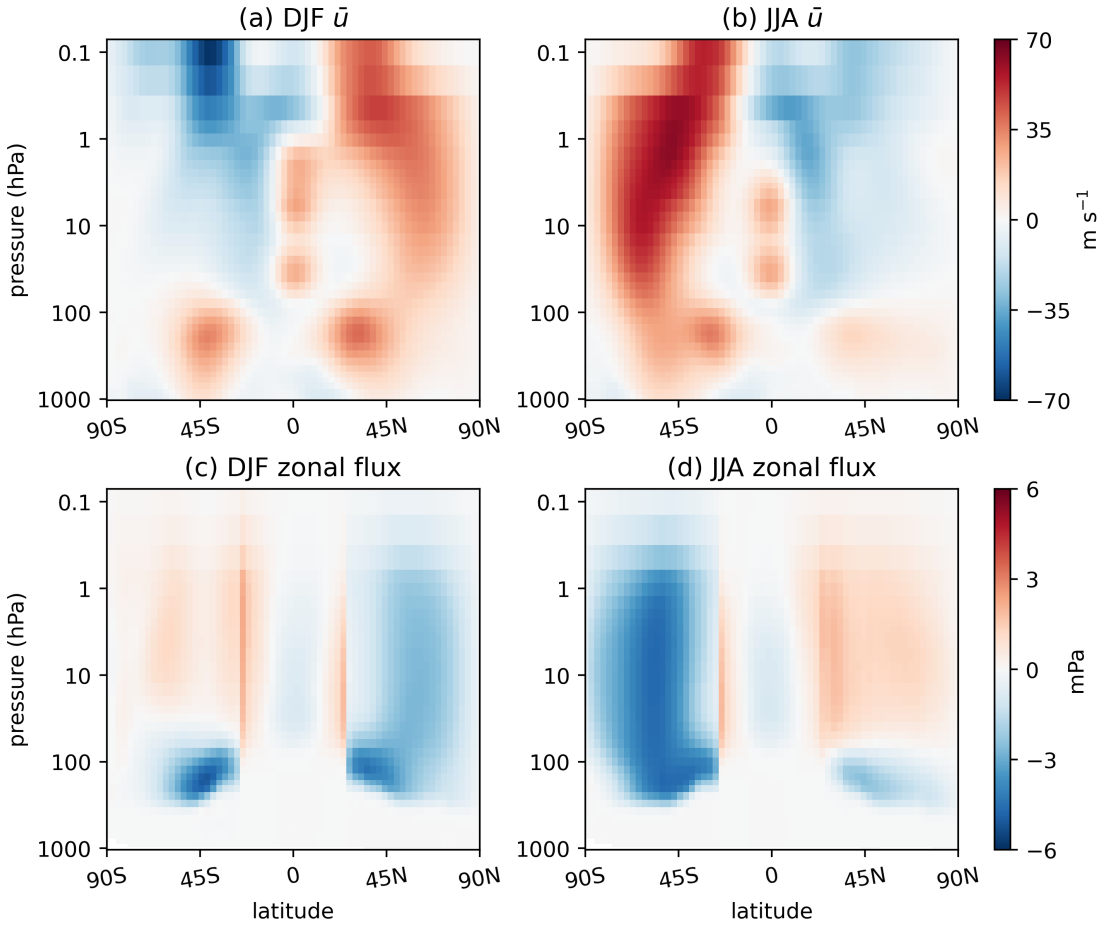

Figure 4 Climatologies of zonal-mean zonal wind (a-b) and zonal gravity wave momentum flux (c-d) in boreal winter (left) and austral winter (right), from the control MiMA integration with $N_{\max} = 2500$.

both horizontal components of the wind, the density ρ, and the buoyancy frequency N^2. The left columns of Figure 5 and Figure 6 show example zonal and meridional wind time series, respectively, from three scenarios. These examples exhibit characteristics seen in many others. In particular, the zonal wind is dominated by persistent (usually westerly) jets, while the meridional flow varies more substantially on daily time scales. As a result, the effects of transience will be more visible in the meridional component.

(We will henceforth refer to the MiMA data saved in these test scenarios interchangeably as "large-scale" or "mean state" variables, since in the parameterization context they represent the mean to which the waves are a perturbation. But note that these values are, from the MiMA perspective, the instantaneous values of the resolved state.)

Location	Coordinates	Month
Anchorage	61.1°N 150.2°W	Jan.
Copenhagen	55.6°N 12.5°E	Jan.
New York	40.7°N 74.0°W	Jan.
Lisbon	39.1°N 9.3°W	Jan.
Miami	25.4°N 80.1°W	Apr.
Maldives	2.0°N 73.7°E	Apr.
Singapore	1.3°N 103.9°E	Oct.
Brisbane	27.0°S 153.7°E	Oct.
Perth	32.0°S 115.9°E	Jul.
Buenos Aires	34.8°S 58.4°W	Jul.
Weddell Sea	69.8°S 42.9°W	Jul.
Amundsen Sea	70.5°S 110.0°W	Jul.

Table 1 Test scenarios extracted from the 25-year MiMA integration.

2.5 Stand-alone ray tracer

While the ultimate aim of this study is to calibrate the behavior of MS-GWaM in MiMA, we use a stand-alone model to rapidly evaluate and compare different configurations. This stand-alone implementation of MS-GWaM is the same as the MiMA version, except for a few minor practical considerations.

First, the stand-alone model uses a prescribed large-scale state, rather than an interactive one. That is, the large-scale wind, density, and buoyancy frequency are read in from the test scenario, rather than being passed as part of an atmospheric model time step. The gravity wave momentum fluxes and accelerations are diagnosed as in (7), but they are not used to drive the mean flow, which evolves purely according to the loaded time series. (A version where the large-scale winds do respond interactively to the waves is available, but it produces artificial results if the large-scale forcing is absent.)

Second, the large-scale variables in the stand-alone model are stored on a vertical grid that is equally spaced in height, rather than on the hybrid-pressure grid used by MiMA. We convert from the latter to the former via linear interpolation using the instantaneous height time series for each scenario. The resulting artifacts when comparing MiMA and stand-alone model fluxes are small, but nonetheless preclude bit-for-bit agreement. The stand-alone model vertical grid is also much finer than that of MiMA, so that we can assess the accuracy of different ray tracing configurations without concerning ourselves with the resolution of the mean state. The stand-alone model domain extends from 5 km to 60 km, and therefore always includes the full ghost layer.

2.6 High-resolution reference integrations

To evaluate the accuracy of subsequent configurations of the stand-alone model, we first integrate each test scenario at very high resolution, with $\delta z = 100\,\mathrm{m}$ and $\delta c_\mathrm{p} = 1\,\mathrm{m\,s^{-1}}$. Notably, we also take $N_\mathrm{max} = \infty$ in

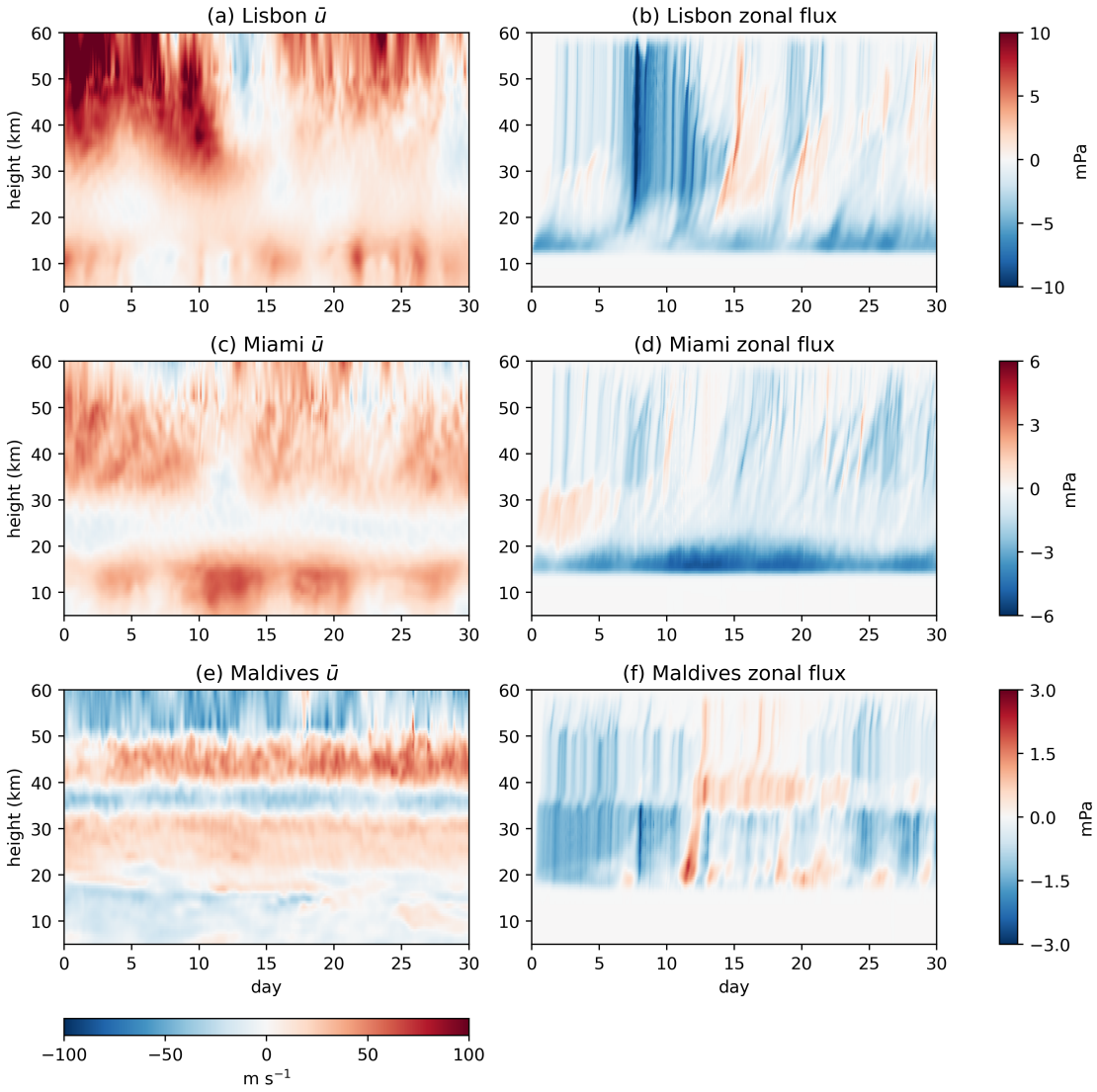

Figure 5 Time series of large-scale zonal wind (left column) and parameterized gravity wave zonal momentum flux (right column) from three of the test scenarios in Table 1. The momentum fluxes are those from the $N_{\max} = \infty$ integration in the single-column ray tracer. For display purposes, the fluxes shown are smoothed by a Gaussian filter with half-width 3 h in time and 1 km in the vertical. The momentum flux scales vary due in part to the different source amplitudes (9).

these integrations; that is, only those rays that have actually exited the domain or dissipated to numerical zero are pruned. (Practically, with fine resolution, this requires 200,000 – 300,000 ray volumes, depending on the scenario.)

This configuration would be far too compute- and memory-intensive to use in MiMA (or, in all likelihood, any other atmospheric model), but it allows us to approximate how the ray tracer would behave absent the practical considerations imposed by numerical computing. The right columns of Figure 5 and Figure 6 show the momentum flux time series from the reference integrations with the wind data on the left.

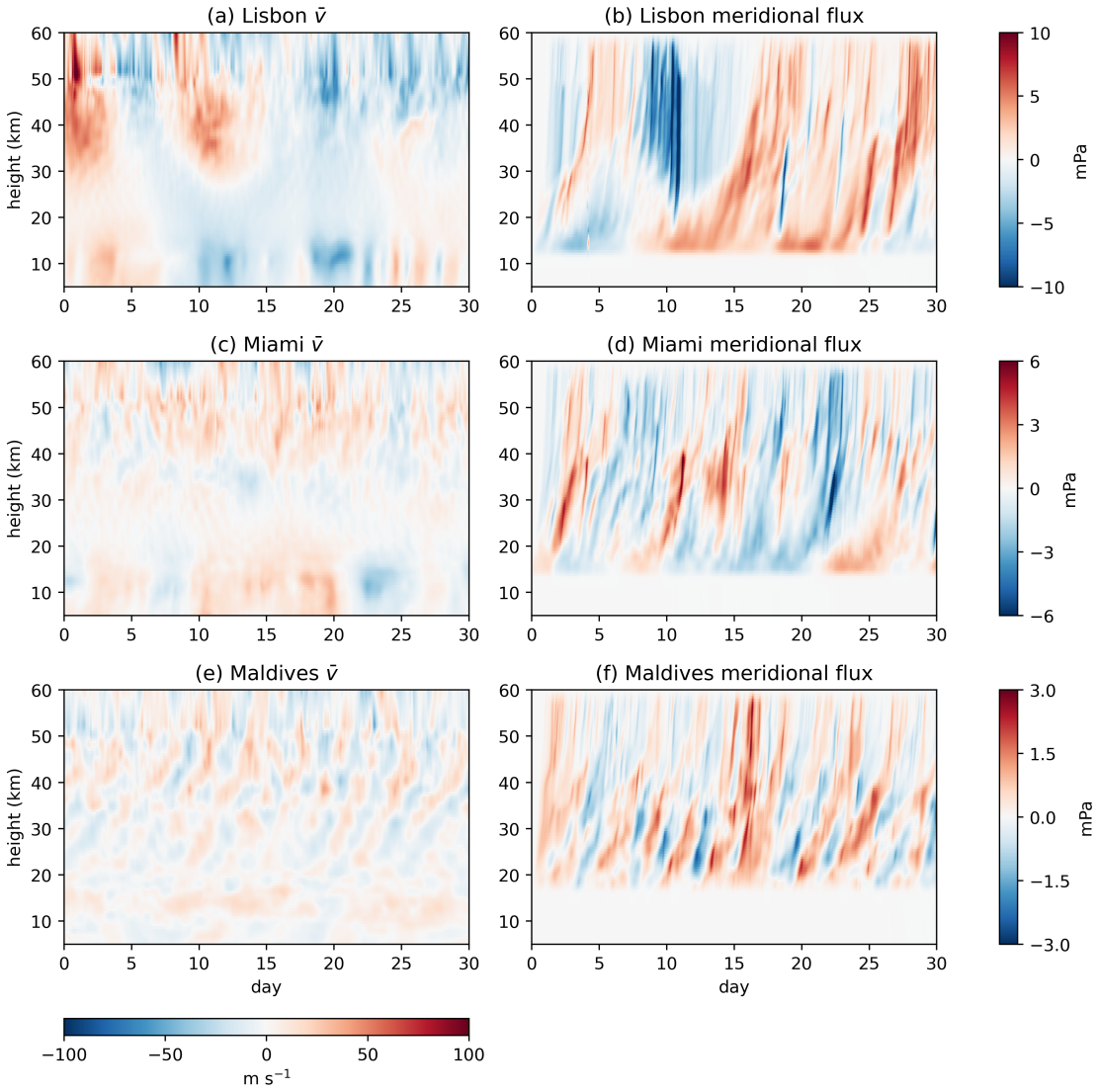

Figure 6 As in Figure 5, but for meridional winds and momentum fluxes.

3 Ray tracing on a budget

Treating the momentum fluxes from the extremely high-resolution integrations (Section 2.6) as ground truth, we can determine the accuracy of various MS-GWaM configurations with our stand-alone model. This section describes the two kinds of low-cost ($N_{\max} = 250$) configurations we evaluate in this manner: an optimal coarse resolution obtained via an extensive parameter search, and a stochastic modification that trades resolution in time for enhanced physical and spectral resolution. We then discuss the accuracy of integrations using these low-cost configurations. These tests are online in that the ray tracer state is persistent, so past errors can affect future time steps, but offline in that the mean wind is still prescribed.

3.1 An optimal coarse resolution

Dramatically reducing $N_{\max}$ without also incurring extreme pruning error requires a corresponding coarsening of the source resolution parameters δz and δc_{p}. However, it is not *a priori* obvious what the appropriate choices for those values are. We expect to need $N_{\max} \sim \left(\delta z \delta c_{\mathrm{p}}\right)^{-1}$, since the launch rate of new ray volumes is inversely proportional to both resolution parameters. Moreover, we might expect the relationship between these parameters to go as

$$\frac{\delta c_{\mathrm{p}}}{\delta z} \sim \Lambda \tag{11}$$

where Λ is a characteristic scale of the wind shear $\partial \bar{u}/\partial z$. The scaling relationship (11) follows by noting that the phase speed of wave packets filtered at each level (i.e. those hitting critical levels) varies with $\partial \bar{u}/\partial z$. However, the precise relationships between these parameters remain opaque.

For each test scenario (Table 1), we integrate the stand-alone model with $N_{\max} = 250$ over a grid of possible source resolutions. Figure 7 shows the result of the grid search for the Maldives test case. As it is the net forcing on synoptic (and larger) that matters for the evolution of the large-scale flow, we apply a Gaussian filter with with half-width one day to the error time series before averaging. Notably, we find local error minima in the interior of the search space. Finer resolutions (bottom, left) suffer from pruning artifacts at this value of $N_{\max}$, while coarser resolutions (top, right) are simply too poorly-resolved to be accurate. We observe similar patterns across all MiMA test cases (not shown), though the precise locations of the error minima vary.

With these grid search results in hand for each test scenario, we select a single optimal coarse resolution by averaging the normalized error for each coarse resolution (e.g. Figure 7b and Figure 7d) across all test scenarios. The zonal and meridional error minima are similar (Figure 8); since the climate features dependent on the GWP are largely zonal, we set the optimal coarse resolution to $\delta z = 2500\,\mathrm{m}$ and $\delta c_{\mathrm{p}} = 12.5\,\mathrm{m\,s^{-1}}$. By (2.6), this suggests a wind shear scale $\Lambda \sim 5\,\mathrm{m\,s^{-1}\,km^{-1}}$, which is reasonable given the wind time series in the test cases.

Equal-flux discretizations In the coarse grid searches shown above, ray volumes are not instantiated with a uniform spectral width δc_{p}. Rather, we discretize the source into the same number of intervals, but with widths chosen so that each ray volume has equal flux according to the Gaussian spectrum (Figure 9). Thus reported Δc_{p} values in this context refer to *average* spectral widths of ray volumes at the source.

Preliminary tests found this approach to achieve higher accuracy and couple more stably with MiMA, likely because it allows volumes from the tails of the spectrum more time to propagate before being pruned. The difference is minimal for fine discretizations, but becomes noticeable as the number of source phase

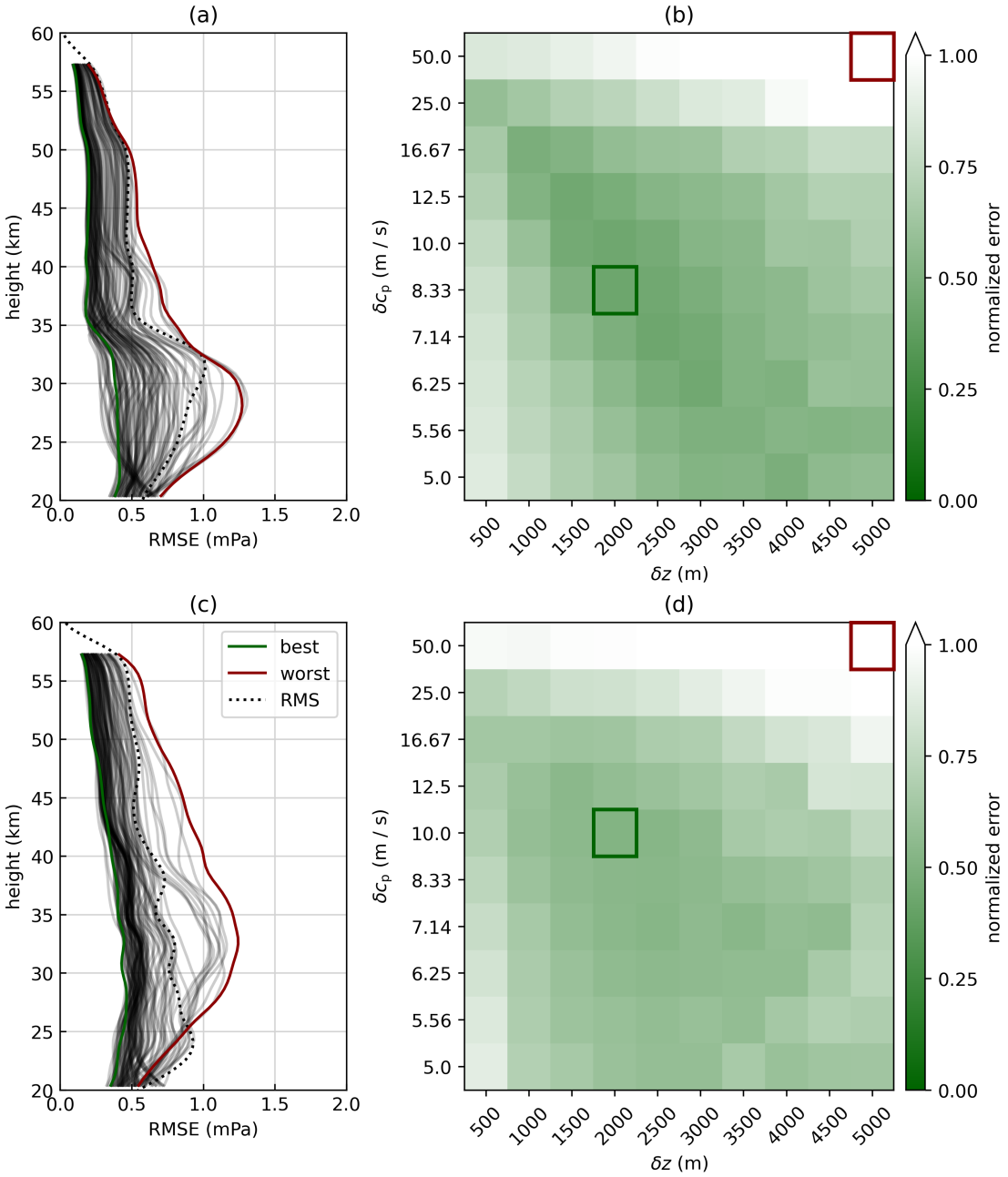

Figure 7 Results from the grid search for the optimal source resolution at $N_{\max} = 250$ for the Maldives test case. Profiles in (a) are the zonal momentum flux RMS (in time) errors for various source resolutions after application of a Gaussian filter with half-width one day. Normalized errors are computed by dividing the error by the RMS flux in the high resolution (dotted line) at each level, and then averaging over the column. Each cell in (b) corresponds to a profile in (a), with the best and worst configurations identified. Panels (c) and (d) are the same for the meridional flux. Levels in the sponge layer are excluded, as are those below 20 km, to allow fair comparison between scenarios with different source altitudes (Figure 3).

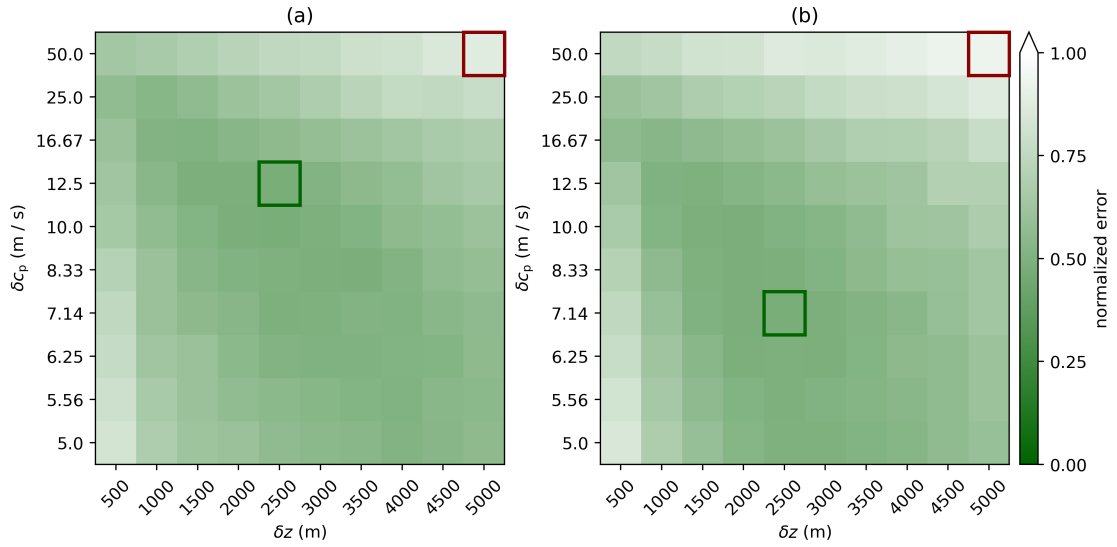

Figure 8 As in panels (b) and (d) of Figure 7, but with normalized errors averaged over all the test scenarios in Table 1.

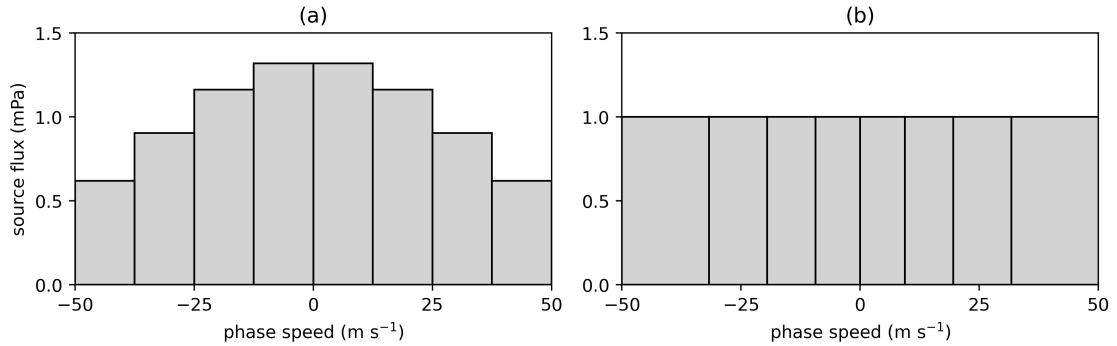

Figure 9 Equal-phase speed (a) and equal-flux (b) discretizations of one component of the source spectrum, corresponding to $\delta c_\mathrm{p} = 12.5\,\mathrm{m\,s^{-1}}$. The total source flux in both cases is 8 mPa.

speeds is lowered.

Alternative pruning strategies Thus far, when a new ray volume is to be instantiated but there are already N_max volumes propagating, we have pruned the one with the lowest energy. This choice was made to match Bölöni et al. (2021). However, we also use the stand-alone model to explore whether other pruning criteria might produce less noticeable artifacts.

First, we try pruning by the momentum flux itself; since we are evaluating configurations based on errors in their momentum flux time series, it would be reasonable to prune those ray volumes with the smallest immediate contributions to those fluxes. Second, we prune by the absolute value of the vertical group velocity $|c_\mathrm{g}|$. Ray volumes with very small group velocities are more likely to have reached critical levels at which they will dissipate away, and so pruning them could be less likely to cause non-local errors

later in the integration.

We find that pruning by flux is the most accurate strategy, although the difference is extremely minimal. Therefore in the following sections we show results for the optimal coarse resolution with pruning by momentum flux.

3.2 A stochastic configuration

We also test a stochastic configuration that achieves finer ray volume resolution, but sacrifices the constant-flux boundary condition. We are motivated by the fact that it is the *net* forcing on synoptic time scales that matters for the large-scale flow. Instead of instantiating ray volumes so that they enter the domain in a continuous stream (Figure 2), ray volume launches follow a Poisson process. Each time step, the number of ray volumes n_j launched in the j^{th} channel of the source discretization is sampled according to

$$\mathbb{P}(n_j = n) = \frac{\lambda^n e^{-\lambda}}{n!} \quad \text{where} \quad \lambda = \epsilon \frac{\left(c_{\text{g}}\right)_j \Delta t}{\delta z} \tag{12}$$

and Δt is the time step of the ODE solver. The ratio $c_{\text{g}}\Delta t/\delta z$ is the distance the ray volume travels in one time step divided by the distance it would have to cover to fully enter the domain; thus if $\epsilon = 1$, rays launch at the same average rate as in the constant-flux case.

The crux of the stochastic setup, however, is to take $\epsilon < 1$, so that rays launch less frequently, and then launch rays with initial physical and spectral resolutions given by $\epsilon^{1/2}$ multiplied by the optimal coarse resolution values. We then scale each ray volume's contribution to the momentum flux by ϵ^{-1} when performing the projection (7). With this approach, each ray volume stands in for ϵ^{-1} volumes in the corresponding deterministic configuration. Note that this is the opposite of the approach to intermittency used by Alexander and Dunkerton (1999), who sought to approximate an intermittent source with a continuous flux of instantaneously propagating packets.

We tested values $\epsilon^{-1} \in \{25, 64, 100\}$ and found the performance of the stochastic scheme to be insensitive to variation of ϵ within this range. In what follows, we show results for $\epsilon = 64^{-1}$, so that the stochastically-launched ray volumes are eight times finer in each dimension than those of the coarse configuration. While the instantaneous forcings are more abrupt, the hope is that the effective forcing on synoptic scales will be better resolved.

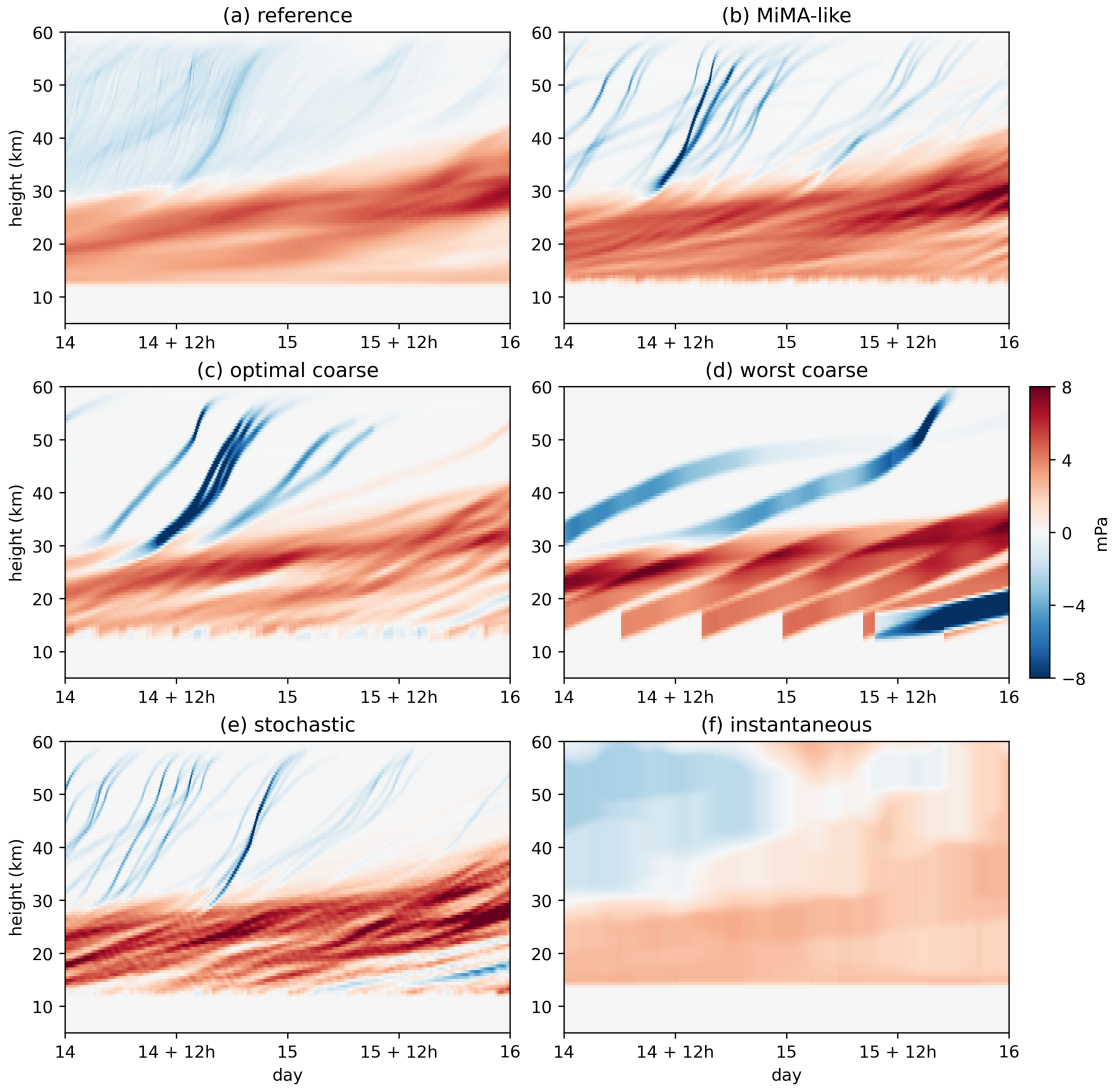

Figure 10 Two-day excerpts of meridional momentum flux time series from stand-alone model integrations in the Lisbon test scenario. The worst coarse configuration, found during grid search, has $\delta z = 5000\,\mathrm{m}$ and $\delta c_\mathrm{p} = 50\,\mathrm{m\,s^{-1}}$.

3.3 A case study

Figure 10 shows part of the meridional flux time series from stand-alone model integrations in the Lisbon scenario using the optimal coarse resolution, the worst coarse resolution found during grid search, and the stochastic configuration. For comparison, we also show the high-resolution reference integration and a "MiMA-like" configuration with $N_\mathrm{max} = 2500$ and resolution matching that of the control integration (Section 2.4). Finally, we include an integration with the instantaneous (not ray tracing) scheme given in Bölöni et al. (2021), similar to most conventional GWPs. The instantaneous scheme is run with a fine phase speed resolution $\delta c_\mathrm{p} = 1\,\mathrm{m\,s^{-1}}$ (and has no vertical resolution to speak of).

The time slice in Figure 10 features a weak northerly jet in the lower half of the domain (below roughly

30 km), the upper edge of which increases in altitude between days 14 and 16 (Figure 6a). Most waves with $\ell < 0$ are therefore filtered, such that the momentum flux in the lower domain is mostly positive. However, those packets with sufficiently negative intrinsic phase velocity can propagate to the upper domain, where the winds are southerly, and so there is negative flux in the upper levels.

The worst coarse resolution suffers from obvious discretization artifacts. The fluxes in both regimes are extremely streaky, because the ray volumes are quite tall at the source, and so the temporal gap between launches is large. Ray volumes with negative phase speed escape into the upper domain (above around 30 km) too infrequently, such that what momentum does break through is overconcentrated in a handful of packets. The optimal coarse resolution fares better: although it exhibits the same kinds of biases, they are much less pronounced. The optimal coarse resolution represents the maximum extent to which the problem of overconcentration can be mitigated at $N_{\max} = 250$ before pruning becomes excessive.

The MiMA-like and stochastic integrations are both able to export flux to the upper half of the domain more frequently and in smaller individual packets, but neither achieves the near-continuous flux field observed in the reference simulation. The stochastic configuration in particular occasionally shows the signal of individual high-amplitude ray volumes, since ray volume amplitudes are scaled by ϵ^{-1} to compensate for the reduced launch rate.

While the instantaneous scheme captures the broadest features of the reference flux time series, it appears much worse than any of the ray tracing configurations (except perhaps the worst coarse one). The lower-domain negative flux is too weak, and too much of it escapes into the upper levels, where it erroneously flips the sign of the total flux for about half the time slice shown.

3.4 RMS error profiles

For a more quantitative description of the accuracy of each configuration, we compute root-mean-square error profiles over all test scenarios, again with Gaussian filtering to highlight synoptic-scale effects (Figure 11). Remarkably, the two $N_{\max} = 250$ configurations perform nearly as well as the $N_{\max} = 2500$ control-like configuration at all levels. The stochastic approach provides a slight but consistent improvement over the optimal coarse resolution. Between roughly 20 km and 35 km, the stochastic scheme is slightly *more* accurate than the control-like configuration in the meridional flux (Figure 11b). The stochastic source resolution is finer than that of the MiMA-like setup, because of the scaling by $\epsilon^{-1/2} = 8$ in each dimension, and evidently this proves useful for capturing precise wave dynamics under certain conditions.

The MiMA-like configuration itself achieves reasonably small absolute errors, on the order of 1 mPa. However, the relative error averaged over all test scenarios is above 0.2 in most of the domain, which is

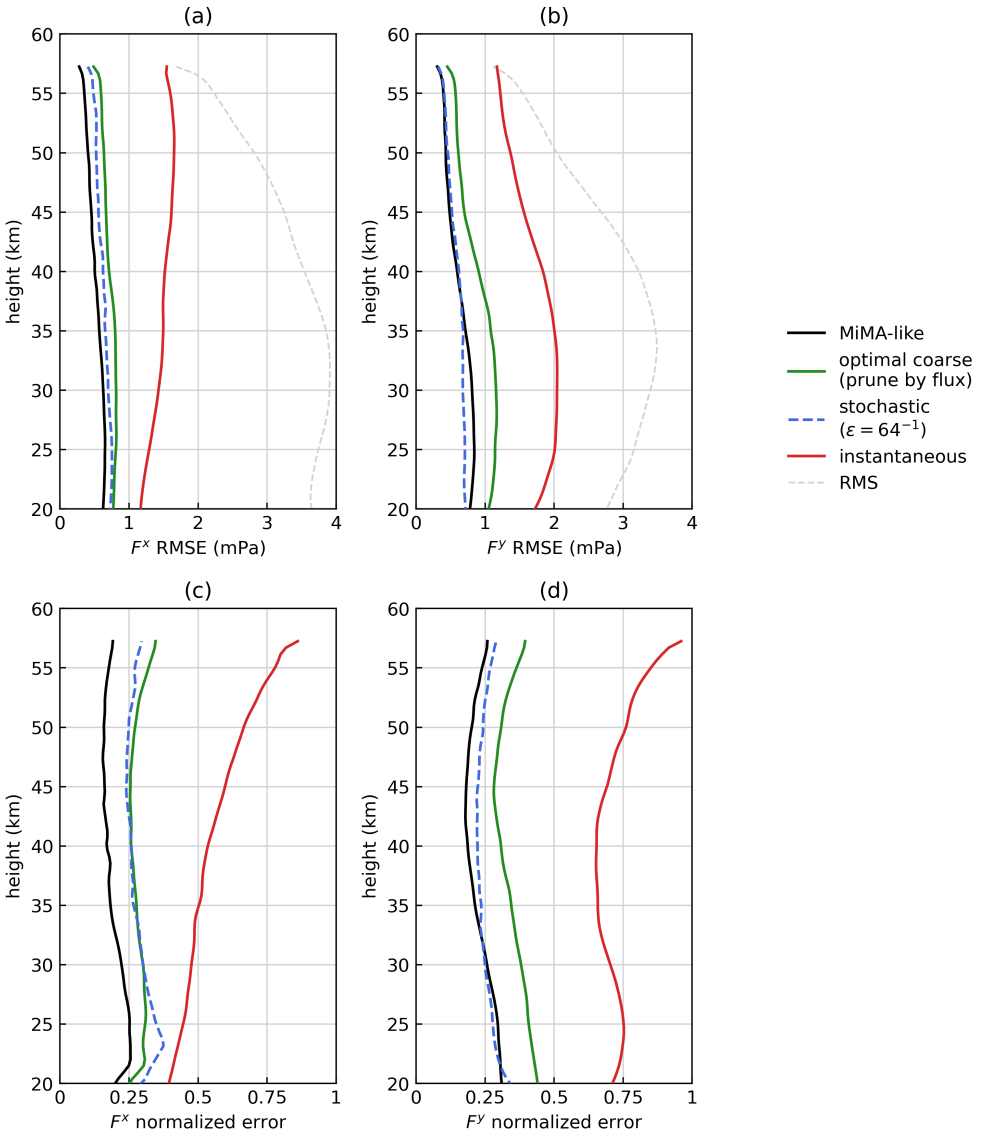

Figure 11 RMS (in time) errors in zonal (a, c) and meridional (b, d) momentum flux for various stand-alone model integrations, averaged over all test scenarios. Panels (c) and (d) show profiles where the RMS errors are normalized by the reference RMS flux profiles in each each scenario before averaging.

small but non-negligible. We note, though, that the source resolution of the MiMA-like configuration was chosen simply to be comparable to that used in ICON by Bölöni et al. (2021). The optimal coarse resolution is not much coarser in the vertical dimension than the value used in MiMA. It is therefore possible that greater accuracy could be obtained at $N_{\max} = 2500$ by more carefully selecting the source resolution with a parameter sweep, as was done in Section 3.1 for $N_{\max} = 250$.

Despite the smoothing, the instantaneous scheme is much less accurate than all the ray tracing schemes throughout the domain, which is unsurprising, since the ground-truth integration itself used a high-resolution ray-tracing scheme. The gap between the instantaneous and ray tracing configurations widens in

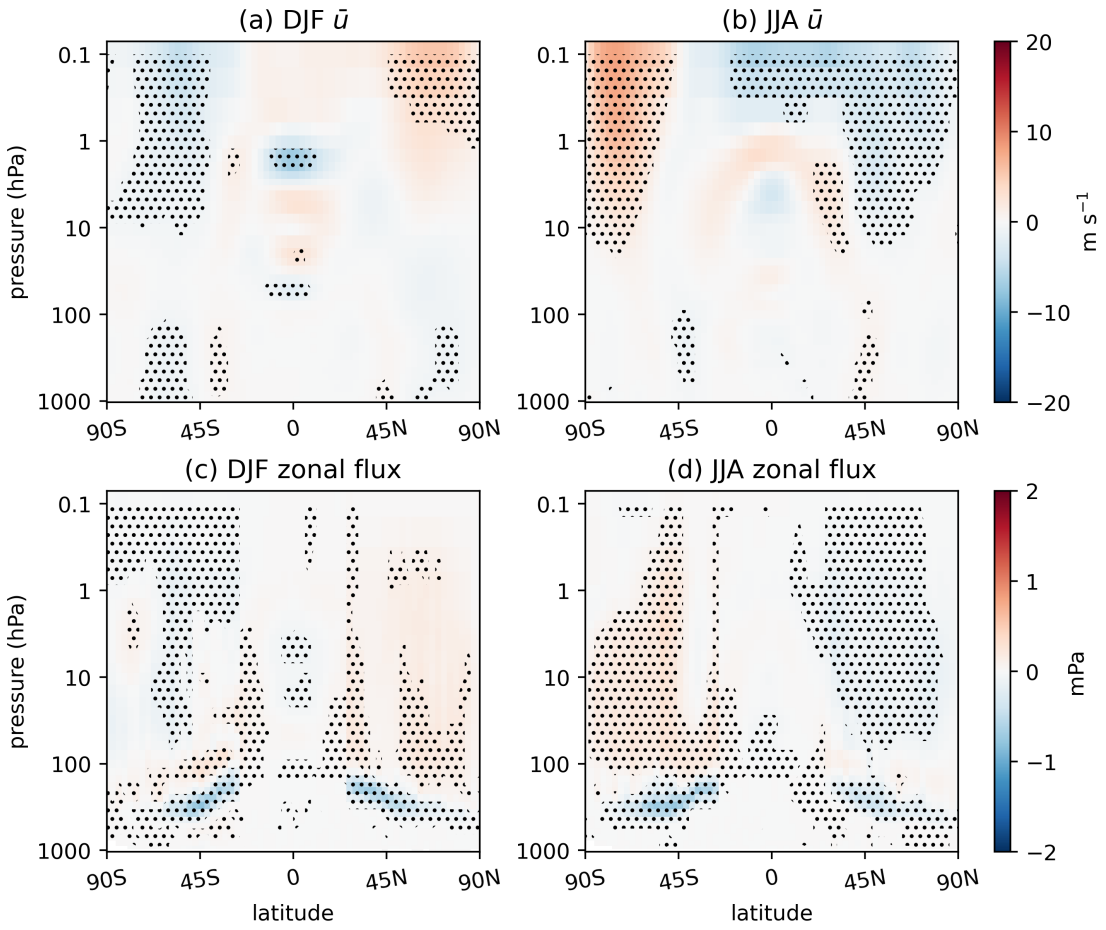

Figure 12 Biases in the coupled MiMA integration using the optimal coarse resolution of MS-GWaM relative to the control integration. The top row shows zonal-mean zonal wind, the bottom zonal-mean gravity wave flux of zonal momentum. Stippling indicates biases significant at the 95% level, treating each year as an independent sample. Note the changes in the color scale relative to Figure 4.

the upper atmosphere, especially in relative terms (Figure 11c and Figure 11d). The instantaneous scheme assumes propagation speeds are infinite, but waves reaching the upper atmosphere require travel times on the order of hours to days, so fluxes aloft are simulated at the wrong times (and sometimes, as Figure 10f shows, with the wrong sign altogether).

4 Fully-coupled evaluation in MiMA

We integrate MiMA with the same setup as the control run (Section 2.4), except that we use the two low-cost configurations of MS-GWaM (optimal coarse and stochastic) developed in Section 3. We focus on measures of the large-scale circulation most directly affected by gravity wave momentum transport: the climatological westerly jets of in the troposphere and stratosphere, SSWs, and the QBO.

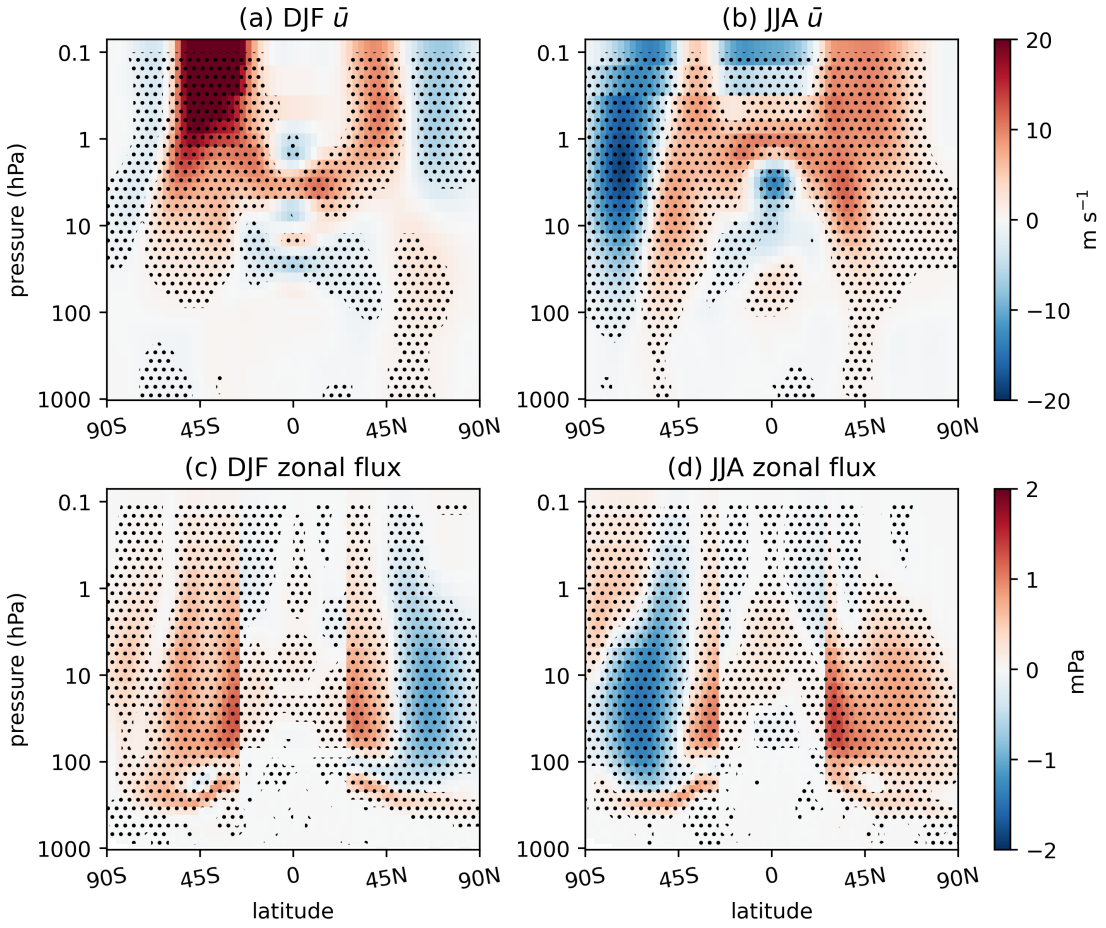

Figure 13 As in Figure 12, but for the MiMA integration using the stochastic configuration of MS-GWaM with $\epsilon^{-1} = 64$.

4.1 Climatological jets

Figure 12 shows biases in the zonal-mean zonal wind and momentum flux from the coupled integration using the optimal coarse resolution. The biases are quite small, with the exception of the southern hemisphere upper stratosphere in austral winter. Overall, there is a slight strengthening of the polar stratospheric vortex in the winter hemisphere during both solstice seasons, paired with a slight strengthening of the summer easterlies. Critically, however, the tropospheric jets remain effectively unchanged, agreeing with the control integration up to sampling uncertainty, as shown in Table 2. We attribute this to the fact that the biases in the optimal coarse integration appear mainly above 10 hPa.

As for the parameterized momentum flux, the largest bias appears around the tropopause, which, particularly in boreal winter, exhibits dipole anomalies, enhanced westerly flux atop enhanced easterly flux. This is likely the result of coarsening the ray volumes in the vertical, which spreads out the momentum deposited by rays reaching critical levels at the top of the subtropical jet. Fortunately, this does not appear

Configuration	Polar night jet speed $(\mathrm{m\,s^{-1}})$		QBO period (months)	SSW frequency ($\mathrm{dec^{-1}}$)
	DJF	JJA		
control	30.4 ± 3.3	56.3 ± 1.7	29.0 ± 2.5	7.7 ± 2.3
optimal coarse	30.3 ± 3.2	57.1 ± 1.5	27.8 ± 2.9	8.5 ± 3.1
stochastic	32.0 ± 4.4	53.4 ± 1.6	22.8 ± 1.2	7.3 ± 4.7

Table 2 Statistics from the control integration and integrations with the low-cost configurations. polar night jet speeds are calculated as the average zonal wind velocity at 60° and 10 hPa in the appropriate hemisphere. QBO periods are calculated by measuring the difference between zero crossings in the 20 hPa wind, averaged from 10°S to 10°N. SSW frequencies are calculated as described in Butler et al. (2017).

to cause any such change in the winds, potentially because momentum transport is dominated by Rossby waves at these altitudes.

Figure 13 shows the same biases for the integration with the stochastic configuration. Surprisingly, given the superior performance of the stochastic configuration in the stand-alone model, there is a notable degradation of the simulation relative to the control. Biases in the polar night jets extend deep into the stratosphere, and the austral winter polar night jet is shifted towards the equator. As quantified in Table 2, the vortex weakens in both winter hemispheres, although it remains within sampling uncertainty in the north. Associated with these stratospheric biases are more substantial changes in the tropospheric jets.

These biases in the circulation are caused by changes in the parameterized momentum flux. It is not trivial to see the direct effect of the change in Figure 13c and Figure 13d, however, as the circulation feeds back on the momentum flux in the coupled runs. Dipolar anomalies appear around the tropopause, but now of opposite sign compared to the coarse integration. This is consistent with the slight increase in low level winds above the subtropical jets, which shifts critical levels higher. In the stratosphere, the biases in both the polar night jets and the summer hemisphere easterlies alter wave propagation, leading to positive anomalies in the summer hemisphere and overall negative anomalies in the winter hemisphere.

We speculate that the larger biases in the MiMA integration with the stochastic source are due to nonlinear dynamics. On average, the stochastic configuration launches the same momentum flux, but it does so sporadically and in large packets that evidently are not sufficiently smoothed out in the mean. Cohen et al. (2013) showed that strong gravity wave forcings are more likely to be compensated for by the resolved circulation. It may be that $\epsilon = 64^{-1}$, despite being the best choice in the stand-alone model, pushes the coupled model too hard. The coarse configuration smooths out fine-scale features of the flux (e.g. Figure 10c). Figure 12 suggests that this lower-variance approximation of the reference flux, despite being slightly less accurate when the winds are prescribed, is useful when they are interactive.

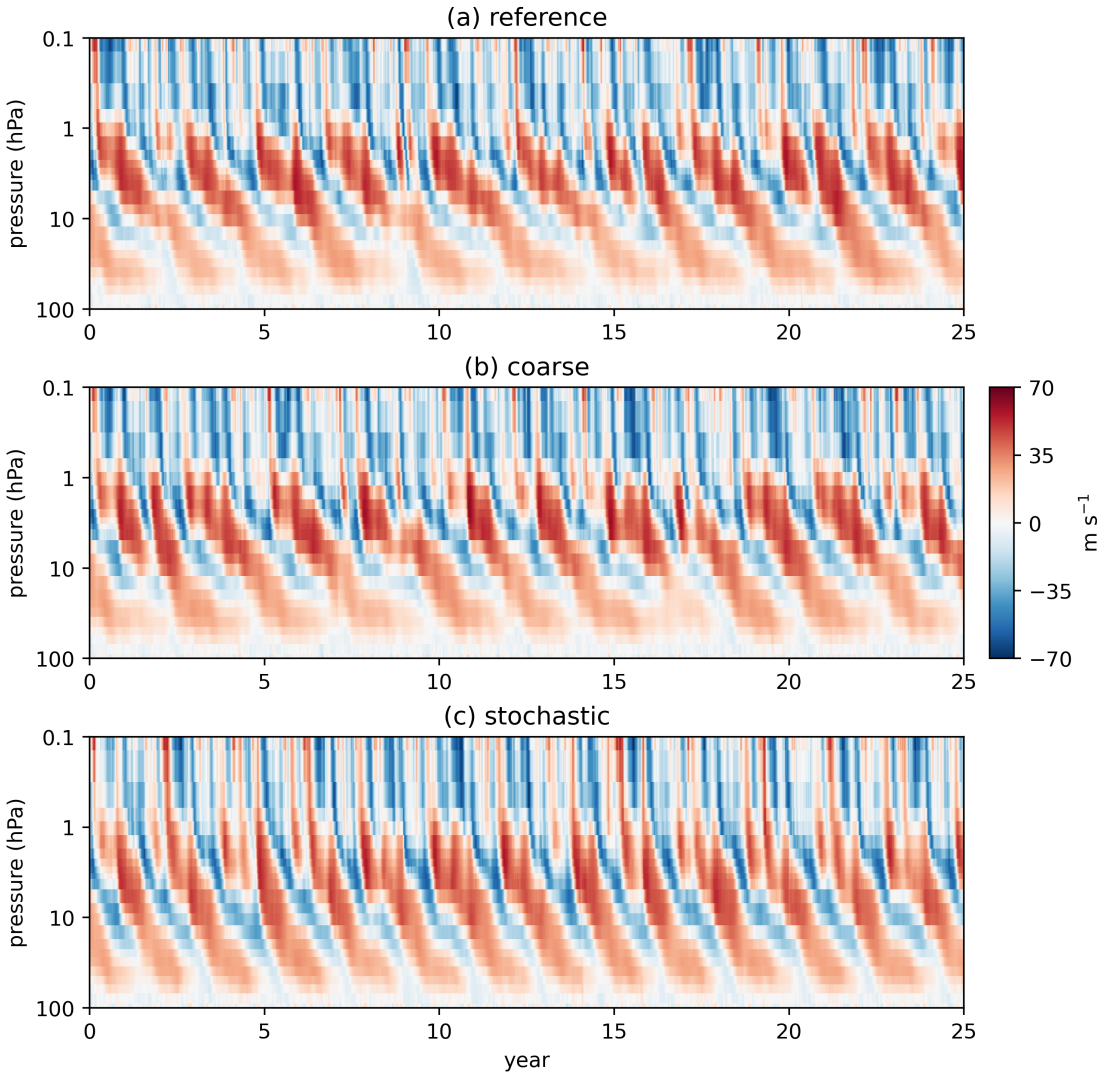

Figure 14 Zonal wind averaged over all grid cells between 10°S and 10°N from the control integration and integrations with the two low-cost configurations of MS-GWaM.

4.2 Stratospheric variability

Figure 14 shows that both the coarse and stochastic configurations of MS-GWaM produce a QBO when coupled to MiMA. The period of the QBO in the optimal coarse integration agrees well with that of the control integration, but the stochastic source configuration drives a QBO that oscillates considerably faster (Table 2). The stochastic configuration produces SSWs with frequency closer to that of the control run than does the optimal coarse setup (Table 2). However, there is considerable overlap between the 95% confidence intervals of all three integrations; 25 years, the limit to which we could integrate the control run, is not long enough to resolve differences in SSW rates of this magnitude. Curiously, the integration with the stochastic scheme exhibited higher interannual variability in the polar vortex: some winters have several SSWs while

Configuration	$N_{\max}$	Time (min)	Speedup
control	2500	52.0	—
optimal coarse	250	8.5	6.1
stochastic	250	10.6	4.9
AD99	—	10.9	4.8

Table 3 Wall time dedicated to MS-GWaM computation per year of model integration. Reported values are averaged over the 25-year integration period. The last column shows the speedup factor for each of the $N_{\max} = 250$ configurations with respect to the $N_{\max} = 2500$ run. Also included is the instantaneous scheme of Alexander and Dunkerton (1999), the other available GWP in MiMA.

many others have none at all (not shown). Given the worse performance of the stochastic configuration in terms of climatological biases and QBO period, as well as the less direct link between the GWP and the SSW frequency, we deem this result consistent with the general finding that the stochastic scheme works poorly online.

It is unclear what causes the stochastic source to drive a faster QBO. Because the number of launches per time step per cell is Poisson-distributed, so too is the cumulative momentum density added over the entire tropics, with a variance that increases linearly in time. (This could explain the heightened interannual variability in the polar vortex.) Lott et al. (2012) found that a stochastic source led to gravity wave momentum flux deposition at lower levels than in a comparable deterministic setup; such an effect would tend to speed the increase of zonal wind anomalies associated with the QBO.

4.3 Computational cost

Table 3 shows the wall time used by MS-GWaM computations in MiMA for each of the three integrations discussed in this chapter. The coarse and stochastic low-cost configurations are roughly six and five times faster than the control integration, respectively. (The slightly increased run time of the stochastic configuration appears to be due to the cost of random number generation.) Both low-cost configurations are slightly faster than the instantaneous GWP of Alexander and Dunkerton (1999) used in previous studies with MiMA (e.g. Connelly, 2023; Garfinkel et al., 2022). Note, however, that those studies called AD99 intermittently; the timings shown in Table 3 are for AD99 invokved each time step, for a fair comparison with MS-GWaM.

Although $N_{\max}$ was lowered from 2500 to 250, the low-cost configurations do not quite achieve an order-of-magnitude speedup. As previously mentioned (Section 3.4), the resolution used in the $N_{\max} = 2500$ integration, selected following Bölöni et al. (2021), is likely not as finely resolved as it could be, since it is only slightly finer in the vertical than the optimal coarse resolution. As a result, we find that the control integration often has fewer than 2500 ray volumes active in a given column, especially in the extratropics,

where persistent jets may efficiently filter whole sections of the source spectrum for extended periods. The runtime is largely determined by the number of *active* rays. As the $N_{\max} = 250$ configurations are nearly always using all available rays, we therefore do not observe a full factor-of-10 speedup.

We took $N_{\max} = 2500$ for the sake of comparison with the ICON configuration of MS-GWaM in Bölöni et al. (2021). However, those integrations set the model top at around 150 km, whereas the model top in MiMA is only around 60 km. As such, ray volume lifetimes are likely shorter in MiMA, such that $N_{\max} = 2500$ is likely more resources than needed for the control resolution. This effect, too, would be compounded towards the poles, where the top of the atmosphere is lower in altitude. We expect that the speedup would be closer to linear if calculated against a control integration forced to use the full 2500 ray volumes per column, either by way of a finer source resolution or a higher model top.

5 Discussion

In this manuscript, we have shown that with a judicious choice of parameters, GWPs based on ray tracing can be made competitive with existing steady-state schemes. We found that MS-GWaM configurations with $N_{\max} = 250$ can compute gravity wave momentum fluxes of comparable accuracy to those produced by much more expensive $N_{\max} = 2500$ configurations (Figure 11). While the $N_{\max} = 2500$ setup is slightly more accurate than the low-cost configurations in tests with prescribed winds, the budget configurations still perform far better than the instantaneous (non-ray tracing) scheme, which is fundamentally unable to capture transient effects or the wave packets' responses to changing mean winds. We consider this evidence that the increased expressiveness of the ray tracing framework can be obtained at much less computational cost than previously achieved.

The difference between the ray tracing configurations and the instantaneous scheme becomes more apparent in the upper atmosphere, because the assumption of instantaneous propagation is more dubious the higher the waves go. Notably, the accuracy of the low-cost ray tracing configurations suffers at worst marginally in the upper atmosphere, suggesting that even low-cost ray tracing is not subject to the same intrinsic limitations as the instantaneous scheme.

However, decreasing $N_{\max}$ without degrading model accuracy required careful selection of the source resolution (Figure 7 and Figure 8). The stand-alone setting furnished the high-resolution ground truth simulation, obtained by integration with $N_{\max}$ functionally infinite, as well as the means to evaluate a wide range of parameter settings across several hundred integrations and suite of test climate regimes. These tasks, especially the former, would almost certainly be too expensive to accomplish in an operational model. A poor choice of coarse resolution can yield a configuration hardly more accurate than predicting zero flux

everywhere (e.g. the high-error profiles in Figure 7a and Figure 7c). Similarly, maintaining high resolution in physical-spectral space while reducing $N_{\max}$ can lead to poor performance due to excessive pruning.

The online behavior of MiMA when driven by various MS-GWaM configurations remains harder to predict. Extratropical climatological features from integrations with $N_{\max} = 250$ matched those from the control integration well. While both low-cost configurations exhibited a QBO, the QBO driven by the stochastic scheme cycled too fast, suggesting systematic excess momentum deposition in the tropial stratosphere. This was surprising, given the stochastic setup's superior performance in the stand-alone model. The effect of stochastic source on the QBO has been studied in simplified models (e.g. Ewetola and Esler, 2024), but future work to better understand the behavior of this scheme in the stand-alone model and in MiMA would be valuable, especially because the wave sources that drive the QBO in reality are likely to be convective, and thus intermittent (Kim et al., 2021).

Refinements of the approaches here might include allowing the source discretization to vary based on latitude or some other indicator of climate regime. Additionally, Figure 11 suggests that the gap between the ray tracing and steady-state schemes is larger in the meridional component of the flux. This is likely because the meridional wind features short-term variability more often than the zonal wind (Figure 5 and Figure 6), and so the ability of ray tracing to capture gravity wave responses to fast changes in the mean flow is more valuable. As a result, it could be useful to set the zonal and meridional source resolutions independently.

Acknowledgements

D.C., E.G. and U.A. acknowledge support for this work through Schmidt Sciences ("DataWave" project). E.G. also acknowledges support form the US NSF through award OAC-2004572. U.A. thanks the German Research Foundation (DFG) for partial support through CRC 301 "TPChange" (Project No. 428312742 and Projects B06 "Impact of small-scale dynamics on UTLS transport and mixing", B07 "Impact of cirrus clouds on tropopause structure", and Z03 "Joint model development and modelling synthesis"). U.A. and G.S.V. thank DFG for support through the CRC 181 "Energy transfers in Atmosphere and Ocean" (Project No. 274762653 and Projects W01 "Gravity-wave parameterization for the atmosphere" and S02 "Improved Parameterizations and Numerics in Climate Models").

References

Achatz, U. (2022). Atmospheric Dynamics. Springer Spektrum, Berlin, Heidelberg.

Alexander, M. J., & Dunkerton, T. J. (1999). A Spectral Parameterization of Mean-Flow Forcing due to Breaking Gravity Waves. *Journal of the Atmospheric Sciences*, *56*, 4167–4182.

Amemiya, A., & Sato, K. (2016). A New Gravity Wave Parameterization Including Three-Dimensional Propagation. *Journal of the Meteorological Society of Japan*, *94*, 237–256.

Betts, A. K. (1986). A new convective adjustment scheme. Part I: Observational and theoretical basis. *Quarterly Journal of the Royal Meteorological Society*, *112*, 677–691.

Betts, A. K., & Miller, M. J. (1986). A new convective adjustment scheme. Part II: Single column tests using GATE wave, BOMEX, ATEX and arctic air-mass data sets. *Quarterly Journal of the Royal Meteorological Society*, *112*, 693–709.

Bölöni, G., Kim, Y.-H., Borchert, S., & Achatz, U. (2021). Toward Transient Subgrid-Scale Gravity Wave Representation in Atmospheric Models. Part I: Propagation Model Including Nondissipative Wave–Mean-Flow Interactions. *Journal of the Atmospheric Sciences*, *78*, 1317–1338.

Borchert, S., Zhou, G., Baldauf, M., Schmidt, H., Zängl, G., & Reinert, D. (2019). The upper-atmosphere extension of the ICON general circulation model (version: ua-icon-1.0). *Geoscientific Model Development*, *12*, 3541–3569.

Butler, A. H., Sjoberg, J. P., Seidel, D. J., & Rosenlof, K. H. (2017). A sudden stratospheric warming compendium. *Earth System Science Data*, *9*, 63–76.

Cohen, N. Y., Gerber, E. P., & Bühler, O. (2013). Compensation between Resolved and Unresolved Wave Driving in the Stratosphere: Implications for Downward Control. *Journal of the Atmospheric Sciences*, *70*.

Connelly, D. S. (2023). *dsconnelly/mubofo [Software]*. Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.10908102

Connelly, D. S., & Gerber, E. P. (2024). Regression Forest Approaches to Gravity Wave Parameterization for Climate Projection. *Journal of Advances in Modeling Earth Systems*, *16*.

Ewetola, M., & Esler, J. G. (2024). The effect of intermittency in wave forcing on the quasi-biennial oscillation. *Journal of Fluid Mechanics*, *988*, A16.

Fritts, D. C., & Alexander, M. J. (2003). Gravity Wave Dynamics and Effects in the Middle Atmosphere. *Reviews of Geophysics*, *41*.

Garfinkel, C. I., Gerber, E. P., Shamir, O., Rao, J., Jucker, M., White, I., & Paldor, N. (2022). A QBO Cookbook: Sensitivity of the Quasi-Biennial Oscillation to Resolution, Resolved Waves, and Parameterized Gravity Waves. *Journal of Advances in Modeling Earth Systems*, *14*.

Garfinkel, C. I., White, I., Gerber, E. P., Jucker, M., & Erez, M. (2020). The Building Blocks of Northern Hemisphere Wintertime Stationary Waves. *Journal of Climate*, *33*, 5611–5633.

Hertzog, A., Souprayen, C., & Hauchecorne, A. (2002). Eikonal simulations for the formation and the maintenance of atmospheric gravity wave spectra. *Journal of Geophysical Research*, *107*.

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

Jochum, F., Chew, R., Lott, F., Voelker, G. S., Weinkaemmerer, J., & Achatz, U. (2025). The Impact of Transience in the Interaction between Orographic Gravity Waves and Mean Flow. *Journal of the Atmospheric Sciences*, *82*, 425–442.

Jucker, M., & Gerber, E. P. (2017). Untangling the Annual Cycle of the Tropical Tropopause Layer with an Idealized Moist Model. *Journal of Climate*, *30*, 7339–7358.

Kim, Y.-H., Voelker, G. S., Bölöni, G., Zängl, G., & Achatz, U. (2024). Crucial role of obliquely propagating gravity waves in the quasi-biennial oscillation dynamics. *Atmospheric Chemistry and Physics*, *24*.

Kim, Y.-H., Bölöni, G., Borchert, S., Chun, H.-Y., & Achatz, U. (2021). Toward Transient Subgrid-Scale Gravity Wave Representation in Atmospheric Models. Part II: Wave Intermittency Simulated with Convective Sources. *Journal of the Atmospheric Sciences*, *78*, 1339–1357.

Kruse, C. G., Alexander, M. J., Hoffmann, L., van Niekirk, A., Polichtchouk, I., Bacmeister, J. T., Holt, L., Plougonven, R., Šácha, P., Wright, C., Sato, K., Shibuya, R., Gisinger, S., Ern, M., Meyer, C. I., & Stein, O. (2022). Observed and Modeled Mountain Waves from the Surface to the Mesosphere near the Drake Passage. *Journal of the Atmospheric Sciences*, *79*.

Lott, F., Guez, L., & Maury, P. (2012). A stochastic parameterization of non-orographic gravity waves: Formalism and impact on the equatorial stratosphere. *Geophysical Research Letters*, *39*.

Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., & Clough, S. A. (1997). Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *Journal of Geophysical Research*, *102*, 16663–16682.

Muraschko, J., Fruman, M. D., Achatz, U., Hickel, S., & Toledo, Y. (2015). On the application of Wentzel-Kramer-Brillouin theory for the simulation of the weakly nonlinear dynamics of gravity waves. *Quarterly Journal of the Royal Meteorological Society*, *141*, 676–697.

Richter, J. H., Butchart, N., Kawatani, Y., Bushell, A. C., Holt, L., Serva, F., Anstey, J., Simpson, I. R., Osprey, S., Hamilton, K., Braesicke, P., Cagnazzo, C., Chen, C.-C., Garcia, R. R., Gray, L. J., Kerzenmacher, T., Lott, F., McLandress, C., Naoe, H., . . . Yukimoto, S. (2020). Response of the Quasi-Biennial Oscillation to a warming climate in global climate models. *Quarterly Journal of the Royal Meteorological Society*.

Song, I.-S., & Chun, H.-Y. (2008). A Lagrangian Spectral Parameterization of Gravity Wave Drag Induced by Cumulus Convection. *Journal of the Atmospheric Sciences*, *65*, 1204–1224.

Voelker, G. S., Bölöni, G., Kim, Y.-H., Zängl, G., & Achatz, U. (2024). MS-GWaM: A Three-Dimensional Transient Gravity Wave Parametrization for Atmospheric Models. *Journal of the Atmospheric Sciences*, *81*, 1181–1200.

Wei, J., Bölöni, G., & Achatz, U. (2019). Efficient Modeling of the Interaction of Mesoscale Gravity Waves with Unbalanced Large-Scale Flows: Pseudomomentum-Flux Convergence versus Direct Approach. *Journal of the Atmospheric Sciences*, *76*(9), 2715–2738.

Zängl, G., Reinert, D., Rípodas, P., & Baldauf, M. (2014). The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society*, *141*, 563–579.