JOURNAL OF GEOPHYSICAL RESEARCH, VOL. ???, XXXX, DOI:10.1002/,

A high-resolution whole-atmosphere model with resolved gravity waves and specified large-scale dynamics in the troposphere and stratosphere

Erich Becker[1], Sharon L. Vadas[1], Katrina Bossert[2], V. Lynn Harvey[3],

Christoph Zülicke[4], Lars Hoffmann[5]

Manuscript submitted to J. Geophys. Res. Atmos.: April 6, 2021

Revised: September 28 and December 9, 2021

Accepted for publication: December 14, 2021

Corresponding author: E. Becker, Northwest Research Associates, 3380 Mitchell Ln, Boulder, CO 80301, USA (erich.becker@nwra.com)

[1]Northwest Research Associates Inc.,

Key Points.

- nudging in spectral space allows for self-consistent simulation of gravity waves up to the thermosphere
- simulated gravity waves in the winter stratosphere agree well with satellite observations
- spontaneous emission of GWs in the winter stratosphere depends critically on vertical wind shear

Abstract. We present a new version of the HIgh Altitude Mechanistic general Circulation Model (HIAMCM) with specified dynamics. We utilize a spectral method that nudges only the large-scale flow to MERRA-2 reanalysis. The nudged HIAMCM simulates gravity waves (GWs) down to hori-

Boulder, CO, USA

[2]School of Earth and Space Exploration, Arizona State University, Tempe, AZ, USA

[3]Laboratory for Atmospheric and Space Physics, University of Colorado Boulder, Boulder, CO, USA

[4]Leibniz Institute of Atmospheric Physics at the University of Rostock, Kühlungsborn, Germany

[5]Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH, Jülich, Germany

zontal wavelengths of about 200 km from the troposphere to the thermosphere like the free-running model, including the generation of secondary and tertiary GWs. Case studies show that the simulated large-scale GWs are consistent with those in the reanalysis, while the medium-scale GWs compare well with observations in the northern winter 2016 stratosphere from the Atmospheric InfraRed Sounder (AIRS). GWs having wavelengths larger than about 1350 km can be described with the nonlinear balance equation. The GWs relevant in the stratosphere, however, have smaller scales and require a different approach. We propose that the GW amplification due to kinetic energy transfer from the large-scale flow combined with GW potential energy flux convergence helps to identify the mesoscale GW sources due to spontaneous emission. The GW amplification is strongest in the region of maximum large-scale vertical wind shear in the mid-stratosphere. Maps of the time-averaged stratospheric GW activity simulated by the HIAMCM and computed from AIRS satellite data show a persistent hot spot over Europe during January 2016. At about 40 km, the average GW amplitudes are maximum in the region of fastest large-scale flow. We argue that refraction of GWs originating in the troposphere, as well as GWs from spontaneous emission in the stratosphere contribute to this effect.

1. Introduction

Atmospheric general circulation models (GCMs) are valuable tools to study large-scale atmospheric variability and its sensitivities to external perturbations [*Garcia et al.*, 2007; *McLandress and Shepherd*, 2009; *Butchart et al.*, 2010; *Schmidt et al.*, 2010; *Marsh et al.*, 2013; *Solomon et al.*, 2019]. GCMs that extend into the thermosphere are useful, for example, to analyze dynamical vertical coupling processes from the lower to the upper atmosphere [*Vadas and Becker*, 2019; *Becker and Vadas*, 2020], or the downward influences induced by energetic particle precipitation [*Sinnhuber et al.*, 2012; *Randall et al.*, 2015; *Funke et al.*, 2017]. Usually, GCMs used for climate modeling are free-running models. That is, they are based on a self-consistent simulation of the internal dynamics using the primitive equations supplemented by a suite of parameterizations and additional prognostic equations for moist processes and chemistry. The external diabatic forcing is due to solar radiation (and other solar influences) and boundary conditions at the surface (e.g., prescribed sea surface temperature or coupling to an ocean model).

This concept is different for so-called nudged GCMs, which use global reanalysis or forecast data to specify the planetary and synoptic-scale dynamical fields in the troposphere and stratosphere. The nudging is imposed by adding artificial terms to the model equations that relax the wind and temperature fields toward the reanalysis/forcecast data [e.g. *McLandress et al.*, 2013; *Jones Jr. et al.*, 2018]. The relaxation rate is gradually reduced with increasing height in the stratosphere such that nudged GCMs are free-running above the stratopause. A general requirement is that the climatology and variability patterns simulated by the corresponding free-running model (i.e., without the artificial relaxation terms) are realistic. Then, adding the nudging is compatible with the dynamics of the model and merely gives rise to small corrections of the actual trajectory in the phase space of the model's prognostic variables. The nudging is not supposed to change the simulated climatology and variability patterns of the model notably.

It is evident that the simulation data from the altitude region where a GCM is nudged reflects the underlying reanalysis/forecast data. Furthermore, a nudged model is supposed to reproduce observed large-scale winds and temperatures in the free-running region (e.g., from the stratopause to the lower thermosphere). This is mainly because the middle and upper atmosphere is strongly dynamically controlled from below through planetary Rossby waves and planetary equatorial waves that are specified via the nudging, and because thermal forcing of tides and the generation of global modes due to barotropic/baroclinic instability above the stratosphere is well represented in GCMs [e.g., *Smith*, 2012; *McLandress et al.*, 2006]. Hence, nudged GCMs extend reanalysis/forecast from the troposphere and stratosphere into the mesosphere and even into the thermosphere. For example, one can analyze particular events (e.g., sudden stratospheric warmings, hereafter: SSWs) in the free-running region of a nudged GCM and compare the simulation data directly to observational data acquired from ground-based or satellite borne instruments [*McLandress et al.*, 2013; *Randall et al.*, 2015]. A major weakness of this reasoning, however, is that gravity waves (GWs) are usually not resolved in GCMs and must be parameterized. Corresponding parameterizations are based on very idealized assumptions and are often not well-constrained by observations. For example, the uncertainty in the mesosphere and lower thermosphere (hereafter: MLT) simulated by a GCM that is nudged at lower altitudes may depend strongly on the representation of the parameterized GWs [*Smith* *et al.*, 2017]. Despite this limitation, the dynamics of the mesosphere of a nudged GCM may nevertheless reflect the dynamics during extreme events like SSWs very well. Moreover, details about the roles of different types of GWs can be inferred [*McLandress et al.*, 2013]. A reliable simulation of the mesosphere requires, however, that the nudged region includes the stratosphere in order to better constrain the mesospheric variations due to internal variability [*Siskind et al.*, 2015].

It is technically feasible to run GCMs at sufficiently high resolution such that a majority of the GW drag required to drive the residual circulation in the middle atmosphere is explicitly simulated [e.g., *Watanabe and Miyahara*, 2009; *Sato et al.*, 2012; *Liu*, 2017; *Becker and Vadas*, 2018]. Even though the resolved GW spectrum in GCMs is limited, the explicit simulation of GWs overcomes the strong assumptions made in existing GW parameterizations, namely the single-column approximation and the assumption of instantaneous response [see discussion in *Becker*, 2017]. These limitations become significant in the upper winter mesosphere, where secondary GWs generated by the body force mechanism [*Vadas et al.*, 2003, 2018] have large amplitudes and significant effects on the mean flow [*Becker and Vadas*, 2018; *Becker et al.*, 2020], and where GW-tidal interactions are crucial for GW dissipation [*Senf and Achatz*, 2011; *Becker*, 2017].

The limitations of conventional GW schemes become severe in the winter thermosphere. Recent modeling and observation-based studies suggest that the majority of GWs in the winter thermosphere are secondary and tertiary GWs, and that horizontal propagation over thousands of kilometers away from the source regions is evident [*Vadas and Becker*, 2019; *Vadas et al.*, 2019; *Becker and Vadas*, 2020]. Furthermore, secondary GWs are also important in the equatorial and summer thermosphere and ionosphere [*Vadas and Crow-*

ley, 2010; *Makela et al.*, 2010; *Vadas and Liu*, 2013; *Vadas et al.*, 2014; *Vadas and Azeem*, 2021]. As discussed in *Becker and Vadas* [2020] (hereafter: BV20), neither secondary and tertiary GWs nor horizontal propagation and GW transience are accounted for in available GW parameterizations. Therefore, in order to construct a nudged GCM that reasonably accounts for GWs in the winter mesopause region and in the thermosphere, GWs need to be simulated explicitly.

The explicit simulation of GWs depends crucially on 1) the numerics of the dynamical core, 2) the effective spatial and temporal resolution, and 3) how the mesoscale cascades of kinetic and available potential energy are balanced by subgrid-scale diffusion. These characteristics vary vastly among models. As a result, the GWs resolved in a particular model may not be compatible with the GWs in the reanalyis/forecast data to which the model is nudged. For example, the mesoscale spectral kinetic energy in the upper troposphere of the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecast System (IFS) T1279L91 was found to be much smaller than that of a free-running high-resolution GCM that was run at a two times coarser resolution than the IFS [*Augier and Lindborg*, 2013]. Furthermore, temperature perturbations in the lower stratosphere simulated by the IFS were found to be a factor of 2-3 smaller than in satellite observations [*Hoffmann et al.*, 2017]. In general, nudging of winds and temperatures of a high-resolution GCM in gridspace constrains the resolved GW dynamics of the GCM to that of the reanalyis/forecast data. This causes either artificial generation or damping of the GWs resolvable by the GCM.

These considerations show that nudging a high-resolution GCM with resolved GWs is not as straight-forward as nudging a GCM with conventional resolution and parameterized GWs. A method to specify only the large-scale dynamical fields of a GW-resolving GCM was proposed by *Shibuya and Sato* [2019]. These authors used reanalysis data with medium resolution to set the initial condition of the Non-hydrostatic Icosahedral Atmospheric Model (NICAM) [*Satoh et al.*, 2014]. In that study, the NICAM extended from the surface to about 80 km and had a sponge layer from 80 to 87 km. *Shibuya and Sato* [2019] assumed that a realistic GW field developed within two model days after initialization. Dynamical fields (including GWs) from the model integration could then be compared to the real atmosphere 3-7 days after the initialization. Beyond day 7, the simulated large-scale dynamics started to deviate significantly from the reanalysis data. *Shibuya and Sato* [2019] generated a longer time series by initializing the NICAM with reanalysis every 5 days and stitched the results together from each simulation for days 3-7, thereby imposing temporal discontinuities every 5 days. Such a method was also used in the study of *Plougonven et al.* [2013]. Note that this method specifies only initial conditions of an otherwise free-running model to perform simulations comparable to observations of the real atmosphere.

In the present study we propose to take advantage of the spectral method to nudge the HIgh Altitude Mechanistic general Circulation Model (HIAMCM) (BV20) to reanalysis continuously in time. The basic idea is to transform a given reanalysis data set into spectral space and then nudge only the large-scale spectral components. Similar spectral methods were used previously for low-resolution climate models [e.g., *von Storch et al.*, 2000; *McLandress et al.*, 2013]. Here we assume that while the large-scale fields follow the trajectory of the reanalysis due to nudging, the resolved mesoscale GWs (including their generation, propagation, and dissipation) are simulated self-consistently like in the free-running model. Even though GW processes are not directly affected by the nudging, we hypothesize that the timing and location of mountain-wave events or GW generation from jets and fronts should be comparable to corresponding events in the real atmosphere, to the extent that 1) the GWs are well resolved by the given spatial resolution, 2) the representation of subgrid-scale processes induces realistic and location-appropriate dissipation of GWs subject to dynamical instability, and 3) the large-scale flow in the reanalysis is accurate.

In this study we will present case studies of GWs generated by spontaneous emission [e.g., *O'Sullivan and Dunkerton*, 1995; *Zülicke and Peters*, 2006, 2008; *Plougonven and Zhang*, 2014; *Dörnbrack et al.*, 2018; *Gassmann*, 2019], and we will compare the simulated GWs to the Atmospheric InfraRed Sounder (AIRS) satellite data, which has previously been used to examine GW hotspots in the stratosphere [e.g., *Gong et al.*, 2012; *Hoffmann et al.*, 2013, 2016; *Bossert et al.*, 2020; *Hindley et al.*, 2020]. Furthermore, we will analyze the GW sources using the transfer of kinetic energy from the large-scale flow to GWs and the GW potential energy flux convergence. This diagnostic tool is derived in Appendix B.

In Sec. 2 and Appendix A we give an updated description of the HIAMCM. Section 3 specifies our nudging technique in detail. We use the three-hourly Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2) for nudging. In Sec. 4 we compare GW results from the nudged model with the free-running model, as well as with the GWs resolved in MERRA-2. We confirm that the simulated GWs in the nudged model are consistent with those in the free-running model. In Sec. 5 we focus on two GW events in the stratosphere over Europe and over the Newfoundland/North Atlantic regions, and we compare the GWs in the HIAMCM or in MERRA-2 with those in the AIRS satellite data. In addition, Sec. 6 presents a comparison of 10-day and monthly averages of the stratospheric GW activity in January 2016 from the HIAMCM and AIRS. Section 7 presents the analysis of the GW events based on the model data. Our results are summarized in Sec. 8.

2. Description of the HIAMCM

The HIAMCM is a GCM based on a standard spectral dynamical core with a terrain-following vertical coordinate and a staggered vertical grid according to *Simmons and Burridge* [1981]. This core is equipped with a correction for non-hydrostatic dynamics, which is important in the thermosphere where many of the resolved GWs have high intrinsic frequencies (BV20). In the present study we employ a triangular spectral truncation at a total horizontal wavenumber of 256 which corresponds to a horizontal grid-spacing of $\sim$52 km and a shortest resolved horizontal wavelength of $\lambda_h \sim 156$ km. The horizontal grid consists of 768 equidistant longitudes and 384 Gaussian latitudes. The vertical level spacing is $\sim$600$-$650 m between the boundary layer and 3×10^{-5} hPa ($z \sim 130$ km). The vertical level spacing increases at higher altitudes to $\sim$ 10 km above $\sim$ 300 km. Using 280 full layers, the model top is at 4×10^{-9} hPa, corresponding to $z \sim 450$ km for temperatures of $T \sim 950$ K above ~ 250 km. We abbreviate this resolution as T256L280. The HIAMCM includes simplified but nevertheless explicit representations of the relevant components of an atmospheric climate model: radiative transfer, water vapor transport, large-scale condensation and moist convection, the full surface energy budget including a slab ocean, macro-turbulent and molecular horizontal and vertical diffusion, and ion drag. The details of these parameterizations are given in BV20. In the current version of the HIAMCM, we use a somewhat higher horizontal resolution and a finer vertical level spacing in the lower thermosphere as compared to BV20. For better compatibility of the simulated stratospheric temperatures with reanalysis, we modified the radiation scheme by including the ozone absorption of reflected UV-A and UV-B radiation, and we adjusted the prescribed ozone mixing ratio and ozone absorption coefficients.

Macro-turbulent vertical and horizontal diffusion is represented by the Smagorinsky scheme, with both diffusion coefficients depending on the Richardson number, R_i, giving rise to strong wave damping in the troposphere for $R_i \leq 0$ and in the mid stratosphere and above for $R_i \leq 0.25$ [*Becker*, 2009]. As in BV20, the diffusion is accomplished by molecular viscosity in both the vertical and horizontal diffusion terms. As a result, the major dissipation mechanism for resolved GWs above about 200 km is molecular viscosity, as it should be, and the model does not need an artificial sponge layer. To better simulate the location of the summer mesopause, as well as GW amplitudes in the stratosphere in comparison with AIRS data, we updated the macro-turbulent diffusion scheme with respect to the horizontal mixing length, the horizontal Prandtl number, and the hyperdiffusion coefficient. Details of the updated horizontal diffusion scheme are given in Appendix A.

3. Nudging in spectral space

In this section we show how the updated HIAMCM can be nudged in spectral space. Since the model is based on a spectral dynamical core, the prognostic variables are represented as a series of spherical harmonics subject to triangular truncation at total horizontal wavenumber $N = 256$. The model employs finite differencing in the vertical direction.

The spherical harmonics used in the HIAMCM are defined as

$$Y_{nm}(\lambda,\phi) \;=\; \begin{cases} \sqrt{\frac{1}{\pi}}\; P_n^m(\sin\phi) & \text{for } \; m=0 \\ \sqrt{\frac{2}{\pi}}\, P_n^m(\sin\phi)\;\cos m\lambda & \text{for } \; m>0 \\ \sqrt{\frac{2}{\pi}}\, P_n^m(\sin\phi)\;\sin|m|\lambda & \text{for } \; m<0\,, \end{cases} \tag{1}$$

where P_n^m are the Legendre functions, n is the total horizontal wavenumber and m is the zonal wavenumber, and λ and ϕ are longitude and latitude, respectively. The relative vorticity and horizontal divergence at the model layer l are written as

$$\xi_l(\lambda,\phi,t) \;=\; \sum_{n=1}^{N}\sum_{m=-n}^{n} \xi_{lnm}(t)\; Y_{nm}(\lambda,\phi) \tag{2}$$

$$D_l(\lambda,\phi,t) \;=\; \sum_{n=1}^{N}\sum_{m=-n}^{n} D_{lnm}(t)\; Y_{nm}(\lambda,\phi)\,, \tag{3}$$

where $\xi_{lnm}(t)$ and $D_{lnm}(t)$ are the spectral expansion coefficients. The horizontal streamfunction and velocity potential corresponding to Eqs. (2) and (3) are

$$\psi_l(\lambda,\phi,t) \;=\; -\sum_{n=1}^{N}\sum_{m=-n}^{n} \frac{a_e^2}{n(n+1)}\,\xi_{lnm}(t)\; Y_{nm}(\lambda,\phi) \tag{4}$$

$$\chi_l(\lambda,\phi,t) \;=\; -\sum_{n=1}^{N}\sum_{m=-n}^{n} \frac{a_e^2}{n(n+1)}\,D_{lnm}(t)\; Y_{nm}(\lambda,\phi)\,, \tag{5}$$

respectively, where a_e denotes the Earth's radius. Hence, the horizontal wind vector becomes

$$\begin{aligned}
\mathbf{v}_l(\lambda,\phi,t) &= u_l(\lambda,\phi,t)\,\mathbf{e}_\lambda(\lambda) \;+\; v_l(\lambda,\phi,t)\,\mathbf{e}_\phi(\lambda,\phi) \qquad (6)\\
&= \mathbf{e}_z(\lambda,\phi)\times\nabla\psi_l(\lambda,\phi,t) \;+\; \nabla\,\chi_l(\lambda,\phi,t)\\
&= -\sum_{n=1}^{N}\sum_{m=-n}^{n} \frac{a_e^2}{n(n+1)}\left(\xi_{lnm}(t)\,\mathbf{e}_z(\lambda,\phi)\times\nabla Y_{nm}(\lambda,\phi) + D_{lnm}(t)\,\nabla Y_{nm}(\lambda,\phi)\right).
\end{aligned}$$

Here, ∇ is the horizontal gradient operator in spherical coordinates, and u_l and v_l are the zonal and meridional wind components, respectively, on the model layer l. The unit vectors in the zonal, meridional, and vertical direction are $\mathbf{e}_\lambda$, $\mathbf{e}_\phi$, and $\mathbf{e}_z$, respectively.

The horizontal momentum equation in gridspace can be written as

$$\partial_t \mathbf{v}_l(\lambda,\phi,t) = \mathbf{f}_l(\lambda,\phi,t) - \nabla\left(\Phi_l(\lambda,\phi,t) + \mathbf{v}_l^2(\lambda,\phi,t)/2\right). \tag{7}$$

Here, $\mathbf{f}_l$ accommodates the Coriolis force, the pressure gradient term (relevant in the lower troposphere due to model surfaces deviating from pressure surfaces), all advection terms other than $-\nabla\,\mathbf{v}_l^2/2$, momentum diffusion, and ion drag (see Eq. (1) in BV20). $\Phi_l(\lambda,\phi,t)$ denotes the sum of the hydrostatic geopotential and the non-hydrostatic correction given in BV20. Equation (7) leads to the following ordinary differential equations for the relative vorticity and horizontal divergence in spectral space:

$$d_t\,\xi_{lnm}(t) \;=\; -\int_{globe} d\Omega\;\mathbf{e}_z(\lambda,\phi)\cdot\Big(\mathbf{f}_l(\lambda,\phi,t)\times\nabla Y_{nm}(\lambda,\phi)\Big) \tag{8}$$

$$d_t\,D_{lnm}(t) \;=\; -\int_{globe} d\Omega\,\Big(\,\mathbf{f}_l(\lambda,\phi,t)\cdot\nabla Y_{nm}(\lambda,\phi) \tag{9}$$

$$+\left(\Phi_l(\lambda,\phi,t)+\mathbf{v}_l^2(\lambda,\phi,t)/2\right)\nabla^2 Y_{nm}(\lambda,\phi)\Big), \tag{10}$$

for $l = 1\ldots 280$, $n = 1\ldots 256$, and $m = -n,\ldots n$, and where $d\Omega = d\lambda\, d\sin\phi$. The spectral representations of the temperature, surface pressure, and surface temperature are

$$T_l(\lambda,\phi,t) \;=\; \sum_{n=0}^{N}\sum_{m=-n}^{n} T_{lnm}(t)\,Y_{nm}(\lambda,\phi) \tag{11}$$

$$p_s(\lambda,\phi,t) \;=\; p_{ref} + \sum_{n=1}^{N}\sum_{m=-n}^{n} p_{s\,nm}(t)\,Y_{nm}(\lambda,\phi) \tag{12}$$

$$T_s(\lambda,\phi,t) \;=\; \sum_{n=0}^{N}\sum_{m=-n}^{n} T_{s\,nm}(t)\,Y_{nm}(\lambda,\phi)\,, \tag{13}$$

respectively, where $p_{ref} = 986\,\mathrm{hPa}$ is the global-mean surface pressure. The grid-space representations of the partial differential equations for T_l, p_s, and T_s give rise to the following ordinary differential equations in spectral space:

$$d_t\, T_{l\,nm}(t) \;=\; \int_{globe} d\Omega\ \partial_t T_l(\lambda,\phi,t)\ Y_{nm}(\lambda,\phi) \tag{14}$$

$$d_t\, p_{s\,nm}(t) \;=\; \int_{globe} d\Omega\ \partial_t p_s(\lambda,\phi,t)\ Y_{nm}(\lambda,\phi) \tag{15}$$

$$d_t\, T_{s\,nm}(t) \;=\; \int_{globe} d\Omega\ \partial_t T_s(\lambda,\phi,t)\ Y_{nm}(\lambda,\phi)\,. \tag{16}$$

Note that in the framework of *Simmons and Burridge* [1981], a spectral model is mass conserving by definition, that is, $\dot{p}_{s\,nm} = 0$ for $n = m = 0$. This constraint is fulfilled in the HIAMCM since we expand the surface pressure in a series of spherical harmonics. Then, p_{ref} is a predefined model constant. Other spectral GCMs expand the logarithm of the surface pressure, thereby allowing spurious changes of the global-mean surface pressure. Also note that we do not nudge the water vapor. Therefore, the water vapor budget and its representation in spectral space is not mentioned further in this paper.

We use the Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2) for nudging. MERRA-2 is a NASA atmospheric reanalysis for the satellite era using the Goddard Earth Observing System Model, Version 5 (GEOS-5) with its Atmospheric Data Assimilation System (ADAS), version 5.12.4 [*Bosilovich et al.*, 2015]. For our purpose we use the "M2I3NVASM: MERRA-2 inst3_3d_asm_Nv: 3d, 3-Hourly, Instantaneous, Model-Level, Assimilation, Assimilated Meteorological Fields V5.12.4". These fields are provided at the model's terrain-following 72 atmospheric levels on a $0.5° \times 0.5°$ longitude-latitude grid. The highest model layer in MERRA-2 is at 0.015 hPa (corresponding to $z \sim 75$ km). In addition, MERRA-2 includes the surface pressure and the orography.

Using the surface pressure of MERRA-2, we construct a terrain following model grid that is identical to that of HIAMCM. We then interpolate the MERRA-2 atmospheric wind and temperature fields to this grid and compute the spectral representations of relative vorticity, horizontal divergence, and temperature using

$$\xi_{lnm}^{X}(t) \;=\; -\int_{globe} d\Omega\; \mathbf{e}_z(\lambda,\phi)\cdot\Big(\mathbf{v}_l(\lambda,\phi,t)^X \times \nabla Y_{nm}(\lambda,\phi)\Big) \tag{17}$$

$$D_{lnm}^{X}(t) \;=\; \int_{globe} d\Omega\; \mathbf{v}_l(\lambda,\phi,t)^X \cdot \nabla Y_{nm}(\lambda,\phi) \tag{18}$$

$$T_{lnm}^{X}(t) \;=\; -\int_{globe} d\Omega\; T_l(\lambda,\phi,t)^X\, Y_{nm}(\lambda,\phi)\,, \tag{19}$$

respectively. Here, X represents MERRA-2 or another data set to which the model can be nudged, and l extends up to the highest stratospheric layer in the HIAMCM where the pressure is larger than 0.015 hPa (for MERRA-2).

The aforementioned atmospheric MERRA-2 reanalysis data sets do not contain the surface temperature. An estimate of T_s is obtained by extrapolating the MERRA-2 atmospheric temperature to the surface using the hydrostatic formula as follows:

$$T_s = w_s^{-1}\left(\frac{g}{R}\,\frac{(w_s\,p_s + w_1\,p_1)\,(z_1 - z_s)}{p_s - p_1} - w_1\,T_1\right) \;-\; 5\,\mathrm{K} \tag{20}$$

Here, w_s and w_1 are weighting factors with $w_s + w_1 = 1$, and z_s and z_1 are the heights above sea level of the surface and the lowest atmospheric layer of MERRA-2, respectively. Furthermore, p_1 and T_1 are the pressure and temperature at the lowest atmospheric layer of MERRA-2. We find that $w_s = 0.67$ and $w_1 = 0.33$ together with an offset of $-5\,\mathrm{K}$ generates a reasonable surface temperature field giving rise to boundary-layer fluxes in the nudged HIAMCM comparable to those in the free-running model. Computation of the spectral representation of the MERRA-2 surface temperature is straightforward:

$$T_{s\,nm}^{X}(t) \;=\; \int_{globe} d\Omega\; T_s(\lambda,\phi,t)^X\, Y_{nm}(\lambda,\phi)\,. \tag{21}$$

The nudging of the HIAMCM is performed by supplementing the spectral tendencies of the prognostic variables in Eqs. (8), (9), (14), and (16) with relaxation towards the MERRA-2 reanalysis according to:

$$d_t\,\xi_{lnm}(t) \;\rightarrow\; d_t\,\xi_{lnm}(t) \,-\, (1/\tau)_{ln}\,\big(\,\xi_{lnm}(t) \,-\, \xi^X_{lnm}(t)\,\big) \tag{22}$$

$$d_t\,D_{lnm}(t) \;\rightarrow\; d_t\,D_{lnm}(t) \,-\, (1/\tau)_{ln}\,\big(\,D_{lnm}(t) \,-\, D^X_{lnm}(t)\,\big) \tag{23}$$

$$d_t\,T_{lnm}(t) \;\rightarrow\; d_t\,T_{lnm}(t) \,-\, (1/\tau)_{ln}\,\big(\,T_{lnm}(t) \,-\, T^X_{lnm}(t)\,\big) \tag{24}$$

$$d_t\,T_{s\,nm}(t) \;\rightarrow\; d_t\,T_{s\,nm}(t) \,-\, (1/\tau)^{T_s}_{n}\,\big(\,T_{s\,nm}(t) \,-\, T^X_{s\,nm}(t)\,\big) \quad . \tag{25}$$

Here, $1/\tau$ is a relaxation rate that depends on the horizontal scale (the total horizontal wavenumber n), as well as on the height (level index l). The same relaxation rate is used for ξ, D, and T. The instantaneous spectral amplitudes from the MERRA-2 reanalysis are computed by linear interpolation between the three-hourly snapshots at which these amplitudes are precalculated from the original data sets. Note that p_s is not nudged, but is computed self-consistently from the vertical integral of the horizontal divergence (see Eqs. (6) in BV20).

As discussed in the introduction, our goal is that the nudging does not directly affect the dynamics of the resolved GWs. For this purpose the dependence of the relaxation rate on the horizontal wavenumber is crucial. Figure 1a shows the relaxation rate as a function of the model layer and wavenumber. The relaxation rate gradually approaches zero from $n = 20$ to $n = 28$ in the troposphere. As a result, horizontal wavelengths shorter than $\sim 1400\,$km are not nudged at all. The shortest relaxation time in the troposphere is 6 hours. To determine these parameters we performed test simulations and shifted the spectral tail of the relaxation rate to the smallest possible wavenumbers that ensured that the planetary-scale and synoptic-scale flow in the troposphere still followed the reanalysis. That same empirical method was also used to specify the relaxation rate. The atmospheric relaxation rate is reduced somewhat in the boundary layer because the dynamical fields close to the surface are strongly controlled by the boundary layer parameterization, which is different in the HIAMCM from that used in the model to generate the reanalysis. The relaxation rate for the surface temperature (Fig. 1b) uses the same spectral profile as the atmospheric relaxation rate in the troposphere. Since the tendency of the surface temperature is generally much smaller than that of the atmospheric temperature, the shortest relaxation time for T_s is set to 37 hours.

From the lower stratosphere on, the relaxation rate decreases with height and approaches zero towards the uppermost layers where the nudging is applied. Also note that the relaxation rates are more concentrated at larger horizontal scales in the stratosphere. The reason is that MERRA-2 applies larger scale-selective horizontal damping in the stratosphere than in the troposphere. Our intension is to nudge only the scales that specify the polar vortex, but not to nudge any large-scale intertia GWs that may develop from imbalance of the vortex and which may be different in the HIAMCM and in MERRA-2.

4. Validation of the nudged HIAMCM

We integrated the free-running HIAMCM for December. We then took 30 December at 0 UT as an initial condition and performed a nudged simulation to 1 February 2016. The free-running simulation was also continued to 1 February. To avoid the free-running simulation deviating too much from the nudged simulation as a result of internal variability associated with the polar vortex in the northern winter hemisphere, we reset the initial condition of the free-running simulation on 2 January and on 19 January at 0 UT, using the corresponding snapshots from the nudged simulation. We used the same parameters for the solar heating and ion drag in the thermosphere that correspond to moderate solar maximum conditions as in BV20. Snapshots of the two simulations were output every 10 minutes.

Figures 2a and b illustrate the temporal evolution of the large-scale upper tropospheric flow in the free running HIAMCM in terms of the zonal wind at 300 hPa ($\sim$10 km) at 0 UT on 30 December and on 1 January, respectively. As expected, the Rossby-wave structures at middle latitudes move slowly to the east. This is clearer in the southern hemisphere because of weaker stationary planetary Rossby waves than in the northern hemisphere. The MERRA-2 reanalysis on 1 January 2016 at 0 UT is shown in Fig. 2c. While the overall wind pattern looks qualitatively similar to that of the free-running HIAMCM, confirming that the model produces realistic large-scale tropospheric dynamics, the winds at a particular location may differ strongly. (Such differences would also be observed if we compared snapshots of MERRA-2 winds for different meteorological situations.) Figure 2d shows the results of the nudged HIAMCM on 1 January 2016 at 0UT. After only two days of nudging, the large-scale tropospheric flow has adjusted to the reanalysis. The minor differences between panels c and d result from finite relaxation rates and the fact that only the large scales are nudged.

When the nudging is initially imposed, the relaxation of the large-scale flow in the troposphere and lower stratosphere causes imbalances in the model equations that artificially generate large-scale GWs. Even though these artificial GWs dissipate in the troposphere within less than a day and are no longer generated later on during the nudged simulation, these waves can occur in the thermosphere for 1-2 days after initialization of the nudging.

This is because GWs have typical vertical group velocities of ∼2-10 $\mathrm{km\,h^{-1}}$, resulting in delays of about 10-50 hours by which the perturbations induced by imposing the nudging reach the thermosphere via multi-step vertical coupling. The spin-up in the thermosphere is illustrated in terms of snapshots from the free-running and the nudged HIAMCM at 0 UT on January 1 in Fig. 3. The upper (lower) two panels show north-polar (south-polar) projections at 300 km geometric height of the relative temperature perturbations due to horizontal wavenumbers $n > 30$ ($\lambda_h < 1350\,\mathrm{km}$, colors) and the large-scale ($n \leq 30$) horizontal wind (white arrows). This wind is largely determined by the diurnal tide, which is equatorward near local time midnight. Importantly, this large-scale wind is roughly the same in both simulations in either hemisphere. Furthermore, the thermosphere shows no artificial GWs in the nudged simulation 2 days after initializing the nudging. The equatorward GWs on the dayside look very similar in both simulations. The main difference is that a pronounced concentric ring GW structure centered over eastern Europe in panel a is hardly visible in panel b. This ring-structure is presumably due to tertiary GWs that result from multi-step vertical coupling over Europe. Differences in the timing of this coupling between the nudged and the free-running model are expected.

Once the model has adjusted to the nudging, we can switch off the nudging without the model generating artificial imbalances and artificial GWs. The reason is that the resultant tendencies from the nudging that keep the large-scale dynamics close to that of the reanalysis/forecasts are very small. These tendencies are only large when the nudging is initiated, but not when the nudging is switched off after the model has adjusted.

Figure 4 shows the zonal-mean climatology of the nudged and the free-running model averaged from 1 to 31 January. This comparison demonstrates that the zonal-mean zonal winds and temperatures are very similar in the two simulations. This is not necessarily expected for the mesosphere and thermosphere because this region is strongly controlled by GWs from below, and because the GW dynamics in the troposphere and stratosphere could be affected by the nudging. The fact that the mesosphere and lower thermosphere look very similar in the left and right columns of Fig. 4 indicates that the mean-flow effects from GWs and thermal tides must be similar too. This conclusion is supported by the residual mass streamfunction (contours in panel a and b), which is very similar in the two simulations. Note that the winter polar vortices are similar in the two simulations because we re-initialized the free-running simulation with snapshots from the nudged simulations on January 2 and 19. Also note that January 2016 was a period with a comparatively strong polar vortex. Therefore, the zonal-mean zonal wind is either eastward or close to zero in the winter polar mesopause region. Such a wind structure is also found in observations [e.g., *Hoffmann et al.*, 2010; *Smith*, 2012; *Harvey et al.*, 2019], particularly in the southern hemisphere [*Stober et al.*, 2021]. On the other hand, conventional models usually simulate significant westward winds in the winter polar mesopause region [e.g., *Marsh et al.*, 2013; *Pedatella et al.*, 2014].

The thin white contours in Fig. 4a,c show the zonal-mean temperature and zonal wind from MERRA-2 reanalysis averaged from 1 to 31 January 2016. From the lower troposphere up to about 1 hPa, the nudged simulation reproduces the MERRA-2 results, as expected. The differences in the lower mesosphere result from the fact that the nudging rate is very small here (see Fig. 1). Uncertainties in the polar vortex in the free-running region of models that were nudged at lower altitudes were analyzed by *Sassi et al.* [2008]. They showed that additional nudging in the MLT significantly reduces the resulting variability in the winter MLT. In addition, *Siskind et al.* [2015] reported higher model fidelity when nudging was extended to higher altitudes. It is therefore likely that additional nudging of the large-scale flow in the MLT would enhance the reliability of the simulated multi-step vertical coupling in the nudged HIAMCM.

Figure 5 illustrates the wave driving in the two simulations. The colors in panels a and b show the complete Eliassen-Palm flux (EPF) divergence which is computed from the resolved flow subject to triangular spectral truncation at a total horizontal wavenumber of $n = 256$. We use the formulation of *Zülicke and Becker* [2013] to compute the EPF divergence. The colors in panels c and d represent the resolved GW drag, which is defined by subtracting the EPF divergence due to planetary-scale waves (black contours in panel c and d) from the complete EPF divergence. The EPF divergence due to planetary-scale waves is computed by retaining only total horizontal wavenumbers $n \leq 30$ and zonal wavenumbers $m \leq 6$. The EPF divergences in panels a and b reproduce the well-known pattern in the lower and middle atmosphere, with westward wave driving in the upper troposphere and in the winter stratosphere and mesosphere, and strong eastward wave driving in the summer mesopause region [*Smith*, 2012]. The thermosphere exhibits strong westward EPF divergence. The ion drag (contours in panels a and b) in the thermosphere above about 150 km is much stronger than the EPF divergence and gives rise to a summer-to-winter-pole circulation (see the residual mass streamfunction in Fig. 4).

The black contours in Figs. 5c and d confirm that the EPF divergence in the winter stratosphere and in the thermosphere above about 150 km is mainly due to planetary-scale waves. While these are Rossby waves at lower altitudes, thermal tides give the predominant contribution to the zonal-mean EPF divergence above the mesopause *Becker* [2017, BV20]. The GW drag (colors in panels c and d) is consistent with conventional wisdom and exhibits a strong eastward drag in the summer mesopause region and a westward GW drag in the winter mesosphere. The GW drag is eastward (westward) in the winter lower (upper) thermosphere as a result of secondary (tertiary) GWs [*Becker and Vadas*, 2018; *Vadas and Becker*, 2019, BV20]. The eastward GW drag in the summer upper mesosphere is somewhat too weak compared to estimates using a GW parameterization, as was also found in BV20. The westward GW drag in the summer lower thermosphere is likely due to secondary GWs generated in the regime of the eastward summer mesospheric GW drag by the body-force mechanism [*Vadas et al.*, 2018]. The westward thermospheric GW drag is superposed with a westward EPF divergence from thermal tides, driving a reversed residual circulation in the summer lower thermosphere [*Becker*, 2017]. There is a partial cancellation between the EPF divergence from planetary-scale waves and GWs in the winter mesopause region and lower thermosphere. As shown by *Becker and Vadas* [2018], the wintertime eastward drag from secondary GWs is necessary to avoid an unrealistic reversal from eastward to westward mean zonal flow at middle and high latitudes in the winter upper mesosphere. The HIAMCM also simulates the well-known westward quasi-geostrophic EPF divergence in the summer upper mesosphere that is due to westward propagating planetary waves (such as the 2-day wave, see white contours in Figs. 5c and d). All these wave-related features compare quantitatively well between the nudged and free-running simulations, except for minor differences in the winter hemisphere that are likely due to the slightly different polar vortices. Overall, the results from our zonal-mean diagnostics suggest that the resolved GW dynamics in the nudged HIAMCM is quite similar to that in the free-running model.

This conclusion is further confirmed by the global kinetic energy spectra shown in Fig. 6. These spectra were computed as in *Brune and Becker* [2013] and were temporally averaged from 19 to 24 January. Both the nudged and the free-running HIAMCM simulate the Nastrom-Gage spectrum in the upper troposphere (panel a) with approximate -3 and $-5/3$ exponential spectral slopes at synoptic scales and in the mesoscales, respectively [e.g., *Augier and Lindborg*, 2013]. The absolute energies in the nudged and free-running simulations compare quantitatively well, even though the free-running model appears to exhibit somewhat larger energies in the mesoscales at all altitudes. The MERRA-2 reanalysis strongly underestimates the energy in the mesoscales and does not capture the mesoscale branch of the Nastrom-Gage spectrum at all. Compared to the HIAMCM, the MERRA-2 reanalysis also dramatically underestimates the mesoscale spectral energy in the stratosphere (see panel b at 1 hPa). On the other hand, both the nudged and the free-running models agree well with MERRA-2 reanalysis at planetary and synoptic scales in the upper troposphere, as well as at planetary scales in the stratosphere. The mesoscale spectral slope in the HIAMCM flattens in the stratosphere and is clearly less than -5/3 there (see panel b). Such a result was also found in *Becker and Brune* [2014]. This behavior may be due the fact that the forward energy cascade is weak in the stratosphere, and that upward propagating inertia GWs having small vertical and comparatively large horizontal scales are strongly damped, while GWs from below having small horizontal wavelengths energize the GW spectrum in the stratosphere.

According to *Becker et al.* [2020], the mesopause region exhibits maximum GW activity in the winter hemisphere due to secondary GWs. This is also the region where the secondary GWs dissipate from dynamic instability, giving rise to tertiary GWs [*Vadas and*

Becker, 2019]. The dynamic instability of the secondary GWs leads to a forward macro-turbulent energy cascade that is partly resolved in the HIAMCM. The approximate -5/3 exponential spectral slope over a wide range of scales in Fig. 6c supports this interpretation. Figure 6d shows the kinetic energy spectrum in the thermosphere at about 250 km. Here, the molecular viscosity is the predominant dissipation mechanism for GWs [*Vadas*, 2007, BV20] (see also Fig. 16). Accordingly, the exponential slope of the energy spectrum is significantly steeper than -5/3, indicating that a macro-turbulent energy cascade is of minor importance compared to the direct dissipation of resolved GWs by molecular viscosity.

Even though the zonal-mean GW effects and the mesoscale spectral kinetic energy in the nudged and free-running model are very similar, the question remains as to what extent the resolved GWs in the nudged model are realistic. Noting that large-scale inertia GWs should be well represented in MERRA-2 reanalysis, we can compare these GWs to that in the nudged HIAMCM, for example, in the upper troposphere. Figures 7a,b show snapshots at 200 hPa ($z \sim 12$ km) on 12 January 2016 at 0 UT from the nudged HIAMCM and MERRA-2 reanalysis. Colors show the temperature perturbations due to horizontal wavenumbers $n > 30$, corresponding to horizontal wavelengths smaller than 1350 km. The horizontal streamfunction (see Eq. (4)) due to wavenumbers $n \leq 30$ is shown as white contours and is essentially the same in both panels, confirming the correct nudging of the large scales. At middle and high latitudes, this streamfunction represents the large-scale (quasi-geostrophic) flow. This flow is parallel to the streamfunction contours, and the distance between contours is a measure of the wind speed. Note that the temperature perturbations for $n > 30$ are not nudged in the HIAMCM (see Fig. 1). They represent tropospheric GWs generated mainly by spontaneous emission and flow over orography. The large-to-medium-scale portion of these GWs (λ_h greater than $\sim$500 km) is resolved in MERRA-2. These GWs agree well with the large-to-medium-scale GWs in the HIAMCM. Wave packets of medium-to-small-scale GWs (λ_h smaller than $\sim$ 500 km) are simulated by the HIAMCM, for example, in the jet exit region over the Pacific, over Alaska, and over eastern Siberia (white arrows in Fig. 7a). Such GW packets are not captured by MERRA-2, which corresponds to the aforementioned deficiency of MERRA-2 regarding the mesoscale branch of the Nastrom-Gage spectrum (Fig. 6a).

Figures 7c,d show temperature perturbations and the horizontal streamfunctions in the lower stratosphere at 20 hPa ($z \sim 25$ km). Again, the large-scale streamfunctions in the two plots are nearly identical. The temperature perturbations in the HIAMCM and MERRA-2 differ significantly at this altitude, with the HIAMCM exhibiting significantly larger GW amplitudes. Again, medium-to-small scale GWs are not captured by MERRA-2. However, both data sets agree by indicating enhanced GW activity over Europe on 12 January 2016.

The results presented in this section show that our method of nudging only the large scales preserves the self-consistent simulation of GWs in the HIAMCM. Moreover, the large-to-medium-scale GWs in the upper troposphere seen in MERRA-2 reanalysis are reproduced by the nudged HIAMCM. This strongly suggests that the model can be used for comparisons of the simulated meso-scale flow in the middle and upper atmosphere with GWs in observations. This requires, however, that the large-scale flow at these altitudes is also simulated in a realistic fashion. Here we use MERRA-2 reanalysis up to about 70 km for nudging (albeit with large relaxation times above about 30 km, see Fig. 1). In the following two sections we compare the simulated GWs with satellite data and analyze the underlying dynamics for a few events.

5. Comparison of simulated stratospheric GW events in January 2016 with MERRA-2 reanalysis and AIRS satellite data

In this section we compare GWs in the stratosphere as simulated by the nudged HIAMCM with GWs in MERRA-2 and in AIRS satellite data during January 2016 [e.g., *Bossert et al.*, 2020]. AIRS temperature perturbations were derived using the high-resolution temperature retrieval method described in *Hoffmann and Alexander* [2009]. Derived temperatures have a vertical resolution which varies from ~ 7 km near 20 km altitude to a resolution of $\sim 12-14$ km near 55 km altitude. Figure 8 shows snapshots of a GW event over Northern Europe at 1:30 UT on 11 January 2016 . The left and middle columns show results from the nudged HIAMCM and from MERRA-2. As in Fig. 7, the GW temperature perturbations are computed from the wavenumber decomposition in terms of spherical harmonics, where T' includes only total horizontal wavenumbers from 31 to 256, corresponding to horizontal wavelengths smaller than ~ 1350 km. This way we compare the same GW scales from the HIAMCM and MERRA-2. The MERRA-2 snapshot at 1:30 UT is computed by linear interpolation between 0 UT and 3 UT, which is justified because the GWs resolved in MERRA-2 change slowly in time. Figures 8a,b show horizontal cross sections at $z = 33$ km, while the panels in the second and third rows are longitude-height plots at 56°N and latitude-height plots at 25°E, respectively. The grey lines mark the longitudes 0° and 25°E, the latitude 56°N, and the height 33 km. These lines are included for better comparison of the different panels.

Figures 8a and b exhibit a strong similarity regarding an inertia GW packet that extends from the Atlantic south of Ireland to the Baltic states, with negative temperature anomalies over the North Sea and the Baltic Sea. Note that this agreement of the HIAMCM with MERRA-2 is not a direct result of the nudging, because these scales are significantly smaller than the scales that are nudged (see Fig. 1). Note that the amplitudes of the inertia GWs are significantly larger in the HIAMCM than in MERRA-2. Figure 8b also shows a long strip of a negative temperature anomaly extending from the Pyrenees to Russia, as well as positive temperature anomalies farther south that maximize over Ukraine. This structure is also visible in Fig. 8a, but is superposed with medium-scale GWs that are not resolved in MERRA-2. The horizontal-height cross-sections in Figs. 8d,e and Figs. 8g,h illustrate again that the large-scale GW patterns resolved in MERRA-2 are reproduced by the HIAMCM with larger amplitudes, and that the HIAMCM shows additional smaller-scale structures not resolved in MERRA-2. In particular, the region around $25-35$ km height, $15°-35°$E, and $50°-60°$N is likely a region of GW generation, as is suggested by GW phases that emanate from this region and extend both upward and downward. The underlying generation mechanism for these GWs will be further analyzed in Sec. 7.

The right column in Fig. 8 shows the corresponding results from AIRS for the January 11 case (1:30 UT). The aforementioned inertia GW packet from the Atlantic south of Ireland to the Baltic states is also observed by AIRS, albeit with amplitudes that exceed ± 20 K. Such amplitudes are larger than that in many wintertime measurements at this ~ 33 km altitude using ground-based instruments [e.g., *Kaifler et al.*, 2015; *Chen et al.*, 2016]. On the other hand, such amplitudes can occur in the stratosphere during strong mountain wave events [*Heale et al.*, 2020]. Also note that the inertia GW packet in AIRS extends to northern Scandinavia, while its amplitude decreases with latitude north of $\sim 60°$N in the HIAMCM and in MERRA-2, and that it shows a different phase behavior in part as compared to the HIAMCM and MERRA-2. Furthermore, the difference between the absolute temperatures in AIRS and MERRA-2 is about ± 20 K in the northern Scandinavian region (not shown). On the other hand, the AIRS results exhibit medium-scale GWs south of $\sim 55°$N in Fig. 8c that are not resolved in MERRA-2 (Fig. 8b), but which resemble the medium-scale GWs in the HIAMCM (Fig. 8a) regarding amplitudes, scales, and phase orientation. These are GWs excited by orographic forcing. The phases of these GWs in AIRS are not captured by the HIAMCM. Opposite or different phases between the model and AIRS results have also been shown in a recent paper by *Hindley et al.* [2020] who investigated GW events during the wintertime in the region of the island of South Georgia using a regional model with very high resolution and driven by reanalysis at its lateral boundaries. Also note that the HIAMCM shows medium-scale GWs in the stratosphere over northern Europe at 33 km (Fig. 8a) and at lower altitudes (Fig. 8d,e) that are neither captured by MERRA-2 nor by AIRS.

From the comparison of Figs. 8f,i to Figs. 8e,h we can conclude that along 56°N and 25°E, the large-scale GWs in AIRS are qualitatively well captured by MERRA-2 between about 25 and 50 km, but that their large amplitudes and poleward extent are not. It is likely that MERRA-2 underestimates these amplitudes. The same holds for the comparison of HIAMCM with AIRS (Fig. 8d,g), even though there is improved agreement between the HIAMCM and AIRS in the stratopause region. Medium-scale GWs in the stratosphere over middle and southern Europe that are likely caused by orographic forcing are observed by AIRS. These GWs are not captured in MERRA-2, but are qualitatively well simulated by the HIAMCM.

We now show results for a GW event on January 14 (2016) at 5, 7 and 16 UT from eastern Canada to the western North Atlantic. Figure 9 shows horizontal cross-sections at 35 km for the nudged HIAMCM, MERRA-2 and AIRS data. As in the previous case, the large-scale GW structures are very similar in the HIAMCM and in the MERRA-2 reanalysis. Part of these large-scale structures are also seen in AIRS, although the AIRS amplitudes are larger at 5 and 7 UT. In particular, there is a large-scale GW packet that extends from Montreal to the Atlantic northeast of Newfoundland at 5 UT, 7 UT, and 16 UT. The AIRS data at 7 UT (panel f) shows a large negative temperature anomaly over Newfoundland and a positive anomaly farther to the West. This structure is also visible in the HIAMCM and in MERRA-2 (panel d and e). In addition, the MERRA-2 data exhibits long negative and positive "stripes" (i.e., inertial GWs) farther to the South that are aligned more zonally (panel b,e,h). These structures are captured by the HIAMCM, where they are superposed with medium-scale GWs not visible in MERRA-2 (panel a,d,g). The T' from AIRS (panel f) also exhibits some medium-scale GW activity in this region that is reminiscent of the corresponding HIAMCM result in panel d. By 16 UT, the GW structure has changed significantly (bottom row). Again, the large-scale GW pattern over eastern Canada and the North Atlantic is consistent between the HIAMCM and MERRA-2 (panel g and h). The AIRS data show some medium-scale GWs over the North Atlantic that look similar in amplitude and scale to the medium-scale GWs in the HIAMCM in that region. The curvature of the corresponding GW phases (ring-like structures in the HIAMCM data) are, however, not consistent.

From these comparisons we conclude that the HIAMCM nudged to MERRA-2 reanalysis simulates medium-scale GWs in the stratosphere reasonably well. For larger-scale GWs there is quantitative agreement between the HIAMCM and MERRA-2 regarding the GW phases, while the GW amplitudes are larger in the HIAMCM. Often these waves have even larger amplitudes in AIRS satellite data, and have different behaviors with latitude and longitude. Medium-scale GWs not resolved in MERRA-2 but resolved in the HIAMCM mostly bear a strong similarity with the corresponding GW structures in AIRS. However, this agreement does not hold everywhere, presumably because AIRS filters GWs having small vertical wavelengths. As a result, only medium-scale GWs having vertical wavelengths in excess of about 9 km are captured by the AIRS data, which is expected from the AIRS measurements [*Hoffmann and Alexander*, 2009]. This may partly explain, for example, why the medium-scale GWs seen in Fig. 8d,g do not agree with the corresponding AIRS results (Fig. 8f,i). It does not, however, explain the discrepancy in the amplitudes of the large-scale GWs in these panels.

6. Stratospheric GW activity near the Arctic vortex edge in January 2016

The comparison of GWs from the HIAMCM simulation and AIRS satellite data in Sec. 5 indicates that the amplitudes of the large-scale GWs in the HIAMCM (and in MERRA-2) are underestimated. Furthermore, the fact that the summer mesopause and the reversal from westward to eastward flow in the summer MLT are too high in altitude (Fig. 4) suggests that also medium-scale GWs resolved in the HIAMCM have amplitudes that are too small. This is because GWs with smaller amplitudes dissipate from dynamical instability at higher altitudes than GWs with larger amplitudes. On the other hand, we saw in Section 5 that wintertime medium-scale GWs simulated in the HIAMCM appear to have amplitudes similar to those in the AIRS satellite data (Figs. 8, 9). However, this comparison did not consider that the AIRS temperatures are subject to vertical averaging [*Hoffmann and Alexander*, 2009], and therefore obscure medium-scale GWs having shorter vertical wavelengths that may be resolved by the HIAMCM.

To get a better picture of the performance of the HIAMCM when compared to AIRS satellite data, we consider north polar projections of temporal averages for January 1-10, 11-20, 21-31, and 1-31 in Fig. 10. The left column shows temperature perturbations for horizontal wavenumbers $n > 30$ at a pressure surface of 2.4 hPa ($z \sim 40\,\mathrm{km}$) from the nudged HIAMCM. The right column shows the AIRS temperature perturbations. The temperature variances from the HIAMCM are larger than those from AIRS by 1-2 magnitudes (note the different color scales). However, AIRS can only see certain GWs with vertical wavelengths greater than about 9 km *Hoffmann et al.* [2014]. In order to mimic this effect, we filter the temperature perturbations from the HIAMCM via

$$\tilde{T} = \int_{p_1}^{p_2} T'(p)\, w(p)\, \frac{dp}{p} \Bigg/ \int_{p_1}^{p_2} w(p)\, \frac{dp}{p} \qquad (26)$$

with $p_1 = 0.16\,\mathrm{hPa}$ and $p_2 = 37\,\mathrm{hPa}$. $T'(p)$ denotes the local and instantaneous temperature perturbation from the HIAMCM as a function of pressure. The weighting function, $w(p)$, is shown in Fig. 11 and is similar to the kernel function used by *Hoffmann et al.* [2014] (see Fig. 4 in their paper). This function is centered at an altitude of about 40 km and extends from about 20 to 60 km. The middle column in Fig. 10 shows time averages of the filtered HIAMCM GW variances using Eq. (26). These filtered temperature variances have about the same magnitudes as in AIRS. Moreover, the HIAMCM roughly reproduces the geographical distribution seen in AIRS. The most prominent example is the stratospheric GW hot spot over Europe, which is persistent throughout the month in both data sets, and which is also evident from the unfiltered HIAMCM results. Such a hot spot is also seen during other years [*Hoffmann et al.*, 2014]. Furthermore, during January 1-10 (2016), all panels in the first row of Fig. 10 show additional centers of GW activity over northeastern Asia and over northern Alaska. An additional center of GW activity is seen over eastern North America in panels a and c. For the time period ten days later (January 11-20), the HIAMCM and AIRS agree on the intensified GW activity over Newfoundland and the North Atlantic. Furthermore, all three plots in the second row are consistent regarding reduced GW activity from about 90°E to 90°W during that period. For the January 21-31 period, the GW activity over Newfoundland and the North Atlantic is reduced, and there is an intensification of GW activity over Siberia. These features are visible in all three plots of the third row of Fig. 10. Overall, there is good quantitative agreement of the simulated time-averaged temperature variance subject to Eq. (26) with the corresponding AIRS satellite data. This suggests that the mesoscale GWs in the winter stratosphere resolvable by AIRS are simulated with reasonably realistic amplitudes by the HIAMCM. Note, however, that the AIRS data generally underestimates these amplitudes because of incomplete temporal coverage.

The HIAMCM and AIRS results agree on the fact that the strongest stratospheric GW activity is roughly coincident with the wind maximum associated with the polar vortex (see the white contours in Fig. 10 that encircle wind speeds of $90\,\mathrm{m\,s^{-1}}$ and higher). Such a feature is well known for the southern hemisphere [e.g., *Sato et al.*, 2012; *Hendricks et al.*, 2014]. The most likely explanation for this finding is that GWs generated in the troposphere have favorable vertical propagation conditions (are less prone to dissipation) if their horizontal wave vector is opposite to the mean wind and the difference between the wind speed and the horizontal phase speed is large. The reason is that the vertical wavelengths become quite long under these conditions, which helps the GWs to avoid dynamical instability and wave breaking. Furthermore, if the wind speed increases with height, the vertical group velocity of GWs propagating against the mean flow increases with height, and their amplitude growth factor with height due to the decreasing background density is less than $\exp(z/(2H))$. The latter effect is because, for a conservative monochromatic GW, the increase of the horizontal wind and temperature amplitudes with height is proportional to $|\lambda_z|^{-1/2}\exp(z/2\mathrm{H})$ [*Lindzen*, 1981], where the $|\lambda_z|^{-1}$ factor accounts for the conservation of vertical energy and momentum flux densities. The wind speed typically increases with height in the lower part of the polar vortex. Hence, this additional effect from vertical refraction also helps to avoid dissipation for GWs propagating against the mean flow at the edge of the polar vortex. We therefore expect that wintertime stratospheric GW amplitudes are strongest around the wind maximum partly as a result of vertical refraction.

Another contributing factor is horizontal refraction. This means that the horizontal wave vector of a GW that propagates oblique to the polar vortex is refracted due to horizontal wind shear in a way that the wavevector tends to be opposite to the wind in the vicinity of the wind maximum. Thereby, GW are focussed into the wind maximum [*Senf and Achatz*, 2011]. A third factor is the in-situ generation of GWs from imbalance of the vortex, which is discussed in the next two sections.

7. Analysis of stratospheric GWs in January 2016

Next we analyze the stratospheric GW events over northern Europe and over eastern Canada/North Atlantic in more detail. Figure 12 shows simulated temperature variations due to horizontal wavenumbers $n > 30$ (horizontal wavelengths shorter than ~ 1350 km) from the nudged HIAMCM over northern Europe at 1:30 UT on January 11. The upper two panels show the temperature perturbations plus the horizontal streamfunction (white contours) at two pressure surfaces in the stratosphere, while the lower two panels show longitude-height and latitude-height cross-sections at 56°N and 25°E, respectively, using pressure as the vertical coordinate and scaling the temperature perturbation with $(p/5\mathrm{hPa})^{1/2}$. This scaling would result in a constant GW amplitude with height in the absence of refraction and dissipation. Figure 12a (20 hPa, $z \sim 25$ km) features GW packets that range 1) from eastern Spain to the western Mediterranean, which presumably are orographic GWs (OGWs) forced mostly by eastward flow over the Central and Iberian Mountains in Spain, 2) from eastern France to the Adriatic Sea, which presumably are OGWs formed by flow over the Alps, and 3) from northern Germany to Russia east of the Baltic states. The latter GWs (#3) have phase fronts that are aligned southwest to northeast, and are composed of the inertia GW packet discussed in the previous subsection. The situation in the upper stratosphere (panel b, 2.5 hPa, $z \sim 40$ km) yields a more blended and uniform picture, which suggests that there is a single, large GW packet propagating over Europe which includes both medium and large-scale GWs.

Although the blended nature of Fig. 12b suggests that all of the (European) GWs are OGWs, some of which could be trailing far north and east of their excitation location over the Alps as recently argued by *Dörnbrack* [2021], Figs. 12c,d reveal that the medium and large-scale GWs over northeastern Europe in panels a and b cannot, in fact, be a GW packet with a tropospheric (e.g., orographic) origin. The pressure-scaled temperature variations in Fig. 12c show a constant amplitude with height at about $30-2$ hPa ($z \sim 25$

to 45 km) and 15°–35°E. Furthermore, these GWs have larger pressure-scaled amplitudes than the GWs in the lower stratosphere, which would not make sense if the GWs were upward propagating, for example, from 50 to 10 hPa. Therefore, these GWs appear to emanate from a source region that is located at 30–5 hPa ($z \sim 25-35$ km) and 15°–35°E in Fig. 12c. Figure 12d suggests a similar altitude regime for GW generation at about 54°–58°N.

From the inclination of the GW phases in Fig. 12c and assuming upward GW propagation above about 10 hPa, we can conclude that the zonal wavenumber component of the GWs at 56°N over northeastern Europe (west of 30°E) is westward (relative to the large-scale flow). Similarly, the GW phases above 10 hPa in Fig. 12d indicate a northward wavenumber component. The GW phases in the lower stratosphere in panel c slope from west to east with increasing height below 50–30 hPa and for 15°–35°E, which is consistent with downward propagating westward GWs. Farther above, the GWs phases slope from east to west, which is consistent with upward propagating westward GWs. This indicates that the GW source region reaches somewhat farther into the lower stratosphere than is suggested by the scaled GW amplitudes. From Fig. 12d we can infer that north of 56°N and below about 20 hPa, most of the GW phases slope southward with increasing height. These GWs presumably propagate north-westward and downward, which is consistent with a GW source around 20 hPa and 56°N. South of 56°N and between about 50 and 10 hPa, most of the GW phases are consistent with downward and southward propagation. Note that there are no continuous phase lines extending from the upper troposphere to the mid stratosphere in panel d, even not south of 50°N. Given all these considerations, the GWs in the stratosphere over northern Europe at 1.30 UT on 11 January 2016 seem to emanate mainly from the 30 to 10 hPa altitude region.

The partly "X-shaped" patterns of GW phases seen in Figs. 12c,d are characteristic of the GWs excited by local body forces [*Vadas et al.*, 2003, 2018]. A local body force refers to a spatially and temporally localized momentum deposition created by the dissipation of a GW packet, which results into an imbalance of the ambient flow. Therefore, GWs that are generated in-situ from the polar vortex due to spontaneous emission should bear some similarity with GWs generated by the body-force mechanism [see also discussion in *Bossert et al.*, 2020]. GW generation in the upper troposphere and in the winter stratosphere from imbalances of the quasi-geostrophic (QG) flow is well known [e.g., *O'Sullivan and Dunkerton*, 1995; *Zhang*, 2004; *Zülicke and Peters*, 2006; *Sato and Yoshiki*, 2008; *Synder et al.*, 2009]. This generation process is often referred to as "spontaneous emission" [*Plougonven and Zhang*, 2014]. While mathematical solutions for the flow response to local body forces were derived by *Vadas et al.* [2003], a corresponding mathematical theory is not available for spontaneous emission. A widely used method is to use criteria that detect imbalances of the QG flow, such as the nonlinear balance equation (NBE). A more advanced theory for a general decomposition of balanced and imbalanced flow was recently proposed by *Gassmann* [2019].

In the present study, we apply the NBE to the large-scale flow to help to interpret the generation of GWs from unbalanced flow. While previous studies employed this theory in Cartesian coordinates [e.g., *Zhang*, 2004], we hereby derive the NBE in spherical coordinates and with pressure as vertical coordinate for better applicability to meteorological data. This derivation is given in Appendix B (see Eq. (B19)) and yields the result of

Zhang [2004] in the f-plane approximation and when the geostrophic horizontal wind is plugged into the Jacobian used in Eq. (2) of *Zhang* [2004]. Note that the inclusion of spherical geometry leads to additional terms that are ignored when the usual formula in Cartesian coordinates is applied. For planetary-scale flows like the polar vortex, these additional terms can be important. ΔNBE represents the lowest order of the non-balanced tendency of horizontal divergence. According to QG scaling for the atmosphere, this interpretation is restricted to large horizontal wavelengths (e.g., larger than 1350 km, see Appendix B). Since QG theory does not apply to the mesoscales, the nonlinear balance equation is considered to be only an indicator of the phases of synoptic-scale GWs that result from imbalance, with the possibility that mesoscale GWs may also be generated.

In addition to the NBE, we derive the mesoscale kinetic energy budget in Appendix B, assuming that the large-scale vortical flow is the mean flow. This allows for the detection of regions where GWs are amplified due to kinetic energy transfer from the mean flow to the GWs (positive mesoscale kinetic energy source, MKS > 0, see Eq. (B22)). Ideally, such a GW source region should also show negative mesoscale potential energy flux convergence (MPC < 0, see Eq. (B21)). Thus, our formalism consists of two significant parts: 1) regions where the flow is unbalanced and likely creates GWs as indicated by ΔNBE, and 2) regions where those created GWs can grow significantly in amplitude by extracting energy from the mean flow. To our knowledge, this second part (MKS > 0 and $MPC < 0$) has not been previously studied.

Figure 13 shows ΔNBE (Eq. (B19)) for the same cross-sections as in Fig. 12. The pattern of ΔNBE corresponds to large-scale GWs that are not included in the temperature perturbations shown in Fig. 12. The overall horizontal pattern of ΔNBE in the upper panels of Fig. 13 indicates stronger large-scale imbalances in the stratosphere over northern than southern Europe, which is consistent with the upper panels of Fig. 8. Furthermore, ΔNBE in Fig. 13a is reminiscent of the large-scale GW packet over Scandinavia seen in AIRS (Fig. 8c). By definition, ΔNBE does not describe the predominant GW scales visible in Figs. 8 and 12. Moreover, comparison of Figs. 13c,d and Figs. 12c,d indicates that also the propagation directions of the synoptic-scale GWs described by ΔNBE can be different from the propagation directions of the medium-scale GWs.

Figure 14 allows for an interpretation of the GW generation from spontaneous emission in terms of kinetic energy transfer from the background flow to the GWs and GW potential energy flux convergence. The colors in Figs. 14a,b show the GW temperature perturbations as in Figs. 12c,d. The black contours show the horizontal wind speed, indicating that the latitude of the assumed stratospheric GW sources coincides approximately with the latitude of the maximum wind speed associated with the polar vortex (panel b). Figures 14c-f show the pressure-weighted kinetic energy transfer (MKS) and the mesoscale potential energy convergence (MPC). To diagnose these quantities from the model data, we first computed the MKS and MPC fields on the model grid and transformed these quantities into series of spherical harmonics. Horizontal averaging as indicated on the right-hand sides of Eqs. (B22) and (B21) is defined by using a triangular truncation at wavenumber 30 when transforming the spectral representations of MKS and MPC back into physical space. From Figs. 14c and d it is apparent that the MKS is positive and maximum in the area of the assumed GW source: at about $15°-35°$E, $50°-60°$N, and $30-5\,$hPa. Figures 14e,f show the mesoscale potential energy flux convergence. The pronounced minima around 10 hPa indicate maximum flux divergence where the mesoscale kinetic energy source is maximum. Thus, the combination of MKS > 0 and MPC < 0 suggests that there is a GW source around $15° - 35°$E, $50° - 60°$N, and $30 - 5$ hPa.

Regions with significant MKS and MPC are also visible in Figs. 14c-f in the stratopause region from about 3 to 0.3 hPa. These regions are presumably indicative of either GW amplification or damping due to transient interaction with the mean flow. A region of GW dissipation (MKS < 0 and MPC > 0) is visible in the lower mesosphere above 0.3 hPa. This altitude region coincides with the onset the maximum westward GW drag in Figs. 5c and d.

This example for GW generation in the northern winter stratosphere suggests that, in addition to secondary GWs generated in the upper stratosphere and lower mesosphere by the body force mechanism from wave breaking/dissipation, the in-situ generation of GWs due to imbalances of the QG flow associated with the polar vortex in the mid stratosphere and the subsequent amplification through interaction with the large-scale flow may play a significant role for the GW effects in the northern winter mesosphere and thermosphere.

The amplification of GW amplitudes through energy transfer from the mean flow to the GWs (Eqs. (B22)) is different from the usual vertical refraction effect, whereby a non-dissipating vertically propagating GW exhibits amplitude growth larger than $e^{z/2H}$ (where H is the density scale height) when approaching a critical level, and amplitude growth weaker than $e^{z/2H}$ when propagating against a background wind that increases with height. According to this strictly linear reasoning, the westward and upward propagating GWs between about 50 and 5 hPa in Fig. 14a should show pressure-scaled amplitudes that decrease with altitude because the eastward zonal wind increases with altitude there, thus refracting the GWs to longer vertical wavelength and enhanced vertical group velocity, requiring smaller energy density for constant vertical energy flux density in the non-dissipative case. Equation (B22), on the other hand, describes a nonlinear mechanism that, in our example, has a much stronger effect on the GW amplitudes than the refraction effect.

Comparing the colors with the contours in Figs. 14c-f yields that MKS and MPC are largely determined by the vertical advection and vertical convergence terms (last terms on the right-hand sides of Eqs. (B22) and (B21)), even though both vertical and horizontal terms are required for a quantitative assessment of the mesoscale kinetic energy budget. This suggests that vertical wind shear is crucial for the amplification of GWs generated by spontaneous emission.

Figure 15 shows an analysis of the GW event over the exit region of the North American upper tropospheric jet on January 14 at 7 UT. This event began on January 11 and persisted through to January 22 (see also previous section and Fig. 10). The GW packet in the tropopause region over Newfoundland and the western North Atlantic in Fig. 15a is an example of a GW generation in the troposphere by the baroclinic jet–front system, with positive ΔNBE in the exit region of the upper tropospheric jet, as was shown to be typical for such events by *Zhang* [2004, see his Fig. 10] and which is confirmed by Fig. 15b.

Another example is found farther to the South. Two GW packets in the tropopause region can be seen southeast of Newfoundland ($\sim$45°N, $\sim$50°W) and over the northeastern US ($\sim$40°N, $\sim$80°W). In a longitude-height plot along 42.5°N (panel c), these GWs appear to extend into the stratosphere, and their phase inclination indicates westward propagation relative to the mean flow, as expected. At these altitudes, these GWs have smaller horizontal scales than the GWs in the tropopause region. This is presumably because of selective transmission into the stratosphere, whereby the GWs with smaller horizontal wavelengths have larger vertical wavelengths and larger vertical group velocities, and are therefore less prone to dissipation (see the flattening of the horizontal energy spectra from the upper troposphere to the stratosphere in Fig. 6). Above about 100 hPa, the largest pressure-scaled GW amplitudes in Fig. 15c occur between about 30 and 3 hPa. This suggests that these GWs are amplified in this region, as is confirmed in panel d which shows by MKS > 0 and MPC < 0 from 80°W to 40°W and from about 30 to 3 hPa. This GW amplification is difficult to distinguish from GW generation due to imbalance. We speculate that in this example, spontaneous emission acts to amplify the GWs propagating upward from the troposphere. Again we found (not shown in the figure) that the vertical terms in Eqs. (B22) and (B21) give the predominant contributions to the energy conversion terms.

8. Summary and conclusions

We presented a new version of the HIgh Altitude Mechanistic general Circulation Model (HIAMCM) with nudging to MERRA-2 reanalysis in the troposphere, stratosphere, and lower mesosphere. The free-running HIAMCM is a high-resolution, whole-atmosphere GCM with resolved GWs up to an altitude of about 450 km (depending on the thermospheric temperature) and was described in detail in *Becker and Vadas* [2020]. Its dynamical core is based on the spectral-transform method for the primitive equations using a terrain-following vertical coordinate. The HIAMCM includes a correction for nonhydrostatic dynamics and a consistent extension of the underlying thermodynamic relationships into the thermosphere. The explicit simulation of the generation, propagation, and dissipation of gravity waves (GWs) is achieved by combining high spatial resolution with an advanced macro-turbulent horizontal and vertical diffusion scheme that consistently includes molecular viscosity. A sponge layer is not required because resolved GWs dissipate mainly from molecular viscosity above $z \sim 200\,\mathrm{km}$. In the updated HIAMCM we use a triangular spectral truncation at total horizontal wavenumber $n = 256$, corresponding to a gridspacing of 52 km, and 280 full vertical levels with a level spacing of $\sim 600-650\,\mathrm{m}$ below $z \sim 130\,\mathrm{km}$, which increases with altitude to about 10 km at $z \sim 400\,\mathrm{km}$. The HIAMCM is considered to be a mechanistic model because the computations of radiative transfer and moist processes are simplified compared to comprehensive models. Furthermore, it does not include chemistry, and the only parameterization of ionospheric processes is ion drag.

When nudging a GW-resolving model to reanalysis it is important to retain the model's properties regarding the simulated GW dynamics. To this end, nudging can not be applied in gridspace, as is usually done in models with parameterized GWs, because this would artificially either damp or generate GWs, subject to the resolved mesoscales in the underlying reanalysis. We therefore applied the nudging in spectral space such that only horizontal wavelengths longer than $\sim 1500\,\mathrm{km}$ ($\sim 2000\,\mathrm{km}$) in the troposphere (stratosphere) are relaxed to reanalysis (Fig. 1). We demonstrated that the simulated GW activity in the nudged HIAMCM is equivalent to that in the free-running model by comparing snapshots in the thermosphere, effects from GWs in the zonal-mean momentum budget, and global horizontal kinetic energy spectra (Figs. 3, 5, and 6).

Case studies for the Arctic winter in January 2016 showed that simulated GWs having horizontal wavelengths of about 500-1000 km were very similar to those in MERRA-2 reanalysis, even though these scales were not nudged (Figs. 7). In addition, the HIAMCM simulated medium-to-smale-scale GWs not resolved in MERRA-2 (Figs. 8 and 9). The temperature perturbations due to these GWs exhibited reasonable similarity with corresponding AIRS satellite data. We applied vertical filtering to the simulated stratospheric temperature perturbations to mimic the kernel function applied in the AIRS data product of *Hoffmann et al.* [2014], and we computed maps of the time-averaged stratospheric temperature variance centered around $z \sim 40$ km (Fig. 10). The HIAMCM results showed roughly the same GW amplitudes and spatial distribution as AIRS. In particular, we found that the strongest wintertime stratospheric GW activity occurs roughly where the wind speeds are strongest. We argued that vertical and horizontal refraction of GWs contributes to this behavior.

The spatial distribution of the stratospheric GW activity during January 2016 showed a persistent GW hot spot over Europe. Furthermore, this period was characterized by a relatively strong polar vortex, as well as by weather systems from the Atlantic penetrating into Europe, causing GW generation from spontaneous emission and flow over orography [e.g., *Bossert et al.*, 2020; *Heale et al.*, 2020]. The aforementioned simulation results with the nudged HIAMCM motivated us to analyze a case on January 11 over Northern Europe where vertically resolved AIRS satellite data were available. We identified GW generation by spontaneous emission in the stratosphere in the HIAMCM simulation nudged to MERRA-2 reanalysis. We applied the nonlinear balance equation in spherical geometry and analyzed the GW kinetic energy budget, specifically the transfer for kinetic energy from the large-scale vortical flow to the mesoscale GWs and the associated mesoscale potential energy flux convergence (see Appendix, Eqs. (B19) and (B20)-(B22)). While the nonlinear balance equation indicates only synoptic-scale GW structures, the transfer of kinetic energy from the large-scale flow to the GWs allowed us to identify the regions where mesoscale GWs are generated or amplified via energy transfer (MKS > 0). We found that the GW amplification is mainly due to vertical momentum flux combined with vertical wind shear. Since the same region also showed significant GW potential energy divergence (negative convergence, MPC < 0), we concluded that this was a source region for medium-scale GW generated by spontaneous emission. Moreover, negative energy transfer combined with positive convergence (corresponding to positive energy deposition in the classical single column picture) allowed us to identify a region of GW dissipation in the lower mesosphere.

A second case for January 14 showed GW generation in the upper troposphere southwestward of Newfoundland and over the northeastern US. These jet-generated waves propagated into the stratosphere. In the lower stratosphere, they were either amplified by energy transfer from the mean flow or were superposed with GWs generated in situ by spontaneous emission. Again, the combination of kinetic energy transfer from the mean flow to the GWs combined with negative potential energy flux convergence confirmed the stratospheric GW amplification or GW source.

The implications from these case studies are: 1) Though it is difficult to see stratospheric GW sources in AIRS satellite data because of its the limited vertical resolution, the combination with GW-resolving model data allows for the analysis of observed GWs regarding GW generation and dissipation. 2) The energy transfer from the large-scale vortical flow to the GWs combined with the GW potential energy flux convergence is a valuable diagnostic tool to identify GW generation or amplification due to imbalance. Without such a diganostic method, GWs in the upper stratosphere can be misinterpreted as trailing mountain waves if the corresponding primary OGWs are also present, as was the case during the investigated January 2016 period [e.g. *Dörnbrack*, 2021]. Whether this new diagnostic tool is also useful to identify GW sources related to the body force mechanism [e.g., *Vadas et al.*, 2018] remains to be investigated.

The formula for the energy transfer term (Eq. (B22)) can explain why the source region of GWs generated by spontaneous emission in the middle atmosphere lies typically in the lower to mid stratosphere and at the edge of the polar vortex where the wind is maximum in a horizontal cross-section. The likely reason is that the vertical shear of the large-scale horizontal wind, dU/dz, is largest at an altitude below where U is maximum. Since this altitude (for example, $z \sim 40\,\mathrm{km}$) is below the wind maximum associated with the polar vortex, the regime of maximum wind in a horizontal cross-section at this altitude is roughly also the regime of maximum vertical wind shear. Maximum vertical wind shear facilitates the amplification of in-situ generated GWs that propagate against the mean flow according to Eq. (B22). Therefore, GWs generated by spontaneous emission in the lower and mid stratosphere may also contribute to the observation of maximum GW activity around the wind maximum of the polar vortex in horizontal cross-sections (Fig. 10). Furthermore, GWs generated in the winter stratosphere by spontaneous emission will dissipate in the upper mesosphere and thermosphere, and the associated body forces will lead to secondary GWs that propagate higher up into the thermosphere. Therefore, these GWs also contribute to multi-step vertical coupling [*Vadas and Becker*, 2019; *Becker and Vadas*, 2020].

This paper demonstrates that the HIAMCM can successfully be nudged to reanalysis while retaining its ability to explicitly simulate the generation, propagation, and dissipation of GWs up to the thermosphere. This allows for comparison of the simulated GW events in the winter hemisphere with GW observations and to study the underlying mechanisms. Future applications of the nudged HIAMCM include, for example, the relative contribution of the different GW sources in the winter troposphere and stratosphere to multi-step vertical coupling.

Appendix A: Macro-turbulent horizontal diffusion

The scheme for macro-turbulent and molecular diffusion in the HIAMCM is described in detail in BV20. Here, we mention the modifications introduced in the updated HIAMCM regarding the macro-turbulent horizontal diffusion only.

The tendencies of the horizontal wind and sensible heat from the macro-turbulent horizontal diffusion (mthd) can be written as (see Sec. 2 in BV20):

$$\left(\partial_t \mathbf{v}\right)_{mthd} = \frac{1}{\partial_\eta p}\nabla \left(\partial_\eta p \left(\left(K_h \, \mathsf{S}_h + K_{hf} \, \mathsf{S}_{hf} \right)\right)\right) \tag{A1}$$

$$\begin{aligned}\left(c_p \, \partial_t T\right)_{mthd} &= \frac{1}{\partial_\eta p}\nabla \left(\partial_\eta p \left(Pr_h^{-1} \left(K_h \, \nabla T + K_{hf} \, \nabla T_f \right)\right)\right) \\ &\quad + K_h \left(\mathsf{S}_h \nabla\right) \cdot \mathbf{v} + K_{hf} (\mathsf{S}_{hf} \nabla) \cdot \mathbf{v}\end{aligned} \tag{A2}$$

Here, p is pressure, η is the model's vertical coordinate, and Pr_h is a (macro-turbulent) horizontal Prandtl number. The horizontal shear tensors are

$$\mathsf{S}_h = \left(\left(\nabla + \mathbf{e}_z/a_e\right) \circ \mathbf{v} \right) + \left(\left(\nabla + \mathbf{e}_z/a_e\right) \circ \mathbf{v} \right)^T - \mathsf{E} \, D \tag{A3}$$

$$\mathsf{S}_{hf} = \left(\left(\nabla + \mathbf{e}_z/a_e\right) \circ \mathbf{v}_f \right) + \left(\left(\nabla + \mathbf{e}_z/a_e\right) \circ \mathbf{v}_f \right)^T - \mathsf{E} \, D_f \, , \tag{A4}$$

where $\mathbf{e}_z$ is the unit vector in the vertical direction, a_e is the earth radius, E is the unit tensor, $D = \nabla \cdot \mathbf{v}$ is the horizontal divergence, and the symbol $\circ$ denotes the tensor product. Furthermore, $\mathbf{v}_f$ and D_f are the filtered horizontal wind and its divergence, while T_f is the filtered temperature. The filtering is with respect to the total horizontal wavenumber, n, and selects only horizontal wavelengths smaller than $\sim$200 km. The filter function in the spectral representation of winds and temperature has the form

$$F_n \;=\; \begin{cases} (n-n_f)^2 \,/\, (N-n_f)^2 & \text{for } n \;>\; n_f \\ 0 & \text{else}\,, \end{cases} \tag{A5}$$

where $n_f = 200$ and $N = 256$. The horizontal diffusion terms in Eqs. (A1) and (A2) that involve the filtered components extend the harmonic horizontal diffusion scheme by a stress-tensor-based hyperdiffusion.

The classical Smagorinsky scheme specifies the horizontal diffusion coefficient with the mixing-length concept of Ludwig Prandtl. Using the symbol l_h for the horizontal mixing length, we write the macro-turbulent horizontal diffusion coefficient as [*Becker*, 2009]

$$K_h \;=\; l_h^2 \, (\, |\, \mathsf{S}_h \,|^2 + S_{hmin}^2 \,)^{1/2} \left(1 + \alpha \, F(R_i - R_{i0})\right) \tag{A6}$$

$$F(R_i) \;=\; \begin{cases} \sqrt{1 - 18\,R_i} & \text{for } R_i \;\leq\; 0 \\ 1/(1 + 9\,R_i) & \text{for } R_i \;>\; 0\,. \end{cases} \tag{A7}$$

Here, $S_{hmin}^2 = 4 \times 10^{-12}\,\mathrm{s}^{-2}$ is the minimum squared horizontal wind shear, ensuring that the spatial derivatives of K_h are always defined, and R_i is the Richardson number. The Richardson number criterion is included in the definition of K_h such that scale-selective horizontal damping is increased for $R_i < R_{i0}$. As in BV20, we account for the linear criterion of GW instability using $R_{i0} = 0.25$ in the middle atmosphere and lower thermosphere.

In BV20, we followed the method of *Brune and Becker* [2013] and used a linear hyperdiffusion, that is, K_{hf} was specified as a function of η. In the updated version of the HIAMCM we introduce a dependence of the hyperdiffusion coefficient on the horizontal shear and dynamic instability using

$$K_{hf} \;=\; K_{hf\,0} \;+\; 4.9\,K_h \tag{A8}$$

Test simulations showed that this nonlinear method improves the effective resolution of the model (see also Fig. 6).

In order to provide complete information about the updated macro-turbulent horizontal diffusion scheme, Figs. 16a-c show the prescribed vertical profiles of the Richardson number offset, the scaling factor for the Richardson number criterion, the squared horizontal mixing length, and the inverse horizontal Prandtl number. In addition, the simulated global-mean hyperdiffusion and Smagorinsky-type diffusion coefficients are shown in panel c and d, respectively. Note that the new hyperdiffusion coefficient is mainly due to the nonlinear term (second term on the right-hand side of Eq. (A8)) from stratopause to the mesopause region (panel c). Also note that K_h and $Pr_h^{-1}K_h$ are completed by the molecular viscosity and heat conduction, respectively, as is described in BV20. The blue curve in Fig. 16d demonstrates that molecular viscosity is the dominant horizontal diffusion coefficient in the upper thermosphere.

Appendix B: Gravity-wave generation due to deviations from quasi-geostrophic balance

To provide the context for our diagnostic method we first recapitulate some basics of quasi-geostrophic (QG) theory [e.g., *Holton*, 1994]. QG theory approximately describes the dynamics of geostrophic flow. The underlying assumptions apply only to the large horizontal scales ($L > 1000$–$2000\,\mathrm{km}$). Furthermore, QG theory is limited to the extratropics and to heights above the boundary layer up to about $p \sim 0.01\,\mathrm{hPa}$ or $z \sim 80\,\mathrm{km}$. Here, we outline QG theory in spherical geometry, as is necessary for application to meteorological data.

Using pressure as the vertical coordinate, the geostrophic wind, $\mathbf{v}_g$ is defined via geostrophic balance according to

$$0 \,=\, \mathbf{v}_g \times f_0\,\mathbf{e}_z \,-\, \nabla\Phi_g\,. \tag{B1}$$

where f_0 is a fixed Coriolis parameter (e.g., an average over a latitude band), $\mathbf{e}_z$ is the unit vector in the vertical direction, and Φ_g is the geostrophic geopotential. The order of the geostrophic wind is

$$O(\mathbf{v}_g) \,=\, U \,\sim\, 30\,\mathrm{m\,s^{-1}}\,. \tag{B2}$$

The relation of inertial forces and the Coriolis force is measured by the Rossby number, which is defined as

$$Ro \,=\, O(\xi_g)\,/\,f_0 \,=\, U\,/\,(\,L\,f_0\,)\,, \tag{B3}$$

where $O(\xi_g) \;=\; U/L$ for the geostrophic relative vorticity and $Ro \sim 0.1$ for QG flow. The temporal evolution of the geostrophic flow can be computed from the QG potential vorticity (PV) equation which is obtained as follows: 1) We derive the relative vorticity equation from the horizontal momentum equation (see Sec. 3),

$$\partial_t\mathbf{v} \,=\, \mathbf{v}\times(f+\xi)\,\mathbf{e}_z - \dot{p}\,\partial_p\mathbf{v} - \nabla\mathbf{v}^2/2 - \nabla\Phi + \mathbf{R}\,, \tag{B4}$$

where $\dot{p}$ denotes the material rate of change of the pressure, Φ is the hydrostatic geopotential, and $\mathbf{R}$ represents turbulent friction; 2) we expand this vorticity equation in powers of Ro, yielding

$$(\partial_t + \mathbf{v}_g\cdot\nabla)\,(f+\xi_g) \,=\, f_0\,\partial_p\,\dot{p} \,+\, \mathbf{e}_z\cdot(\nabla\times\mathbf{R}) \,+\, O(Ro\,U^2/L^2)\,; \tag{B5}$$

3) we substitute $\partial_p \dot{p}$ from the sensible heat equation in the QG approximation. The final result is:

$$(\partial_t + \mathbf{v}_g \cdot \nabla)\, q \;=\; \delta \,+\, O(Ro\, U^2/L^2) \tag{B6}$$

$$q \;=\; f + \nabla^2\, \Psi_g + \partial_p \left(g^2\, f_0^2\, \rho_r^2\, N_r^{-2}\, \partial_p \Psi_g \right)$$

$$\delta \;=\; \mathbf{e}_z \cdot (\nabla \times \mathbf{R}) \;-\; f_0\, \partial_p(\rho_r\, Q)$$

Here, Ψ_g is the streamfunction of the geostrophic wind, q denotes the QG PV, Q is the diabatic heating, and ρ_r and N_r denote the density profile and the buoyancy frequency of the reference state, respectively.

The horizontal divergence equation related to Eq. (B5) plays a passive role in QG theory, because the balanced ageostrophic flow can be deduced from the geostrophic flow. Expansion of the horizontal divergence equation with respect to powers of Ro leads to the so-called nonlinear balance equation. The complete horizontal divergence equation related to Eq. (B4) and in spherical geometry can be written as:

$$\partial_t D \;=\; -(\mathbf{v} \cdot \nabla \,+\, \dot{p}\partial_p)D \,+\, f\,\xi \,-\, \nabla^2\Phi_h \,-\, u\,\partial_y f \,-\, D^2 \,-\, \partial_p \mathbf{v} \cdot \nabla\dot{p} \,+\, \mathbf{v}^2/a_e^2 \tag{B7}$$

$$+\,2\,\big(\,(D - \partial_y v)\,\partial_y v \,-\, (\xi + \partial_y u)\,\partial_y u\,\big) \,+\, \nabla \cdot \mathbf{R}$$

Here, u and v are the zonal and meridional wind components, respectively, $\partial_y = a_e^{-1}\,\partial_\phi$ is the derivation in the latitudinal direction, and a_e denotes the Earth radius. Expanding each term in Eq. (B7) with respect to the Rossby number according to the usual QG scaling, we can derive the following relations:

$$O\big(\tfrac{U^2}{L^2\,Ro}\big): \quad 0 \;=\; f_0\,\xi_g \,-\, \nabla^2\Phi_g \tag{B8}$$

$$O\big(\tfrac{U^2}{L^2}\big): \quad 0 \;=\; (f - f_0)\,\xi_g \,+\, f_0\,\xi_{ag} \,-\, u_g\,\partial_y f \,-\, \nabla^2\Phi_{ag} \tag{B9}$$

$$+\,2\,\big(\,-(\partial_y v_g)^2 - \,(\xi_g + \partial_y u_g)\,\partial_y u_g\,\big)\,.$$

Terms of order $Ro\,U^2/L^2$ or higher give rise to a complicated tendency equation for the horizontal divergence that is not further used in this study. While Eq. (B8) corresponds to geostrophic balance, Eq. (B9) is a constraint for QG balance. Here, ξ_{ag} and Φ_{ag} are the balanced ageostrophic relative vorticity and geopotential, respectively. Since it is difficult in meteorological data to distinguish between ξ_g and ξ_{ag} or Φ_g and Φ_{ag}, one can combine Eqs. (B8) and (B9) into a single constraint that is known as the nonlinear balance equation,

$$\Delta\text{NBE} \;=\; f\,\xi \,-\, \nabla^2\Phi \,-\, u_g\,\partial_y f - \,2\,(\partial_y v_g)^2 -\, 2\,(\xi_g + \partial_y u_g)\,\partial_y u_g\,, \qquad \text{(B10)}$$

with $\Delta\text{NBE} = 0$ being the constraint for QG balance. In that case, ξ and Φ in Eq. (B10) include only balanced components. For the sake of convenience, we have added $(f - f_0)\,\xi_{ag}$ on the right-hand side of Eq. (B10) which is of order $O\big(\,Ro\,U^2/L^2\big)$. Equation (B10) is equivalent to Eq. (2) in *Zhang* [2004] if we assume the f-plane approximation and use $\partial_x u_g = -\partial_y v_g$ as well as $\xi_g = \partial_x v_g - \partial_y u_g$ (both of which are incomplete in spherical coordinates). Also note that the geostrophic horizontal wind must be plugged into the Jacobian used in Eq. (2) of *Zhang* [2004]. As noted by *Zhang* [2004], the deviation of ΔNBE from zero marks the regions where QG balance is violated. Such regions are thought to be indicative of GW generation by spontaneous emission, which typically results from large nonlinearities of the QG flow. More specifically, ΔNBE is the leading order tendency of the unbalanced ageostrophic horizontal divergence. Therefore, it indicates the large-scale (inertia) GWs generated by spontaneous emission.

In the following we derive an expression that explicitly describes the amplification of mesoscale ageostrophic flow. We start again with the horizontal momentum equation (B4) and assume a decomposition of the flow into large scales (superscript ^{ls}) and mesoscales (superscript ^{ms}). This decomposition can be applied to meteorological data when we assume the spectral decomposition described in Sec. 3. Here we assume that the large-scale components include total horizontal wavenumbers from $n = 0$ to $n = 30$, corresponding to a minimum horizontal wavelength of $\sim 1350\,\mathrm{km}$, and that the mesoscales include all smaller scales contained in the data (up to wavenumber $n = 256$ or down to horizontal wavelengths of $\sim 156\,\mathrm{km}$ in the case of the current HIAMCM version). For the sake of feasibility, the large-scale vortical wind is denoted as $\mathbf{v}_{\tilde{g}}$ and the large-scale relative vorticity as $\xi_{\tilde{g}}$, and we assume that these large-scale components include only geostrophic and balanced ageostrophic components. Hence, the ageostropic flow is defined as the mesoscale vortical flow plus all components related to horizontal divergence, part of which is in QG balance for the large scales. This ageostrophic flow is denoted as $\mathbf{v}_{\widetilde{ag}} = \mathbf{v}^{ls}_{\widetilde{ag}} + \mathbf{v}^{ms}$ for the horizontal wind and $\xi_{\widetilde{ag}} = \xi^{ms}$ for the mesoscale relative vorticity. The geopotential is decomposed as $\Phi = \Phi^{ls} + \Phi^{ms}$, where $\Phi^{ls} = \Phi_g + \Phi^{ls}_{\widetilde{ag}}$. The notation for the streamfunction representation of the large-scale vortical flow corresponds to a modified definition of the geostrophic wind: $\mathbf{v}_{\tilde{g}} \times f_0\mathbf{e}_z = \nabla\Phi_{\tilde{g}}$. We now plug this decomposition into Eq. (B4) and sort the individual terms with respect to powers of the Rossby number. The leading order terms determine the dynamics of the geostrophic flow:

$$
\begin{aligned}
O\left(\frac{U^2}{L\,Ro} + \frac{U^2}{L}\right): \quad \partial_t \mathbf{v}_{\tilde{g}} \;&=\; \mathbf{v}_{\tilde{g}} \times (f + \xi_{\tilde{g}})\,\mathbf{e}_z + \mathbf{v}^{ls}_{\widetilde{ag}} \times f\mathbf{e}_z \;-\; \nabla\,\mathbf{v}^2_{\tilde{g}}/2 \;-\; \nabla\,\Phi_{\tilde{g}} \qquad \text{(B11)} \\
&\quad + \overline{\mathbf{v}^{ms} \times \xi^{ms}\mathbf{e}_z} \;-\; \frac{1}{2}\nabla\,\overline{(\mathbf{v}^{ms})^2} \;-\; \overline{\dot{p}^{ms}\,\partial_p \mathbf{v}^{ms}}\,.
\end{aligned}
$$

Here, the second row includes wave-mean flow interaction of the mesoscales acting on the large-scale geostrophic flow, and horizontal averaging over the GW scale (e.g., 1350 km times 1350 km) of a quantity X is indicated by $\overline{X}$. The Stokes drift from GWs is neglected for the sake of simplicity. Furthermore, we assume that subgrid-scale diffusion affects only the mesoscales and can therefore can be neglected for the large scales. The large-scale ageostrophic horizontal momentum equation in this decomposition is

$$O\left(\frac{Ro\,U^2}{L}\right): \quad \partial_t \mathbf{v}_{\widetilde{ag}}^{ls} = \mathbf{v}_{\widetilde{ag}}^{ls} \times \xi_{\tilde{g}} \mathbf{e}_z + \mathbf{v}_{\tilde{g}} \times \xi_{\widetilde{ag}}^{ls} - \nabla\left(\mathbf{v}_{\tilde{g}} \cdot \mathbf{v}_{\widetilde{ag}}^{ls}\right) - \dot{p}\,\partial_p \mathbf{v}_{\tilde{g}} - \nabla\Phi_{\widetilde{ag}}^{ls} \tag{B12}$$

and is not of further importance for our purpose. The remaining momentum equation for the (ageostrophic) mesoscales is analogous to Eq. (B12), but includes in addition the Coriolis force for the mesoscales:

$$O\left(\frac{Ro\,U^2}{L}\right): \quad \partial_t \mathbf{v}^{ms} = \mathbf{v}^{ms} \times (f_0 + \xi_g)\,\mathbf{e}_z + \mathbf{v}_{\tilde{g}} \times \xi^{ms}\mathbf{e}_z - \nabla\left(\mathbf{v}_{\tilde{g}} \cdot \mathbf{v}^{ms}\right) - \dot{p}^{ms}\,\partial_p \mathbf{v}_{\tilde{g}} - \nabla\Phi^{ms} + \mathbf{R}^{ms}. \tag{B13}$$

This equation yields the usual linear horizontal momentum equation for GWs if we apply the f-plane approximation and assume that $\mathbf{v}_{\tilde{g}}$ is uniform and constant. Note that Eq. (B13) does not include the interaction with the large-scale ageostrophic flow. It includes, however, the advection of the large-scale geostrophic flow by the mesoscales. These terms are usually neglected when computing the GW dispersion and polarization relation from the f-plane version of Eq. (B13), but must be retained to derive the correct mesoscale kinetic energy budget. This budget follows upon multiplication of Eq. (B13) with $\mathbf{v}^{ms}$ and averaging over the GW scale. The mesoscale kinetic energy budget then yields after several manipulations (invoking the continuity equation and hydrostatic balance for the mesoscale flow):

$$\partial_t \overline{(\mathbf{v}^{ms})^2}/2 \,+\, \mathbf{v}_{\tilde{g}} \cdot \nabla \overline{(\mathbf{v}^{ms})^2}/2 \tag{B14}$$

$$= -\partial_p \,(\, \overline{\Phi^{ms}\dot{p}^{ms}} \,) \,-\, \nabla \cdot (\, \overline{\Phi^{ms}\mathbf{v}^{ms}} \,)$$

$$-\big(\, \overline{(v^{ms})^2} - \overline{(u^{ms})^2} \,\big)\, \partial_y v_{\tilde{g}} \,-\, \overline{u^{ms}v^{ms}} \,(\, \xi_{\tilde{g}} + 2\,\partial_y u_{\tilde{g}} \,) \,-\, (\, \overline{\mathbf{v}^{ms}\dot{p}^{ms}} \,) \cdot \partial_p \mathbf{v}_{\tilde{g}}$$

$$-\, R\,p^{-1}\, \overline{T^{ms}\dot{p}^{ms}} \,+\, \overline{\mathbf{v}^{ms} \cdot \mathbf{R}^{ms}} \,.$$

When we neglect all horizontal derivatives (single-column approximation) in Eq. (B14) and substitute the friction term by the corresponding negative mechanical dissipation rate, ϵ^{ms}, we arrive at the GW kinetic energy equation given in, for example, *Becker and McLandress* [2009, their Eq. (9)] or *Becker* [2017, his Eq. (7), see also references therein]:

$$\partial_t \overline{(\mathbf{v}^{ms})^2}/2 \,=\, -\,\partial_p \overline{\Phi^{ms}\dot{p}^{ms}} \,-\, \overline{\mathbf{v}^{ms}\dot{p}^{ms}} \cdot \partial_p \mathbf{v}_{\tilde{g}} \,-\, R\,p^{-1}\, \overline{T^{ms}\dot{p}^{ms}} \,-\, \overline{\epsilon^{ms}} \,. \tag{B15}$$

The only differences of Eq. (B15) to the previous forms of the GW kinetic energy equation in the single-column approximation are that we assume the geostrophic flow as the background flow and therefore neglect the vertical advection of mesoscale kinetic energy, and that the kinetic energy equation is transformed into the pressure vertical coordinate system. The sum of the first two terms on the right-hand side of Eq. (B15) is known as the energy deposition of gravity waves (GWs) [*Hines and Reddy*, 1967]. In the quasi-stationary limit, the energy deposition is positive definite and is balanced by the buoyancy production of mesoscale kinetic energy and the mechanical dissipation (third and last term on the right-hand side of Eq. (B15)). The buoyancy production is either zero for conservative GWs, or negative in the dissipative case. In the quasi-stationary dissipative case, the buoyancy production equals the negative thermal dissipation of GWs [*Becker*, 2017, his Eq. (12)]. The leading term of the energy deposition (first term on the right-hand side of Eq. (B15)) is the convergence of the vertical potential energy flux. This term is positive for dissipating GWs. The second term is the shear production of mesoscale kinetic energy, which is usually negative for dissipating GWs [e.g., *Becker and McLandress*, 2009].

Equation (B14) holds in the general case where we do not resort to the single-column or steady-state approximation. We rewrite this equation in the following way:

$$\partial_t \overline{(\mathbf{v}^{ms})^2}/2 + \mathbf{v}_{\tilde{g}} \cdot \nabla \overline{(\mathbf{v}^{ms})^2}/2 = \mathrm{MPC} + \mathrm{MKS} - R p^{-1} \overline{T^{ms} \dot{p}^{ms}} - \overline{\epsilon^{ms}} \tag{B16}$$

$$\mathrm{MPC} = -\nabla \cdot \left(\overline{\Phi^{ms} \mathbf{v}^{ms}} \right) - \partial_p \left(\overline{\Phi^{ms} \dot{p}^{ms}} \right) \tag{B17}$$

$$\mathrm{MKS} = -\left(\overline{(v^{ms})^2} - \overline{(u^{ms})^2} \right) \partial_y v_{\tilde{g}} - \overline{u^{ms} v^{ms}} \left(\xi_{\tilde{g}} + 2\, \partial_y u_{\tilde{g}} \right) - \left(\overline{\mathbf{v}^{ms} \dot{p}^{ms}} \right) \cdot \partial_p \mathbf{v}_{\tilde{g}} \,. \tag{B18}$$

Here, MPC is the 3D mesoscale potential energy flux convergence and MKS denotes the mesoscale kinetic energy source (which equals the three-dimensional shear production). MKS is the only term by which the mesoscale kinetic energy can increase due to interaction with the mean flow. Given the aforementioned properties of MPC and MKS in the steady state and single-column approximation for dissipating GWs, it is plausible to assume that MKS is positive in regions of GW generation from spontaneous emission. Furthermore, potential energy flux is expected to emanate from a GW source region, which is therefore expected to be associated with negative MPC (equivalent to positive potential energy flux divergence). In Sec. 7 of this paper we use MKS and MPC to identify GW sources from spontaneous emission.

For the sake of technical feasibility, we apply the same flow decomposition made to derive Eqs. (B20)-(B22) to the nonlinear balance equation. This is leads to the following approximate formula to identify deviations from QG balance in the tendency of the large-scale horizontal divergence:

$$\Delta\mathrm{NBE} \;=\; f\,\xi_{\tilde{g}} \,-\, \nabla^2\Phi^{ls} \,-\, u_{\tilde{g}}\,\partial_y f - 2\,(\partial_y v_{\tilde{g}})^2 - \,2\,(\xi_{\tilde{g}} + \partial_y u_{\tilde{g}})\,\partial_y u_{\tilde{g}}\,. \tag{B19}$$

When evaluating MKS and MPC using z as vertical coordinate, several terms in Eqs. (B20)-(B22) need to be substituted by the corresponding expressions in the z-system. Using the anelastic approximation according to *Becker* [2017], the GW kinetic energy equation in the z-system corresponding to Eqs. (B20)-(B22) can be written as:

$$\partial_t\,\overline{(\mathbf{v}^{ms})^2}/2 \,+\, \mathbf{v}_{\tilde{g}}\cdot\nabla\,\overline{(\mathbf{v}^{ms})^2}/2 \;=\; \mathrm{MPC} \,+\, \mathrm{MKS} \,+\, \frac{g}{T^{ls}}\,\overline{T^{ms}w^{ms}} \,-\, \overline{\epsilon^{ms}} \tag{B20}$$

$$\mathrm{MPC} = -\frac{1}{\rho^{ls}}\,\nabla\cdot(\,\overline{p^{ms}\mathbf{v}^{ms}}\,) \,-\, \frac{1}{\rho^{ls}}\,\partial_z\,(\,\overline{p^{ms}w^{ms}}\,) \tag{B21}$$

$$\mathrm{MKS} = -\big(\,\overline{(v^{ms})^2} - \overline{(u^{ms})^2}\,\big)\,\partial_y v_{\tilde{g}} - \overline{u^{ms}v^{ms}}\,(\,\xi_{\tilde{g}} + 2\,\partial_y u_{\tilde{g}}\,) - (\,\overline{\mathbf{v}^{ms}w^{ms}}\,)\cdot\partial_z\mathbf{v}_{\tilde{g}}\,. \tag{B22}$$

Here, p^{ms} and w^{ms} are the mesoscale perturbations of the pressure and vertical wind.

Acknowledgments. Model simulations were performed by the authors. Model documentations can be found in Sec. 2, Appendix A and in *Becker and Vadas* [2020]. The MERRA-2 reanalysis used in this study are publicly available at https://disc.gsfc.nasa.gov/dataset. AIRS data are available at
`https://datapub.fz-juelich.de/slcs/airs/gravity_waves/data/variance_4mu/`
All data shown in this paper are available via NWRA's website under
`https://www.cora.nwra.com/~erich.becker/Becker-etal-JGRA-2021-files`. EB and SLV were supported by DARPA contract 140D6319C0032, by NSF grant AGS-1832988 and by NASA grants 80NSSC19K0836 and 80NSSC20K0628. EB was additionally supported by NASA grant 80NSSC19K0834 and by the Leibniz Institute of Atmospheric Physics at the University of Rostock (IAP), which provided the HPC facility used for this study. SLV was additionally supported by NSF grant 1822867. VLH ac-

knowledges NASA grants 80NSSC18K1046, 80NSSC19K0834, and 80NSSC20K0628. CZ was supported by the Deutsche Forschungsgemeinschaft through grant ZU 120/2-2 (FOR 1898). The valuable comments of three anonymous reviewers are gratefully acknowledged.

References

Augier, P., and E. Lindborg (2013), A new formulation of the spectral energy budget of the atmosphere, with application to two high-resolution general circulation models, *J. Atmos. Sci.*, *70*, 2293–2308, doi:10.1175/JAS-D-12-0281.1.

Becker, E. (2009), Sensitivity of the upper mesosphere to the Lorenz energy cycle of the troposphere, *J. Atmos. Sci.*, *66*, 648–666, doi:10.1175/2008JAS2735.1.

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.

Becker, E., and S. Brune (2014), Reply to "Comments on 'Indications of stratified turbulence in a mechanistic GCM'", *J. Atmos. Sci.*, *71*, 858–862, doi:10.1175/JAS-D-13-0281.1.

Becker, E., and C. McLandress (2009), Consistent scale interaction of gravity waves in the doppler spread parameterization, *J. Atmos. Sci.*, *66*, 1434–1449.

Becker, E., and S. L. Vadas (2018), Secondary gravity waves in the winter mesosphere: Results from a high-resolution global circulation model, *J. Geophys. Res. Atmos.*, *123*, doi:10.1002/2017JD027460.

Becker, E., and S. L. Vadas (2020), Explicit global simulation of gravity waves in the thermosphere, *J. Geophys. Res. Space Phys.*, doi:10.1029/2020JA028034.

Becker, E., M. Grygalashvyly, and G. R. Sonnemann (2020), Gravity wave mixing effects on the OH*-layer, *Advances in Space Research*, *65*, 175–188, doi: 10.1016/j.asr.2019.09.043.

Bosilovich, M. G., S. Akella, L. Coy, R. Cullather, C. Draper, R. Gelaro, R. Kovach, Q. Liu, A. Molod, P. Norris, K. Wargan, W. Chao, R. Reichle, L. Takacs, Y. Vikhliaev, S. Bloom, A. Collow, S. Firth, G. Labow, G. Partyka, S. Pawson, O. Reale, S. D. Schubert, , and M. Suarez (2015), MERRA-2: Initial evaluation of the climate, *NASA Tech. Rep. Series on Global Modeling and Data Assimilation , NASA/TM-2015-104606/Vol. 43*, Goddard Space Flight Center.

Bossert, K., S. L. Vadas, L. Hoffmann, E. Becker, V. L. Harvey, and M. Bramberger (2020), Observations of stratospheric gravity waves over Europe on 12 January 2016: The role of the polar night jet, *J. Geophys. Res. Atmos.*, doi:10.1029/2020JD032893.

Brune, S., and E. Becker (2013), Indications of stratified turbulence in a mechanistic GCM, *J. Atmos. Sci.*, *70*, 231–247, doi:10.1175/JAS-D-12-025.1.

Butchart, N., I. Cionni, V. Eyring, T. G. Shepherd, D. W. Waugh, H. Akiyoshi, J. Austin, C. Brühl, M. P. Chipperfield, E. Cordero, M. Dameris, R. Deckert, S. Dhomse, S. M. Frith, R. R. Garcia, A. Gettelman, M. A. Giorgetta, D. E. K. F. Li, E. Mancini, C. McLandress, S. Pawson, G. Pitari, D. A. Plummer, E. Rozanov, F. Sassi, J. F. Scinocca, K. Shibata, B. Steil, and W. Tian (2010), Chemistry-climate model simulations of twenty-first century stratospheric climate and circulation changes, *J. Climate*, *23*, 5349–5374, doi:10.1175/2010JCLI3404.1.

Chen, C., X. Chu, J. Zhao, B. R. Roberts, Z. Yu, W. Fong, X. Lu, and J. A. Smith (2016), Lidar observations of persistent gravity waves with periods of 3−10 h in the Antarctic middle and upper atmosphere at McMurdo (77.83°S, 166.67°E), *J. Geophys. Res. Space Physics*, *121*, 1483–1502, doi:10.1002/2015JA022127.

Dörnbrack, A. (2021), Stratospheric mountain waves trailing across Northern Europe, *J. Atmos. Sci.*, *78*, 2835–2857, doi:10.1175/JAS-D-20-0312.1.

Dörnbrack, A., S. Gisinger, N. Kaifler, T. C. Portele, M. Bramberger, M. Rapp, M. Gerding, M. Söder, N. Zagar, and D. Jelic (2018), Gravity waves excited during a minor sudden stratospheric warming, *Atmos. Chem Phys.*, *18*, 12,915–12,931, doi:10.5194/acp-18-12915-2018.

Funke, B., W. Ball, S. Bender, A. Gardini, V. L. Harvey, A. Lambert, M. López-Puertas, D. R. Marsh, K. Meraner, H. Nieder, S.-M. Päivärinta, K. Pérot, C. E. Randall, T. Reddmann, E. Rozanov, H. Schmidt, A. Seppälä, M. Sinnhuber, T. Sukhodolov, G. P. Stiller, N. D. Tsvetkova1, P. T. Verronen, S. Versick, T. von Clarmann, K. A. Walker, and V. Yushkov (2017), HEPPA-II model-measurement intercomparison project: EPP indirect effects during the dynamically perturbed NH winter 2008-2009, *Atmos. Chem. Phys.*, *17*, 3573–3604, doi:10.5194/acp-17-3573-2017.

Garcia, R. R., D. R. Marsh, D. E. Kinnison, B. A. Boville, and F. Sassi (2007), Simulation of secular trends in the middle atmosphere, 1950-2003, *J. Geophys. Res.*, *112*, doi: 10.1029/2006JD007485.

Gassmann, A. (2019), Analysis of large-scale dynamics and gravity waves under shedding of inactive flow components, *Mon. Wea. Rev.*, *47*, 2861–2876, doi:10.1175/MWR-D-18-0349.1.

Gong, J., D. L. Wu1, and S. D. Eckermann (2012), Gravity wave variances and propagation derived from AIRS radiances, *Atmos. Chem. Phys.*, *12*, 1701–1720, doi:doi:10.5194/acp-

12-1701-2012.

Harvey, V. L., C. E. Randall, E. Becker, A. K. Smith, C. G. Barden, J. A. France, and L. P. Goncharenko (2019), Evaluation of the mesospheric polar vortices in waccm, *J. Geophys. Res. Atmos.*, *124*, doi:10.1029/2019JD030727.

Heale, C. J., K. Bossert, S. L. Vadas, L. Hoffmann, A. Dörnbrack, G. Stober, J. B. Snively, and C. Jacobi (2020), Secondary gravity waves generated by breaking mountain waves over europe, *J. Geophys. Res. Atmos.*, *125*, doi:10.1029/2019JD031662.

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

Hindley, N. P., C. J. Wright, A. M. Gadian, L. Hoffmann, J. K. Hughes, D. R.Jackson, J. C. King, N. J. Mitchell, T. Moffat-Griffin, A. C. Moss, S. B.Vosper, and A. N. Ross (2020), Stratospheric gravity-waves over the mountainous island of southgeorgia: testing a high-resolution dynamical model with 3-d satellite observations and radiosondes, *Atmos. Chem. Phys.*, doi:10.5194/acp-2020-465.

Hines, C. O., and C. A. Reddy (1967), On the propagation of atmospheric gravity waves through regions of wind shear, *J. Geophys. Res.*, *72*, 1015–1034.

Hoffmann, L., and M. J. Alexander (2009), Retrieval of stratospheric temperatures from Atmospheric Infrared Sounder radiance measurements for gravity wave studies, *J. Geophys. Res.*, *114*(D07105), doi:10.1029/2008JD011241.

Hoffmann, L., X. Xue, and M. J. Alexander (2013), A global view of stratospheric gravity wave hotspots located with Atmospheric Infrared Sounder observations, *J. Geophys. Res. Atmos.*, *118*, 416–434, doi:10.1029/2012JD018658.

Hoffmann, L., M. J. Alexander, C. Clerbaux, A. W. Grimsdell, C. I. Meyer, T. Robler, and B. Tournier (2014), Intercomparison of stratospheric gravity wave observations with AIRS and IASI, *Atmos. Meas. Tech.*, *7*, 4517–4537, doi:10.5194/amt-7-4517-2014.

Hoffmann, L., A. W. Grimsdell, and M. J. Alexander (2016), Stratospheric gravity waves at southern hemisphere orographic hotspots: 2003-2014 AIRS/Aqua observations, *Atmos. Chem. Phys.*, *16*, 9381–9397, doi:10.5194/acp-16-9381-2016.

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

Hoffmann, P., E. Becker, W. Singer, and M. Placke (2010), Seasonal variation of mesospheric waves at northern middle and high latitudes, *J. Atmos. Sol.-Terr. Phys.*, *72*, 1068 – 1079.

Holton, J. R. (1994), *Introduction to Circulating Atmospheres*, Cambridge University Press.

Jones Jr., M., D. P. Drob, D. E. Siskind, J. P. McCormack, A. Maute, S. E. McDonald, and K. F. Dymond (2018), Evaluating different techniques for constraining lower atmospheric variability in an upper atmosphere general circulation model: A case study during the 2010 sudden stratospheric warming, *Journal of Advances in Modeling Earth Systems*, *10*, 3076–3102, doi:10.1029/2018MS001440.

Kaifler, N., , F.-J. Lübken, J. Höffner, R. J. Morris, and T. P. Viehl (2015), Lidar observations of gravity wave activity in the middle atmosphere over davis (69°s, 78°e), antarctica, *J. Geophys. Res. Atmos.*, *120*, doi:10.1002/2014JD022879.

Lindzen, R. S. (1981), Turbulence and stress owing to gravity wave and tidal breakdown, *J. Geophys. Res.*, *86*, 9707–9714.

Liu, H.-L. (2017), Large wind shears and their implications for diffusion in regions with enhanced static stability: The mesopause and the tropopause, *J. Geophys. Res. Atmos.*, *122*, doi:10.1002/2017JD026748.

Makela, J. J., S. L. Vadas, R. Muryanto, T. Duly, and G. Crowley (2010), Periodic spacing between consecutive equatorial plasma bubbles, *Geophys. Res. Lett.*, *37*(L14103), doi: 10.1029/2010GL043968.

Marsh, D. R., M. J. Mills, D. E. Kinnison, J. F. Lamarque, N. Calvo, and L. M. Polvani (2013), Climate change from 1850 to 2005 simulated in CESM1(WACCM), *J. Clim.*, *26*, 7372–7391, doi:10.1175/JCLI-D-12-00558.1.

McLandress, C., and T. G. Shepherd (2009), Simulated anthropogenic changes in the Brewer-Dobson circulation, including its extension to high latitudes, *J. Clim.*, *22*, 1516–1540.

McLandress, C., W. E. Ward, V. I. Fomichev, K. Semeniuk, S. R. Beagley, N. A. McFarlane, and T. G. Shepherd (2006), Large-scale dynamics of the mesosphere and lower thermosphere: An analysis using the extended Canadian Middle Atmosphere Model, *J. Geophys. Res.*, *111*, doi:10.1029/2005JD006776.

McLandress, C., J. F. Scinocca, T. G. Shepherd, M. C. Reader, and G. L. Manney (2013), Dynamical control of the mesosphere by orographic and nonorographic gravity wave drag during the extended northern winters of 2006 and 2009, *J. Atmos. Sci.*, *70*(7), 2152–2169, doi:10.1175/JAS-D-12-0297.1.

O'Sullivan, D., and T. J. Dunkerton (1995), Generation of inertia-gravity waves in a simulated life-cycle of baroclinic instability, *J. Atmos. Sci.*, *52*, 3695–3716.

Pedatella, N. M., T. Fuller-Rowell, H. Wang, H. Jin, Y. Miyoshi, H. Fujiwara, H. Shinagawa, H.-L. Liu, F. Sassi, H. Schmidt, V.Matthias, and L. Goncharenko (2014), The neutral dynamics during the 2009 sudden stratosphere warming simulated by different whole atmosphere models, *J. Geophys. Res. Space Phys.*, *119*, 1306–1324, doi: 10.1002/2013JA019421.

Plougonven, R., and F. Zhang (2014), Internal gravity waves from atmospheric jets and fronts, *Rev. Geophys.*, *52*, 33–76, doi:10.1002/2012RG000419.

Plougonven, R., A. Hertzog, and L. Guez (2013), Gravity waves over Antarctica and the Southern Ocean: consistent momentum fluxes in mesoscale simulations and stratospheric balloon observations, *Q. J. Roy. Meteor. Soc.*, *139*, 101–118, doi: 10.1002/qj.19651.

Randall, C. E., V. L. Harvey, L. A. Holt, D. R. Marsh, D. Kinnison, B. Funke, and P. F. Bernath (2015), Simulation of energetic particle precipitation effects during the 2003-2004 Arctic winter, *J. Geophys. Res. Space Physics*, *120*, 5035–5048, doi: 10.1002/2015JA021196.

Sassi, F., D. E. Siskind, J. L. Tate, H.-L. Liu, and C. E. Randall (2008), Simulations of the boreal winter upper mesosphere and lower thermosphere with meteorological specifications in SD-WACCM-X, *J. Geophys. Res. Atmos.*, *123*, 3791–3811, doi: 10.1002/2017JD027782.

Sato, K., and M. Yoshiki (2008), Gravity wave generation around the polar vortex in the stratosphere revealed by 3-hourly radiosonde observations at Syowa Station, *J. Atmos.*

Sci., *65*, 3719–3735, doi:10.1175/2008JAS2539.1.

Sato, K., S. Tanteno, S. Watanabe, and Y. Kawatani (2012), Gravity wave characteristics in the southern hemisphere revealed by a high-resolution middle-atmosphere general circulation model, *J. Atmos. Sci.*, *69*, 1378–1396, doi:10.1175/JAS-D-11-0101.1.

Satoh, M., H. Tomita, H. Yashiro, H. Miuram, C. Kodama, T. Seiki, A. T. Noda, Y. Yamada, D. Goto, M. Sawada, T. Miyoshi, Y. Niwa, M. Hara, T. Ohno, S. Iga, T. Arakawa, T. Inoue, and H. Kubokawa (2014), The Non-hydrostatic Icosahedral Atmospheric Model: Description and development, *Progress in Earth and Planetary Science*, *1*(18), doi:10.1186/s40645-014-0018-1.

Schmidt, H., G. P. Brasseur, and M. A. Giorgetta (2010), Solar cycle signal in a general circulation and chemistry model with internally generated quasi-biennial oscillation, *J. Geophys. Res.*, *115*(D00I14), doi:10.1029/2009JD012542.

Senf, F., and U. Achatz (2011), On the impact of middle-atmosphere thermal tides on the propagation and dissipation of gravity waves, *J. Geophys. Res.*, *116*, doi: 10.1029/2011JD015794.

Shibuya, R., and K. Sato (2019), A study of the dynamical characteristics of inertia-gravity waves in the Antarctic mesosphere combining the PANSY radar and a non-hydrostatic general circulation model, *Atmos. Chem. Phys.*, *19*, 3395–3415, doi:10.5194/acp-19-3395-2019.

Simmons, A. J., and D. M. Burridge (1981), An energy and angular momentum conserving vertical finite-difference scheme and hybrid vertical coordinates, *Mon. Wea. Rev.*, *109*, 758–766.

Sinnhuber, M., H. Nieder, and N. Wieters (2012), Energetic particle precipitation and the chemistry of the mesosphere / lower thermosphere, *Surv. in Geophys.*, *33*, 1281–1334, doi:10.1007/s10712-012-9201-3.

Siskind, D. E., F. Sassi, C. E. Randall, V. L. Harvey, M. E. Hervig, and S. M. Bailey (2015), Is a high-altitude meteorological analysis necessary to simulate thermosphere-stratosphere coupling?, *Geophys. Res. Lett.*, *42*, 8225–8230, doi:10.1002/2015GL065838.

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

Smith, A. K., N. M. Pedatella, D. R. Marsh, and T. Matsuo (2017), On the dynamical control of the mesosphere-lower thermosphere by the lower and middle atmosphere, *J. Atmos. Sci.*, *74*, 933–947, doi:10.1175/JAS-D-16-0226.1.

Solomon, S. C., H.-L. Liu, D. R. Marsh, J. M. McInerney, L. Qian, and F. M. Vitt (2019), Whole atmosphere climate change: Dependence on solar activity, *J. Geophys. Res. Space Phys.*, *124*, 3799–3809, doi:10.1029/2019JA026678.

Stober, G., D. Janches, V. Matthias, D. Fritts, J. Marino, T. Moffat-Griffin, K. Baumgarten, W. Lee, D. Murphy, Y. H. Kim, N. Mitchell, and S. Palo (2021), Seasonal evolution of winds, atmospheric tides, and reynolds stress components in the southern hemisphere mesosphere-lower thermosphere in 2019, *Ann. Geophys.*, *39*, 1–29, doi:10.5194/angeo-39-1-2021.

Synder, C., R. Plougonven, and D. J. Muraki (2009), Mechanisms for spontaneous gravity wave generation within a dipole vortex, *J. Atmos. Sci.*, *66*, 3464–3478, doi:10.1175/2009JAS3147.1.

Vadas, S. L. (2007), Horizontal and vertical propagation and dissipation of gravity waves in the thermosphere from lower atmospheric and thermospheric sources, *J. Geophys. Res.*, *112*, doi:10.1029/2006JA011845.

Vadas, S. L., and I. Azeem (2021), Concentric secondary gravity waves in the thermosphere and ionosphere over the continental united states on 25-26 march 2015 from deep convection, *J. Geophys. Res. Space Phys.*, doi:10.1029/2020JA028275.

Vadas, S. L., and E. Becker (2019), Numerical modeling of the generation of tertiary gravity waves in the mesosphere and thermosphere during strong mountain wave events over the Southern Andes, *J. Geophys. Res. Space Phys.*, doi:10.1029/2019JA026694.

Vadas, S. L., and G. Crowley (2010), Sources of the traveling ionospheric disturbances observed by the ionospheric TIDDBIT sounder near Wallops Island on October 30, 2007, *J. Geophys. Res. Space Physics*, *115*(A07324), doi:10.1029/2009JA015053.

Vadas, S. L., and H.-L. Liu (2013), Numerical modeling of the large-scale neutral and plasma responses to the body forces created by the dissipation of gravity waves from 6 h of deep convection in Brazil, *J. Geophys. Res.*, *118*, 2593–2617, doi:10.1002/jgra.50249.

Vadas, S. L., D. C. Fritts, and M. J. Alexander (2003), Mechanisms for the generation of secondary waves in wave breaking regions, *J. Atmos. Sci.*, *60*, 194–214, doi: 10.1029/2004JD005574.

Vadas, S. L., H.-L. Liu, and R. S. Lieberman (2014), Numerical modeling of the global changes to the thermosphere and ionosphere from the dissipation of gravity waves from deep convection, *J. Geophys. Res.*, *119*, doi:10.1002/2014JA020280.

Vadas, S. L., J. Zhao, X. Chu, and E. Becker (2018), The excitation of secondary gravity waves from local body forces: Theory and observation, *J. Geophys. Res. Atmos.*, *123*, 9296–9325, doi:10.1029/2017JD027970.

Vadas, S. L., S. Xu, J. Yue, K. Bossert, E. Becker, and G. Baumgarten (2019), Characteristics of the quiet-time hotspot gravity waves observed by GOCE over the Southern Andes on 5 July 2010, *J. Geophys. Res. Space Phys.*, doi:10.1029/2019JA026693.

von Storch, H., H. LANGENBERG, and F. FESER (2000), A spectral nudging technique for dynamical downscaling purposes, *Mon. Ea. Rev.*, *123*, 3664–3673.

Watanabe, S., and S. Miyahara (2009), Quantification of the gravity wave forcing of the migrating diurnal tide in gravity wave-resolving general circulation model, *J. Geophys. Res.*, *114*(D07110), doi:10.1029/2008JD011218.

Zhang, F. (2004), Generation of mesoscale gravity waves in upper-tropospheric jet-front systems, *J. Atmos. Sci.*, *61*, 440–457.

Zülicke, C., and E. Becker (2013), The structure of the mesosphere during sudden stratospheric warmings in a global circulation model, *J. Geophys. Res.*, *118*, 1–17, doi:10.1002/jgrd.50219.

Zülicke, C., and D. H. W. Peters (2006), Simulation of inertia-gravity waves in a poleward breaking rossby wave, *J. Atmos. Sci.*, *63*, 3253 – 3276, doi:10.1175/JAS3805.1.

Zülicke, C., and D. H. W. Peters (2008), Parameterization of strong stratospheric inertia gravity waves forced by poleward breaking rossby waves, *Mon. Wea. Rev.*, *136*, 98–119, doi:10.1175/2007MWR2060.1.

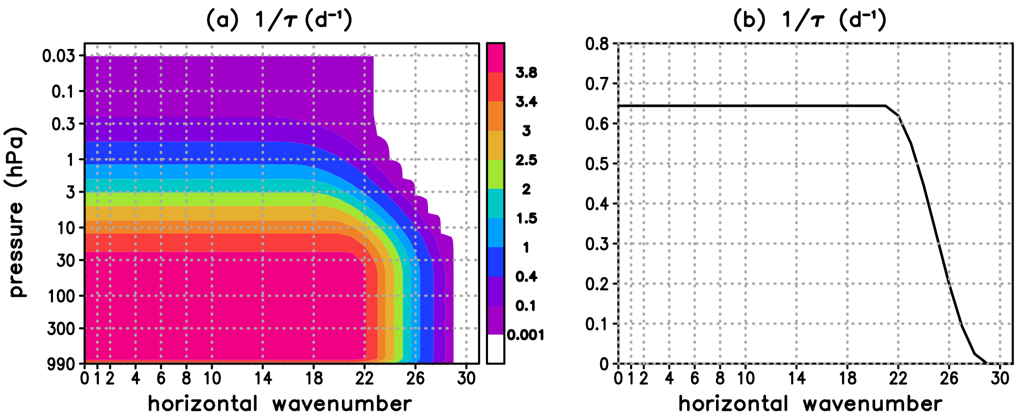

Figure 1. (a) Atmospheric relaxation rate as a function of the total horizontal wavenumber and the model's hybrid vertical coordinate times 1013 hPa. The shortest relaxation time is 6 hours. (b) Relaxation rate used to nudge the surface temperature. The shortest relaxation time is ~ 37 hours.

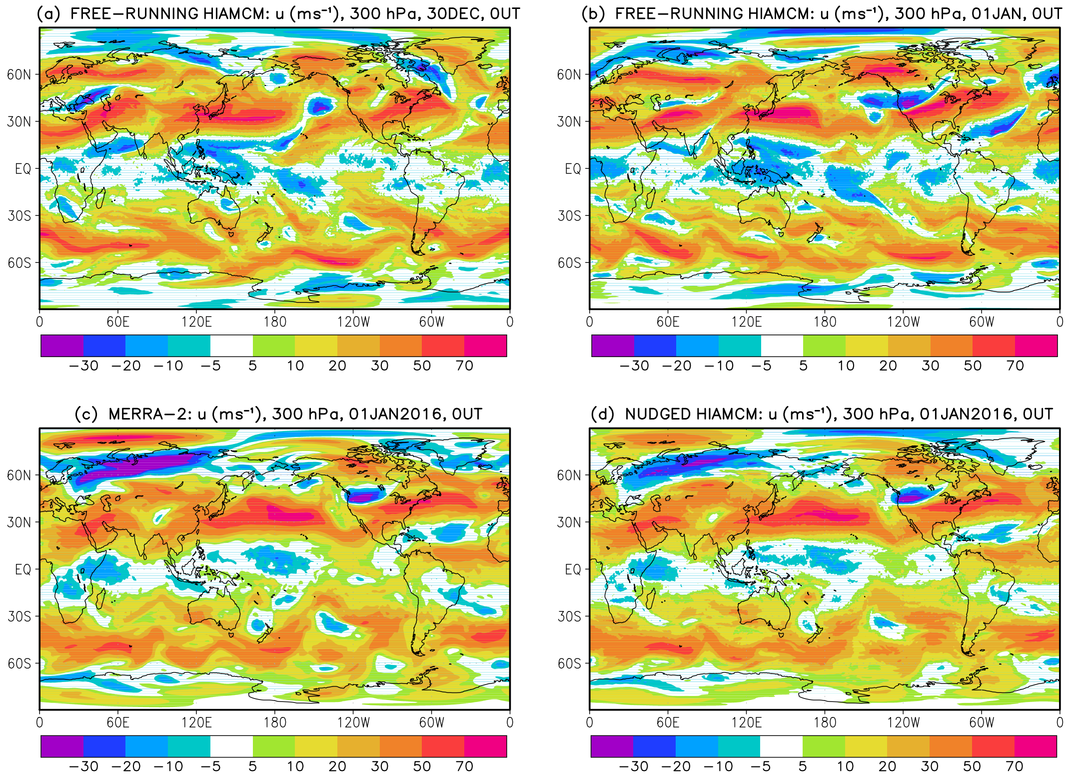

Figure 2. Simulated upper tropospheric zonal wind (at 300 hPa, $z \sim 10\,\mathrm{km}$). Free-running HIAMCM at 0 UT on (a) 30 December and (b) 1 January. (c) MERRA-2 reanalysis at 0 UT on 1 January 2016. (d) Nudged HIAMCM at 0 UT on 1 January 2016; the nudging was started at 0 UT on 30 December 2015).

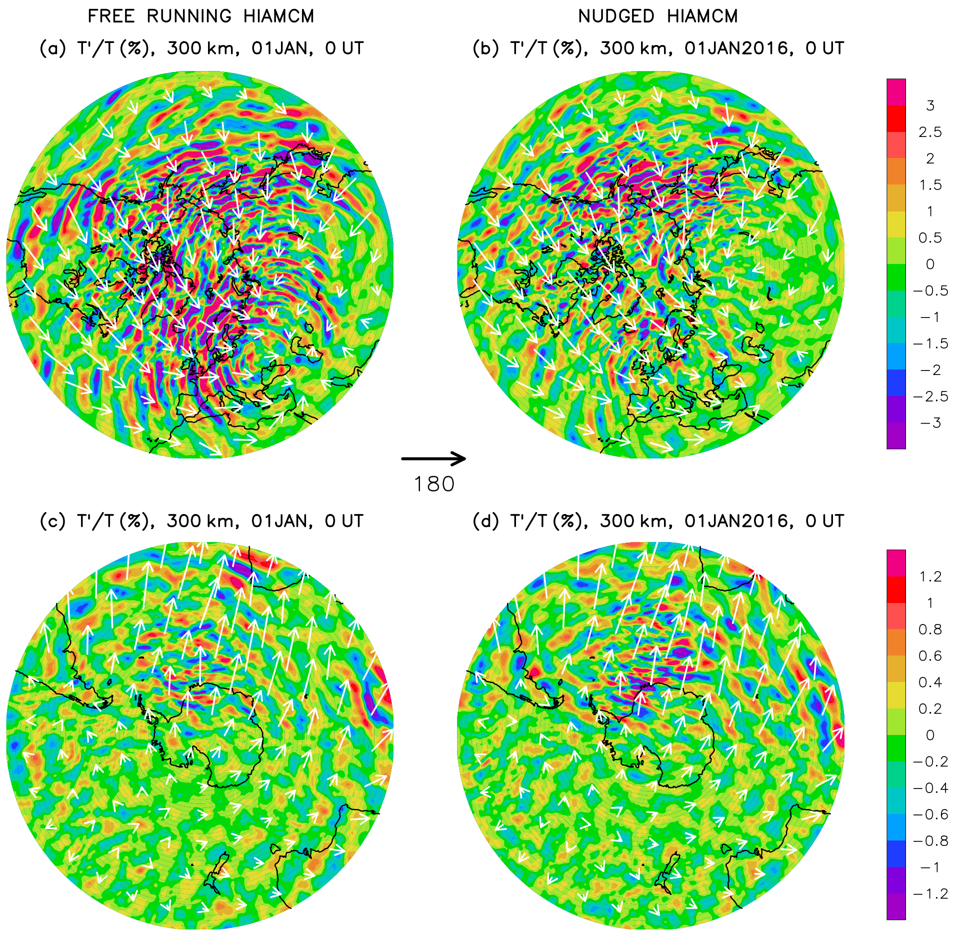

Figure 3. Relative temperature perturbations (horizontal wavenumbers $n > 30$ or $\lambda_h < 1350$ km, colors) and large-scale horizontal wind ($n \leq 30$, white arrows) at 300 km on 1 January 2016 at 0 UT. (a) North-polar projection (25°-90°N) based on the free-running HIAMCM. (b) Same as (a) but for the HIAMCM nudged to MERRA-2 reanalysis, with the nudging started at 0 UT on 30 December 2015. (c),(d) Same as (a),(b) but for a south-polar projection (90°-25°S).

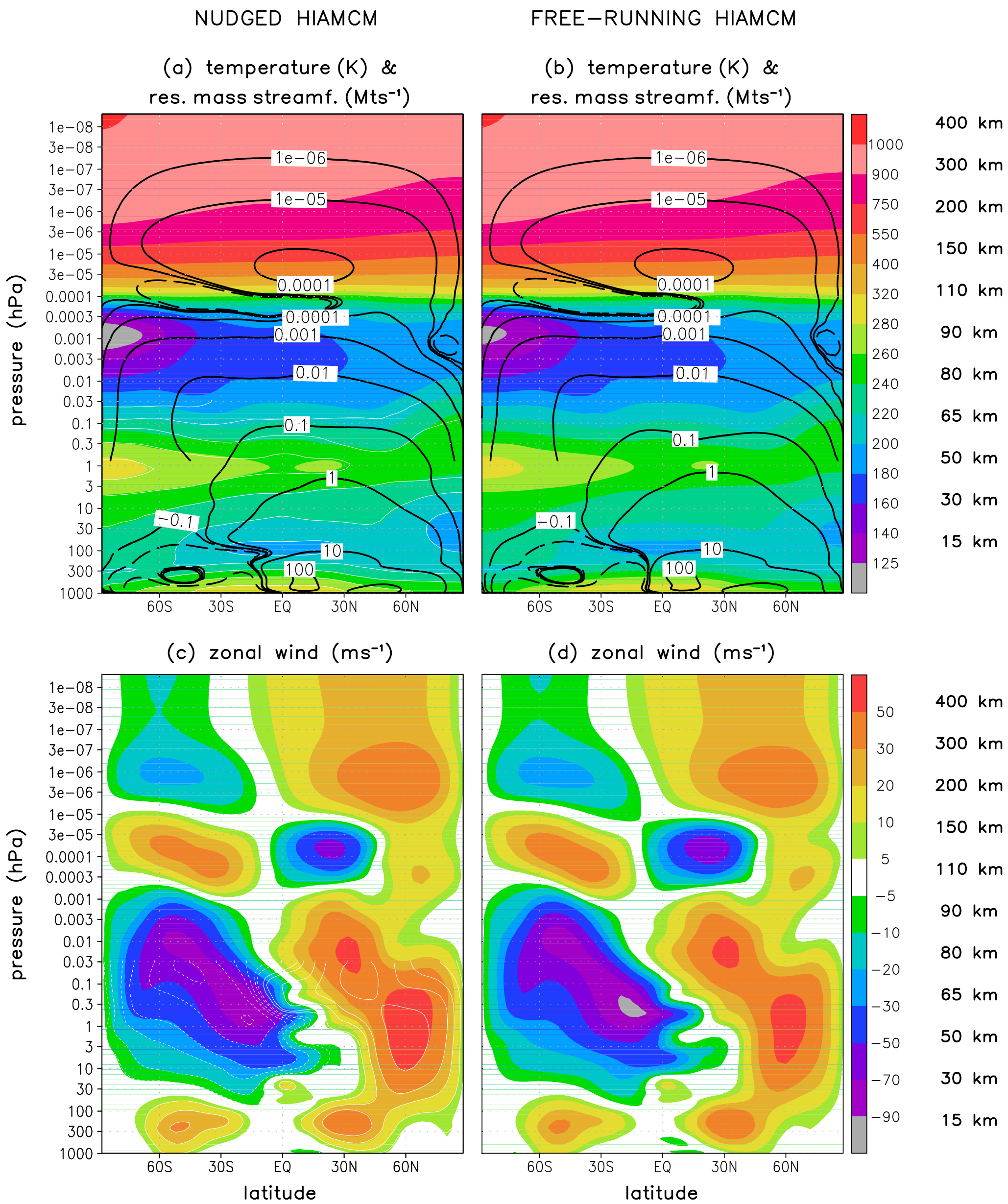

Figure 4. Simulated zonal-mean temperature (first row, colors) and zonal wind (second row, colors) during 1-31 January 2016 from the HIAMCM nudged to MERRA-2 reanalysis (left column) and from the free running HIAMCM. The black colors in the upper panels show the residual mass streamfunction (plotted for $+10^{-6}, \pm 10^{-5}, +10^{-4}, +10^{-3}, +10^{-2}$ Mts^{-1} above 1 hPa, and for $\pm 0.1, \pm 1, \pm 10, +100$ Mts^{-1} below 0.03 hPa). White contours in panel a and c show the zonal-mean temperature and zonal wind from MERRA-2. The vertical coordinate is the hybrid-vertical coordinate of the HIAMCM times 1013 hPa. Approximate geometric heights are given on the right-hand side of panel b and d.

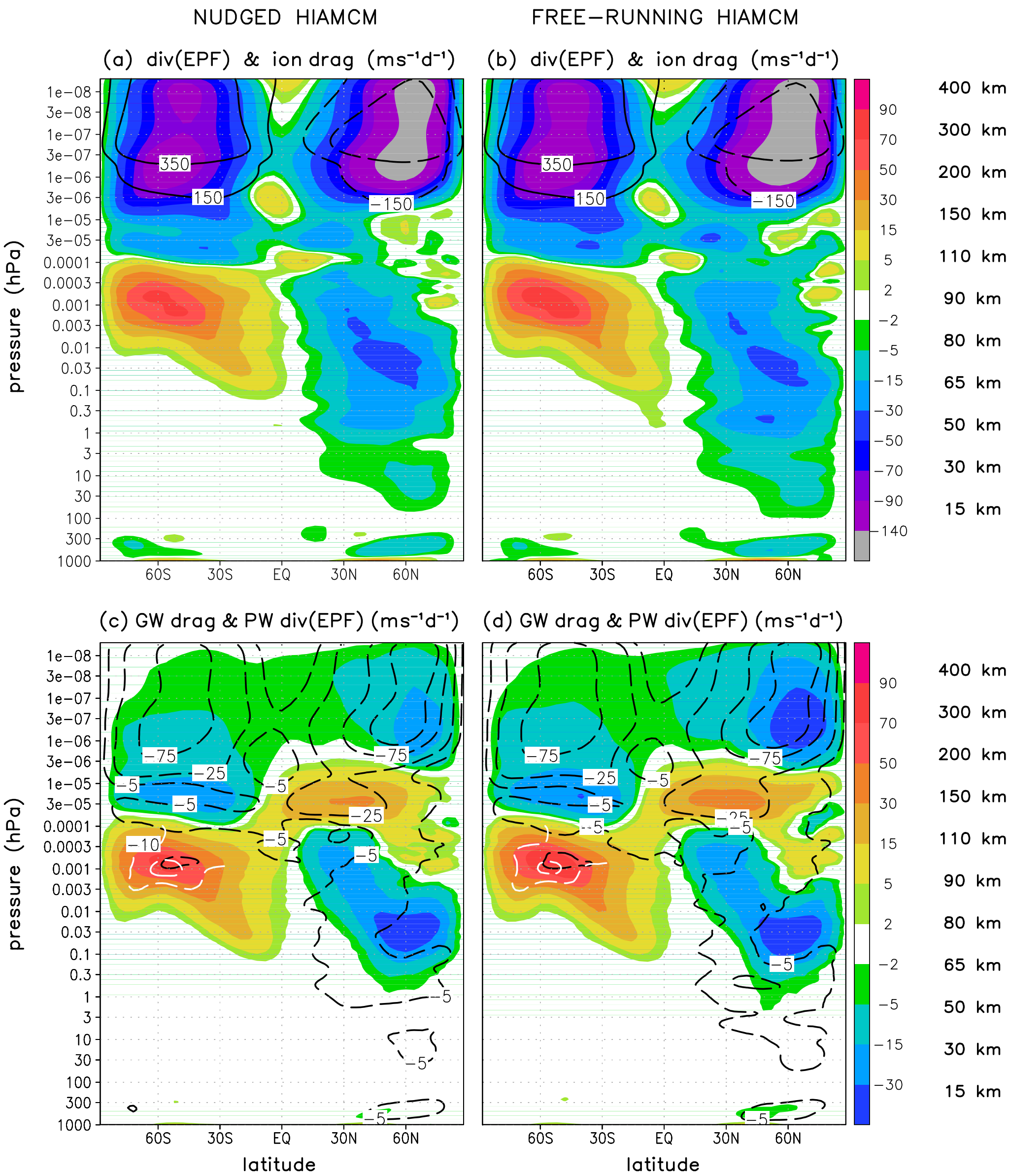

Figure 5. Same as Fig. 4, but for the wave driving per unit mass. Colors in the upper row show the complete Eliassen-Palm flux (EPF) divergence and contours show the zonal ion drag (for $\pm150, \pm350\,\mathrm{m\,s^{-1}d^{-1}}$). Colors in the lower row show the resolved GW drag, which is defined as the complete EPF divergence minus the EPF divergence that is due to planetary-scale waves (PWs). The EPF divergence due to PWs is shown by black contours in panels c and d for $-125, -75, -25, -5\,\mathrm{m\,s^{-1}d^{-1}}$. It is defined as the EPF divergence that is due to total horizontal wavenumbers $n \leq 30$ and zonal wavenumbers $m \leq 6$. The quasi-geostrophic contributation to this planetary-wave (PW) wave driving is indicated by white contours for -10 and $-30\,\mathrm{m\,s^{-1}d^{-1}}$ in the region of the summer mesopause ($0.01-0.0001\,\mathrm{hPa}$, $90^\circ-30^\circ$S).

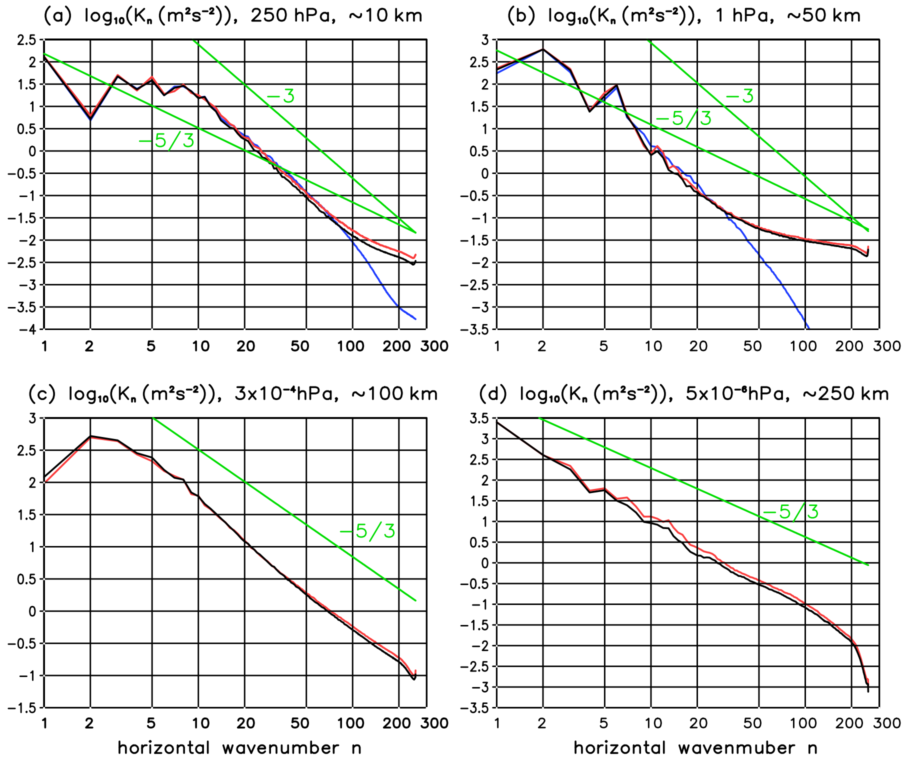

Figure 6. Logarithm of the spectral kinetic energy as a function of the total horizontal wavenumber (n=100 corresponds to a horizontal wavelength of 400 km) at different pressure levels. The black curves are from the nudged simulation from 19 to 24 January 2016. The red curves are from the free-running simulation that was initialized with a snapshot from the nudged simulation on 19 January at 0 UT. The blue curves in the upper two panels give the corresponding results from MERRA-2 reanalysis. The $-5/3$ and -3 exponential slopes are indicated by green lines, as labelled.

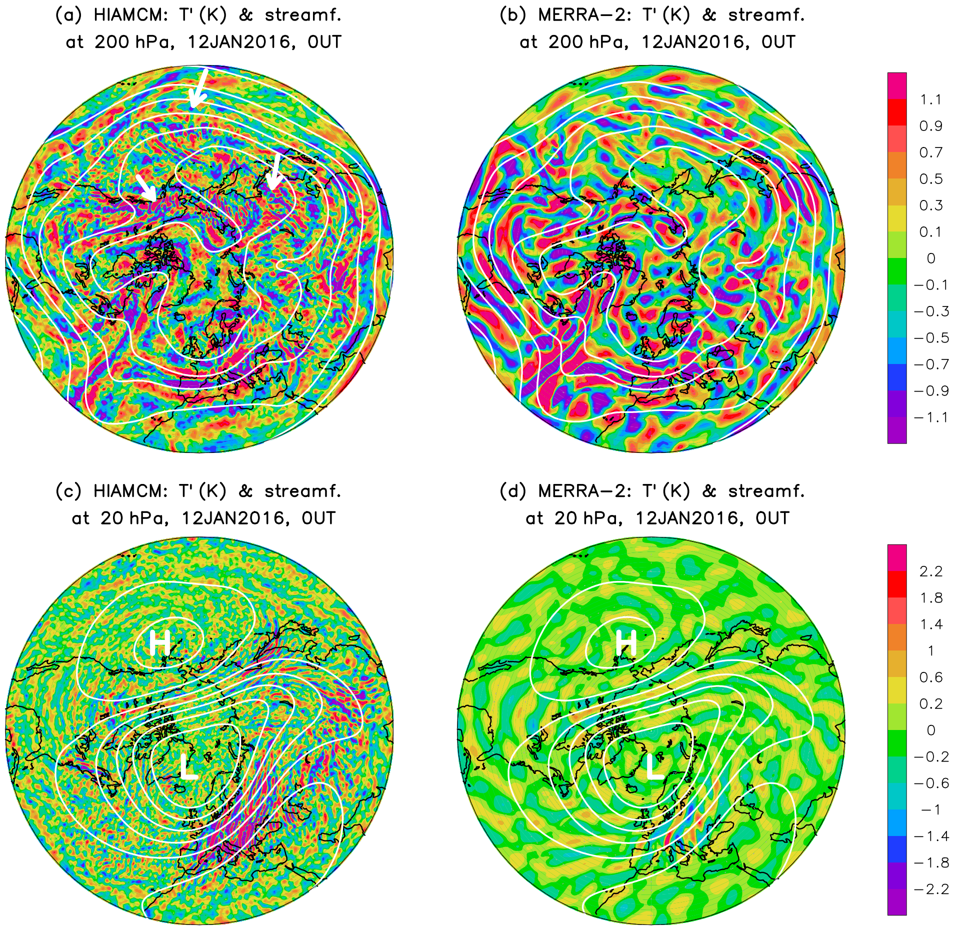

Figure 7. Northpolar projection of temperature perturbations (colors) for horizontal wavenumbers $n > 30$ (λ_h smaller than ~ 1350 km) and of the horizontal streamfunction (white contours) for $n \leq 30$ (λ_h larger than ~ 1350 km) in the HIAMCM (left) and MERRA-2 reanalysis (right) for 12 January 2016 at 0 UT. (a),(b) Upper troposphere at 200 hPa ($z \sim 12$ km). The large-scale flow is counterclockwise along the streamlines. (e),(f) Same as (a),(b) but at 20 hPa ($z \sim 25$ km). The large-scale flow is counterclockwise (clockwise) along the streamlines around the lows (highs) marked by the white letters L (H). White arrows in (a) indicate packets of medium-to-small-scale GWs. The horizontal streamfunction contour interval is $3 \times 10^7 \mathrm{m}^2\mathrm{s}^{-1}$.

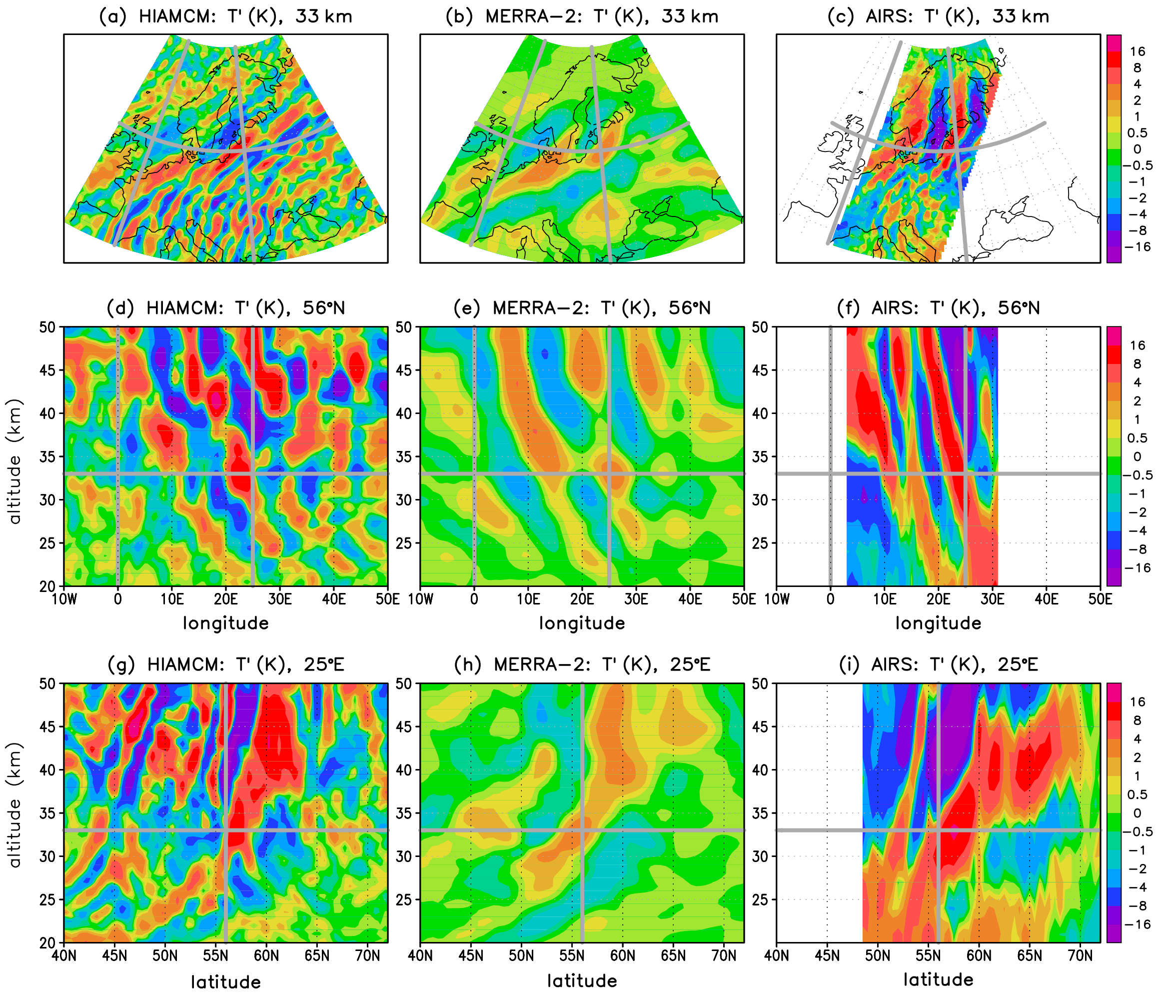

Figure 8. Instantaneous temperature perturbations during a GW event over Northern Europe at 1:30 UT on January 11, 2016 from the nudged HIAMCM (left), MERRA-2 reanalysis (middle), and AIRS (right). First row: horizontal map segments at 33 km height from 10°W to 50°E and from 40°N to 72°N. Second row: longitude-height cross-sections at 56°N. Third row: latitude-height cross-sections at 25°E. The grey lines mark the longitudes 0° and 25°E, the latitude 56°N, and the height 33 km. These lines are included for better comparison of the different panels.

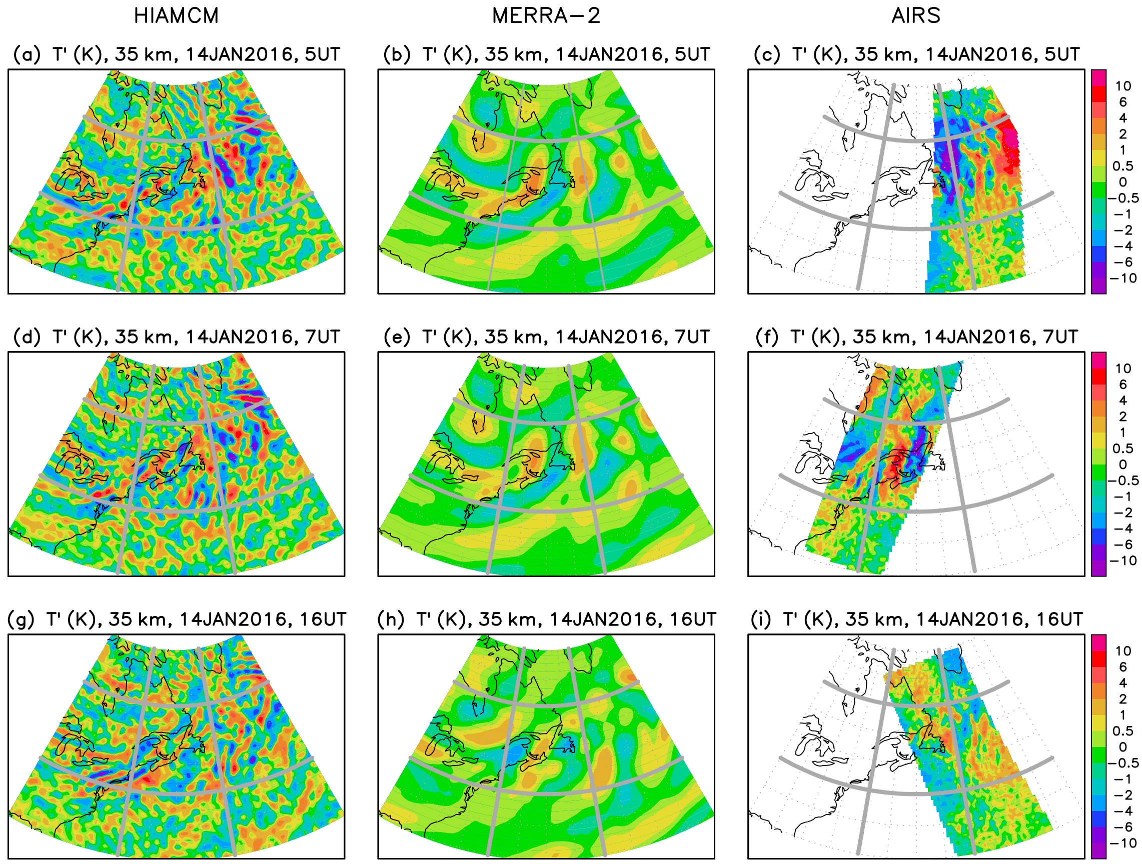

Figure 9. Instantaneous temperature perturbations during a GW event over eastern North America and the northwest Atlantic on January 14, 2016 at 5 UT (upper row), 7 UT (middle row), and 16 UT (lower row) from the nudged HIAMCM (left), MERRA-2 reanalysis (middle), and AIRS (right). The horizontal map segments are at 35 km height and extend from 90°W to 30°W and from 30°N to 65°N. The grey lines show the longitudes 70°W and 50°W and the latitudes 40°N and 55°N.

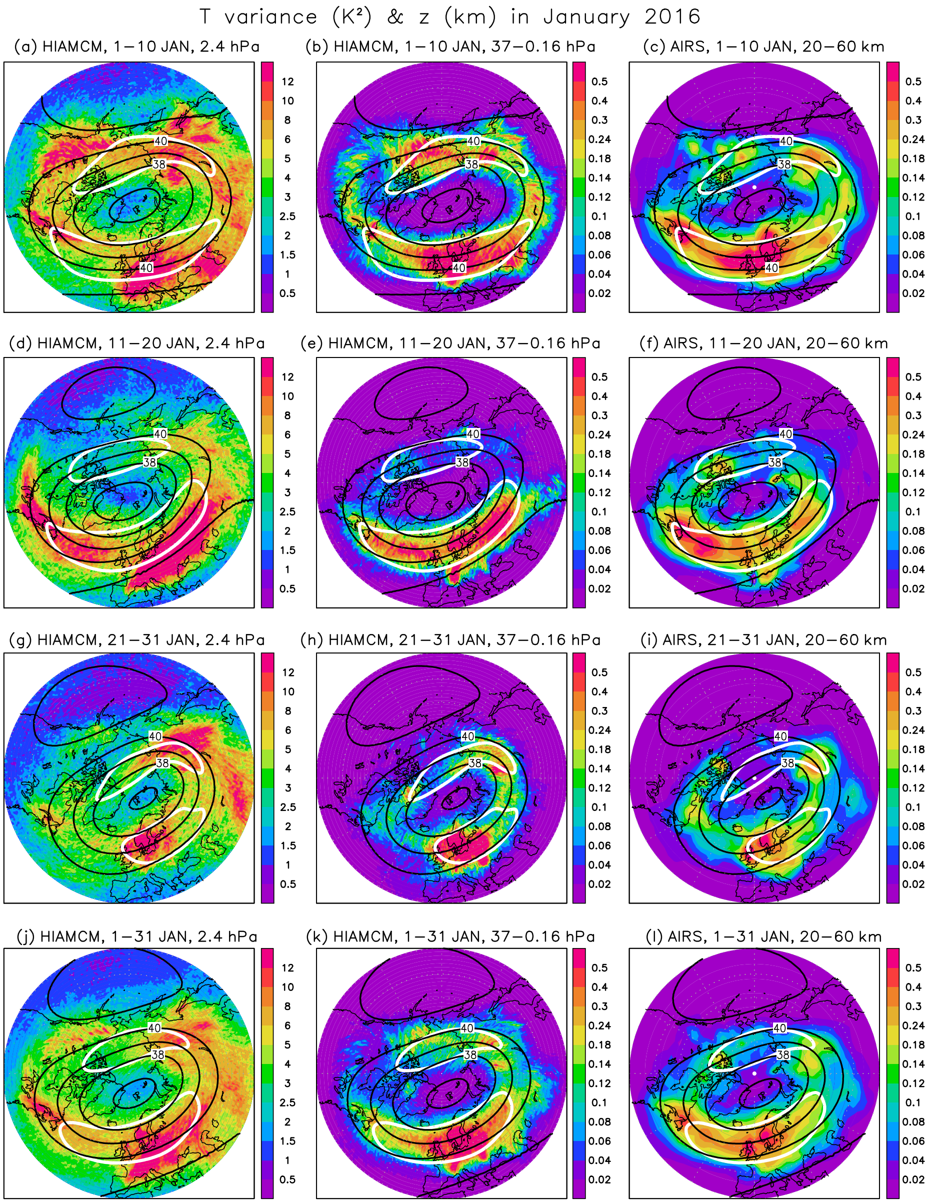

Figure 10. Stratospheric temperature variances due to GWs simulated by the HIAMCM nudged to MERRA-2 reanalysis (first and second columns) and corresponding result from the AIRS satellite data (third column) in January 2016. The left column shows HIAMCM results at 2.4 hPa. The middle column shows the same HIAMCM results but with a vertical filter applied to the temperature perturbation before computing the variance (see Eq. (26) and Fig. 11). The temperature perturbation in the HIAMCM is defined from an expansion in spherical harmonics, retaining only wavenumbers $n > 30$ (horizontal wavelength smaller than $\sim$1350 km). The four rows refer to temporal averages as indicated in the title of each panel. Black contours show the geometric height at 2.4 hPa in intervals of 1 km. A large-scale horizontal wind speed of $90\,\mathrm{m\,s^{-1}}$ is indicated by a white contour in each panel.

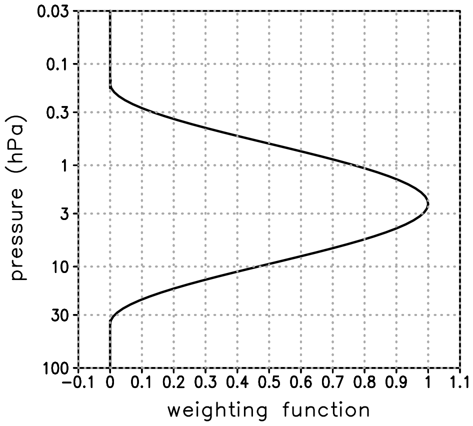

Figure 11. Weighting function $w(p)$ for the computation of height-averaged temperature perturbations from the HIAMCM according to Eq. (26).

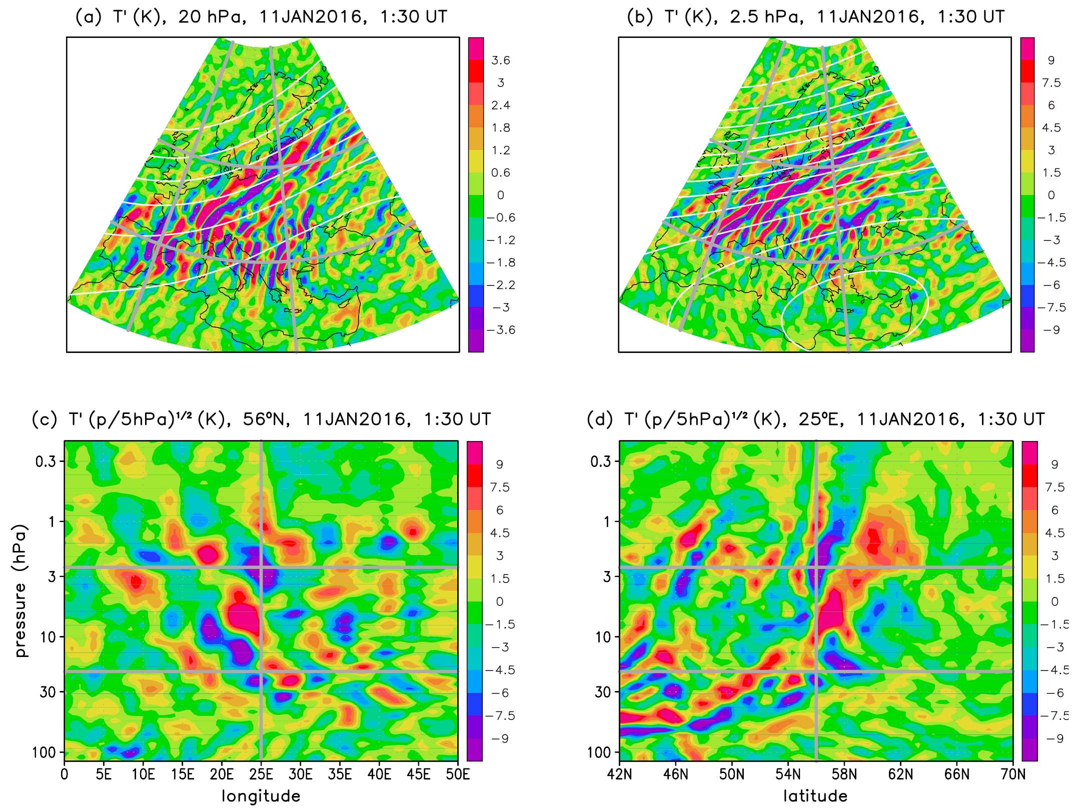

Figure 12. Temperature perturbation, T', due to horizontal wavenumbers $n > 30$ (λ_h smaller than $\sim$1350 km) on 11 January 2016, 1:30 UT. (a),(b) Northpolar projections at 20 and 2.5 hPa ($z\sim$25 and 39 km, respectively). The white contours show the horizontal streamfunction (see Eq. (4)) for $n \leq 30$ with a contour interval of $3 \times 10^7 \mathrm{m}^2\mathrm{s}^{-1}$. The grey lines mark 42°N, 56°N, 0°E, and 25°E. (c),(d) Longitude-height cross-section at 56°N and latitude-height cross-section at 25°E of T' scaled by $\sqrt{p\,/\,5\mathrm{hPa}}$. The grey lines mark the longitude 25°E, the latitude 56°N, and the pressure surfaces 20 hPa and 2.5 hPa.

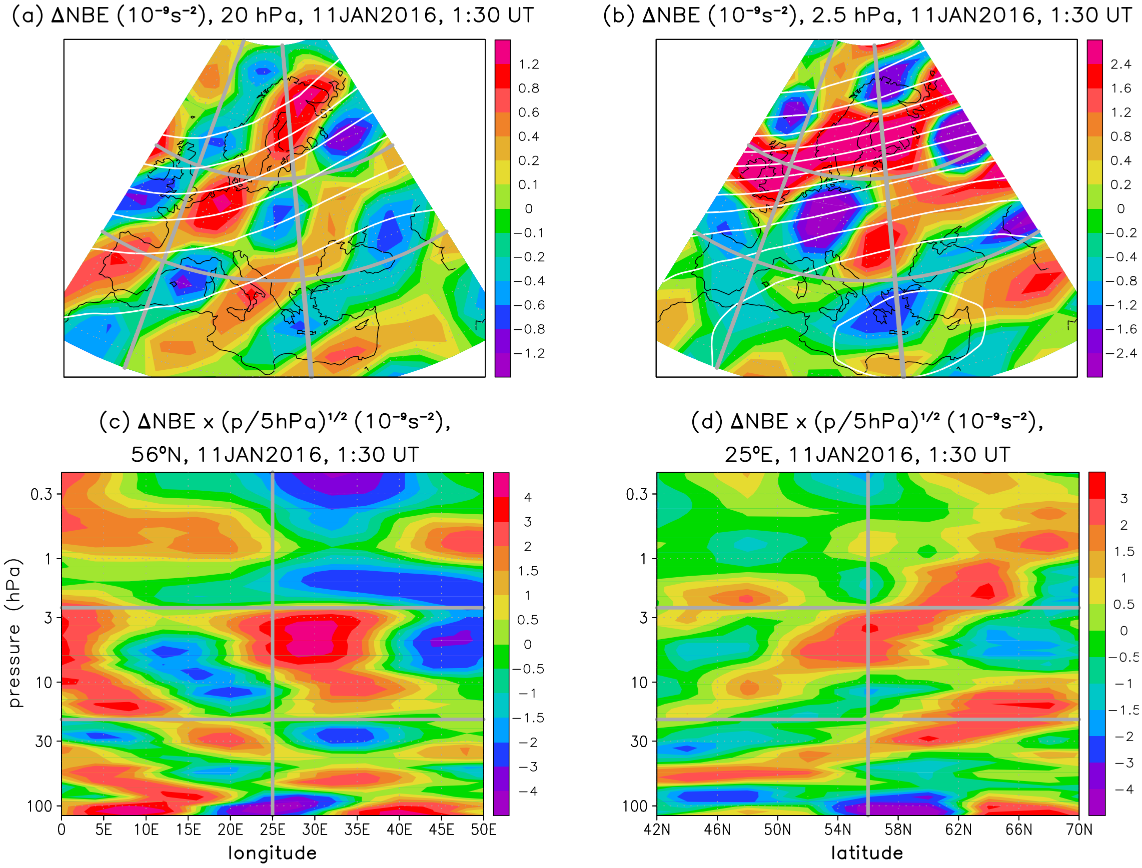

Figure 13. Same as Fig. 12, but for the nonlinear balance equation (Eq. (B19)) in units of $10^{-9}\mathrm{s}^{-2}$ and scaled by $\sqrt{p\,/\,5\mathrm{hPa}}$ in (c),(d).

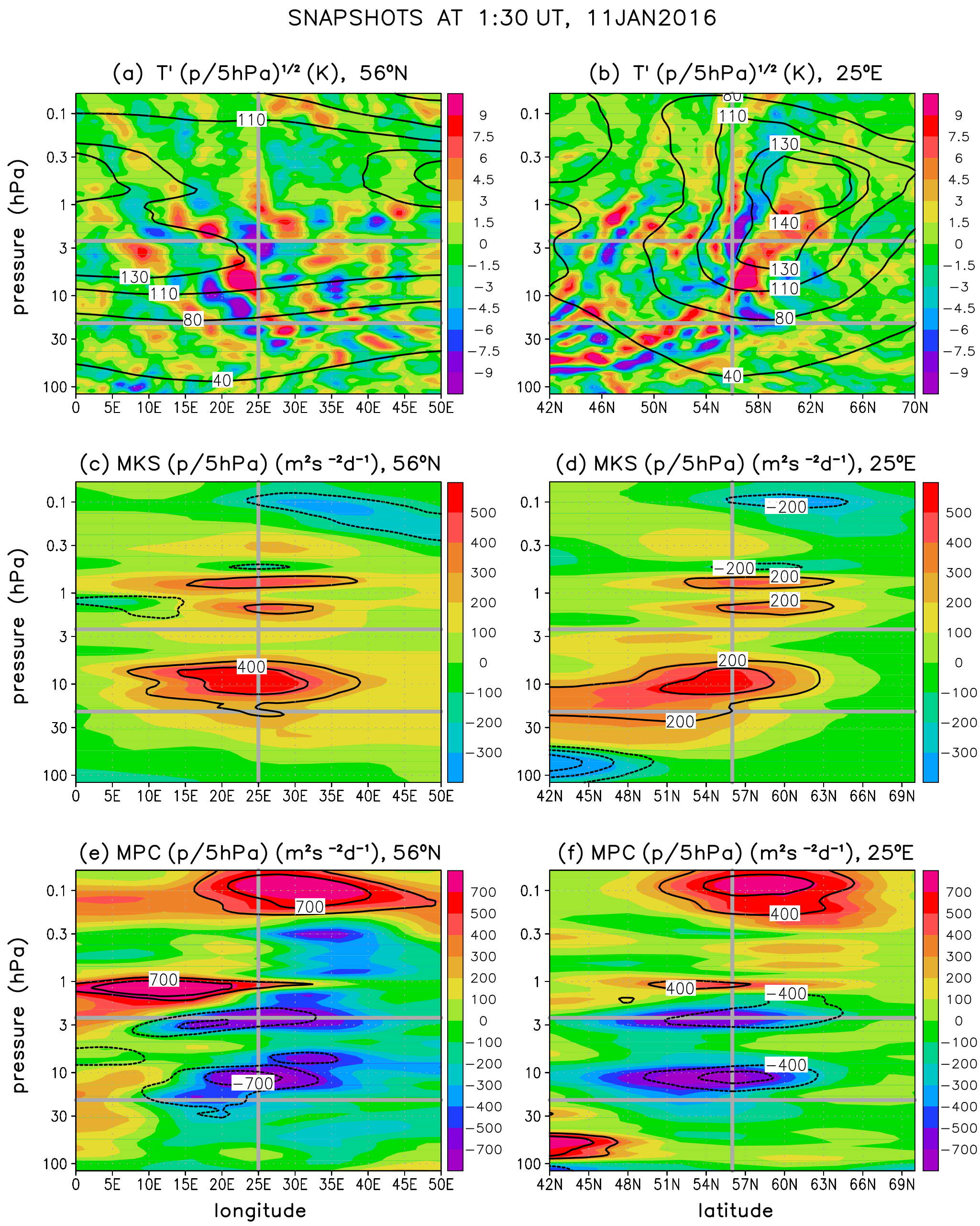

Figure 14. (a),(b) Scaled temperature perturbation as in Fig. 12c,d, but extending up to 0.06 hPa. Black contours show the large-scale horizontal wind speed for $40, 80, 110, 130, 140\ \mathrm{m\,s^{-1}}$. (c),(d) Mesoscale kinetic energy source (Eq. (B22)) in units of $\mathrm{m^2s^{-2}d^{-1}}$ and scaled by $p\,/\,5\mathrm{hPa}$. Black contours show the corresponding contribution from the vertical wind shear (last term on the right-hand side of Eq. (B22)) for $\pm 200, \pm 400\,\mathrm{m^2s^{-2}d^{-1}}$. (e),(f) Same as (c),(d), but for the mesoscale potential energy flux convergence (Eq. (B21)). Contours from the vertical convergence (last term on the right-hand side of Eq. (B21)) are plotted for $\pm 400, \pm 700\,\mathrm{m^2s^{-2}d^{-1}}$.

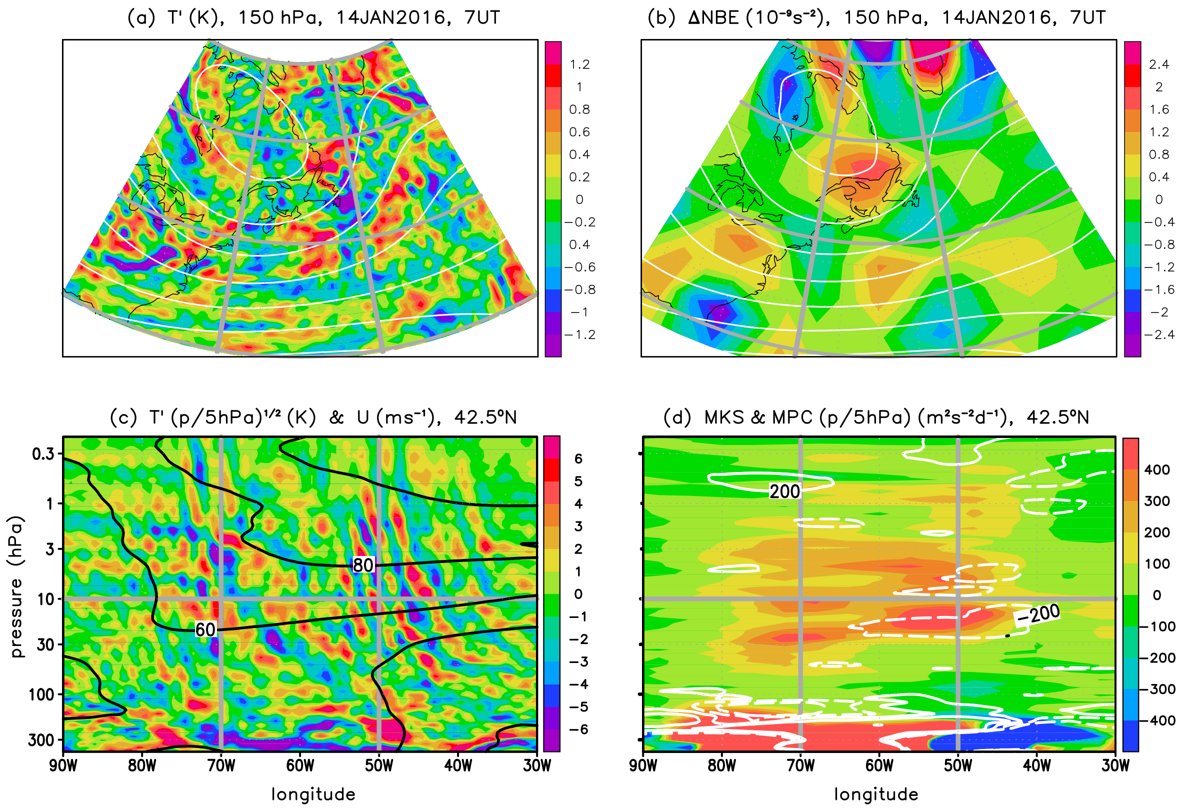

Figure 15. (a) Temperature perturbation (colors) and horizontal streamfunction (white contours, interval $2 \times 10^7 \mathrm{m}^2\mathrm{s}^{-1}$) on 14 January 2016 (7 UT) at 150 hPa ($z \sim 15$ km). The horizontal cross-section extends from 90°W to 30°W and from 30°N to 65°N. Grey lines mark the latitudes 30°N, 42.5°N, 55°N, and 65°N, as well as the longitudes 70°W and 50°W. (b) Same as (a) for but for the nonlinear balance equation (colors, in units of $10^{-9}\mathrm{s}^{-2}$). (c) Longitude-height cross-section of the scaled temperature perturbation (colors) at 42.5°N on 14 January 2016 (7 UT). Black contours show the large-scale horizontal speed ($U = |\mathbf{v}_{\tilde{g}}|$, see Appendix A) for $20, 40, 60, 80, 100\,\mathrm{m\,s}^{-1}$. (d) Mesoscale kinetic energy source (colors, Eq. (B22)) and GW potential energy flux convergence (white contours, Eq. (B21)) in units of $\mathrm{m}^2\mathrm{s}^{-2}\mathrm{d}^{-1}$ and scaled by $p\,/\,5$hPa. Contours of MPC are drawn for $\pm 200, \pm 600\,\mathrm{m}^2\mathrm{s}^{-2}\mathrm{d}^{-1}$. The grey lines in (c),(d) mark 70°W, 50°W, and 10 hPa.

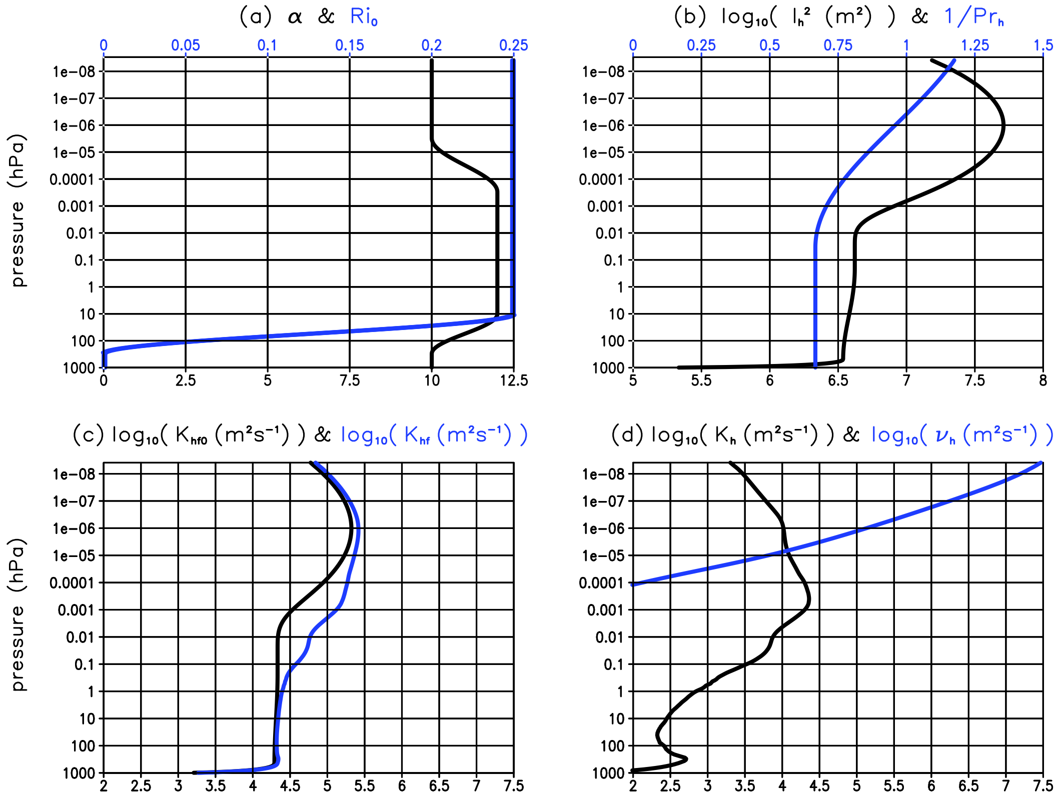

Figure 16. Parameters of the horizontal diffusion scheme. (a) Vertical profiles of the Richardson number offset, Ri_0 (blue curve), and the scaling factor, α (black curve), for the Richardson number criterion in Eq. (A6). (b) Logarithm of the squared horizontal mixing length, l_h^2 (black curve), and of the inverse horizontal Prandtl number, $1/Pr_h$ (blue curve). (c) Logarithm of the linear hyperdiffusion coefficient, K_{hf0} (black curve), and of the complete globally and temporally averaged hyperdiffusion coefficient, $K_{hf0} + 4.9\,K_h$ (blue curve, see Eq. (A8)). (d) Logarithm of the global and temporal averages of the Smagorinsky-type horizontal diffusion coefficient, K_h (black curve, see Eq. (A6)), and of the molecular viscosity (blue curve, see Eqs. (A17) and (A18) in BV20).