Toward transient subgrid-scale gravity wave representation in atmospheric models. Part I: Propagation model including non-dissipative direct wave-mean-flow interactions

Gergely Bölöni*, Young-Ha Kim

Institut für Atmosphäre und Umwelt, Goethe-Universität Frankfurt am Main, Frankfurt am Main, Germany

Sebastian Borchert

Deutscher Wetterdienst, Offenbach, Germany

Ulrich Achatz

Institut für Atmosphäre und Umwelt, Goethe-Universität Frankfurt am Main, Frankfurt am Main, Germany

**Corresponding author*: Gergely Bölöni, boeloeni@uni-frankfurt.de

ABSTRACT

Current gravity-wave (GW) parameterization (GWP) schemes are using the steady-state assumption, where an instantaneous balance between GWs and mean flow is postulated, thereby neglecting transient, non-dissipative direct interactions between the GW field and the resolved flow. These schemes rely exclusively on wave dissipation, by GW breaking or near critical layers, as a mechanism leading to forcing of the mean flow. In a transient GWP, without steady-state assumption, non-dissipative direct wave-mean-flow interactions are enabled as an additional mechanism. Idealized studies have shown that this is potentially important, so that the transient GWP Multi-Scale Gravity-Wave Model (MS-GWaM) has been implemented into a state-of-the-art weather and climate model. In this implementation, MS-GWaM leads to a zonal-mean circulation well in agreement with observations, and increases GW momentum-flux intermittency as compared to steady-state GWPs, bringing it into better agreement with super-pressure balloon observations. Transient effects taken into account by MS-GWaM are shown to make a difference even on monthly time-scales: in comparison with steady-state GWPs momentum fluxes in the lower stratosphere are increased and the amount of the missing drag at Southern Hemispheric high latitudes is decreased to a modest but non-negligible extent. An analysis of the contribution of different wavelengths to the GW signal in MS-GWaM suggests that small scale GWs play an important role down to horizontal and vertical wavelengths of $50km$ (or even smaller) and $200m$ respectively.

1. Introduction

Gravity waves (GW) play an important role in atmospheric dynamics. They are excited mostly in the troposphere e.g. by flow over orography, convection, and jets and front systems. In the course of their propagation they affect the momentum and energy balance in the atmosphere everywhere up to the thermosphere (see e.g. Kim et al. (2003)). The direct impact of GWs on the large-scale circulation is largest in the middle atmosphere, however they also affect tropospheric weather and climate significantly (e.g. Scaife et al. (2005, 2012)).

In GCMs[1] and NWP[2] models, effects of GWs must be parameterized, given the wide spatial and temporal spectrum they act on, part of it being far below the effective resolution of global model applications. Wentzel-Kramer-Brillouin (WKB) theory (Bretherton 1966; Grimshaw 1975; Achatz et al. 2017) is the basis of most GW parameterizations (GWP) in climate simulations and weather predictions (Lindzen 1981; Medvedev and Klaassen 1995; Warner and McIntyre 1996; Hines 1997a,b; Lott and Miller 1997; Alexander and Dunkerton 1999; Scinocca 2003; Orr et al. 2010; Lott and Guez 2013). There is, however, an increasing appreciation that the present handling of this technique needs improvements: a simplification typically used is the neglect of (1) horizontal GW propagation (single-column approximation) and (2) transient effects such as non-dissipative direct GW-mean-flow interactions (steady-state approximation). The former has been shown to be an important weakness of state-of-the-art parameterizations, by e.g. Sato et al. (2009), Ribstein et al. (2015), Ribstein and Achatz (2016) and Ehard et al. (2017), while Bölöni et al. (2016), Muraschko et al. (2015), and Wilhelm et al. (2018) propose improvements with regard to the latter aspect. Another drawback of GWPs in current climate and weather codes is that their applicability outside of the tropics (where Coriolis effects are non-negligible) relies on

[1]General Circulation Models

[2]Numerical Weather Prediction

the assumption of balanced (hydrostatic, geostrophic) resolved flows, which might not be valid with the increasing spatial resolutions applied nowadays. If, however, the resolved flow is not balanced, additional forcing terms due to the GW dynamics appear both in the momentum and the entropy equation representing e.g. elastic effects (Achatz et al. 2017; Wei et al. 2019). Potential triad wave-wave interactions in the atmosphere are also not taken into account in current GWPs, although their neglect has never, to the best of our knowledge, been justified explicitly. In addition to the propagation issues listed above, faithful representation of GW sources is a key to success, and is another area where one finds room for improvement: theory and applications for orographic (Palmer et al. 1986; Bacmeister et al. 1994; Lott and Miller 1997) and convective GW sources (Beres et al. 2005; Song and Chun 2005) are relatively well-developed, but the representation of GW emissions by jets and fronts - in spite of the efforts of Charron and Manzini (2002); Richter et al. (2010); de la Cámara and Lott (2015) - remains difficult.

This paper is focusing on the issues of GW propagation. In a novel framework, transient effects are incorporated by removing the steady-state approximation. This work is an extension of the study by Bölöni et al. (2016), where effects of the transient, non-dissipative direct GW-mean-flow interactions have been assessed in an idealized set-up, while here the same is done in a more complex framework, where the proposed transient GWP has been implemented into a state-of-the-art GCM/NWP model. The single-column approximation has been kept for sake of simplicity, with the intention to give it up in a later step of our developments.

Section 2 motivates the implementation of a transient GWP to a state-of-the-art GCM and recalls the necessary theoretical background for the rest of the manuscript. This is followed by the actual implementation details in section 3, and by the presentation of the GCM-simulation results in section 4. Finally a summary of the most important findings is given in section 5.

2. Theory

In the following we outline the theoretical basis of MS-GWaM. In section 2.a we do so for locally monochromatic GWs together with the simplifying assumptions applied and a comparison to standard parameterization approaches. In section 2.b the monochromatic perspective is generalized to full GW spectra.

a. Locally monochromatic GW fields

In this section we first sketch the general WKB theory that MS-GWaM is built on, (section 2.a.1), then describe the simplifying pseudomomentum-flux approach and single-column approximation which are used in the current study (sections 2.a.2 and 2.a.3). Finally, our transient formulation is compared to the one with the steady-state approximation on which present-day GW parameterizations are based.

1) General WKB

Following WKB theory as applied, e.g., by Grimshaw (1975) and Achatz et al. (2017), the spatio-temporal structure of a locally monochromatic small-scale GW field in a larger-scale flow is characterized by a local wavenumber $\mathbf{k}(\mathbf{x},t) = \mathbf{e}_x k + \mathbf{e}_y l + \mathbf{e}_z m$ and local frequency $\omega(\mathbf{x},t)$, while its amplitude can be deduced from its wave-action density $\mathcal{A}(\mathbf{x},t)$, all depending on position $\mathbf{x} = \mathbf{e}_x x + \mathbf{e}_y y + \mathbf{e}_z z$ and time. Frequency and wavenumber are connected by the dispersion relation

$$\hat{\omega} = \omega - \mathbf{k}_h \cdot \mathbf{U} = \pm\sqrt{\frac{N^2 K_h^2 + f^2 m^2}{K^2}} \quad (1)$$

where $\mathbf{U}(\mathbf{x},t)$ is the local horizontal wind of the large-scale flow, $N^2(z)$ the squared Brunt-Väisälä frequency, f the Coriolis frequency and $K = \sqrt{k^2 + l^2 + m^2}$ and $K_h = \sqrt{k^2 + l^2}$ the magnitude of the total and horizontal wavenumber, respectively. Wave-action density $\mathcal{A} = E_{gw}/\hat{\omega}$ is the ratio between the GW energy density E_{gw} and the intrinsic frequency $\hat{\omega}$. Because all fields are real-valued, amplitudes corresponding to the negative branch in (1) can be determined directly from those corresponding to the positive branch, so that henceforth only the latter will be considered. The local group velocity then is

$$\mathbf{c}_g(\mathbf{x},t) = \nabla_{\mathbf{k}}\Omega \qquad \text{with} \qquad \Omega(\mathbf{x},\mathbf{k},t) = \mathbf{U}(\mathbf{x},t)\cdot\mathbf{k}_h + \sqrt{\frac{N^2(z)K_h^2 + f^2(y)m^2}{K^2}} \tag{2}$$

where Ω expresses the local frequency so that explicit space and time dependencies are only due to the large-scale (mean) flow, vertical variations of stratification, and horizontal variations of the Coriolis parameter. In the absence of dissipation, the development of the GW field, given the mean flow, is determined by

$$\left(\frac{\partial}{\partial t} + \mathbf{c}_g\cdot\nabla\right)\mathbf{k} = -\nabla_{\mathbf{x}}\Omega \tag{3}$$

$$\frac{\partial \mathcal{A}}{\partial t} + \nabla\cdot(\mathbf{c}_g\mathcal{A}) = 0 \tag{4}$$

while the GW effect on the mean flow is described by

$$\left(\frac{\partial \mathbf{U}}{\partial t}, \frac{\partial \Theta}{\partial t}\right)_{gw} = \left[-\frac{1}{\bar{\rho}}\nabla\cdot\left(\bar{\rho}\overline{\mathbf{v}'\mathbf{u}'}\right) + \frac{f}{\bar{\theta}}\mathbf{e}_z\times\overline{\mathbf{u}'\theta'}, -\nabla_h\cdot\overline{\mathbf{u}'\theta'}\right]. \tag{5}$$

Here Θ is the mean-flow potential temperature deviation from the reference-atmosphere potential temperature $\bar{\theta}(z)$, while $\bar{\rho}(z)$ is the reference-atmosphere density, $\mathbf{v}$ and $\mathbf{u}$ denote the full and the horizontal wind vector respectively and ∇_h stands for the horizontal components of ∇. The GW momentum fluxes $\bar{\rho}\overline{\mathbf{v}'\mathbf{u}'}$ and horizontal potential temperature flux $\overline{\mathbf{u}'\theta'}$ can be calculated from $\mathbf{k}$ and $\mathcal{A}$. Clearly, Eq. (5) does not account for the energy deposition by GWs, which has an important thermal effect in the mesosphere and lower thermosphere (MLT) (e.g. Becker 2017). It also implies that the representation of this effect is left out from the current study and will have to be incorporated in the future.

2) Pseudomomentum approximation

In the spirit of a step-wise implementation of the most general theory, for the time being, MS-GWaM does not use Eq. (5) in its full complexity. Instead, resting on considerations by Andrews and McIntyre (1976, 1978a) and following the procedure of all present-day GWPs, Eq. (5) is replaced by

$$\left(\frac{\partial \mathbf{U}}{\partial t}, \frac{\partial \Theta}{\partial t}\right)_{gw} = \left[-\frac{1}{\bar{\rho}}\nabla \cdot \left(\hat{\mathbf{c}}_g \mathbf{k}_h \mathcal{A}\right), 0\right] \tag{6}$$

where $\mathbf{k}_h$ is the horizontal part of the wavenumber vector, $\hat{\mathbf{c}}_g = \mathbf{c}_g - \mathbf{U}$ the intrinsic group velocity, and $\hat{\mathbf{c}}_g \mathbf{k}_h \mathcal{A}$ the GW Eliassen-Palm or pseudomomentum flux. An advantage of this approximation is that no GW potential temperature fluxes are required. The latter would enter via their horizontal convergence, which one is inclined to avoid in single-column GWPs. Wei et al. (2019) discussed this approximation in detail. Eq. (6) basically assumes that the large-scale flow is in geostrophic and hydrostatic balance. When this is not the case, errors can occur outside of the tropics whenever near-inertial GWs are involved with $\hat{\omega}$ close to f. In future work it is intended to drop both the pseudomomentum- and the single-column approximation.

3) Single-column approximation

The single-column approximation is taken in present-day GWPs for the sake of efficiency, and we do so here as well. One neglects in the GW-mean-flow interaction all horizontal derivatives and one also neglects in the wave-action equation all horizontal group-velocity components, so that, using the pseudomomentum approximation as well, the approximated dynamics is described by

$$\left(\frac{\partial}{\partial t} + c_{gz}\frac{\partial}{\partial z}\right)(\mathbf{k}_h, m) = \left(0, -\frac{\partial \Omega}{\partial z}\right) \tag{7}$$

$$\frac{\partial \mathcal{A}}{\partial t} + \frac{\partial}{\partial z}\left(c_{gz}\mathcal{A}\right) = 0 \tag{8}$$

$$\left(\frac{\partial \mathbf{U}}{\partial t}\right)_{gw} = -\frac{1}{\bar{\rho}}\frac{\partial}{\partial z}\left(c_{gz}\mathbf{k}_h \mathcal{A}\right) \tag{9}$$

where c_{gz} is the vertical group velocity. This approximation neglects all effects of horizontal GW propagation. Note that the pseudomomentum-flux convergence in the righ-hand side of Eq. (9) can be written as (e.g. Achatz et al. 2017):

$$-\frac{1}{\bar{\rho}}\frac{\partial}{\partial z}\left(c_{gz}\mathbf{k}_h\mathcal{A}\right) = -\frac{1}{\bar{\rho}}\left[\frac{\partial}{\partial z}\left(\bar{\rho}\overline{\mathbf{u}'w'}\right) - f\mathbf{e}_z\times\frac{\partial}{\partial z}\left(\frac{\bar{\rho}\overline{\mathbf{u}'\theta'}}{d\bar{\theta}/dz}\right)\right] \tag{10}$$

as is also known from derivations from GLM theory (Andrews and McIntyre 1978b).

4) Steady-state approximation and its implications

The final step taken in present-day GWPs for the sake of efficiency is the assumption that the GW field adjusts instantaneously to a given mean-flow distribution. This way GW effects are propagating within one time step from a source to the model top and bottom. One neglects in the prognostic equations for the GW field all time derivatives so that (7) - (8) are replaced by

$$\frac{\partial}{\partial z}(\mathbf{k}_h, m) = \left(0, -\frac{1}{c_{gz}}\frac{\partial\Omega}{\partial z}\right) \tag{11}$$

$$\frac{\partial}{\partial z}\left(c_{gz}\mathcal{A}\right) = S \tag{12}$$

Here we have introduced a source or sink S on the right-hand side of the wave-action-density equation. This is decisive. One sees that in the steady-state approximation the horizontal wave number is a constant so that without any source or sink there would be no GW forcing of the mean flow in (9). Hence in this approximation, GW dissipation, e.g. by GW breaking or close to critical layers, is indispensable for a GW effect on the mean flow, while the explicit description of GW transience as in (7) - (8) also allows GW impacts on the mean flow via non-dissipative direct wave-mean-flow interactions.

Consequences of applying the steady-state approximation instead of the transient GW-model Eqs. (7) - (9) - and thus neglecting non-dissipative direct GW-mean-flow interactions - have been studied by Bölöni et al. (2016) in a highly idealized setup using wave-resolving simulations as a reference.

They achieved a reliable evolution of the GW energy and the mean flow only using the transient model. In case of using the steady-state equations, important features of the GW-mean-flow interactions were not captured: the GW packet propagated way too fast until static instability set in and its induced mean flow did not agree with the results from wave-resolving simulations. Using a Fourier-ray model (Broutman et al. 2006) and high-resolution WRF (Skamarock et al. 2019) simulations, Kruse and Smith (2018) found that in the interaction of mountain waves with the mean flow, both dissipative and non-dissipative forcing of the mean flow seem to play an important role. The natural question - how important non-dissipative direct GW-mean-flow interactions are in the context of global dynamics - have motivated the present study.

b. Spectral treatment of transient GW distributions

Although the consideration of locally monochromatic GW fields is helpful for deriving the prognostic system (3) - (5) or its single-column pseudomomentum approximation (7) - (9), real-world GW fields are made up of a full spectrum of components. Even if one starts out from a locally monochromatic initial condition, GW-mean-flow interactions tend to quickly lead to caustics where more than one wavenumber is observed at a single location. Correspondingly, attempts to solve the above discussed equation sets directly on a computer most often fail due to numerical instabilities near caustics. As shown by Muraschko et al. (2015), this can be avoided by considering a spectral wave-action density in wavenumber-position phase space (e.g. Hertzog et al. 2002) instead.

$$\mathcal{N}(\mathbf{x},\mathbf{k},t) = \int_{\mathbb{R}^3} d^3\beta \mathcal{A}_\beta(\mathbf{x},t)\delta(\mathbf{k}-\mathbf{k}_\beta) \tag{13}$$

where β is a three-dimensional index field and each combination of $\mathcal{A}_\beta$ and $\mathbf{k}_\beta$ satisfies (3) - (4) or (7) - (8) separately. If the corresponding wave amplitudes are weak enough, a superposition of these solutions is a WKB solution of the basic dynamical equations as well, assuming that the required scale separation between the various spectral components and the large-scale mean flow still holds, and one can derive the prognostic equation

$$\frac{D_r\mathcal{N}}{Dt} \equiv \frac{\partial \mathcal{N}}{\partial t} + \mathbf{c}_g \cdot \nabla_\mathbf{x}\mathcal{N} + \dot{\mathbf{k}} \cdot \nabla_\mathbf{k}\mathcal{N} = 0 \tag{14}$$

Here $\mathbf{c}_g(\mathbf{x}, \mathbf{k}, t) = \nabla_\mathbf{k}\Omega$ is again the group velocity defined in (2) for wavenumber $\mathbf{k}$, and $\dot{\mathbf{k}}(\mathbf{x}, \mathbf{k}, t) = -\nabla_\mathbf{x}\Omega$ is the rate of change of the wavenumber $\mathbf{k}$ as it appears on the right-hand side of (3). D_r/Dt is a material derivative along trajectories in phase space, so called rays, tangential to the phase-space velocity $(\mathbf{c}_g, \dot{\mathbf{k}})$. Along these rays the phase-space wave-action density is conserved. The GW impact on the mean flow is the sum of the impact of all spectral components so that, with the pseudomomentum approximation,

$$\left(\frac{\partial \mathbf{U}}{\partial t}\right)_{gw} = -\frac{1}{\bar{\rho}}\nabla \cdot \int d^3k\, \hat{\mathbf{c}}_g \mathbf{k}_h \mathcal{N} \tag{15}$$

with $d^3k = dkdldm$. Similar expressions can be formulated also without pseudomomentum approximation (Wei et al. 2019). We note in passing that in the absence of background winds, (14) agrees with the radiative transfer equation without wave-wave interactions that has been used in the oceanic context for GWPs (Olbers et al. 2019, and references therein). There, however, the shape of the GW spectrum is prescribed, while in our implementation it develops without constraints.

In the single-column approximation, one again neglects all horizontal derivatives as well as the horizontal group velocity in (14), resulting with

$$\frac{D_r}{Dt} = \frac{\partial}{\partial t} + c_{gz}\frac{\partial}{\partial z} + \dot{m}\frac{\partial}{\partial m} \qquad \text{and} \qquad \left(c_{gz}, \dot{m}\right) = \left(\frac{\partial \Omega}{\partial m}, -\frac{\partial \Omega}{\partial z}\right) \tag{16}$$

in the system

$$\frac{D_r\mathcal{N}}{Dt} = 0 \tag{17}$$

$$\left(\frac{\partial \mathbf{U}}{\partial t}\right)_{gw} = -\frac{1}{\bar{\rho}}\frac{\partial}{\partial z}\int d^3k\, c_{gz}\mathbf{k}_h\mathcal{N} \tag{18}$$

The corresponding rays, along which $\mathcal{N}$ is conserved, are given by

$$\frac{D_r}{Dt}(\mathbf{x}_h, z) = (0, c_{gz}) = \left[0, -\frac{m(\hat{\omega}^2 - f^2)}{\hat{\omega}K^2}\right] \tag{19}$$

$$\frac{D_r}{Dt}(\mathbf{k}_h, m) = (0, \dot{m}) = \left(0, -\mathbf{k}_h \cdot \frac{\partial \mathbf{U}}{\partial z} - \frac{N K_h^2}{\hat{\omega}K^2}\frac{\partial N}{\partial z}\right) \tag{20}$$

Moreover, partial integration of (17), using (16), also yields

$$\frac{\partial}{\partial t}\int dm\,\mathcal{N} + \frac{\partial}{\partial z}\int dm\,c_{gz}\mathcal{N} = 0 \tag{21}$$

which is the equivalent to (8).

In a steady-state approximation, one again neglects the time derivatives in the wave-action density equation (17) and in the ray equations (19) and (20). Hence $\mathbf{k}_h$ is again a constant and (18) yields together with the steady-state version of (21) the non-acceleration result $\partial\mathbf{U}/\partial t = 0$, i.e. the mean flow is unaffected by GWs, unless (21) is supplemented by sources or sinks.

3. Implementation in a high-top atmosphere model

Our single-column pseudomomentum-approximation subgrid-scale GW model applying the transient Eqs. (17) - (20), extended by a saturation scheme, has been named MS-GWaM[3]. It has been implemented into the ICON[4] model (Zängl et al. 2015) in its upper-atmosphere configuration UA-ICON (Borchert et al. 2019), allowing numerical studies over a wide altitude range from the Earth's surface to the lower thermosphere. For the sake of simplicity and clear traceability of causes and consequences, the current orographic GWP in UA-ICON, based on Lott and Miller (1997), has been left untouched, and MS-GWaM only replaces the non-orographic GWP there, based on Orr et al. (2010).

As a reference and a representative of currently available GWP schemes, in addition to the transient implementation, two steady-state versions of MS-GWaM have also been implemented

[3]Multi-Scale Gravity-Wave Model

[4]ICOsahedral Non-hydrostatic model

to UA-ICON. The first one excludes non-dissipative direct GW-mean-flow interactions through the steady-state approximation but shares all other parameterization components with the transient MS-GWaM, such as GW sources and the saturation scheme. The other one differs from MS-GWaM in its saturation scheme as well, i.e. instead of an integrated treatment of the GW breaking criterion, it applies a monochromatic approach (see the details in section 3.b). Throughout the paper, the implementation of the transient MS-GWaM into UA-ICON will be referred to as TR, while the two steady-state implementations will be called ST and STMO respectively.

a. Transient scheme

In a global implementation the interaction equations would have to be rewritten in spherical coordinates. The single-column approximation, however, eliminates any horizontal changes of the GW field and all metric terms, which amounts to treating the parameterization equations in local Cartesian coordinates on an f-plane.

1) GW propagation and interaction with the mean flow

Following Muraschko et al. (2015), we define Lagrangian ray volumes as carriers of the GW fields' wave-action density and simply trace their positions in phase space. Due to (17), their spectral wave-action density is conserved, unless wave dissipation is active. Each ray volume is six-dimensional, and its horizontal cross-section is given by that of the corresponding ICON column. In the single-column approximation, it does not change so that we suppress it in the following notation. Likewise, horizontal wavenumbers do not change either, but due to the source formulation below we keep track of the ray-volume extent in the corresponding directions.

As illustrated in Figs.1a and 1b, each ray volume has an extent Δz in z direction and extents Δk, Δl, Δm in the three-dimensional wavenumber space. They move, expand or shrink in the z and m

directions. Due to $\partial c_{gz}/\partial z + \partial \dot{m}/\partial m = 0$, the phase-space content of each ray volume, and hence in our discretization $V_p = \Delta z \Delta m$, do not change. To achieve this, first, changes in the vertical extent of ray volumes are calculated via Eq. (19) as $\dot{\Delta} z = c^t_{gz} - c^b_{gz}$ where c^t_{gz} and c^b_{gz} stand for the vertical group velocities at the top ($z = z_t$) and the bottom ($z = z_b$) of the ray volume, respectively. Second, the displacement of the ray-volume center-point is calculated via Eq. (19) as $\dot{z}_c = (c^t_{gz} + c^b_{gz})/2$ in the vertical direction and, via Eq. (20), as $\dot{m}_c = \dot{m}(\mathbf{k}_h, m_c)$ in m-direction. For the latter, the resolved dynamical fields $N, \partial_z \mathbf{U}$ and $\partial_z N$ are interpolated to the ray-volume center-point $z = z_c$. Finally the ray-volume extent in m-direction is updated by $\Delta m = V_p/\Delta z$.

Next, the acceleration of the resolved horizontal wind is calculated via Eq. (18). The pseudo-momentum fluxes (PMF) are calculated on the half-levels of the ICON vertical grid, i.e. at the half-level at $z = z_{i+1/2}$, the integral on the right-hand-side of Eq. (18) is approximated as

$$PMF_{z_{i+1/2}} = \left(\int d^3 k c_{gz} \mathbf{k}_h \mathcal{N} \right)_{z_{i+1/2}} \approx \left(\sum_{j=1}^{N_i} \frac{\Delta z^j_i}{\delta z_i} c^j_{gz} \mathbf{k}^j_h \mathcal{N}^j \Delta k^j \Delta l^j \Delta m^j \right)_{z_{i+1/2}} \tag{22}$$

where N_i is the number of ray volumes overlapping the vertical layer $z_{i+1} < z \leqslant z_i$ with a thickness of δz_i and j is the index over those ray volumes. As can be seen in Fig. 1a, the ratio $\Delta z^j_i/\delta z_i$ represents the pseudomomentum-flux fraction of the j-th ray volume contributing to the layer centered at $z = z_{i+1/2}$. After the calculation of the pseudomomentum fluxes at half-levels, the mean-flow tendency $(\Delta \mathbf{U}/\Delta t)_{gw_{z_i}}$ at the full level at $z = z_i$ is calculated via (18) by centered finite differences. The resolved horizontal wind is updated as $\mathbf{U}^{new}_{z_i} = \mathbf{U}_{z_i} + \delta t (\Delta \mathbf{U}/\Delta t)_{gw_{z_i}}$ where a time-step $\delta t = 60s$ is used. In order to ensure that the ray volumes do not jump over strong, shallow shear-layers - and thus describe reflection and critical-layer filtering properly - , a 4^{th} order Runge-Kutta sub-time-stepping is used for the integration of Eqs. (19) and (20) with a time-step of $\delta t_s = 30s$. Note that the development of z and m via Eqs. (19) and (20) depends on the stratification and the wind shear, i.e. on $\mathbf{U}_{z_i}$, which - among others - includes the GW induced wind contribution.

2) GW breaking

The phase-space wave-action density $\mathcal{N}$ is conserved along rays until GWs break. In that case turbulence is generated that damps the GWs via turbulent viscosity and diffusivity. Hence $\mathcal{N}$ decreases, which generates pseudomomentum-flux convergence additionally to the non-dissipative direct GW-mean-flow interactions described in the previous section. In MS-GWaM, following Lindzen (1981), this is taken into account by diagnosing whether the GW field can turn the flow into a statically unstable state. Once this is the case in a given layer, the wave-action density of all overlapping ray volumes is reduced so that static instability cannot occur anymore. Following Bölöni et al. (2016), the static instability criterion for a quasi-monochromatic wave is given by $m^2|B|^2 > N^4$ where $|B|^2 = 2\mathcal{A}N^4K_h^2/(\bar{\rho}\hat{\omega}K^2)$ is the squared complex GW buoyancy amplitude and $|\cdot|$ denotes the modulus. Applying the phase-space concept to represent the full spectrum, this reads

$$\frac{2N^4}{\bar{\rho}}\int d^3k\, m^2 \mathcal{N}\frac{K_h^2}{\hat{\omega}K^2} > N^4. \tag{23}$$

The single-column discretization of this, e.g. in the layer with thickness δz_i centered at $z = z_{i+1/2}$, is

$$\left(\frac{2N^4}{\bar{\rho}}\sum_{j=1}^{N_i}\mathcal{S}^j > N^4\right)_{z_{i+1/2}} \qquad \text{with} \qquad \mathcal{S}^j = \frac{\Delta z_i^j}{\delta z_i}\mathcal{N}^j\frac{{m^j}^2{K_h^j}^2}{\hat{\omega}^j{K^j}^2}\Delta k^j \Delta l^j \Delta m^j \tag{24}$$

where again, all variables with a j -index denote known properties of ray volumes overlapping the layer, and those without are resolved variables at $z = z_{i+1/2}$. Whenever static instability is diagnosed via Eq. (24), the saturation scheme is called, and the turbulent diffusivity and viscosity are determined so that they exactly counteract the amplitude increase of the contributing GWs that would cause (23) or (24) to be satisfied. In the case of buoyancy, e.g., the turbulence effect is then captured via $\partial_t b = ... + \mathcal{K}\left(\partial_x^2 b + \partial_y^2 b + \partial_z^2 b\right)$ with the turbulent diffusivity $\mathcal{K}$. After a Fourier transformation in space the corresponding buoyancy change over a short time Δt is

$\Delta|\tilde{b}|^2 = \ldots - 2\mathcal{K}\Delta t|\tilde{b}|^2(K_h^2 + m^2)$ or for the amplitude

$$\Delta\left(\frac{d|B|^2}{dm}dm\right) = \frac{2N^2}{\bar{\rho}}\Delta(m^2\hat{\omega}\mathcal{N}dm) = -2\mathcal{K}\Delta t\frac{2N^2}{\bar{\rho}}m^2(K_h^2 + m^2)\hat{\omega}\mathcal{N}dm. \tag{25}$$

After simplifying (25), requiring the diffusivity and viscosity to be just strong enough to prevent (23) and (24) to be satisfied, and returning to discretized variables, one is left with

$$\mathcal{N}_{sat}^j = \mathcal{N}^j(1 - \mathcal{K}K^{j\,2}) \qquad (j = 1, \ldots, N_i), \tag{26}$$

where $\mathcal{N}^j$ is the phase-space wave-action density one would have directly from wave-action conservation without the turbulence impact, and $\mathcal{N}_{sat}^j$ is the saturated wave-action density corresponding to the equality sign in (23) and (24). The local turbulent diffusion coefficient hence is

$$\mathcal{K}(z) = \frac{\sum_{j=1}^{N_i}\mathcal{S}^j - \bar{\rho}/2}{\sum_{j=1}^{N_i}\mathcal{S}^j K^{j\,2}}. \tag{27}$$

The vertical wavenumber dependence of the saturation equation (26) is such that small-scale GWs are damped most strongly. The diffusivity estimated as above could be used to predict corresponding frictional heating as well as modifications of the GW energy deposition (e.g. Becker 2017). For the time being, however, these effects are here not taken into account. Furthermore, the effect of GW damping due to the heat diffusion (i.e. downward heat flux) is also ignored. We also note that the above described saturation scheme, similarly to Lindzen (1981), assumes that the Prandtl number is ~ 1, i.e. that momentum and potential temperature are equally effected by turbulent diffusion. This assumption is to be revisited in the future as e.g. Fritts and Dunkerton (1985) suggest that the Prandtl number should be very large for breaking GWs.

3) GW source representation

A simple representation of the non-orographic GW sources has been chosen. Instead of parameterizing the GW sources associated e.g. with convection and jets and fronts, it is assumed that the superposition of all non-orographically emitted GWs obeys the universal Desaubies spectrum (VanZandt 1982; Fritts and VanZandt 1993). Following Scinocca (2003), the corresponding GW launch momentum-flux spectrum, projected onto the horizontal propagation direction of each spectral component, defined by the azimuth angle $\phi \in [0, 2\pi)$ so that $k = K_h \cos\phi, l = K_h \sin\phi$, is

$$\bar{\rho} F_0(\tilde{c}, \tilde{\omega}, \phi) = \bar{\rho} C m_*^3 \frac{\tilde{c} N^3 \tilde{\omega}^{1-p}}{N^4 + m_*^4 \tilde{c}^4} \tag{28}$$

where $\tilde{\omega} = N K_h / |m|$ is the non-rotational and hydrostatic approximation of the intrinsic frequency, $\tilde{c} = N/|m|$ is the respective intrinsic phase speed. Note that upward GW propagation corresponds to $m < 0$, implying $\tilde{\omega} = -N K_h / m$ at the source. In the current study, four azimuthal angles have been used defining GW propagation directions towards east, north, west and south, i.e. $\phi = (0, \pi/2, \pi, 3\pi/2) \Rightarrow k = K_h(1, 0, -1, 0), l = K_h(0, 1, 0, -1)$. The launch spectrum is characterized in terms of intrinsic phase speeds as $\tilde{c} \in (\tilde{c}_{min}, \tilde{c}_{max}] = (0, 36] m s^{-1}$ in each of the four directions, with an equidistant spacing and thus equally large spectral elements $\Delta\tilde{c} = (\tilde{c}_{max} - \tilde{c}_{min})/n_{\tilde{c}}$, where $n_{\tilde{c}} = 6$ is the number of spectral elements. In terms of intrinsic frequency, a range $\tilde{\omega} \in [\tilde{\omega}_{min}, \tilde{\omega}_{max}] = [10^{-4}, 5 * 10^{-4}] s^{-1}$ is considered again with an equidistant spacing and equally large spectral elements $\Delta\tilde{\omega} = (\tilde{\omega}_{max} - \tilde{\omega}_{min})/n_{\tilde{\omega}}$ where $n_{\tilde{\omega}} = 2$ is the number of elements. A characteristic vertical wavenumber $m_* = 2\pi/2km$ is used, while the value of p is set to $5/3$ based on Warner and McIntyre (1996) and Fritts and Lu (1993). The factor C is a tuning parameter enabling to set a desired launch-level pseudomomentum-flux magnitude M, so that $\int_{\tilde{c}_{min}}^{\tilde{c}_{max}} \int_{\tilde{\omega}_{min}}^{\tilde{\omega}_{max}} \bar{\rho} F_0(\tilde{c}, \tilde{\omega}, \phi) d\tilde{\omega} d\tilde{c} = M$ for each azimuthal direction ϕ. In order to account for the seasonal variability of non-orographic GW sources emitted by jets and fronts, time and latitudinal dependence of M has been introduced as

$$M(\varphi, t) = M_{bs}(\varphi) + \beta(t)[M_{bw}(\varphi) - M_{bs}(\varphi)] \tag{29}$$

where φ is the latitude in degrees, $\beta(t) = \{1 + \cos[2\pi(t - t_0)]\}/2$ is a time dependent function with t_0 being 00 UTC 22 December of the given year, and M_{bw} (M_{bs}) is the boreal winter (summer) flux-magnitude profile given as

$$M_{bw}(\varphi) = [1-\alpha(\varphi)]M_{min} + \alpha(\varphi)M_{max} \tag{30}$$

$$M_{bs}(\varphi) = [1-\alpha(\varphi)]M_{max} + \alpha(\varphi)M_{min} \tag{31}$$

with

$$\alpha(\varphi) = [1+\tanh(\varphi/s)]/2 \tag{32}$$

being a function with a smooth transition between 0 and 1 with $s = 11°$. The resulting meridional launch flux profile is plotted in Fig.2 for the northern summer and winter solstices with the choice of $(M_{min}, M_{max}) = (1.5, 2.5)mPa$, which are the actual values chosen for the implementation.

In order to express the spectral distribution of the source in terms of the wavenumbers (k, l, m) as needed by MS-GWaM, the $\tilde{c}, \tilde{\omega}, \phi$ -dependent pseudomomentum-flux spectrum (28) is transformed via the sequence of Jacobian transformations

$$\bar{\rho}F_1(m, \tilde{\omega}, \phi) = J_0\bar{\rho}F_0(\tilde{c}, \tilde{\omega}, \phi) \qquad \text{with} \qquad J_0 = \frac{\partial \tilde{c}}{\partial m} = \frac{N}{m^2}, \tag{33}$$

$$\bar{\rho}F_2(m, K_h, \phi) = J_1\bar{\rho}F_1(m, \tilde{\omega}, \phi) \qquad \text{with} \qquad J_1 = \frac{\partial \tilde{\omega}}{\partial K_h} = -\frac{N}{|m|}, \tag{34}$$

$$\bar{\rho}F_3(m, k, l) = J_2\bar{\rho}F_2(m, K_h, \phi) \qquad \text{with} \qquad J_2 = \frac{\partial(K_h, \phi)}{\partial(k, l)} = \frac{1}{K_h}, \tag{35}$$

where the magnitude of the horizontal wave vector is always calculated through the dispersion relation as $K_h = \tilde{\omega}|m|/N$. After the transformation, using a typical N at launch level, the launch spectrum spans $\lambda_z \in [0.8, 8]km$ with an increasing resolution in m towards large vertical wave lengths (corresponding to large group velocities) and $\lambda_{x,y} \in [47, 1036]km$ with an increasing resolution in k, l towards large horizontal wave scales. Because $\int d^3k\bar{\rho}F_3\mathbf{k}_h/K_h = \int d^3k\mathbf{k}_h c_{gz}\mathcal{N}$, the phase-space wave-action density of ray volumes at launch level $z = z_l$ (= $300hPa$ in this study) is initialized as

$$\left(\mathcal{N}^j = \frac{\bar{\rho}F_3(m^j, k^j, l^j)}{K_h^j c_{gz}^j}\right)_{z_l}, \tag{36}$$

where $j = 1, ..., N_l$ is the ray-volume index with $N_l = 4n_c n_\omega$ being the total number of spectral elements launched at a time in the four azimuthal directions and m^j, k^j, l^j denoting the wavenumber values at the ray-volume centers. The spectral extent of the j-th ray volume in m-direction is calculated as $\Delta m^j = \Delta\tilde{c}^j {m^j}^2/N$, which results in decreasing ray volume extents towards large vertical wave lengths. As shown in Fig.3, the spectral extents in k- and l-directions are defined as $\Delta k^j = \Delta K_h^j$, $\Delta l^j = K_h^j \Delta\phi$ ($\Delta k^j = K_h^j \Delta\phi$, $\Delta l^j = \Delta K_h^j$) for eastward and westward (northward and southward) propagating waves where $\Delta K_h^j = \Delta\omega^j |m^j|/N$ and $\Delta\phi = \pi/2$ is the azimuthal angle difference between propagation directions.

In the transient framework discussed in this section, the GW emission by non-orographic sources is implemented as a lower boundary condition for Eqs. (17) - (20). This requires that the GW ray volumes are emitted continuously at the launch level for the whole spectrum, so that the total pseudomomentum flux $\int d^3k\, \bar{\rho} F_3(\mathbf{k})\mathbf{k}_h/K_h$ is maintained at all times. To achieve this, an elaborate ray-volume launching procedure is applied as demonstrated by Fig.4. Below the launch level z_l, a ghost layer is defined with a thickness δz_g, and at launch time $t = t_0$ the ray volumes are initialized via Eq. (36) with their top matching the bottom of the ghost layer at $z = z_l - \delta z_g$ (see Fig.4a). The five ray volumes sketched in the figure are located in adjacent spectral positions in m. At time $t = t_0 + \Delta t$ all ray volumes have propagated vertically, but to different extents. (see Fig.4b). In order to preserve the spectral shape of the spectrum until the launch level is reached, all ray volumes with a center-point below $z = z_l$ propagate without refraction, i.e. with $c_{gz} = const., \dot{m} = 0, \Delta z = const., \Delta m = const.$ As soon as a ray volume is completely in the ghost layer (i.e. its bottom has passed the height $z = z_l - \delta z_g$), a new ray volume is launched, so that its top matches the bottom of the previous ray volume in the same spectral position. In Fig.4b, this happens at the spectral positions 4 and 5, where the "old" ray volumes launched at $t = t_0$ are denoted by a black center-point and boundaries, while the "new" ray volumes launched at $t = t_0 + \Delta t$

have red center- points and boundaries. Two additional features are demonstrated at the spectral position 5: i) here, the "old" ray volume has traveled so much within Δt, that even the "new" ray volume's bottom ends up at $z = z_l - \delta z_g$, which allows the emission of a second "new" ray volume right away at the same launch time $t = t_0 + \Delta t$, ii) the "old" ray volume's center-point has passed $z = z_l$, so that the full set of GW-mean-flow interaction Eqs. (17) - (20) begins to act, leading to refractions (displacement of the center-point in m-direction) and deformations (changes of Δz and Δm). It is repeated here that the above described GW source is kept simple and non-intermittent on purpose in order to allow a clear separation of transient propagation effects from those implied by intermittent sources.

b. Steady-state schemes

In this section the steady-state implementations of the single-column GW-mean-flow interaction Eqs. (17) - (20) are presented. In the steady-state context it is assumed that the GWs propagate instantaneously from any source to model bottom and model top and that they instantaneously assume an equilibrium with the resolved mean flow and the source distribution. This equilibrium remains unchanged until source or resolved flow change, when the GW distribution again adjusts instantaneously. As a consequence of this, GWs cannot influence the resolved flow, unless wave dissipation is active. The mean-flow acceleration by GWs is hence realized exclusively via GW breaking and critical-layer filtering, i.e. by diagnosing at what height the equilibrium breaks down due to dissipative processes, leading to corresponding pseudomomentum-flux convergences. The next few subsections describe the steady-state implementations of MS-GWaM in detail.

1) GW source representation

The spectral characteristics and the magnitude of the non-orographic GW sources are identical to the transient implementation presented in section 3. a3, i.e. the GW launch-level pseudomomentum flux is distributed among monochromatic spectral elements characterized in the very same way in spectral space as in the transient case ($\lambda_{x,y} \in [47, 1036]km, \lambda_z \in [0.8, 8]km$), with the very same values $\bar{\rho}F_3(m,k,l)$, so that the total GW pseudomomentum flux at each launch time is $\int d^3k\, \bar{\rho}F_3(\mathbf{k})\mathbf{k}_h/K_h$. The corresponding wave-action contribution by each spectral element at the launch level $z = z_l$ can be calculated as $\mathcal{A}^i(z_l) = \mathcal{A}(z_l, \mathbf{k}^i) = \bar{\rho}F_3(\mathbf{k}^i)/(K_h^i c_{gz}^i)\Delta k^i \Delta l^i \Delta m^i$, where $i = 1, ..., N_l$ is the spectral-element index with $N_l = 4n_c n_\omega$.

2) Equilibrium profile

One can easily convince oneself that the steady-state version of (21) holds also component-wise so that we have for each spectral element $c_{gz}^i(z)\mathcal{A}^i(z) = c_{gz}^i(z_l)\mathcal{A}^i(z_l) = const.$, where both $c_{gz}^i(z_l)$ and $\mathcal{A}^i(z_l)$ are known from the GW source. In this way, one is left with equilibrium profiles of the wave-action flux, which entirely determine the GW dynamics after adjustment to the resolved flow. In order to diagnose GW breaking altitudes, critical-, or reflection layers, it is not sufficient to have the products $c_{gz}^i(z)\mathcal{A}^i(z)$ as known quantities, but in addition, one needs the corresponding $\mathcal{A}^i(z)$ and $c_{gz}^i(z)$ profiles individually. The vertical group velocity profiles are obtained from the dispersion relation (1) as

$$c_{gz}^i(z) = -\frac{m^i(z)(\hat{\omega}^{i2}(z) - f^2)}{\hat{\omega}^i(z)({K_h^i}^2 + {m^i}^2(z))} \quad , \quad (i = 1, ..., N_l) \tag{37}$$

with

$$m^i(z) = -\sqrt{\frac{{K_h^i}^2(N^2(z) - \hat{\omega}^{i2}(z))}{\hat{\omega}^{i2}(z) - f^2}} \quad , \quad (i = 1, ..., N_l) \tag{38}$$

where $\hat{\omega}^i(z)$ are the intrinsic frequency profiles of the adjusted GW field's spectral elements. The key to calculate $\hat{\omega}^i(z)$, needed in Eqs. (37)- (38), is the Eikonal frequency equation $D_r\Omega/Dt = \partial\Omega/\partial t = \mathbf{k}\cdot\partial\mathbf{U}/\partial t$ that one can derive directly from the definitions $\mathbf{c}_g = \nabla_\mathbf{k}\Omega$ and $\dot{\mathbf{k}} = -\nabla_\mathbf{x}\Omega$. Hence, in the steady-state case extrinsic frequencies $\omega^i(z_l)$ are unchanged along a ray. This means that after diagnosing $\omega^i(z_l)$ at the launch level, the intrinsic frequency profiles can be calculated based on the known wind profile as $\hat{\omega}^i(z) = \omega^i(z_l) - \mathbf{k}^i\cdot\mathbf{U}(z)$. Using this in (37), the wave-action profile of each spectral element can be calculated as

$$\mathcal{A}^i(z) = \frac{c^i_{gz}(z_l)\mathcal{A}^i(z_l)}{c^i_{gz}(z)} \quad , \quad (i = 1, \ldots, N_l). \tag{39}$$

3) Critical layer filtering and reflection

At critical layers, the intrinsic frequency approaches f and the vertical wave number diverges, see e.g. (1) or (38). With decreasing vertical wavelength a GW eventually becomes unstable and dissipates. In the steady-state picture, critical layers are diagnosed at the lowest altitude $z = z_c$ where $\hat{\omega}(z_c) = \omega(z_l) - \mathbf{k}\cdot\mathbf{U}(z_c) \leq |f|$, i.e., where the Doppler-shift term turns the wave intrinsic frequency to a smaller value than the Coriolis frequency. Accordingly, for each spectral element i, we set the pseudomomentum-flux profile $PMF^i(z)$ to zero at $z \geq z_c$.

When wave reflection occurs, the intrinsic frequency approaches N and m changes sign so that the group velocity is reverted. In the steady-state versions of MS-GWaM this is taken into account by diagnosing the height of potential reflection by finding the lowest altitude $z = z_r$ where $\hat{\omega}(z_r) \geq N$. If a reflection layer is diagnosed at $z = z_r$ for a spectral element, its corresponding pseudomomentum-flux profile $PMF^i(z)$ is set to zero above z_r. Unless GW breaking has changed the equilibrium profile below the reflection layer, the pseudomomentum-flux profile $PMF^i(z)$ vanishes also for $z < z_r$ as well, because under reflection the pseudomomentum-flux changes sign so that the contributions from the upward and downward propagating components cancel each other when the mean flow is in a steady state. However, if GW breaking takes place (see next section) at any altitude $z = z_b < z_r$, the pseudomomentum fluxes carried by the upward and downward propagating spectral elements do not completely cancel, thus, at altitudes $z < z_b$ a residual $PMF^i(z) = PMF^i_{up}(z) - PMF^i_{down}(z)$ is maintained.

4) GW BREAKING

In the steady-state setups of MS-GWaM the instability criterion (23) is used as well. The two steady-state implementations (ST and STMO) differ, however, in the way this is done and how the GW amplitudes are adjusted whenever wave breaking is diagnosed.

A simple approach - most often applied in present-day non-spectral GWPs - is to treat each spectral element independently from each other, i.e. applying (23) for each element separately, leading to a saturation amplitude

$$\mathcal{A}^i_{sat}(z) = \frac{\bar{\rho}}{2}\hat{\omega}\left(\frac{1}{m(z)^{i^2}} + \frac{1}{K_h^{i\,2}}\right) \quad , \quad (i = 1, ..., N_l). \tag{40}$$

Wave breaking is diagnosed at an altitude $z = z_b$ if $\mathcal{A}^i(z_b) > \mathcal{A}^i_{sat}(z_b)$ and static stability is then reinforced by setting $\mathcal{A}^i(z_b) = \mathcal{A}^i_{sat}(z_b)$. Given the monochromatic treatment of the saturation process, this implementation is called steady-state monochromatic MS-GWaM, or shortly STMO.

An integrated treatment of wave breaking proceeds completely in line with the treatment in the transient MS-GWaM, i.e. wave breaking is diagnosed at an altitude $z = z_b$ if (23) is fulfilled there, with the integral taken over all spectral components, so that

$$\frac{2N^4(z_b)}{\bar{\rho}(z_b)}\sum_{i=1}^{N_l}\mathcal{P}^i(z_b) > N^4(z_b) \qquad \text{with} \qquad \mathcal{P}^i(z_b) = \mathcal{A}^i(z_b)\frac{m^{i^2}(z_b)K_h^{i\,2}}{\hat{\omega}^i(z_b)K^{i^2}(z_b)}. \tag{41}$$

Then static stability is reinforced by reducing the wave-action densities via

$$\mathcal{A}^i(z_b) \rightarrow \mathcal{A}^i_{sat}(z_b) = \mathcal{A}^i(z_b)\left(1 - \frac{\tilde{\mathcal{K}}}{c^i_{gz}}K^{i^2}(z_b)\right) \qquad (i = 1, ..., N_l) \tag{42}$$

with

$$\tilde{\mathcal{K}}(z_b) = \frac{\sum_{i=1}^{N_l} \mathcal{P}^i(z_b) - \bar{\rho}(z_b)/2}{\sum_{i=1}^{N_l} \mathcal{P}^i(z_b) K^{i^2}(z_b)/c_{gz}^i}, \tag{43}$$

which is proportional to an altitude dependent turbulent diffusivity. Since diffusivity cannot be applied in terms of time increments in the steady-state framework, wave breaking and the resulting state of stability is reinforced in terms of vertical increments. This in turn introduces c_{gz} in (42) and (43) [cf. (26) and (27)] so that, for a given vertical distance and diffusivity, slowly propagating waves tend to dissipate more than those propagating fast. This implementation is called steady-state MS-GWaM, or shortly ST. Given that it shares the treatment of the GW sources and GW breaking with the transient implementation, the only difference between ST and TR is how the propagation is accounted for, i.e. GW transience.

Both STMO and ST account for the case of multiple wave breaking in the course of the adjustment to the equilibrium profile. This is achieved by calculating the equilibrium profile sequentially from layer to layer, i.e. solving Eq. (39) as

$$\mathcal{A}^i(z_{i-1/2}) = \frac{c_{gz}^i(z_{i+1/2})\mathcal{A}^i(z_{i+1/2})}{c_{gz}^i(z_{i-1/2})} \quad , \quad (i = 1, ..., N_l). \tag{44}$$

This allows for applying the GW saturation Eq. (40) or Eqs. (41)-(26) within the vertical adjustment process and eventually using the replacement $\mathcal{A}^i(z_{i+1/2}) \to \mathcal{A}_{sat}^i(z_{i+1/2})$ if GW breaking is diagnosed at the half-level $z_{i+1/2}$.

As in the transient implementation, the present study does not take corresponding effects on frictional heating and GW energy deposition into account.

5) Mean-flow forcing

After having calculated the equilibrium profile and having taken into account dissipative processes, the total pseudomomentum flux at half-levels is calculated as

$$PMF_{z_{i+1/2}} = \sum_{i=1}^{N_l} \mathbf{k}_h^i c_{gz}^i (z_{i+1/2}) \mathcal{A}^i (z_{i+1/2}), \tag{45}$$

and the GW drag at full-levels is obtained exactly as in the transient case.

c. Stability measures and computational aspects

In order to facilitate GW studies in a large altitude range, our model top within UA-ICON has been set to $150km$. In UA-ICON and ICON in general a sponge layer prevents spurious wave reflections from the model top, based on a Rayleigh damping applied to the vertical wind (Zängl et al. 2015). In the setup used here the bottom of the sponge layer is at $110km$. Several measures had to be taken in MS-GWaM to prevent numerical instabilities in the sponge, due to excessive mean-flow accelerations by insufficiently controlled GW pseudomomentum fluxes.

1) Molecular viscosity

Molecular viscosity, inversely proportional to density, is taken into account by

$$\mathcal{N}^j (t+\Delta t) \quad = \quad \mathcal{N}^j (t) \exp\left(-2K^{j^2}(t)\Delta t \frac{\eta}{\bar{\rho}}\right) \tag{46}$$

with the temperature dependent dynamic viscosity $\eta = \eta(T(z,t))$ based on Sutherland's viscosity law (see e.g., Atkins and Escudier 2013). In the steady-state implementations the same prescription is used, but proceeding from layer to layer, and with $\Delta t = \Delta z / c_{gz}^i$, i.e.,

$$\mathcal{A}^i (z+\Delta z) \quad = \quad \mathcal{A}^i (z) \exp\left(-2K^{i^2}(z) \frac{\Delta z}{c_{gz}^i} \frac{\eta}{\bar{\rho}}\right). \tag{47}$$

2) Scale height correction

The WKB theory applied by Achatz et al. (2017) predicts that in case of a clear scale separation between GWs and a resolved flow, the former obey the Boussinesq GW dispersion relation (1). In the numerical implementation (i.e. TR, ST, STMO), however, the scale separation does not always hold. Vertical GW wavelengths can grow by refraction and eventually assume values similar to the scales of vertical variations of the resolved mean flow. An ideal treatment of such a situation would be to somehow "transfer" the large scale GW to the resolved flow and stop treating it as a subgrid-scale wave. A theory for such a procedure, however, is not known to us, and the problem is complicated further by the possibility of such a wave still being unresolved in the horizontal.

Based on our experience during the implementation of TR, the primary problem arising from large vertical GW scales is that using (1) the vertical group velocity gets too large and leads to excessively strong pseudomomentum fluxes $PMF = \int d^3k c_{gz} \mathbf{k}_h \mathcal{N}$. This is now avoided by using

$$\hat{\omega} = \pm\sqrt{\frac{N^2 K_h^2 + f^2(m^2 + \Gamma^2)}{K^2 + \Gamma^2}} \tag{48}$$

in all calculations, where $\Gamma = \frac{1}{2H_\rho} - \frac{1}{H_\theta}$ is the pseudo-incompressible scale height with $H_\rho = -(\frac{1}{\bar{\rho}}\frac{d\bar{\rho}}{dz})^{-1}$, $H_\theta = (\frac{1}{\theta}\frac{d\theta}{dz})^{-1}$ being the density scale height and the potential temperature scale height respectively. By assuming that the atmosphere is locally isothermal, the pseudo- incompressible scale height in TR, ST, STMO is further simplified to $\Gamma = \frac{1}{2H_\rho}\left(\frac{1}{2} - \frac{R}{c_p}\right)$ with R being the ideal gas constant and c_p being the specific heat capacity at constant pressure. From the modified dispersion relation one obtains for the vertical group velocity

$$c_{gz} = -\frac{m(\hat{\omega}^2 - f^2)}{\hat{\omega}(K^2 + \Gamma^2)} \tag{49}$$

and the prognostic equation for m becomes

$$\frac{D_r m}{Dt} = -\mathbf{k}_h \cdot \frac{\partial \mathbf{U}}{\partial z} - \frac{1}{\hat{\omega}(K^2 + \Gamma^2)}\left[N K_h^2 \frac{\partial N}{\partial z} + (f^2 - \hat{\omega}^2)\Gamma\frac{\partial \Gamma}{\partial z}\right]. \tag{50}$$

The scale height correction begins to matter at vertical GW scales where the squared vertical wavenumber m^2 becomes small enough so that it is only an order of magnitude larger than Γ, i.e. when $0.1 \times m^2 \sim \Gamma^2$. By substituting $m = 2\pi/\lambda_z, H_\rho \approx 5-8km, R/c_p \approx 2/7$, one arrives at a vertical wavelength of $\lambda_{zcorr} \approx 45-75km$. Although, in the launch spectrum used in this paper, the vertical wavelength is not larger than $\lambda_z = 8km$, the transient implementation has been observed to lead to vertical GW scales as large as λ_{zcorr}. Thus, to be on the safe side, and for being fully consistent between the transient and steady-state schemes, all implementations of MS-GWaM have been based on the discretized versions of Eqs. (48)-(50) to describe the evolution of the GW field.

3) Pseudomomentum-flux smoothing

In the TR implementation, due to unavoidable local under-sampling of ray volumes, pseudomomentum-flux profiles can get noisy so that the GW impact on the resolved flow can exhibit undesired spikes. Thus, a crucial numerical aspect to stabilize TR simulations has been to apply a vertical smoothing on the pseudomomentum fluxes after the projection Eq. (22) and before calculating the resolved wind tendencies. The smoothing is using the zeroth-order filter of Shapiro (1975), which removes noise with length scales of $2\delta z$ but leaves larger-scale structures rather unaffected.

4) Controlling the total number of ray volumes

In order to prevent excessive computational costs, the total number N_c of ray volumes per column is limited to a value N_{cmax}. It has been found that in terms of the time-averaged zonal-mean circulation, a numerical convergence of the TR simulations has been achieved by using $N_{cmax} = 2500$ with $N_l = 4 \times n_c \times n_\omega = 4 \times 6 \times 2 = 48 \sim 50$ (see section 4.b.4). The practical implementation is simple: each time step before the call to the ray-volume emission at the launch level, it is checked column-wise whether $N_c > N_{cmax}$. If this is the case in a column, $N_c - N_{cmax}$ of lowest-energy ray volumes are removed.

5) Computational costs

Table 1 shows the computational costs of TR, ST, STMO and the operational GW drag scheme used in ICON for NWP purposes (Orr et al. 2010). The computational costs are presented in terms of i) t_{tot}, i.e. total run-times of 1-month simulations with UA-ICON using the different GW schemes (see Table caption for the grid spacing) and ii) t_{av}, i.e. average time spent on a single call of the subroutines corresponding to the different parameterizations. The TR scheme is ~ 5 times more expensive than ST in terms of t_{av}, which leads to about a factor ~ 2.5 of overhead costs in terms of t_{tot}. This is what transience costs. If the wave breaking scheme is monochromatic (STMO), and as such simpler, accelerations by a factor ~ 2.3 in terms of t_{av}, and by a factor ~ 1.3 in terms of t_{tot} can be achieved. There is a further acceleration by the factors ~ 4.5 and ~ 1.2 between STMO and the operationally used (Orr et al. 2010) scheme in terms of t_{av} and t_{tot}, respectively. Hence TR is ~ 50 (~ 4.1) more costly than the operational scheme in terms of t_{av} (t_{tot}).

Regarding the costs in memory, TR simulations use by 2% more memory than ST simulations, where 100% stands for the memory cost of the ST simulations. This is not negligible but small.

4. Results

a. Experimental setup

The first step in order to validate the implementation of MS-GWaM was to reproduce the idealized 1D cases of Bölöni et al. (2016) in UA-ICON. This technical step has been followed by global simulations using TR, ST, STMO, with a horizontal grid spacing of $\sim 160km$ (R2B4 grid[5]). A stretched vertical grid has been used with layer thicknesses gradually increasing with height, with a typical thickness of a few tens of meters in the boundary layer, $700-1500m$ in the stratosphere and a maximum of $\sim 4km$ in the lower thermosphere. Similarly to Borchert et al. (2019), a model top at $150km$ has been used with a sponge layer acting above $110km$. As initial condition, operational IFS/ECMWF[6] analyses have been used. They have been interpolated to the ICON grid at altitudes covered by IFS/ECMWF and extrapolated towards a simple climatology above. The first few weeks simulations after each initialization have been discarded from any scientific analysis in order to make sure that the adjustment process from the climatology towards the actual realization of the circulation at higher altitudes is excluded.

b. Mean circulation

The first proof of concept for MS-GWaM in a global modeling framework was to validate the zonal-mean circulation it generates as coupled to UA-ICON. For this validation the URAP data (Swinbank and Ortland 2003) have been used as a reference, because this data set involves zonal-mean climatologies up to rather high altitudes, i.e. $85km$ for temperature and $110km$ for zonal wind.

1) Zonal-mean wind and temperature

Simulations with TR, ST and STMO have been run for the 8 URAP years (1991–1998) for December and June (initialized on the 1st of November and 1st of May respectively). Figure 5 (figure 6) shows the time averaged zonal-mean zonal winds (temperatures) from the TR and ST

[5]for the ICON model, by RnBk a global grid is denoted that originates from an icosahedron whose edges have been initially divided into n parts, followed by k subsequent edge bisections

[6]Integrated Forecasting System of the European Centre for Medium-Range Weather Forecasts, https://www.ecmwf.int/en/publications/ifs-documentation

simulations as well as the reference URAP data. In general, both the TR and ST simulations produce a very similar zonal-mean circulation (results from the STMO simulations are not shown due to their similarity to ST), which compares reasonably well with URAP. Both models capture the reversal of the summer mesospheric jet although somewhat too low in altitude both in December and June and too weak in June. The corresponding summer mesopause is by $\sim 20-30K$ too cold, which might be explained by the fact that the thermal effects of energy deposition (e.g. Becker 2017) of GWs are ignored for the time being. The polar night jet is reasonably well placed, but its magnitude is overestimated in both the TR and ST simulations in both months, especially in June. The stratospheric easterly jet cores are placed too much equatorward in both models and both months. Based on this qualitative comparison, the similarity between the TR and ST suggests that transience does not play a very important role in terms of seasonal-mean and zonal-mean circulation. This does not come as a surprise, as indeed, spatial and time averaging should hide local, short-lived transient effects and eventually reflect a quasi steady-state circulation of the respective months. At a second look, a sharp eye will spot already non-negligible differences between TR and ST simulations in Figs.5 and 6. For instance, the magnitude of the June polar night jet is overestimated to a larger extent, while the June lower-thermospheric jet magnitude in Northern Hemisphere (NH) is underestimated to a larger extent in the ST simulation.

2) Residual circulation and zonal-mean GW drag

The residual circulation of UA-ICON with MS-GWaM is presented in Fig.7 by plotting the residual-mean mass streamfunction along with the corresponding meridional velocity (v^*) in the transformed Eulerian mean (TEM) equations (Andrews and McIntyre 1978b; Hardiman et al. 2010, Eq. (19)). Both TR and ST simulations result in a qualitatively similar circulation as presented by, e.g., Smith (2012) and Becker (2017). It appears that ST simulations lead to a stronger v^* in the upper mesosphere in comparison with TR simulations, implying that the vertical branch of the residual circulation near the poles is also stronger in ST simulations. This is in line with the somewhat colder temperatures at the summer mesopause regions in ST simulations as compared to TR simulations, because a stronger residual circulation corresponds to stronger cooling in the summer upper mesosphere and heating in the winter lower mesosphere. The difference in the residual circulations of TR and ST can be explained by the zonal-mean GW drag (Fig.8) from both simulations. The structure of GW drag is well matched with that of v^* in the mesosphere, demonstrating the impact of GWs on the residual circulation. The GW drag of ST in the MLT is larger than that of TR by $\sim 80ms^{-1}day^{-1}$ ($\sim 160ms^{-1}day^{-1}$) in December (June). This corresponds well to the differences found in the strength of the residual circulation between TR and ST and reflects that adding transience to a GWP has important implications on the mean circulation and the heat budget.

3) Perpetual runs

A more comprehensive appreciation of the differences between the simulations with the transient and the steady-state GW schemes has been enabled by running a perpetual December simulation with TR, ST and STMO for a longer duration. The perpetual run has been achieved by imposing a constant radiational and surface forcing, corresponding to December 22nd 1992, including a diurnal cycle. The simulations have been run for 24 months of which the last 12 months have been used for comparison. Mean wind differences between the TR and ST simulations (Fig.9a) are larger in magnitude, and more statistically significant, than those between the two steady-state simulations ST and STMO (Fig.9b). This shows that the impact of GW transience is somewhat larger than that of the change in the saturation scheme between ST and STMO (see section 3.b.4), even in the context of the time averaged zonal-mean circulation.

4) Numerical convergence

As a validation of the employed maximum number of ray volumes per column $N_{cmax} = 2500$, we show in figure 9c the mean-wind difference between perpetual December TR simulations using $N_{cmax} = 2500$ and $N_{cmax} = 5000$. These differences are clearly lower in magnitude and less statistically significant than those between the ST and TR. This demonstrates that the effect of transience is much larger than the effect of doubling the amount of ray volumes in the TR simulations. It confirms both that the TR simulations using $N_{cmax} = 2500$ are numerically converged and that the difference due to transient GW propagation (ST-TR) is robust, i.e. it reflects a physical feature and not a numerical uncertainty.

c. GW pseudomomentum fluxes

Apart from the time averaged zonal-mean circulation, temporal- and spatial variability of the GW pseudomomentum fluxes is of interest. As will be shown, the modulation of the GW spectrum through transient propagation leads to fundamentally different pseudomomentum-flux magnitudes and spatial structures as compared to the steady-state GW schemes.

1) Intermittency and variability

A simple quantification of GW intermittency is the histogram of pseudomomentum fluxes, i.e. the probability of occurrence of various pseudomomentum-flux values at given geographical locations. Following Hertzog et al. (2012), histograms of GW absolute zonal pseudomomentum fluxes have been plotted for TR, ST (Fig.10) and STMO (not shown but similar to ST) with a similar spatial and temporal sampling as in the above-mentioned paper (see figure captions). The difference between TR and ST is obvious showing a much better fit of the TR simulations to the observed histograms based on the Vorcore superpressure balloons and the HIRDLS satellite (see

Fig.2 in Hertzog et al. (2012)). The low intermittency of the ST simulations is not surprising, since steady-state schemes with a non-intermittent source - such as used here - are known to underestimate the occurrence of high pseudomomentum fluxes. Due to the fact that in the steady-state approximation only dissipative effects - due to wave breaking or close to critical layers - can lead to pseudomomentum-flux variations, no higher values can occur than the launch-level pseudomomentum-flux magnitudes. With the GW source used in this study, the launch-level absolute zonal pseudomomentum-flux magnitude in October is $\sim 4mPa$. Indeed, in the ST simulations no higher values occur than that. In contrast, in the TR simulations at $z \approx 20km$, pseudomomentum fluxes of $60mPa$ occur with a non-zero probability which means that fluxes happen to grow by a factor 15 at this altitude with respect to their launch values. Figure 10 also shows that up to flux values of $\sim 30mPa$, the probability of large fluxes decreases with altitude, which is in line with the findings of e.g. de la Cámara et al. (2016) in this respect. The probability of occurrence for flux values larger than $\sim 30mPa$ shows a vertical dependence that has never been found in steady-state GWPs: it is increasing with altitude between $z \approx 20km$ and $z \approx 40km$, and then it drops down significantly above.

In order to understand the vertical dependence of GW intermittency in the TR simulations and to further illustrate the large difference between the TR and ST simulations, Hovmöller diagrams of absolute pseudomomentum fluxes are shown in Fig.11. Obviously, in the ST simulation pseudomomentum-flux magnitudes decrease monotonically with altitude, while in the TR simulation slanted stripes of increased values with time and altitude demonstrate that GW packets gain pseudomomentum flux in a non-dissipative manner in the course of their propagation up to the altitude of $50-70km$ and then they dissipate due to saturation. The only way the non-dissipative increase can happen - $\mathbf{k_h}$ being constant - is via variation of $\int dmc_{gz}\mathcal{N}$. This can originate from local variations of $\mathcal{N}$ via Eq. (17), from a local increase of the vertical group velocity via Eq.

(20), or by both effects together. Because the Hovmöller diagrams and the histograms in Fig.10 have been sampled for locations beneath the southern hemispheric polar night jet, it is to be expected that variations described by the non-dissipative direct GW-mean-flow interaction Eqs. (17) - (20) are primarily driven by a resolved wind shear $\partial U/\partial z > 0$ between the launch level and $z \approx 55-60km$. Then, given $m < 0$, this shear tends to shift westward (eastward) propagating GWs towards larger (smaller) vertical wavelengths and thus larger (smaller) vertical group velocities in this altitude range. Hence the effect of local vertical group-velocity increase can only play a role for the westward propagating GWs. Separate Hovmöller diagrams for the westward and eastward pseudomomentum fluxes (not shown), however, reflect similar levels of intermittency as for the absolute values (Fig.11). Thus, the variation of $\int dm\mathcal{N}$ seems to be the dominant cause, which is the process that can act only in transient dynamics, while group velocity variations are also possible in a steady-state framework. A critical reader will note that we are here at the limits of WKB theory. While WKB assumes the time scale of the wave amplitudes to be significantly longer than the GW periods, this is not really the case here. Had we only the derivations of the theory on paper this would be a worry. Fortunately, however, we know from comparisons between wave-resolving simulations and transient MS-GWaM that the WKB theory still works surprisingly well even in this range. Bölöni et al. (2016) show that their WKB code - a "toy-model" version of transient MS-GWaM - can reproduce GW behavior at reflection and critical levels, and it also shows similar short-time-scale variations of GW energy (their Fig. 6) that are strictly beyond the validity of WKB but still in good agreement with the wave-resolving LES. Hence one can have confidence in the simulated wave packets that we are seeing here. The question arises now over which time scales these transient effects survive and thus make a difference as compared to steady-state schemes.

2) Zonal and time mean

An interesting consequence of direct GW-mean-flow interactions is that the non-dissipative pseudomomentum-flux convergence is reflected not only locally and for short periods, but also in the time averaged zonal mean. This is illustrated in Figs.12a-f, where monthly-mean (Octobers of 1991-1998) zonal-mean pseudomomentum fluxes from TR simulations turn out to be larger than those obtained from ST simulations everywhere below $z \sim 40km$. This is the mean effect of the transient flux changes shown in Fig. 11, which - as explained above - should be due to local variations of $\int dm\mathcal{N}$. All this suggests that transient effects do not average out completely even in the zonal-mean over monthly time-scales, or in other words, the steady-state approximation does not really hold even over these time-scales.

The pseudomomentum-flux differences between the TR and ST simulations can be put in the context of the missing drag - a general underestimation of the GW forcing at about 60°S by GCMs (McLandress et al. 2012). In particular, Jewtoukoff et al. (2015) showed that the relatively high-resolution operational IFS/ECMWF analyses are underestimating the GW momentum fluxes by a factor 5 over the Southern Ocean at about $20km$ altitude, as compared to superpressure balloon observations. Additionally, de la Cámara et al. (2016) showed that the parameterized GW fluxes in the LMDz[7] model agree with those resolved by the operational IFS/ECMWF to a good degree, indicating that some state-of-the-art GCMs suffer from an underestimation of GW pseudomomentum fluxes by about a factor 5. Several studies suggested that part of this underestimation originates from the lack of orographic drag due to small islands not represented in the topographic databases of GCMs (McLandress et al. 2012; Alexander et al. 2009; Alexander and Grimsdell 2013; Garfinkel and Oman 2018) and others proposed that some of the underestimation is due to the lack of horizontal GW propagation in GWPs (Sato et al. 2009; Ehard et al. 2017) or

[7]Laboratoire de Météorologie Dynamique Zoom

the misrepresentation of non-orographic sources (Hendricks et al. 2014; de la Cámara et al. 2016). It appears, however, that the lack of transience in present-day GWP might also be responsible for a small but non-negligible fraction of the missing drag. This is demonstrated in Figs.12g-i, where horizontal maps of absolute pseudomomentum fluxes are plotted at $z \approx 20km$ above the Southern Ocean from the TR (Fig.12g) and the ST (Fig.12h) simulations. As expected from the cross sections in Figs.12c,f, the absolute pseudomomentum-flux values from the ST simulations are smaller than those from TR, and as shown in Fig.12i, if ST fluxes are multiplied by a factor of 1.3, a relatively close match with TR is achieved. Transience thus brings an increase of 30% in terms of absolute momentum fluxes, which is however still very far from the missing 500% reported by e.g. Jewtoukoff et al. (2015). The difference between TR and ST simulations can also be expressed in terms of the zonal GW drag. The drag averaged over $\varphi \in [-65°, -55°], z \in [20, 50]km$ and over the Octobers of 1991-1998 is $-0.291 ms^{-1} day^{-1}$ from ST and $-0.476 ms^{-1} day^{-1}$ from TR simulations. Hence transience seems to increase the drag by about 60%. The non-negligible effect discussed above shows that the transience does matter even over monthly time-scales. It is also recalled that all differences presented between ST and TR simulations in this paper are due to the non-orographic fluxes only, in a completely non-intermittent GW source setup.

d. Contribution of different wavelengths to the GW signal

Given that MS-GWaM is a spectral scheme, a decomposition of the GW momentum fluxes and drag into the contributions from different wavelengths is straightforward. Such a decomposition could be of interest for validation purposes against observations if GW sources were realistically taken into account. This is yet not the case here, however even with the simple GW source used in this study, a decomposition by scales is useful to get a simple first guess about the required horizontal and vertical resolutions for GW resolving simulations. The decomposition is based on the TR implementation of MS-GWaM given its additional realism as compared to ST, i.e. given transience and the prognostic treatment of the vertical wavenumber spectrum. The contribution of GWs with different spatial scales to the pseudomomentum fluxes has been diagnosed by calculating Eq. (22) for a subset of the ray volumes $j = 1, \ldots, N_i$ for which certain conditions hold with respect to their horizontal (λ_h) or vertical wavelengths (λ_z). The corresponding drag contribution has been calculated via Eq. (18) (its discretized form) just like for the full drag. These diagnostics have all been achieved in an off-line mode, meaning that the resolved flow has been forced with the total drag imposed by the total pseudomomentum fluxes.

1) Decomposition results

The contribution of GWs with different spatial scales to the total absolute flux and drag is shown in Fig.13. Panel a) shows the zonal-mean total absolute pseudomomentum flux and drag averaged over Junes of 1991-1998. Panels b), c), d), e) suggest that excluding horizontal wavelengths smaller than $50km, 100km$, $200km$ and $250km$ leads to signal losses of $\sim 20\%, 50\%$, 75% and $75-80\%$, respectively, both with respect to fluxes and the drag. The contribution of GWs with different vertical scales can be seen by comparing the total signal with panels f), g), h), i) where vertical wavelengths smaller than $1km, 2km, 5km$ and $10km$ are excluded respectively. Here the drag signal is much less affected, namely no loss can be seen if having contributions from waves with $\lambda_z > 5km$, and only $\sim 25\%$ is lost if waves with $\lambda_z < 10km$ are excluded. In terms of fluxes however, the loss of signal below $z \approx 40km$ is larger than $\sim 50\%$ if GWs with $\lambda_z < 2km$ are excluded, which increases to a loss of $\sim 75-80\%$ if GWs with $\lambda_z < 10km$ do not contribute.

The contribution of GWs with different scales to the total intermittency has been examined as well (not shown). It turns out that at $z = 20km$, occurrence of large momentum fluxes ($\gtrsim 10mPa$) is completely lost if waves with horizontal scales smaller than $100km$ are excluded, leading to similarly unrealistic intermittency curves as obtained from ST simulations. If only small scale GWs with horizontal scales $\lambda_h < 50km$ are left out, most of the total intermittency is reproduced leaving us with a loss of at most $\sim 30\%$ for all flux values. Excluding the smallest vertical scales ($\lambda_z < 2km$) does not affect intermittency but leaving out even larger scale waves ($\lambda_z < 5$ or $10km$) reduces the occurrence of fluxes between 5 and $40mPa$ significantly.

2) Consequences for GW resolving simulations

Explicitly resolving GWs instead of parameterizing them is recently of increasing interest even in global simulations. In the light of the above, a simple estimate of the required spatial resolution can be given: in order to get most of the GW signal one needs to resolve horizontal scales of $50km$ or smaller and vertical scales of $2km$ or smaller. Because in NWP models and GCMs the effective resolution of a given spatial scale λ requires $7-10$ grid points per λ, the necessary horizontal (vertical) grid spacing to be used for GW resolving simulations can be estimated as $\Delta x < 5km$ ($\Delta z < 200m$). This estimate has to be treated with caution as it does not take into account that resolving the generation (GW source mechanisms) and dissipation of GWs might require even higher spatial resolution than suggested by the scale decomposition applied here.

5. Summary and Conclusions

This paper describes the first implementation - to the best of our knowledge - of a transient subgrid-scale GW parameterization into a state-of-the-art GCM and NWP model. This parameterization is called Multi-Scale Gravity-Wave Model (MS-GWaM). It does not rely on the steady-state approximation, and therefore enables both dissipative and non-dissipative direct GW-mean-flow interactions, while standard GW parameterizations assume an instantaneous equilibrium between GWs, mean flow, and sources, thereby leaving room only for dissipative forcing. For an estimate of the GW-transience impact, a steady-state version of MS-GWaM (ST), using exactly the same GW saturation scheme, and coupled to the same GW source, has been implemented and used as reference for the transient GW parameterization (TR). The TR implementation of MS-GWaM differs in several respects from other GWPs in the literature that use ray tracing. Song and Chun (2008) as well as Amemiya and Sato (2016) have implemented related GWPs into state-of-the-art GCMs. They have, however, kept the steady-state assumption in the prediction of the wave amplitudes via wave-action conservation. As compared to earlier transient implementations (Senf and Achatz 2011; Ribstein et al. 2015), one main difference is that TR MS-GWaM allows a feedback from the resolved mean flow to the subgrid-scale GW field through the ray equations, which is especially not the case for Senf and Achatz (2011). Also MS-GWaM applies the phase-space representation (section 2 b), which, so far, is the only viable solution to avoid numerical problems arising due to caustics. Ribstein and Achatz (2016) already used a fully coupled ray-tracer including the phase-space approach, however not in a GCM but in a more simple tidal model, similarly to Senf and Achatz (2011). Last but not least, the wave breaking scheme of TR MS-GWaM is also a point that makes an important difference with respect to other GWPs, in that the saturation is diagnosed with a contribution from the full GW spectrum represented by the parameterization at a given altitude at a given time.

The time averaged zonal-mean circulation turned out to be broadly similar in TR and ST simulations, both of them agreeing reasonably well with observations (URAP data by Swinbank and Ortland (2003)). Closer inspection shows, however, that in some aspects TR yields slightly better results than ST. By excluding interannual variability via perpetual runs, it has also been shown that the effect of transience is larger than that of varying the saturation scheme in the steady-state implementation, especially in the mesosphere and lower thermosphere. That the summer mesopauses are too cold both in TR and ST simulations is likely a consequence of ignoring thermal effects of energy deposition by GWs. Having a leading order thermal effect in the MLT (e.g. Becker 2017), this process will have to be included into MS-GWaM. Another finding in the same context is that temperature errors at summer mesopauses are smaller in TR simulations than in ST simulations, which is explained by the weaker residual circulation driven by weaker zonal-mean net GW drag in the MLT region. This is a sign that transient effects do not average out completely and may have important implications on the mean zonal and meridional circulations.

Even more evident differences between TR and ST simulations are found in terms of GW pseudomomentum-flux variability. As expected from earlier studies (e.g., de la Cámara et al. 2016), ST simulations strongly underestimate the intermittency of GW pseudomomentum fluxes (occasional occurrence of large values), while TR simulations lead to considerably more realistic results. The reason for this is that the steady-state assumption only allows dissipative effects to change GW-pseudomomentum fluxes, and hence only allows them to decrease as compared to the source, while non-dissipative direct GW-mean-flow interactions can also lead to an increase of these fluxes. This effect is not only visible locally and over short time-scales, but it also affects monthly averages of zonal means: mean pseudomomentum fluxes in the lower stratosphere are $\sim 30\%$ larger in TR simulations than in ST. In the SH, this is where a missing GW drag has been diagnosed by several studies (McLandress et al. 2012; Jewtoukoff et al. 2015). Hence the neglect of transient GW-mean-flow interactions in standard GW parameterizations might contribute to this issue in a modest extent, beside the lack of lateral propagation (Sato et al. 2009; Ehard et al. 2017), the misrepresentation of non-orographic sources (Hendricks et al. 2014; de la Cámara et al. 2016) or the lack of orographic drag due to missing islands in the insufficiently detailed model topographies (Alexander et al. 2009; McLandress et al. 2012; Alexander and Grimsdell 2013; Garfinkel and Oman 2018).

Increasing the realism of GW parameterizations by including transient wave-mean-flow interactions is seen by us as only a first step. Lateral GW propagation will have to be included as well, which - based on Senf and Achatz (2011); Kalisch et al. (2014); Ribstein et al. (2015); Amemiya and Sato (2016) - changes several aspects of the GW distribution and its impact on the mean flow. Corresponding work has just begun, after a 6D[8] version of MS-GWaM has been successfully implemented into the same f-plane pseudo-incompressible flow solver as used by Bölöni et al. (2016) and Wei et al. (2019). More realistic source schemes are also an issue. In part II of this study we report on the effects of coupling of MS-GWaM in ICON to a convective GW-source scheme, and more improvements with regard to mountain waves and GWs due to jets and fronts will have to follow. Finally, as pointed out by Plougonven et al. (2019) one should always be aware that a realistically looking large-scale circulation is no proof that the parameterization is correct. Instead the parameterized processes will have to be studied by measurements and wave-resolving simulations as well, and it will have to be made sure that all parts of the parameterization reproduce the properties identified therein. Only then can we have a guarantee that the GWP will be reliable even in a changing climate.

The reader might wonder whether the computational cost of a Lagrangian ray-tracing approach as suggested here is not too overwhelming. As summarized in Table 1 and in section 3.c.5, according to the strictest measure (t_{av}), including transient effects increases the computational costs by a factor ~ 5 with respect to the ST implementation of MS-GWaM, and by a factor ~ 50 with respect to Orr et al. (2010) - the current operational scheme used in the NWP configuration of ICON. The discrepancy in computational costs by a factor ~ 10 between the steady-state scheme ST and Orr et al. (2010) - which should perform calculations of similar complexity - suggests that MS-GWaM's efficiency in general (both TR and ST) could probably be improved by means

[8]3+3 dimensions in physical and spectral space

of code optimization. Based on this assumption, an optimized transient MS-GWaM should be about factor ~ 5 more costly than state-of-the-art GW schemes. For the time being it cannot be excluded that lateral GW propagation might increase the costs further, although there is no reason to expect that more ray volumes will be needed per column than already used in the present MS-GWaM implementation. Keeping in mind other potential overhead costs, e.g. such as the MPI communication of ray volumes, a safe estimate for a 6D version of MS-GWaM is a factor ~ 10 increase of computational costs, as compared to standard steady-state GW parameterizations. This might be seen as a large increase in costs, however, relating it to costs of other alternatives - such as GW-resolving simulations - might quickly change one's perspective. As also suggested in section 4.d, GW resolving simulations would require a horizontal grid-spacing of $5km$ (or smaller, e.g. $1km$) and a vertical grid-spacing of $200m$. If this requirement was satisfied with respect to the horizontal resolution alone, the computational costs (in terms of t_{tot}) would increase by a factor of ~ 30 thousand (~ 5 million) respectively. The vertical resolution increase to $200m$ everywhere above the troposphere would lead to a cost increase of further factor ~ 8 ending up with something between factor 240 thousand and 40 million. Therefore, already in its present state, ICON/MS-GWaM can be a useful tool for research purposes, allowing much less costly simulations than those resolving GWs globally and more realistic than achievable by standard GCM resolutions with classic steady-state parameterizations. Once flow-dependent sources for GWs from orography and jet-frontal systems have been implemented, it will be ready, e.g., to accompany field campaigns and help interpreting their results. The long-term goal of eventually using ICON/MS-GWaM in climate simulations and weather forecasting, however, is also not to be left out of sight.

Data availability statement. The ICON-Software is freely available to the scientific community for non-commercial research purposes under a license of DWD and MPI-M. If you would like to obtain

ICON, please contact icon@dwd.de. The MS-GWaM code and its module for an implementation in ICON have been developed at Goethe-Universität Frankfurt am Main. Please contact Prof. Ulrich Achatz (achatz@iau.uni-frankfurt.de) for these. The URAP wind and temperature data are available at https://www.sparc-climate.org/data-centre/data-access/reference-climatology/urap/ and ERA5 reanalysis data are accessible at https://cds.climate.copernicus.eu.

Acknowledgments. The authors thank the German Research Foundation (DFG) for partial support through the research unit *Multiscale Dynamics of Gravity Waves (MS-GWaves)* and through Grants AC 71/8-2, AC 71/9-2, AC 71/10-2, AC 71/11-2, AC 71/12-2, BO 5071/2-2, BO 5071/1-2, and ZA 268/10-2. Calculations for this research were conducted on the supercomputer facilities of the Center for Scientific Computing (CSC) of the Goethe University Frankfurt. This work also used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bb1097.

References

Achatz, U., B. Ribstein, F. Senf, and R. Klein, 2017: The interaction between synoptic-scale balanced flow and a finite-amplitude mesoscale wave field throughout all atmospheric layers: weak and moderately strong stratification. *Quart. J. Roy. Meteor. Soc.*, **143**, 342–361, doi: 10.1002/qj.2926.

Alexander, M. J., and T. J. Dunkerton, 1999: A spectral parameterization of mean-flow forcing due to breaking gravity waves. *J. Atmos. Sci.*, **56**, 4167–4182, doi:10.1175/1520-0469(1999)056<4167:ASPOMF>2.0.CO;2.

Alexander, M. J., S. D. Eckermann, D. Broutman, and J. Ma, 2009: Momentum flux estimates for South Georgia Island mountain waves in the stratosphere observed via satellite. *Geophys. Res.*

Lett., **36**, L12 816, doi:10.1029/2009GL038587.

Alexander, M. J., and A. W. Grimsdell, 2013: Seasonal cycle of orographic gravity wave occurrence above small islands in the Southern Hemisphere: Implications for effects on the general circulation. *J. Geophys. Res. Atmos.*, **118**, 11 589–11 599, doi:10.1002/2013JD020526.

Amemiya, A., and K. Sato, 2016: A new gravity wave parameterization including three-dimensional propagation. *J. Met. Soc. Japan*, **94**, 237–256, doi:10.2151/jmsj.2016-013.

Andrews, D., and M. McIntyre, 1976: Planetary Waves in Horizontal and Vertical Shear: The Generalized Eliassen-Palm Relation and the Mean Zonal Acceleration. *J. Atmos. Sci.*, **33**, 2031–2048, doi:10.1175/1520-0469(1976)033<2031:pwihav>2.0.co;2.

Andrews, D., and M. McIntyre, 1978a: An exact theory of nonlinear waves on a Lagrangian-mean flow. *J. Fluid Mech.*, **89**, 609–646.

Andrews, D. G., and M. E. McIntyre, 1978b: Generalized eliassen-palm and charney-drazin theorems for waves on axismmetric mean flows in compressible atmospheres. *J. Atmos. Sci.*, **35**, 175–185, doi:10.1175/1520-0469(1978)035<0175:GEPACD>2.0.CO;2.

Atkins, T., and M. Escudier, 2013: A dictionary of mechanical engeneering. *Oxford Univrsity Press*, ISBN: 9780199587438, doi:10.1093/acref/9780199587438.001.0001.

Bacmeister, J., P. Newman, B. Gary, and K. Chan, 1994: An algorithm for forecasting mountain wave-related turbulence in the stratosphere. *Weather and Forecasting*, **9**, 241–253, doi:10.1175/1520-0434(1994)009<0241:AAFFMW>2.0.CO;2.

Becker, E., 2017: Mean-flow effects of thermal tides in the mesosphere and lower thermosphere. *J. Atmos. Sci.*, **74**, 2043–2063, doi:10.1175/JAS-D-16-0194.1.

Beres, J. H., R. R. Garcia, B. A. Boville, and F. Sassi, 2005: Implementation of a gravity wave source spectrum parameterization dependent on the properties of convection in the whole atmosphere community climate model (WACCM). *J. Geophys. Res.*, **110**, D10 108, doi:10.1029/2004JD005504.

Bölöni, G., B. Ribstein, J. Muraschko, C. Sgoff, J. Wei, and U. Achatz, 2016: The interaction between atmospheric gravity waves and large-scale flows: An efficient description beyond the nonacceleration paradig. *J. Atmos. Sci.*, **73**, 4833–4852, doi:10.1175/JAS-D-16-0069.1.

Borchert, S., G. Zhou, M. Baldauf, H. Schmidt, G. Zängl, and D. Reinert, 2019: The upper-atmosphere extension of the ICON general circulation model. *Geosci. Model Dev. Discuss.*, **12**, 3541–3569, doi:10.5194/gmd-2018-289.

Bretherton, F. P., 1966: The propagation of groups of internal gravity waves in a shear flow. *Quart. J. Roy. Meteor. Soc.*, **92**, 466–480, doi:10.1002/qj.49709239403.

Broutman, D., J. Ma, S. D. Eckermann, and J. Lindeman, 2006: Fourier-ray modeling of transient trapped lee waves. *Mon. Wea. Rev.*, **134**, 2849–2856, doi:10.1175/MWR3232.1.

Charron, M., and E. Manzini, 2002: Gravity waves from fronts: Parameterization and middle atmosphere response in a general circulation model. *J. Atmos. Sci.*, **59**, 923–941, doi:10.1175/1520-0469(2002)059<0923:GWFFPA>2.0.CO;2.

de la Cámara, A., and F. Lott, 2015: A parameterization of gravity waves emitted by fronts and jets. *Geophys. Res. Lett.*, **42**, 2071–2078, doi:10.1002/2015GL063298.

de la Cámara, A., F. Lott, V. Jewtoukoff, R. Plougonven, and A. Hertzog, 2016: On the gravity wave forcing during the southern stratospheric final warming in LMDZ. *J. Atmos. Sci.*, **73**, 3213–3226, doi:10.1175/JAS-D-15-0377.1.

Ehard, B., and Coauthors, 2017: Horizontal propagation of largeâĂŘamplitude mountain waves into the polar night jet. *J. Geophys. Res.*, **122**, 1423–1436, doi:10.1002/2016JD025621.

Fritts, D., and T. J. Dunkerton, 1985: Fluxes of heat and constituents due to convectively unstable gravity waves. *J. Atmos. Sci.*, **42**, 549–556, doi:10.1175/1520-0469(1985)042<0549:FOHACD>2.0.CO;2.

Fritts, D., and W. Lu, 1993: Spectral estimates of gravity wave energy and momentum fluxes. part II: Parameterization of wave forcing and variability. *J. Atmos. Sci.*, **50(22)**, 3695–3713, doi:10.1175/1520-0469(1993)050<3695:SEOGWE>2.0.CO;2.

Fritts, D., and T. E. VanZandt, 1993: Spectral estimates of gravity wave energy and momentum fluxes. part I: Energy dissipation, acceleration and constraints. *J. Atmos. Sci.*, **50**, 3685–3694, doi:10.1175/1520-0469(1993)050<3685:SEOGWE>2.0.CO;2.

Garfinkel, C. I., and L. Oman, 2018: Effect of gravity waves from small islands in the Southern Ocean on the Southern Hemisphere atmospheric circulation. *J. Geophys. Res.*, **123(3)**, 1552–1561, doi:10.1002/2017JD027576.

Grimshaw, R., 1975: Nonlinear internal gravity waves in a rotating fluid. *J. Fluid Mech.*, **71**, 497–512, doi:10.1017/S0022112075002704.

Hardiman, S. C., D. G. Andrews, A. A. White, N. Butchart, and I. Edmond, 2010: Using different formulations of the transformed eulerian mean equations and eliassen-palm diagnostics in general circulation models. *J. Atmos. Sci.*, **67**, 1983–1995, doi:10.1175/2010JAS3355.1.

Hendricks, E., J. Doyle, S. Eckermann, Q. Jiang, and P. Reinecke, 2014: What is the source of the stratospheric gravity wave belt in austral winter? *J. Atmos. Sci.*, **71**, 1583–1592, doi: 10.1175/JAS-D-13-0332.1.

Hertzog, A., J. Alexander, and R. Plougonven, 2012: On the intermittency of gravity wave momentum flux in the stratosphere. *J. Atmos. Sci.*, **69**, 3433–3448, doi:10.1175/JAS-D-12-09.1.

Hertzog, A., C. Souprayen, and A. Hauchecorne, 2002: Eikonal simulations for the formation and the maintenance of atmospheric gravity wave spectra. *J. Geophys. Res.*, **107 (D12)**, ACL 4–1–ACL 4–14, doi:10.1029/2001JD000815.

Hines, C. O., 1997a: Doppler spread gravity-wave parametrization of momentum deposition in the middle atmosphere. part 1: Basic formulation. *J. Atmos. Sol. Terr. Phys.*, **59**, 371–386, doi:10.1016/S1364-6826(96)00079-X.

Hines, C. O., 1997b: Doppler spread gravity-wave parametrization of momentum deposition in the middle atmosphere. part 2: Broad spectra and quasi monochromatic spectra, and implementation. *J. Atmos. Sol. Terr. Phys.*, **59**, 387–400, doi:10.1016/S1364-6826(96)00080-6.

Jewtoukoff, V., A. Hertzog, R. Plougonven, A. de la Cámara, and F. Lott, 2015: Comparison of gravity waves in the Southern Hemisphere derived from balloon observations and the ECMWF analyses. *J. Atmos. Sci.*, **72**, 3449–3468, doi:10.1175/JAS-D-14-0324.1.

Kalisch, S., P. Preusse, M. Ern, D. S. Eckermann, and M. Reise, 2014: Differences in gravity wave drag between realistic oblique and assumed vertical propagation. *J. Geophys. Res. Atmos.*, **119**, 10,081–10,099, doi:10.1002/2014JD021779.

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

Kruse, G. C., and R. B. Smith, 2018: Nondissipative and dissipative momentum deposition by mountain wave events in sheared environments. *J. Atmos. Sci.*, **75**, 2721–2740, doi:10.1175/JAS-D-17-0350.1.

Lindzen, R. S., 1981: Turbulence and stress owing to gravity wave and tidal breakdown. *J. Geophys. Res.*, **86**, 9707–9714, doi:10.1029/JC086iC10p09707.

Lott, F., and L. Guez, 2013: A stochastic parameterization of the gravity waves due to convection and its impact on the equatorial stratosphere. *J. Geophys. Res.*, **118**, 8897–8909, doi:10.1002/jgrd.50705,2013.

Lott, F., and M. Miller, 1997: A new subgrid-scale orographic drag parametrization: Its formulation and testing. *Quart. J. Roy. Meteor. Soc.*, **123**, 101–127, doi:10.1002/qj.49712353704.

McLandress, C., T. G. Sheperd, S. Polavarapu, and S. R. Beagley, 2012: Is missing orographic gravity wave drag near 60^o s the cause of the stratospheric zonal winds biases in chemistry–climate models? *J. Atmos. Sci.*, **69**, 802–818, doi:10.1175/JAS-D-11-0159.1.

Medvedev, A. S., and G. P. Klaassen, 1995: Vertical evolution of gravity wave spectra and the parametrization of associated gravity wave drag. *J. Geophys. Res.*, **100**, 25 841–25 854, doi:10.1029/95JD02533.

Muraschko, J., M. Fruman, U. Achatz, S. Hickel, and Y. Toledo, 2015: On the application of Wentzel-Kramer-Brillouin theory for the simulation of the weakly nonlinear dynamics of gravity waves. *Quart. J. Roy. Meteor. Soc.*, **141**, 676–697, doi:10.1002/qj.2381.

Olbers, D., C. Eden, E. Becker, F. Pollmann, and J. Jungclaus, 2019: *The IDEMIX Model: Parameterization of Internal Gravity Waves for Circulation Models of Ocean and Atmosphere,* *In: Energy Transfers in Atmosphere and Ocean*. Isbn: 978-3-030-05704-6 ed., Springer, doi: 10.1007/978-3-030-05704-6_3.

Orr, A., P. Bechtold, J. F. Scinocca, M. Ern, and M. Janiskova, 2010: Improved middle atmosphere climate and forecasts in the ECMWF model through a nonorographic gravity wave drag parameterization. *J. Clim.*, **23**, 5905–5926, doi:10.1175/2010JCLI3490.1.

Palmer, T. N., G. J. Shutts, and R. Swinbank, 1986: Alleviation of a systematic westerly bias in general circulation and numerical weather-prediction models through an orographic gravity wave drag parametrization. *Quart. J. Roy. Meteor. Soc.*, **112**, 1001–1039, doi: 10.1002/qj.49711247406.

Plougonven, R., A. de la Cámara, A. Hertzog, and F. Lott, 2019: How does knowledge of atmospheric gravity waves guide their parameterizations? *Quart. J. Roy. Meteor. Soc.*, **146**, 1529–1543, doi:10.1002/qj.3732.

Ribstein, B., and U. Achatz, 2016: The interaction between gravity waves and solar tides in a linear tidal model with a 4D ray-tracing gravity-wave parameterization. *J. Geophys. Res. Space Physics*, **121**, doi:10.1002/2016JA022478.

Ribstein, B., U. Achatz, and F. Senf, 2015: The interaction between gravity waves and solar tides: Results from 4D ray tracing coupled to a linear tidal model. *J. Geophys. Res.*, **120**, 6795–6817, doi:10.1002/2015JA021349.

Richter, J. H., F. Sassi, and R. R. Garcia, 2010: Towards a physically based gravity wave source parameterization. *J. Atmos. Sci.*, **67**, 136–156, doi:10.1175/2009JAS3112.1.

Sato, K., S. Watanabe, Y. Kawatani, Y. Tomikawa, K. Miyazaki, and M. Takahashi, 2009: On the origin of mesospheric gravity waves. *Geophys. Res. Lett.*, **36**, L19 801, doi:10.1029/2009GL039908.

Scaife, A. A., J. R. Knight, G. K. Vallis, and C. K. Folland, 2005: A stratospheric influence on the winter NAO and North Atlantic surface climate. *Geophys. Res. Lett.*, **32**, (18) L18 715, doi:10.1029/2005GL023226.

Scaife, A. A., and Coauthors, 2012: Climate change projections and stratosphere-troposphere interaction. *Clim. Dyn.*, **38**, 2089–2097, doi:10.1007/s00382-011-1080-7.

Scinocca, J. F., 2003: An accurate spectral nonorographic gravity wave drag parameterization for general circulation models. *J. Atmos. Sci.*, **60**, 667–682, doi:10.1175/1520-0469(2003)060<0667:AASNGW>2.0.CO;2.

Senf, F., and U. Achatz, 2011: n the impact of middle-atmosphere thermal tides on the propagation and dissipationof gravity waves. *J. Geophys. Res.*, **116**, D24 110, doi:10.1029/2011JD015794.

Shapiro, R., 1975: Linear filtering. *Math. of Comp.*, **29**, 1094–1097, doi:10.1090/S0025-5718-1975-0389356-X.

Skamarock, W. C., and Coauthors, 2019: A description of the advanced research WRF version 4. *NCAR Tech. Note,NCAR/TN-556+STR*, 145pp, doi:10.5065/1dfh-6p97.

Smith, A. K., 2012: Global dynamics of the mlt. *Surv. Geophys.*, **33**, 1177–1230, doi:10.1007/s10712-012-9196-9.

Song, I.-S., and H.-Y. Chun, 2005: Momentum flux spectrum of convectively forced internal gravity waves and its application to gravity wave drag parameterization. part I: Theory. *J. Atmos. Sci.*, **62**, 107–124, doi:10.1175/JAS-3363.1.

Song, I.-S., and H.-Y. Chun, 2008: A lagrangian spectral parameterization of gravity wave drag induced by cumulus convection. *J. Atmos. Sci.*, **65**, 1204–1224, doi:10.1175/2007JAS2369.1.

Swinbank, R., and D. A. Ortland, 2003: Compilation of wind data for the (UARS) reference atmosphere project. *J. Geophys. Res.*, **108, D19**, 4615, doi:10.1029/2002JD003135.

VanZandt, T. E., 1982: A universal spectrum of buoyancy waves in the atmosphere. *Geophys. Res. Lett.*, **9**, 575–578, doi:10.1029/GL009i005p00575.

Warner, C. D., and M. E. McIntyre, 1996: On the propagation and dissipation of gravity wave spectra through a realistic middle atmosphere. *J. Atmos. Sci.*, **53**, 3213–3235, doi:10.1175/1520-0469(1996)053<3213:OTPADO>2.0.CO;2.

Wei, J., G. Bölöni, and U. Achatz, 2019: Efficient modelling of the interaction of mesoscale gravity waves with unbalanced large-scale flows: Pseudomomentum-flux convergence versus direct approach. *J. Atmos. Sci.*, **76**, 2715–2738, doi:10.1175/JAS-D-18-0337.1.

Wilhelm, J., T. Akylas, G. Bölöni, J. Wei, B. Ribstein, R. Klein, and U. Achatz, 2018: Interactions between mesoscale and submesoscale gravity waves and their efficient representation in mesoscale-resolving models. *J. Atmos. Sci.*, **75**, 2257–2280, doi:10.1175/JAS-D-17-0289.1.

Zängl, G., D. Reinert, P. Ripodas, and M. Baldauf, 2015: The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quart. J. Roy. Meteor. Soc.*, **141**, 563–579, doi:10.1002/qj.2378.

LIST OF TABLES

Table 1. Computational costs of the different GW parameterizations coupled to UA-ICON on 960 CPUs with a horizontal grid-spacing of $\sim 160km$ (R2B4 grid) and with 120 vertical levels up to $150km$ with the same distribution as described by Borchert et al. (2019). 52

Table 1. Computational costs of the different GW parameterizations coupled to UA-ICON on 960 CPUs with a horizontal grid-spacing of $\sim 160km$ (R2B4 grid) and with 120 vertical levels up to $150km$ with the same distribution as described by Borchert et al. (2019).

Measure	TR	ST	STMO	Orr et al. (2010)
t_{tot} : total runtime for one month simulation [mm:ss]	52:00	20:30	16:00	12:30
t_{av} : average time spent in parameterization subroutines [s]	0,04579	0,00904	0,00389	0,00088

LIST OF FIGURES

Fig. 1. A schematic of the GW field representation by ray volumes in the transient MS-GWaM. a) representation in $z - m$ space (see the text for explanations on how the subgrid-scale GW pseudomomentum fluxes are projected to the vertical grid), b) representation in spectral space, showing that the ray volume center-point m_c and the extent in m-direction changes prognostically. 55

Fig. 2. Latitudinal profile of launch momentum-flux magnitudes $[mPa]$ in the non-orographic GW source at Northern Hemispheric winter (blue) and summer (green) solstices. 56

Fig. 3. A schematic showing the attribution of horizontal wavenumbers to ray volumes in the non-orographic GW source (see the text for more explanation). 57

Fig. 4. A schematic of the representation of the non-orographic GW emission as lower boundary condition in terms of sequential ray-volume launches (see the text for more explanation). . . 58

Fig. 5. Zonal-mean zonal wind $[ms^{-1}]$ averaged over Decembers (a,b,c) and Junes (d,e,f) in the period 1991-1998. URAP data: a,d, TR: b,e, ST: c,f. Solid black isolines are drawn between $-120ms^{-1}$ and $120ms^{-1}$ with a spacing of $30ms^{-1}$. 59

Fig. 6. Zonal-mean temperature $[C^{\circ}]$ averaged over Decembers (a,b,c) and Junes (d,e,f) in the period 1991-1998. URAP data: a,d, TR: b,e, ST: c,f. Solid black isolines are drawn between $-150C^{\circ}$ and $30C^{\circ}$ with a spacing of $30C^{\circ}$. 60

Fig. 7. Residual circulation averaged over Decembers (a,b) and Junes (c,d) in the period 1991-1998. TR: a,c, ST: b,d. Filled contours: residual-mean meridional velocity in the transformed Eulerian mean equations (v^*) $[ms^{-1}]$. Black contour lines: residual mass streamfunction $[kgs^{-1}]$ shown by solid (dashed) lines for positive (negative) values for the magnitudes of $10^4 - 10^{10} kgs^{-1}$ with two contours per 10 times increase. 61

Fig. 8. Zonal-mean net gravity wave drag $[ms^{-1}day^{-1}]$ averaged over Decembers (a,b) and Junes (c,d) in the period 1991-1998. TR: a,c, ST: b,d. Black contours are drawn between the intervals $[-360, -40]$ (dashed) and $[40, 160]$ (solid) with an interval of 40. 62

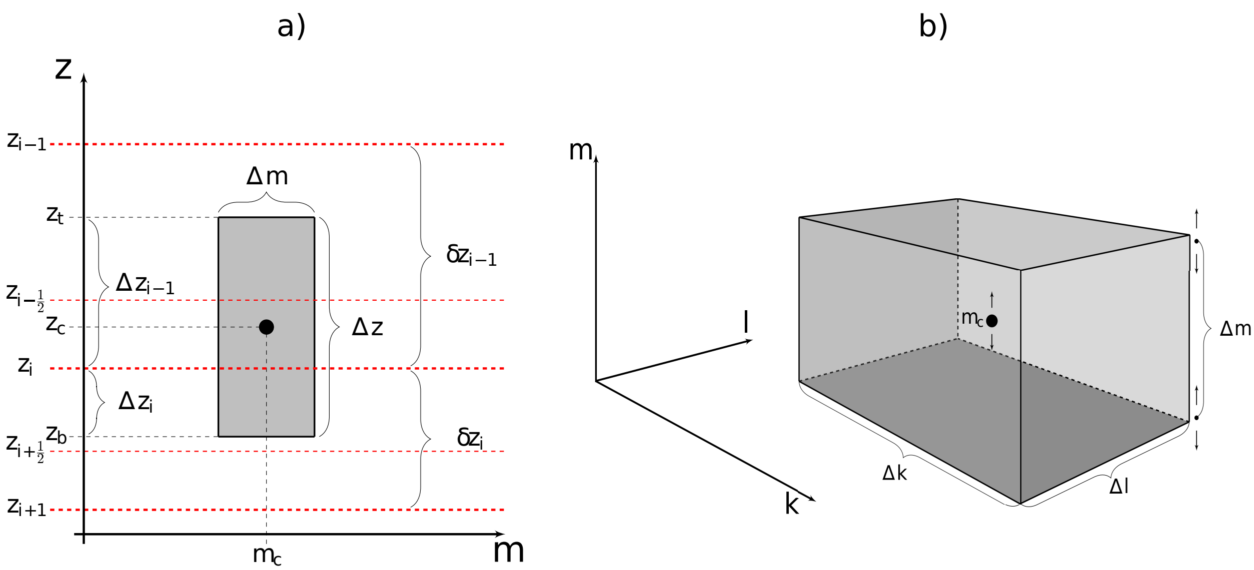
Fig. 9. Zonal-mean zonal wind differences $[ms^{-1}]$ averaged over the last 12 months of a 24-months perpetual December simulation. a: difference ST-TR, b: difference STMO-ST, c: difference TRx2-TR where TRx2 stands for the transient MS-GWaM with a doubled number of ray volumes per column, i.e. $N_{cmax} = 5000$ instead of $N_{cmax} = 2500$. The solid black isolines are drawn for $0ms^{-1}$. Red dots denote points where the differences are statistically significant with a 95% confidence, based on t-tests taking into account time correlation of the sample. The numbers of statistically significant points are 334 for ST-TR, 313 for STMO-ST and 63 for TRx2-TR. 63

Fig. 10. Histogram of absolute zonal pseudomomentum-flux occurrences sampled at $\varphi \in [-65^{\circ}, -50^{\circ}], \lambda \in [-180^{\circ}, 180^{\circ}]$ over the Octobers of 1991-1998 for TR (solid) and ST (dashed) at different altitudes: $z \approx 20km$ (black), $z \approx 40km$ (red), $z \approx 60km$ (blue), $z \approx 80km$ (green). 64

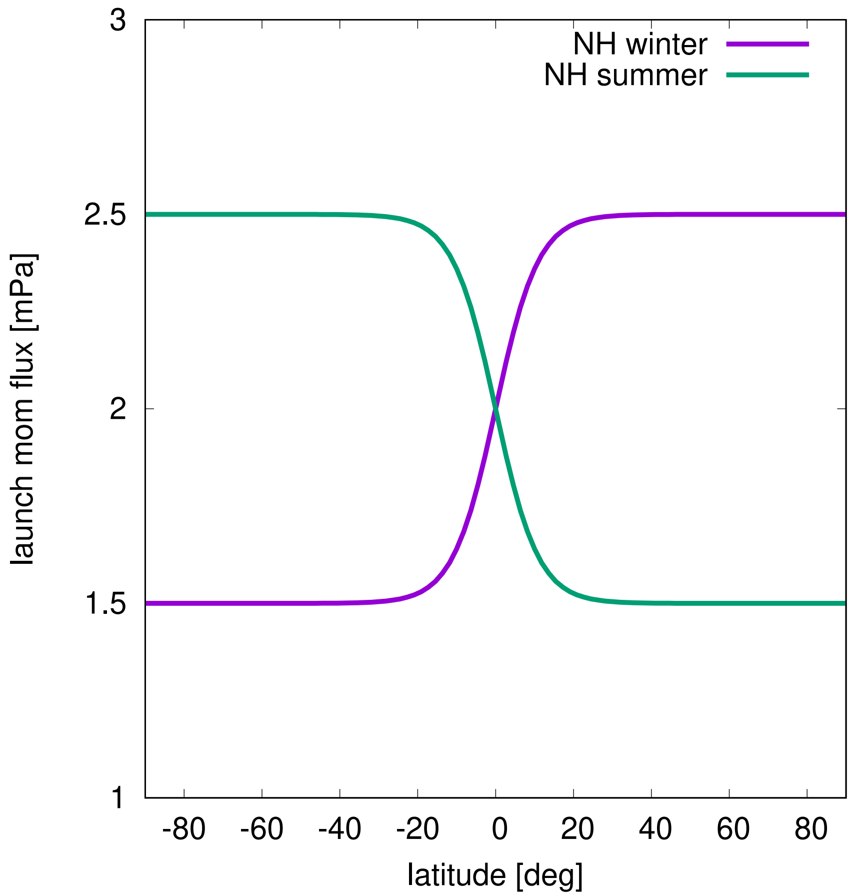
Fig. 11. Hovmöller diagram of absolute pseudomomentum fluxes at $\varphi =\sim -60^{\circ}, \lambda =\sim -150^{\circ}$, 1^{st} of June 1998, a): TR, b): ST. The base-10 logarithm of the fluxes in mPa is presented. 65

Fig. 12. Zonal-mean pseudomomentum fluxes [mPa] averaged over the Octobers of 1991-1998. The 2 upper rows show vertical cross-sections over the SH for eastward (a,d), westward (b,e) and absolute (c,f) fluxes as simulated by TR (a-c) and ST (d-f). The spacing of the solid black lines is $0.25mPa$ in a,b,d,e and $0.5mPa$ in c,f. The bottom row shows horizontal cross-sections of absolute fluxes over the Southern Ocean as simulated by TR (g) and ST (h,i), where in i), the values from the ST simulations are multiplied by 1.3. 66

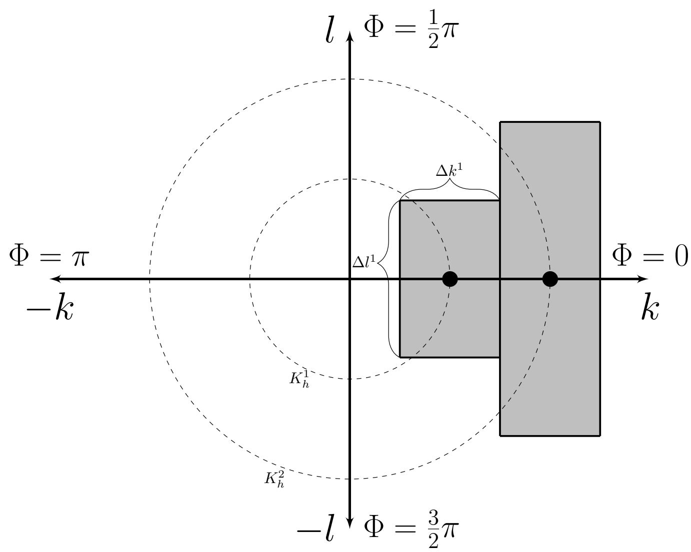
Fig. 13. Zonal-mean absolute pseudomomentum fluxes [mPa] (filled red contours) and zonal GW drag (gray contours) averaged over Junes of 1991-1998. The spacing of fluxes is $0.25mPa$ in the interval $[0,1]mPa$ and $1mPa$ in the interval $[1,9]mPa$. Contours of GW drag are drawn by $20ms^{-1}day^{-1}$ between $[-180,-20]ms^{-1}day^{-1}$ (dashed) and by $10ms^{-1}day^{-1}$ between $[10,80]ms^{-1}day^{-1}$ (solid). Panel a) shows the total, i.e. the contribution from all horizontal and vertical scales. Panels b), c), d), e) show contributions from scales $\lambda_h > 50km, 100km, 200km, 250km$ respectively, while panels f), g), h), i) show contributions from scales $\lambda_z > 1km, 2km, 5km, 10km$ respectively. 67



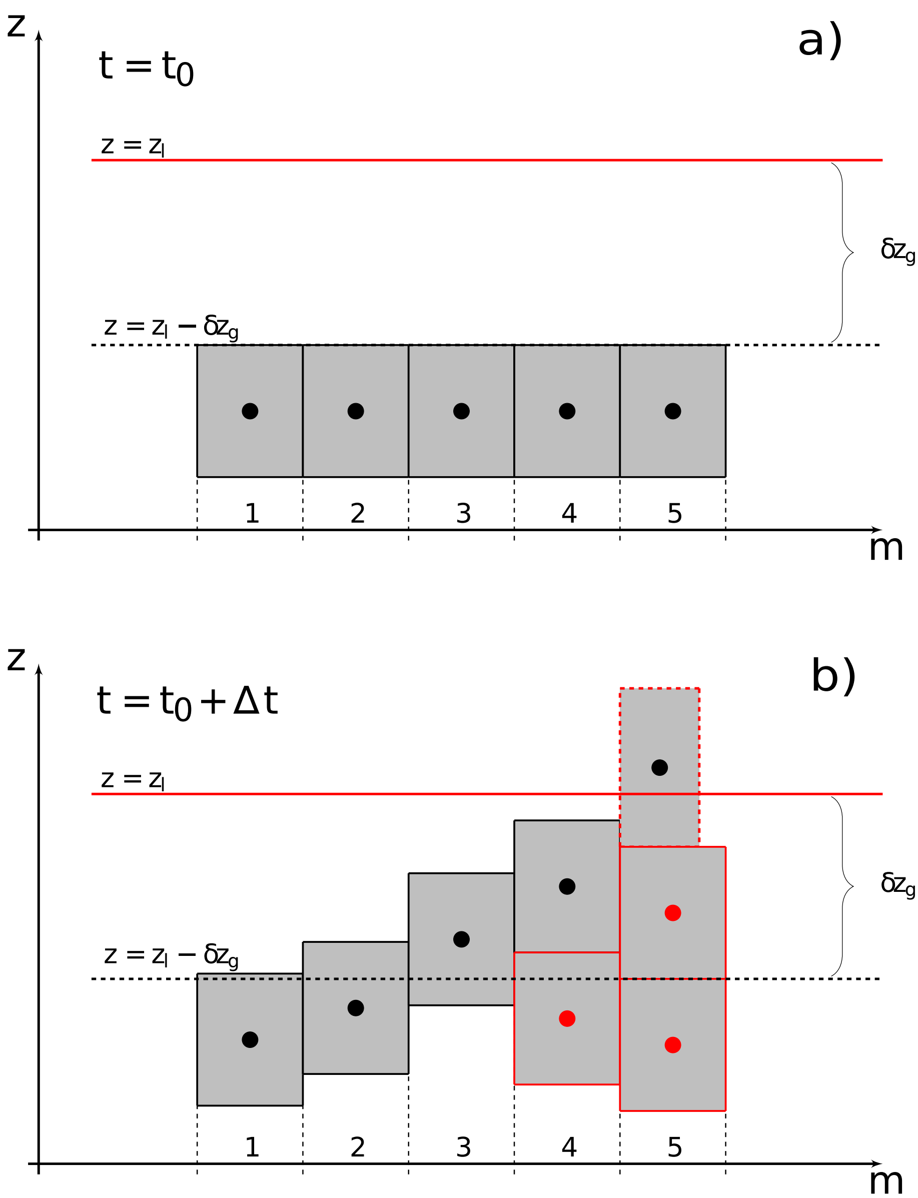
1088 FIG. 1. A schematic of the GW field representation by ray volumes in the transient MS-GWaM. a) representation
1089 in $z-m$ space (see the text for explanations on how the subgrid-scale GW pseudomomentum fluxes are projected
1090 to the vertical grid), b) representation in spectral space, showing that the ray volume center-point m_c and the
1091 extent in m-direction changes prognostically.



1092 Fig. 2. Latitudinal profile of launch momentum-flux magnitudes [mPa] in the non-orographic GW source at
1093 Northern Hemispheric winter (blue) and summer (green) solstices.



1094 FIG. 3. A schematic showing the attribution of horizontal wavenumbers to ray volumes in the non-orographic GW source (see the text for more explanation).



1096 FIG. 4. A schematic of the representation of the non-orographic GW emission as lower boundary condition in
1097 terms of sequential ray-volume launches (see the text for more explanation).

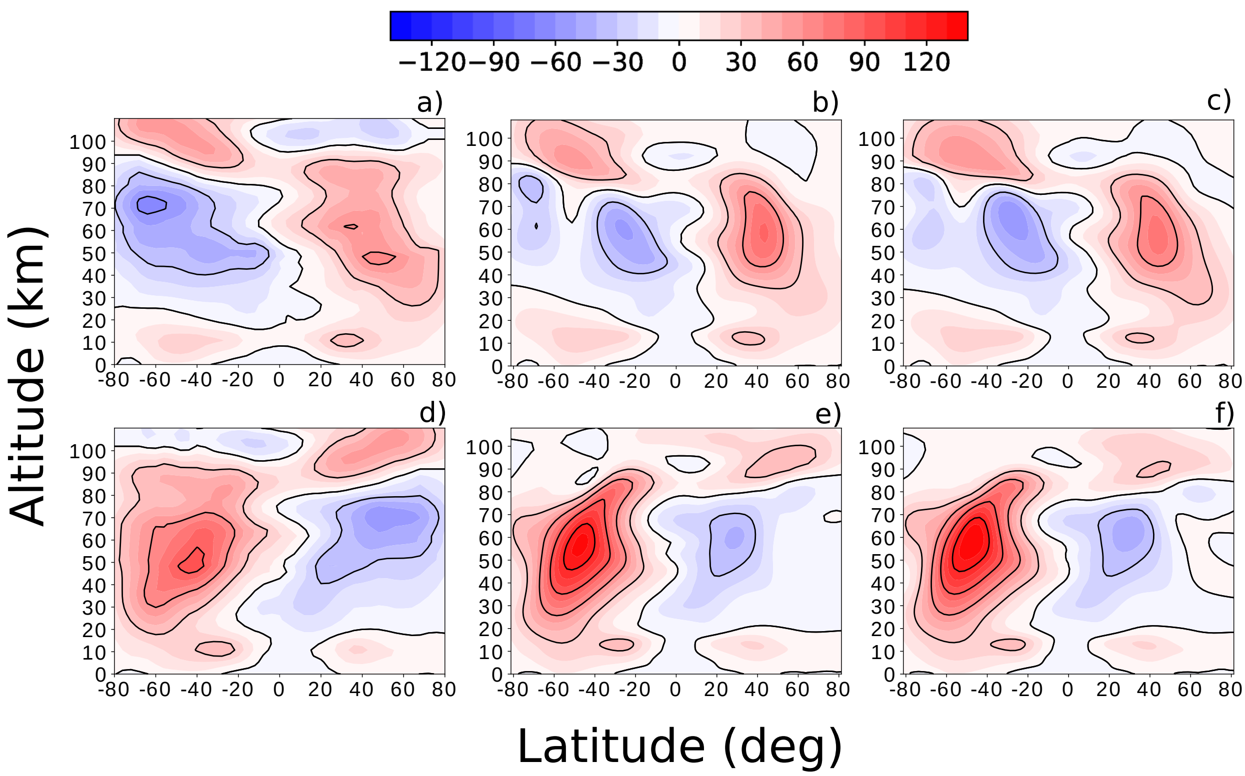


FIG. 5. Zonal-mean zonal wind [ms^{-1}] averaged over Decembers (a,b,c) and Junes (d,e,f) in the period 1991-1998. URAP data: a,d, TR: b,e, ST: c,f. Solid black isolines are drawn between $-120ms^{-1}$ and $120ms^{-1}$ with a spacing of $30ms^{-1}$.

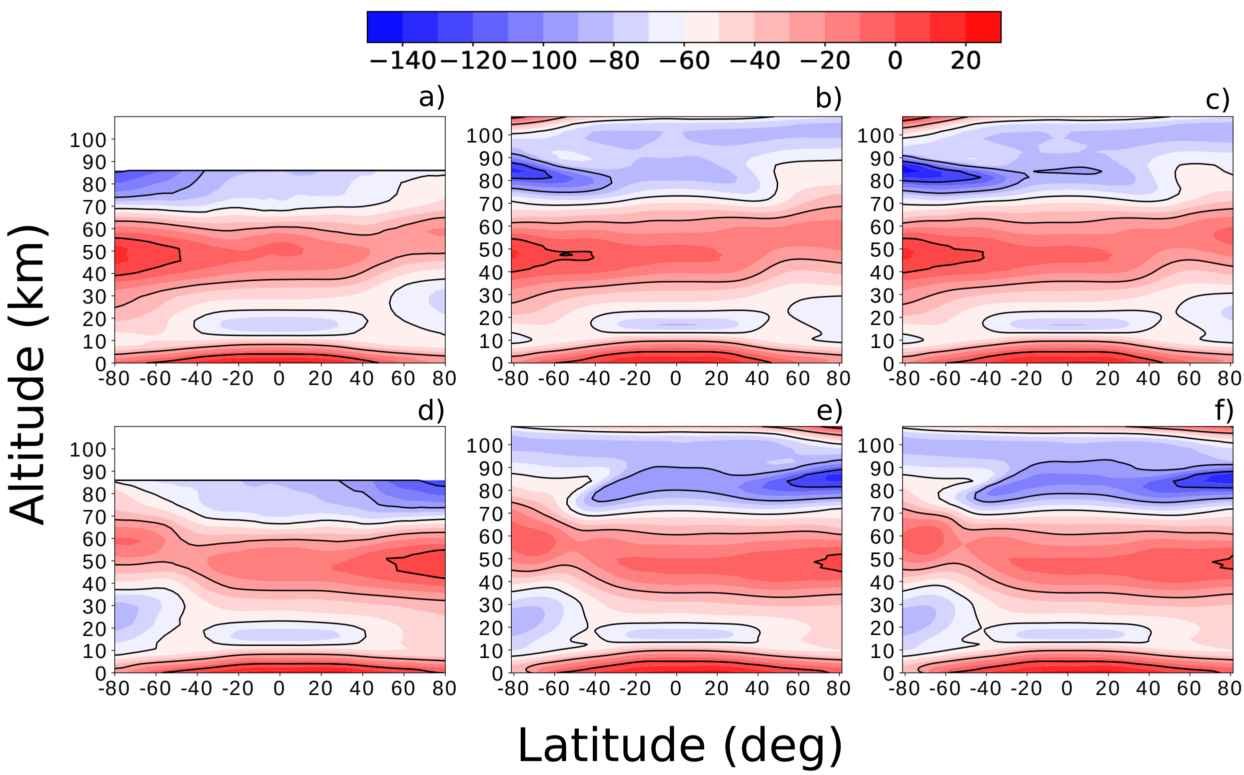
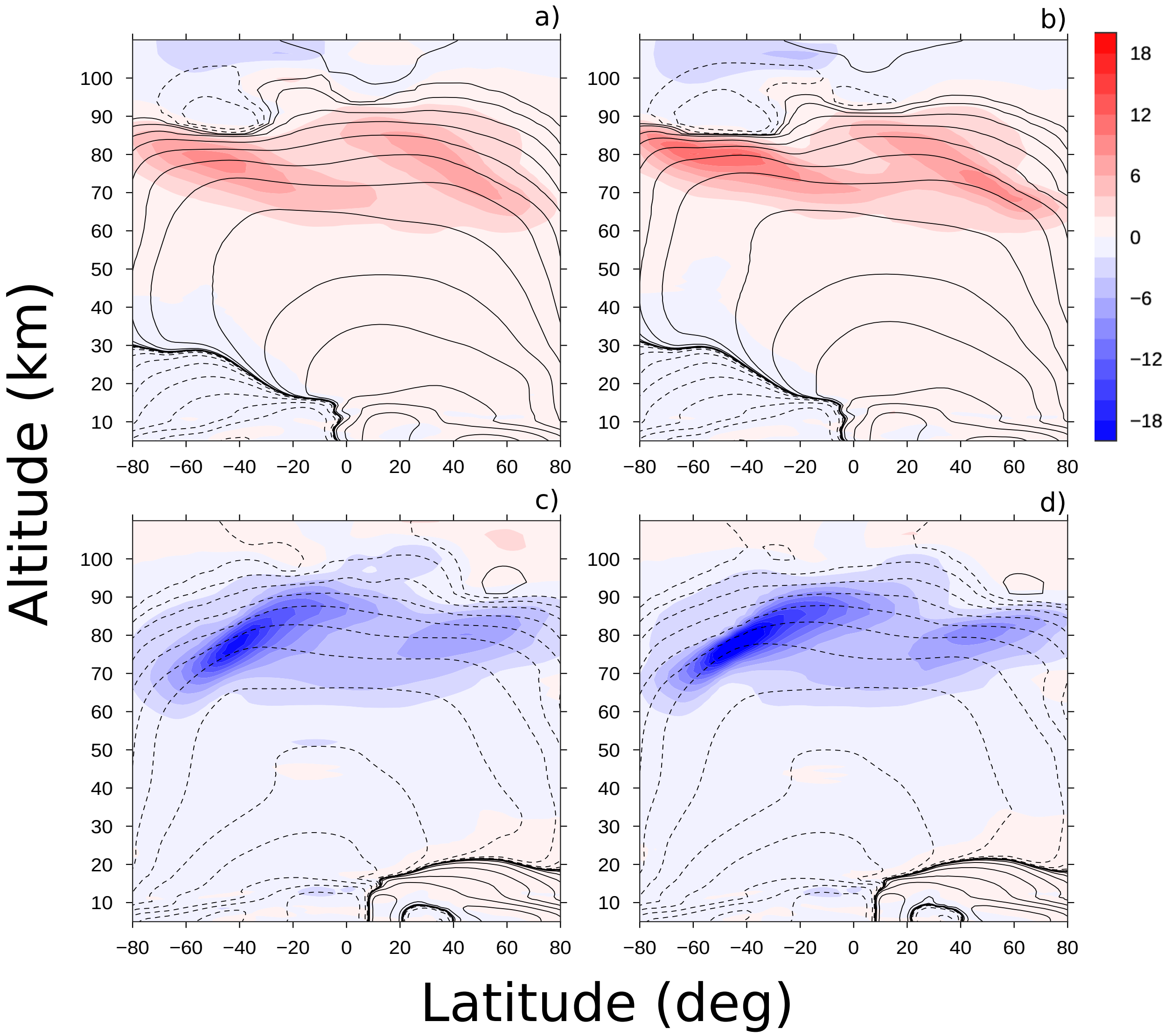
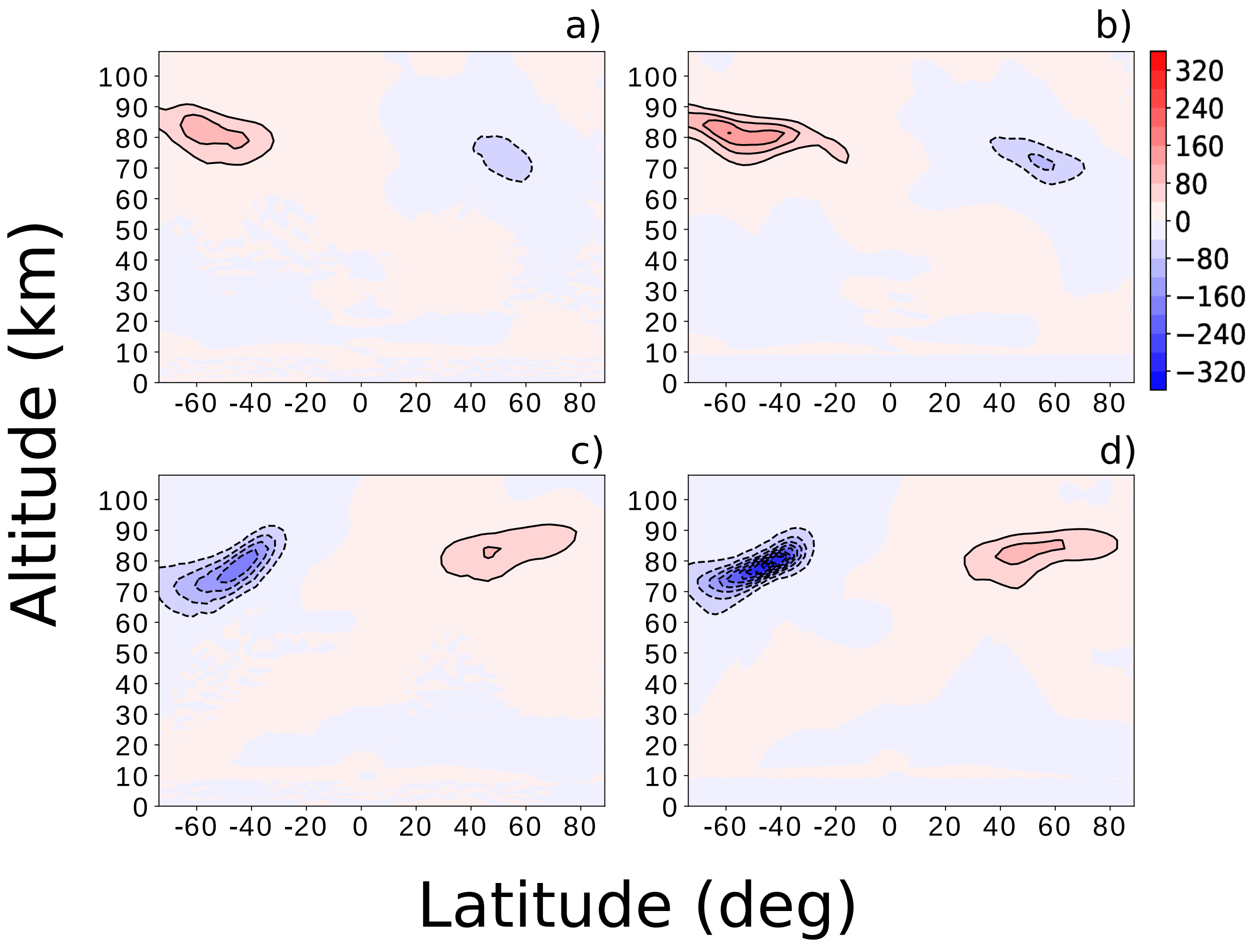


FIG. 6. Zonal-mean temperature [C°] averaged over Decembers (a,b,c) and Junes (d,e,f) in the period 1991-1998. URAP data: a,d, TR: b,e, ST: c,f. Solid black isolines are drawn between $-150C^{\circ}$ and $30C^{\circ}$ with a spacing of $30C^{\circ}$.



1104 FIG. 7. Residual circulation averaged over Decembers (a,b) and Junes (c,d) in the period 1991-1998. TR: a,c,
1105 ST: b,d. Filled contours: residual-mean meridional velocity in the transformed Eulerian mean equations (v^*)
1106 [ms^{-1}]. Black contour lines: residual mass streamfunction [kgs^{-1}] shown by solid (dashed) lines for positive
1107 (negative) values for the magnitudes of $10^4 - 10^{10} kgs^{-1}$ with two contours per 10 times increase.



1108 FIG. 8. Zonal-mean net gravity wave drag [$ms^{-1}day^{-1}$] averaged over Decembers (a,b) and Junes (c,d) in the period 1991-1998. TR: a,c, ST: b,d. Black contours are drawn between the intervals $[-360,-40]$ (dashed) and $[40,160]$ (solid) with an interval of 40.

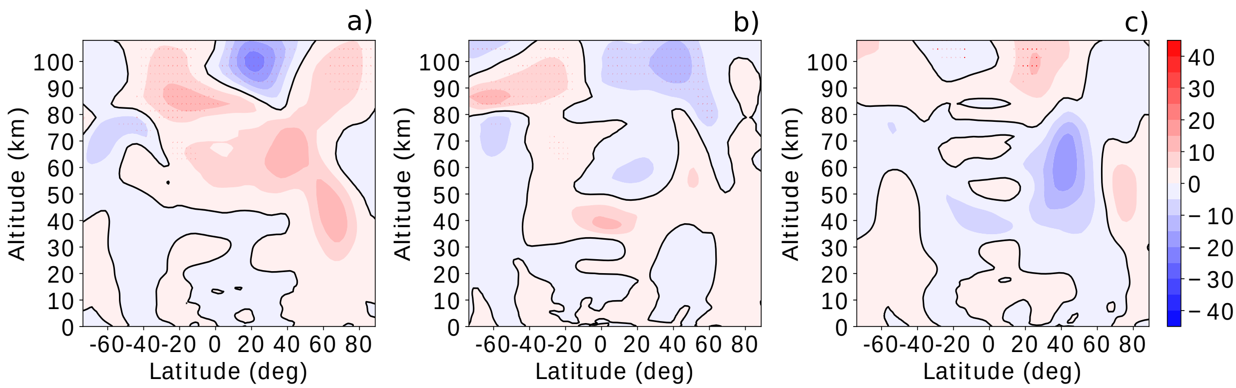


FIG. 9. Zonal-mean zonal wind differences [ms^{-1}] averaged over the last 12 months of a 24-months perpetual December simulation. a: difference ST-TR, b: difference STMO-ST, c: difference TRx2-TR where TRx2 stands for the transient MS-GWaM with a doubled number of ray volumes per column, i.e. N_{cmax} = 5000 instead of N_{cmax} = 2500. The solid black isolines are drawn for $0ms^{-1}$. Red dots denote points where the differences are statistically significant with a 95% confidence, based on t-tests taking into account time correlation of the sample. The numbers of statistically significant points are 334 for ST-TR, 313 for STMO-ST and 63 for TRx2-TR.

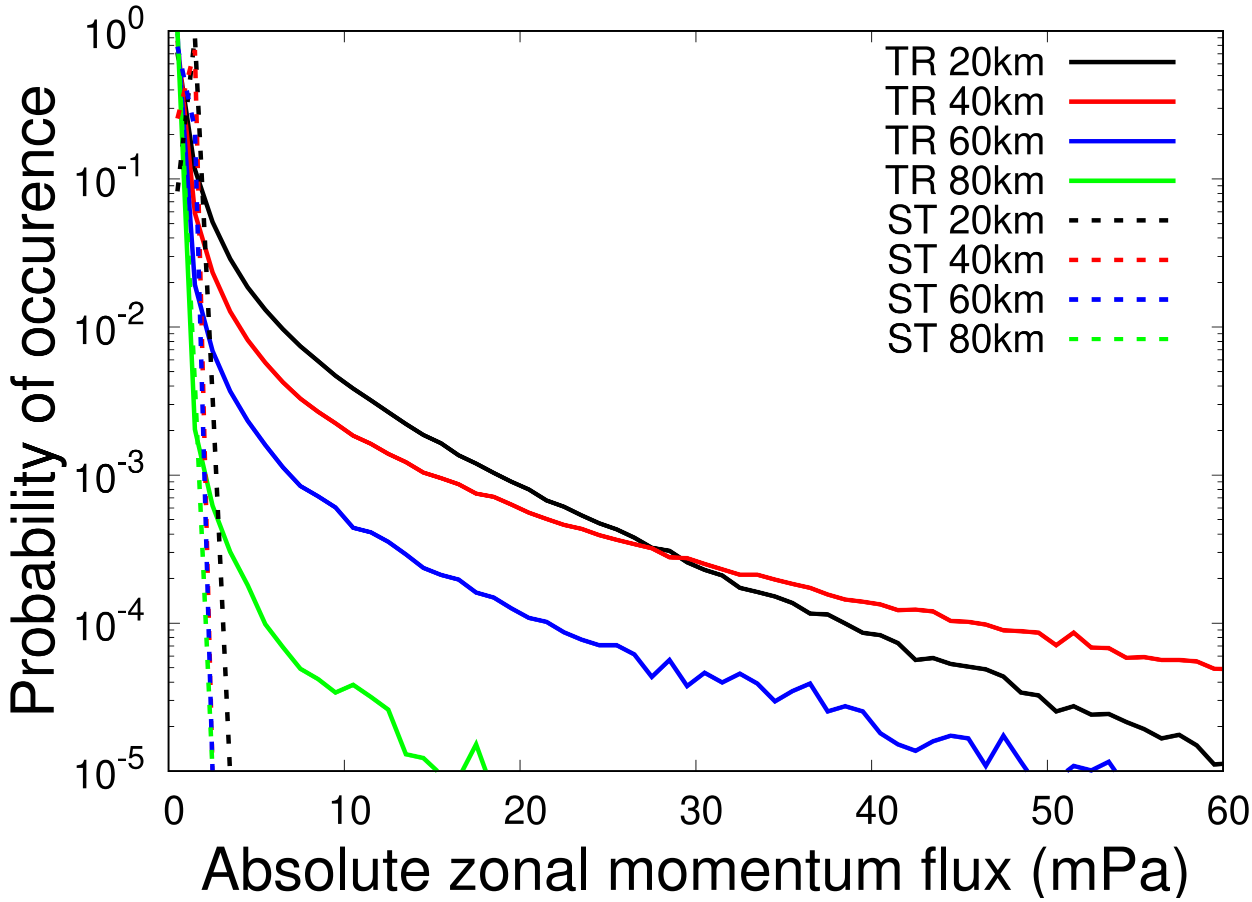
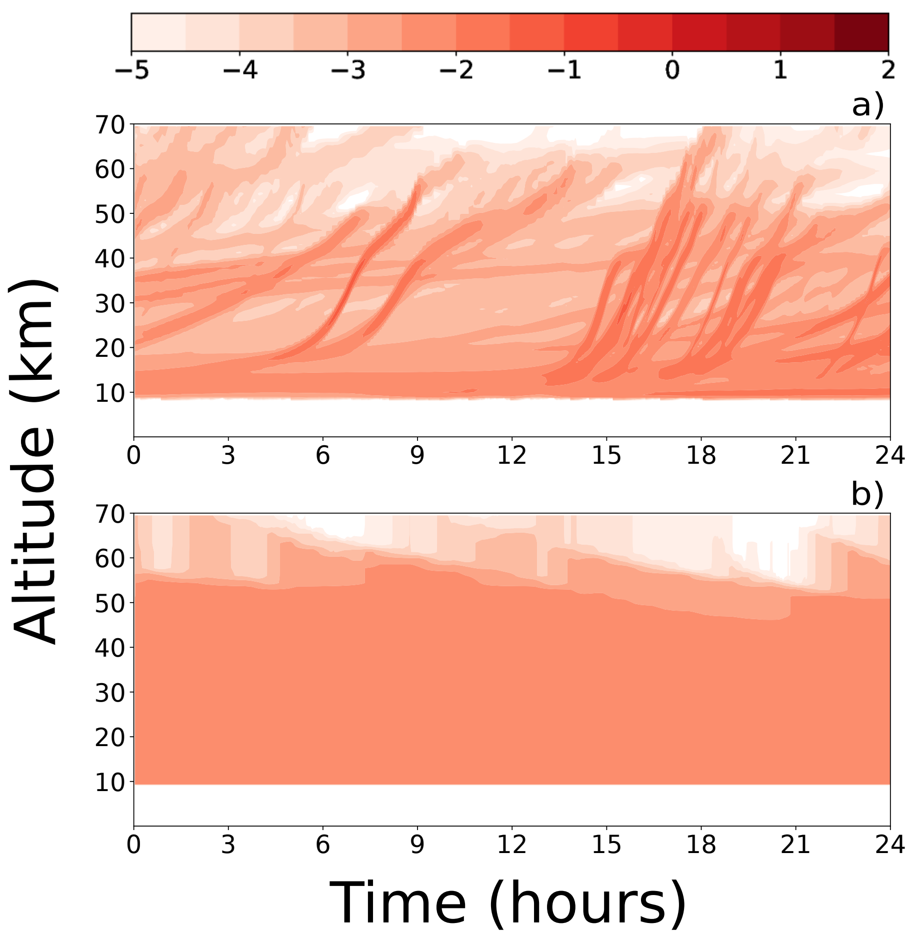


FIG. 10. Histogram of absolute zonal pseudomomentum-flux occurrences sampled at $\varphi \in [-65°, -50°], \lambda \in [-180°, 180°]$ over the Octobers of 1991-1998 for TR (solid) and ST (dashed) at different altitudes: $z \approx 20km$ (black), $z \approx 40km$ (red), $z \approx 60km$ (blue), $z \approx 80km$ (green).



1120 FIG. 11. Hovmöller diagram of absolute pseudomomentum fluxes at $\varphi = \sim -60°, \lambda = \sim -150°$, 1^{st} of June 1998,
1121 a): TR, b): ST. The base-10 logarithm of the fluxes in mPa is presented.

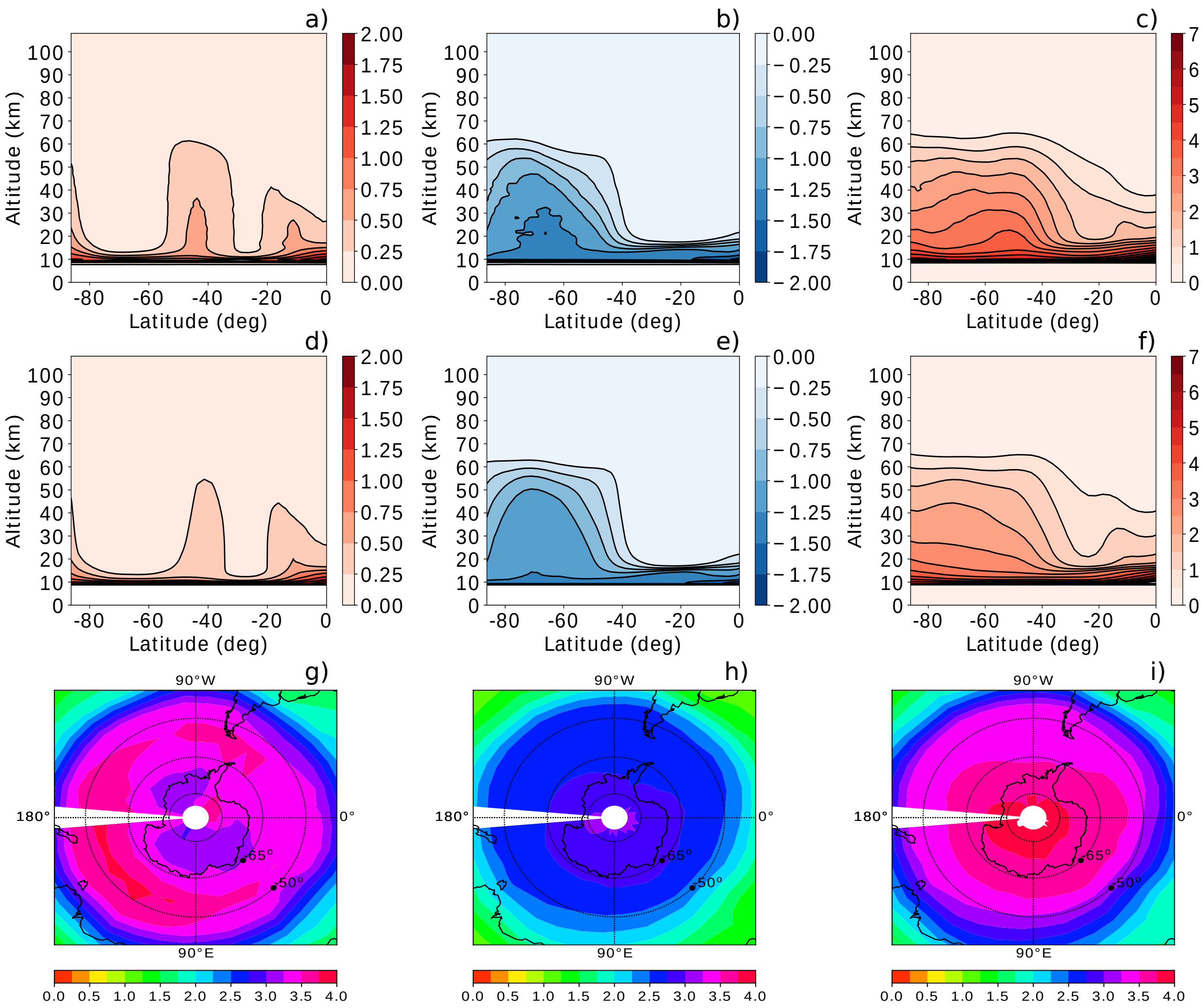


FIG. 12. Zonal-mean pseudomomentum fluxes [mPa] averaged over the Octobers of 1991-1998. The 2 upper rows show vertical cross-sections over the SH for eastward (a,d), westward (b,e) and absolute (c,f) fluxes as simulated by TR (a-c) and ST (d-f). The spacing of the solid black lines is 0.25mPa in a,b,d,e and 0.5mPa in c,f. The bottom row shows horizontal cross-sections of absolute fluxes over the Southern Ocean as simulated by TR (g) and ST (h,i), where in i), the values from the ST simulations are multiplied by 1.3.

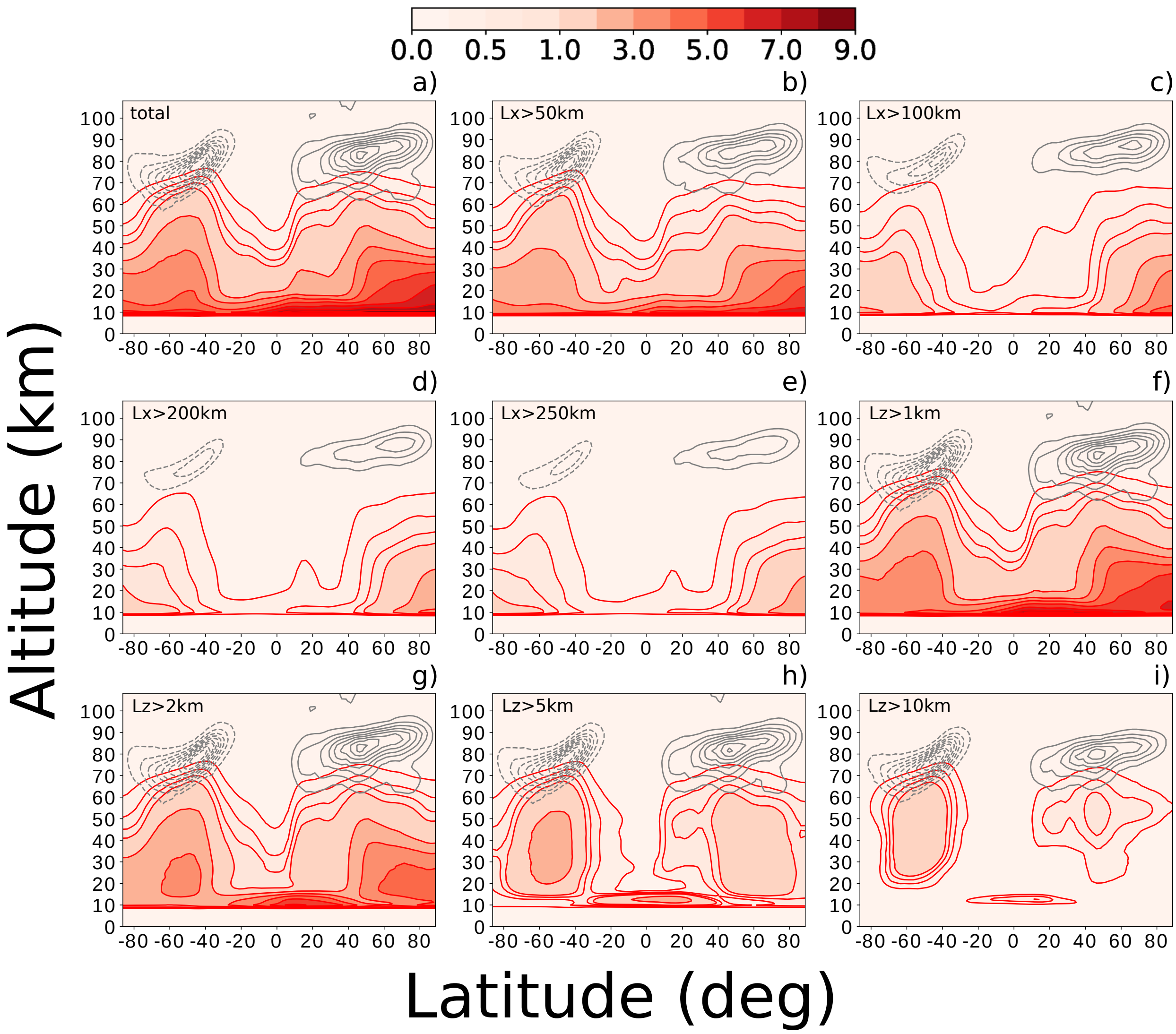


FIG. 13. Zonal-mean absolute pseudomomentum fluxes $[mPa]$ (filled red contours) and zonal GW drag (gray contours) averaged over Junes of 1991-1998. The spacing of fluxes is $0.25mPa$ in the interval $[0,1]mPa$ and $1mPa$ in the interval $[1,9]mPa$. Contours of GW drag are drawn by $20ms^{-1}day^{-1}$ between $[-180,-20]ms^{-1}day^{-1}$ (dashed) and by $10ms^{-1}day^{-1}$ between $[10,80]ms^{-1}day^{-1}$ (solid). Panel a) shows the total, i.e. the contribution from all horizontal and vertical scales. Panels b), c), d), e) show contributions from scales $\lambda_h > 50km, 100km, 200km, 250km$ respectively, while panels f), g), h), i) show contributions from scales $\lambda_z > 1km, 2km, 5km, 10km$ respectively.