Gravity-Wave-Driven Seasonal Variability of Temperature Differences between ECMWF IFS and Rayleigh Lidar Measurements in the Lee of the Southern Andes

Sonja Gisinger[1,1], Inna Polichtchouk[2,2], Andreas Dörnbrack[3,3], Robert Reichert[1,1], Bernd Kaifler[1,1], Natalie Kaifler[1,1], Markus Rapp[4,4], and Irina Sandu[2,2]

[1]Institute of Atmospheric Physics, German Aerospace Center
[2]ECMWF
[3]DLR, Institut für Physik der Atmosphäre
[4]Deutsches Zentrum für Luft- und Raumfahrt (DLR)

November 30, 2022

Abstract

Long-term high-resolution temperature data of the Compact Rayleigh Autonomous Lidar (CORAL) is used to evaluate temperature and gravity wave (GW) activity in ECMWF Integrated Forecasting System (IFS) over R\'io Grande (53.79°S, 67.75°W), which is a hot spot of stratospheric GWs in winter. Seasonal and altitudinal variations of the temperature differences between the IFS and lidar are studied for 2018 with a uniform IFS version. Moreover, interannual variations are considered taking into account updated IFS versions. We find monthly mean temperature differences <2~K at 20-40~km altitude. At 45-55~km, the differences are smaller than 4~K during summer. The largest differences are found during winter (4~K in May 2018 and -10~K in August 2018, July 2019 and 2020). The width of the difference distribution (15th/85th percentiles), the root mean square error, and maximum differences between instantaneous individual profiles are also larger during winter ($>\pm10$~K) and increase with altitude. We relate this seasonal variability to middle atmosphere GW activity. In the upper stratosphere and lower mesosphere, the observed temperature differences result from both GW amplitude and phase differences. The IFS captures the seasonal cycle of GW potential energy (E_p) well, but underestimates E_p in the middle atmosphere. Experimental IFS simulations without damping by the model sponge for May and August 2018 show an increase in the monthly mean E_p above 45~km from only ≈10~\% of the E_p derived from the lidar measurements to 26~\% and 42~\%, respectively. GWs not resolved in the IFS are likely explaining the remaining underestimation of the E_p.

Gravity-Wave-Driven Seasonal Variability of Temperature Differences between ECMWF IFS and Rayleigh Lidar Measurements in the Lee of the Southern Andes

Sonja Gisinger[1], Inna Polichtchouk[2], Andreas Dörnbrack[1], Robert Reichert[1], Bernd Kaifler[1], Natalie Kaifler[1], Markus Rapp [1,3], Irina Sandu[2]

[1]Deutsches Zentrum für Luft- und Raumfahrt, Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany
[2]European Centre for Medium-Range Weather Forecasts, Reading, UK
[3]Ludwig-Maximilians-Universität München, Munich, Germany

Key Points:

- Monthly mean temperature differences between the IFS and ground-based lidar measurements in the middle atmosphere are largest in winter.
- Wintertime temperature differences are related to middle atmosphere gravity wave activity and both amplitude and phase differences are important.
- Damping by the model sponge and unresolved gravity waves reduce gravity wave potential energy in the middle atmosphere in the IFS.

Corresponding author: Sonja Gisinger, `sonja.gisinger@dlr.de`

Abstract

Long-term high-resolution temperature data of the Compact Rayleigh Autonomous Lidar (CORAL) is used to evaluate temperature and gravity wave (GW) activity in ECMWF Integrated Forecasting System (IFS) over Río Grande (53.79°S, 67.75°W), which is a hot spot of stratospheric GWs in winter. Seasonal and altitudinal variations of the temperature differences between the IFS and lidar are studied for 2018 with a uniform IFS version. Moreover, interannual variations are considered taking into account updated IFS versions. We find monthly mean temperature differences < 2 K at 20-40 km altitude. At 45-55 km, the differences are smaller than 4 K during summer. The largest differences are found during winter (4 K in May 2018 and -10 K in August 2018, July 2019 and 2020). The width of the difference distribution (15th/85th percentiles), the root mean square error, and maximum differences between instantaneous individual profiles are also larger during winter ($> \pm 10$ K) and increase with altitude. We relate this seasonal variability to middle atmosphere GW activity. In the upper stratosphere and lower mesosphere, the observed temperature differences result from both GW amplitude and phase differences. The IFS captures the seasonal cycle of GW potential energy (E_p) well, but underestimates E_p in the middle atmosphere. Experimental IFS simulations without damping by the model sponge for May and August 2018 show an increase in the monthly mean E_p above 45 km from only ≈ 10 % of the E_p derived from the lidar measurements to 26 % and 42 %, respectively. GWs not resolved in the IFS are likely explaining the remaining underestimation of the E_p.

1 Introduction

Even now with a growing understanding of stratospheric processes, highly developed numerical models, and increasing computational resources, middle atmosphere temperature (re)analyses have a larger uncertainty than their tropospheric counterparts. Improving the representation of the past (reanalysis), current (analysis), and future (forecast) state of the middle atmosphere in general circulation models (GCMs) is important for the validation and forecasting of tropospheric weather and future climate. It is known that the circulation in the middle and upper atmosphere is strongly influenced by internal gravity waves (GWs) triggered for example by flow over mountains (Fritts & Alexander, 2003). At the same time, processes in the stratosphere such as anomalies in the winter- and spring-time stratospheric polar vortex impact the tropospheric circulation (Baldwin & Dunkerton, 2001; Garfinkel & Hartmann, 2011; Byrne & Shepherd, 2018).

One issue when modelling the middle atmosphere is that there is a limited amount of observations to constrain the current model state (e.g., Eckermann et al., 2018). Above 10 hPa, most of the observations assimilated into the Integrated Forecasting System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF) are from satellites and have limited spatial and temporal resolutions. They mainly provide temperature-related data (e.g., Global Navigation Satellite System Radio Occultation (GNSS-RO), Atmospheric Infrared Sounder (AIRS), Advanced Microwave Sounding Unit (AMSU-A)) and the topmost radiances assimilated peak at approximately 1-2 hPa. The range of sensitivity of the satellite observations to certain horizontal and vertical scales of GWs depends on the instrument and viewing geometry (observational filter, see Alexander, 1998) as can be seen in e.g., Figure 9 of Preusse et al. (2008). To produce the most accurate representation of the atmospheric state, all the observations irregularly distributed in time and space and each having their limitations and uncertainties are combined with the numerical weather prediction model on a global grid. For the (re)analyses at ECMWF, this is achieved by 4-dimensional variational data assimilation (4D-Var).

The analysis is the best guess of the current atmospheric state that is used to initialize forecasts. Many satellite observations in the upper stratosphere are rejected by the 4D-Var in the IFS over the GW hot spot region of the Southern Andes, the Drake Passage, and the Antarctic Peninsula in the Southern Hemisphere extended winter period (April to September), most frequently in May (Tony McNally, personal communication, December 2018). The observations deviate too strongly from the IFS background which is likely due to GW-induced temperature perturbations. Stratospheric GW activity is not homogeneous over the globe but numerous hot spots exist close to mountain ranges, coasts, lakes, deserts, or isolated islands (Hoffmann et al., 2013). For the Southern Hemisphere, backward ray tracing of GWs at 25 km altitude, which are resolved in the IFS in simulated satellite observations imitating an infrared limb imager, revealed the Antarctic Peninsula and the Southern Andes as prominent GW sources (Preusse et al., 2014). Together with GWs generated by storms, these GWs are responsible for large day-to-day variations (factor of two) in the stratospheric GW momentum flux in the Southern Hemisphere (Preusse et al., 2014).

The sparseness and limitations of observations in the middle atmosphere means that the model plays a larger role in determining the atmospheric state in (re)analyses. To represent stratospheric processes, the model top and corresponding sponge layers have to be moved to higher altitudes (Shepherd et al., 1996). This and the enhancement of vertical resolution led to an increase in demand of computational resources that only became available in the past decades. For example, in the IFS the vertical resolution has increased from 31 levels in 2003 to 137 levels in 2013 (still in use today). At the same time the model top has increased from mid-stratosphere at 10 hPa to the mesosphere at 0.01 hPa (i.e. from altitude $z \approx 28$ km to $z \approx 80$ km). Currently the sponge layer, designed to reduce wave reflection at the model top, starts weak at 10 hPa and is strongest above 1 hPa ($z \approx 45$ km) in the IFS. All waves, including GWs, are significantly damped by the sponge. The 4D-Var in the IFS is unstable when large-amplitude GWs are allowed to exist in the mesosphere, which occurs if the sponge layer is too thin. The sponge layer leads to a misrepresentation of GW drag, which can affect the large-scale circulation in the middle atmosphere (Shepherd et al., 1996). Therefore, reducing the depth and the strength of the sponge layer could help to improve the representation of GWs and temperature biases in the middle atmosphere.

Challenges of middle atmosphere modelling that include the representation of physical and dynamical processes, data assimilation, and artificial damping by the sponge layer motivate our study. Local middle atmosphere lidar measurements can be used to evaluate IFS-based (re)analyses and forecasts at altitudes where there is little assimilated data, the influence of the model sponge is large, and the vertical resolution is coarse.

Several studies have already compared lidar observations to ECMWF (re)analyses. Marlton et al. (2021) compared stratospheric temperatures in ERA-Interim and ERA5 reanalyses to ground-based lidar at four sites in the Northern Hemisphere winter for 1990-2017 and found mean temperature differences in the range of ± 5 K. ERA5 temperatures were found to be too low at 1 hPa at all four sites but a different behaviour was found at each site below 1 hPa. Le Pichon et al. (2015) found the largest differences and the highest variability of the differences in winter when comparing nightly-mean lidar wind and temperature data to IFS analyses in Europe for winter 2012/2013 and summer 2013. In 2012/2013 winter, the variability from large-scale planetary waves dominated and a sudden stratospheric warming, accompanied by enhanced GW activity, took place in January 2013. Above altitude $z = 45$ km, the IFS temperatures were found to be over -5 K too cold and the 95 % percentile of the difference distribution was around -30 K (Le Pichon et al., 2015). For $z > 40$ km over northernmost Europe, also Ehard et al. (2018) estimated IFS to be too cold by -8 K to -20 K when compared to lidar measurements in December 2015. For the Southern Island of New Zealand located in the mid-latitude Southern Hemisphere, wintertime-averaged temperature differences (July to September 2014) between lidar and IFS data were between -3 K and 2 K for 45 km $< z <$ 60 km and exceeded -10 K at $z = 70$ km (Appendix B in Gisinger et al., 2017).

The past studies exemplify that differences of model temperatures in the middle atmosphere depend on the season and the location, and can be different compared to global- or zonal-mean bias characteristics (e.g., Simmons et al., 2020, for ERA5). However, the total of all local differences determines the global- or zonal-mean bias. Therefore, understanding and quantifying local differences can help to reduce such biases. For the stratospheric GW hot spot region of the Southern Andes, a detailed quantification of local differences between middle atmosphere temperature measurements and IFS temperatures, their vertical structure, and their seasonal and inter-annual variability is still missing. Further, the contribution of shortcomings in the representation of middle atmosphere GWs in the IFS to site-specific temperature differences can be studied for this region because GWs are dominating the atmospheric state for several months of the year (Hoffmann et al., 2013). In November 2017, the DLR Institute of Atmospheric Physics deployed the ground-based Compact Rayleigh Autonomous Lidar (CORAL) at Río Grande at the southern tip of South America in Argentina (B. Kaifler & Kaifler, 2021). The nightly lidar temperature measurements have high temporal (15 min) and vertical (900 m) resolutions between 15-95 km altitude. Comprehensive analyses of the whole three-year data set including GW characteristics are presented by Reichert et al. (2021).

GW activity can be estimated from lidar temperature measurements via GW potential energy, which is calculated from temperature perturbations relative to the background temperature. GW potential energy is related to the GW momentum flux based on linear theory (Ern et al., 2004), though the momentum flux is a conservative wave property but the wave energy is not. Ehard et al. (2018) found that the IFS is capable of reproducing the overall temporal evolution of the GW activity in the stratosphere at 30 km $< z <$ 40 km over northernmost Europe for a four-months-period, but that GW amplitudes are effectively damped by the sponge layer at higher altitudes. GW potential energy was also found to be lower in reanalysis data (Modern-Era Retrospective analysis for Research and Applications (MERRA), ERA5) in the middle atmosphere compared to multi-year lidar measurements from two European stations at higher mid- and polar latitudes (Strelnikova et al., 2021). For the Southern Hemisphere, a simplified comparison of GW potential energy between the IFS and lidar measurements (i.e., not a one-to-one comparison but different years of IFS and observational data) at two locations in Antarctica (Rothera and South Pole) was presented in Yamashita et al. (2010). The IFS generally captured site-specific seasonal variations of GW potential energy in the stratosphere: These are a winter maximum and a summer minimum at Rothera and continuously low values at the South Pole (Yamashita et al., 2010). Comparisons of three-day averaged GW temperature amplitudes of SABER (Sounding of the Atmosphere Using Broadband Emission Radiometry) and IFS at $z = 30$ km showed that the annual cycle and shorter-term variations dominated by mountain waves are well represented in the IFS also for South America, but that temperature amplitudes are underestimated in the IFS (Schroeder et al., 2009). Prior to 2010, the IFS had 91 vertical layers and a horizontal resolution of approximately 25 km.

In this study, we present a systematic comparison of middle atmosphere temperatures and GW potential energy of the independent (i.e., not assimilated in the IFS), high-resolution CORAL lidar data set and operational and experimental IFS simulations for Río Grande (53.79°S, 67.75°W), which is a hot spot of stratospheric GWs in the Southern Hemisphere winter (Hoffmann et al., 2013), located in the lee of the Southern Andes. Temperature differences between the lidar and IFS and seasonal variability of the differences are investigated. The role of winter-time GW representation by means of wave amplitude and phase in the middle atmosphere in the IFS is discussed. This is only possible due to the high temporal resolution of the lidar data, allowing a one-to-one comparison of quasi-instantaneous values. The annual cycle of GW activity in the middle atmosphere over Río Grande in the IFS is compared to that derived from the lidar observations. The results for temperature differences and GW activity are then combined to investigate the hypothesis that the seasonal variability of the temperature differences over Río Grande is related to the GW activity in the middle atmosphere. For two selected months with enhanced GW activity (May and August 2018), the importance of individual strong GW events for the monthly mean GW potential energy in the middle atmosphere in the observations and the IFS is evaluated (i.e. GW intermittency). Finally, the effect of damping by the sponge on GW potential energy in the middle atmosphere is quantified in experimental IFS simulations without a sponge layer for these two months.

Section 2 describes the lidar system CORAL, its temperature data taken at Río Grande, the IFS model data, and the data analysis methods. Results are presented in section 3 and discussed and summarized in section 4.

2 Data and methods

2.1 Lidar system and data

CORAL (B. Kaifler & Kaifler, 2021) uses a 12-W-laser beam at 532 nm wavelength and a 0.64-m-diameter telescope installed in an 8 ft container for night-time, autonomous atmospheric soundings. Backscattered photons are detected with three height-cascaded elastic detector channels and one Raman channel. Density and temperature profiles on a 100-m vertical grid for altitudes 15 km $< z <$ 95 km are determined by top-down integration of the hydrostatic equation every 5 minutes using an integration window of 15 minutes and 900-m vertical smoothing for an adequate signal to noise ratio. The precision for temperature is better than 1 K for 35 km $< z <$ 60 km and typically better than 4 K for $z <$ 30 km and for $z >$ 65 km. A comparison to radiosonde and satellite observations (SABER) can be found in B. Kaifler and Kaifler (2021). They show that the lidar and radiosonde temperatures closely agree ($\Delta T <$ 0.6 K) for time-synchronized measurements at $z =$ 30 km and that the lidar and SABER temperatures agree well ($\Delta T <$ 3 K) at 45 km $< z <$ 50 km (note that the SABER data was taken at approximately 500 km distance from Río Grande). At times, the lidar measurements at the lowest altitudes are affected by the presence of aerosols. If the aerosol load is too high, temperature is underestimated due to cross-talk between the elastic channel and the Raman channel. Such data are omitted by the retrieval algorithm (most frequently for $z <$ 20 km). To allow for adequate sampling at all altitudes for all months, we limit the lowest altitude to 20 km for our analysis.

Measurements with CORAL started at Río Grande in November 2017. Río Grande is located in the lee of the Southern Andes at the east coast of Argentina at 100-200 km distance from the mountains that are to the south and west and at greater distance north-west of Río Grande (Reichert et al., 2021). The analyses in this study take into account data of the year 2018 which is continuously covered by the lidar measurements and by a uniform version of the IFS (see Sec. 2.2). In addition, data for May and July 2019 and 2020 are analyzed to investigate interannual variability using updated IFS versions. Note that CORAL measurements are taken fully autonomously with the help of IFS cloud forecasts and a cloud monitoring all-sky camera relying on star detection. Measurements are only possible during cloud-free/patchy conditions and during the night, which are the conditions our results are valid for. Night-time hours are between 2 and 7 UTC in mid-summer (December) and between 21 and 12 UTC in mid-winter (July). Figure 1a shows the time series of the nightly mean middle atmosphere temperature measurements from CORAL from 2018 to 2020, averaged over all measurements available each night. The band of highest middle atmosphere temperatures at the stratopause is perturbed by atmospheric waves in the extended winter period (April to September) and at the same time minimum temperatures in the mid-stratosphere are below 200 K (Fig. 1a).

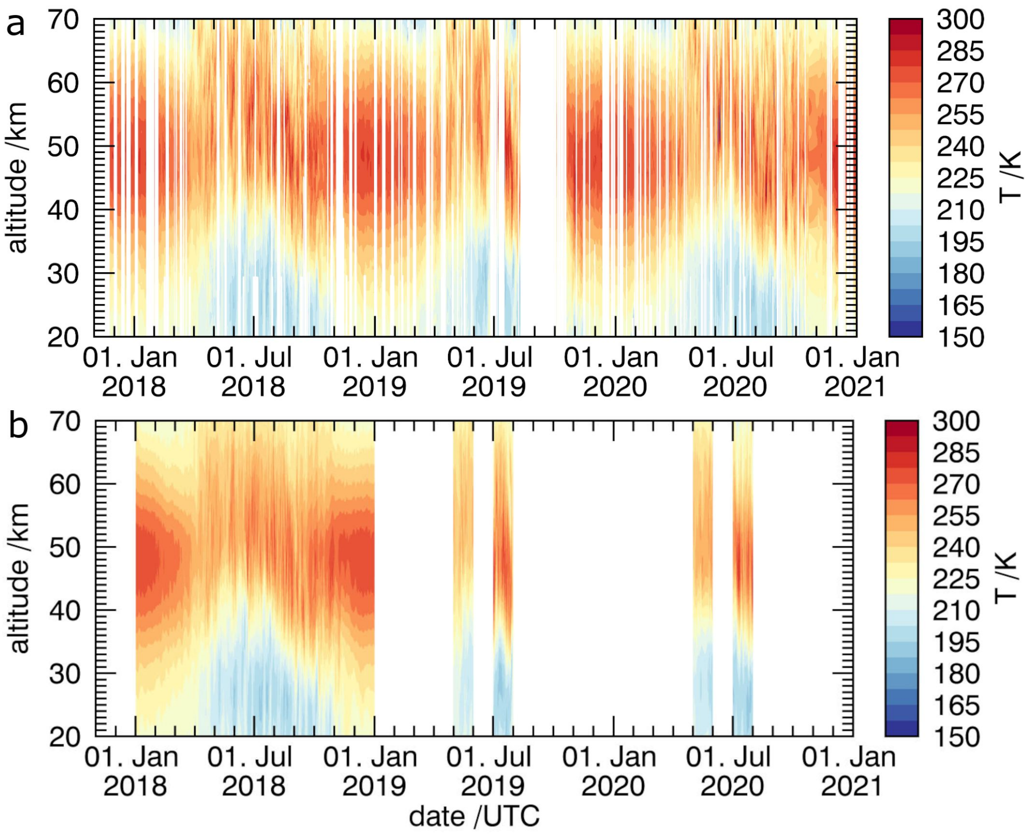

Figure 1. Nightly mean temperatures from (a) CORAL and (b) IFS. Measurement gaps of less than four nights are linearly interpolated in the upper contour plot (a). Bottom panel (b) shows IFS only for periods used in the comparison.

2.2 IFS model and data

IFS cycle 45r1 was already running in pre-operational phase during the first months of 2018 and eventually became operational in June 2018. Therefore, seasonal variations of the temperature differences between the lidar measurements and the IFS can be investigated based on a uniform version of the IFS for 2018. The updated cycles 46r1 and 47r1 became operational in June 2019 and June 2020, respectively. All three cycles have a horizontal grid-spacing of $\approx$ 9 km on the cubic octahedral grid (TCo1279). The model top is located at 0.01 hPa ($z \approx 80$ km) and 137 vertical levels are used. The layer thickness gradually increases from $\sim$ 300 m at $z \approx 10$ km to $\sim$ 400 m at $z \approx 20$ km, and $\sim$ 2 km at $z \approx 60$ km. We only use data up to $z = 70$ km, due to sparse coverage with only three more levels above that altitude. In the sponge layer, vertically propagating waves and the zonal-mean flow are damped above 10 hPa by hyper-diffusion applied on vorticity, divergence, and temperature and by additional strong first-order damping applied on divergence above 1 hPa. The smaller-scale waves are damped more strongly by such sponge formulation in the horizontal direction. Timescales of both damping mechanisms decrease with altitude and result in stronger damping at the higher altitudes (Polichtchouk et al., 2017; Ehard et al., 2018). A more detailed description of the changes in the IFS can be found on the ECMWF website (www.ecmwf.int/en/forecasts/documentation-and-support/changes-ecmwf-model, last access April 2022).

IFS analyses for 0, 6, 12, and 18 UTC are used and gaps are filled with short-lead-time forecasts (+1, +2, ...,+5, +7, +8, ..., +11 h) to get hourly data coverage. In addition, experimental 48 h forecasts without the sponge layer using cycle 45r1 are performed for May and August 2018. These forecasts can be directly compared to the operational forecasts with the sponge (up to +11 h). Further, we briefly investigate the effect of longer lead times (+25, ..., +35 h) on the temperature differences. For best temporal synchronisation, we extract single lidar temperature profiles that are closest in time (max. $\pm$10 min) to each IFS temperature profile at full hour interpolated on the location of Río Grande. The time step of the IFS (7.5 minutes) is close to the integration window of 15 minutes for the lidar profiles which makes this a reasonable one-to-one comparison. This selection results in 17 (summer) to 183 (winter) profiles per month. The profiles contribute 4-25 nights per month (Tab. 1). Especially for February to September above $z = 30$ km, the profiles provide an adequate sample for our study of middle atmosphere temperatures over Río Grande.

In summary, all IFS data for 2018 and May 2019 used here are based on operational high-resolution forecast (HRES) data for cycle 45r1 and hence variability due to fundamental changes in the model itself can be excluded. IFS data for July 2019 and May 2020 are based on cycle 46r1 and for July 2020 on cycle 47r1. Figure 1b shows the timeseries of nightly-mean IFS temperature data, taking into account hourly data between 21 and 12 UTC. Differences between the cycles are not expected to have an impact on the temperature over Río Grande, though it is beyond the scope of this study to quantify this. Such a quantification between

Table 1. Total number of nights with measurements and total number of profiles per months. Numbers for those reaching down to 20 km are also listed.

year	2018											
month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
#nights	8	15	20	21	22	22	22	24	19	14	10	12
#nights (20 km)	8	12	18	21	17	21	19	11	7	4	8	9
#total	19	54	90	117	153	183	162	122	102	69	39	40
#total (20 km)	17	40	73	86	113	170	139	43	28	15	31	33

year	2019		2020	
month	May	Jul	May	Jul
#nights	25	22	15	22
#nights (20 km)	25	21	14	22
#total	176	89	146	163
#total (20 km)	157	69	113	150

different IFS cycles was done in Ehard et al. (2018) for one month in Northern Europe, when IFS experienced a more major upgrade that included an increase in horizontal resolution in 2016.

2.3 Analysis of temperature differences, GW potential energy, and GW intermittency

The first part of the analysis focuses on temperature differences between individual IFS and lidar profiles and their seasonal and altitudinal variability

$$T_{\mathrm{diff}}(z,t) = T_{\mathrm{ECMWF}}(z,t) - T_{\mathrm{lidar}}(z,t), \tag{1}$$

where T_{ECMWF} is the IFS temperature profile, bilinearly interpolated to the horizontal location of the lidar at Río Grande taking into account the four surrounding grid-points, and T_{lidar} is the lidar temperature profile. All data are interpolated to a 100 m equidistant grid in altitude (z) and are available in time (t) at full hour. Afterwards monthly means are calculated

$$\overline{T_{\text{diff}}}(z) = \frac{\sum T_{\text{diff}}(z,t)}{\#\text{total}}, \tag{2}$$

where #total is the number of profiles for each month. In order to show the variability of the temperature differences between the individual profiles and account for the skewness of the difference distributions, the 15th/85th percentiles are also calculated. The number of profiles at the lowest altitudes can be small for individual months because not all measurements reach down to $z = 20$ km due to the presence of high amounts of aerosols (Sec. 2.1). The number of profiles per month and those reaching down to $z = 20$ km are summarized in Table 1 (also included in the relevant figures in Section 3). The numbers give an estimate of the number of profiles that determines the monthly means below and above $z = 30$ km. The number of profiles is largest in the extended winter period (April to September) when the nights are longest and cloud conditions are most favourable. $\overline{T_{\text{diff}}}(z)$ is equivalent to the difference between the monthly mean temperature profiles (i.e. $\overline{T_{\text{ECMWF}}(z)} - \overline{T_{\text{lidar}}(z)}$). $\overline{T_{\text{diff}}}(z)$ is likely dominated by large scale atmospheric features rather than GWs because temperature differences found for individual profiles may cancel out when averaged over a month. However, a systematic misrepresentation of GWs in the models can have an influence on the mean circulation (including temperature) in the middle atmosphere.

Averaged temperature differences for three altitude ranges (25 km $< z <$ 35 km, 35 km $< z <$ 45 km, and 45 km $< z <$ 55 km) are computed

$$\langle T_{\text{diff}} \rangle_{z_1\text{-}z_2}(t) = \frac{\sum_{z_1}^{z_2} T_{\text{diff}}(z,t)}{n_z}, \tag{3}$$

where n_z is the number of data points in each altitude range (z_1 to z_2). The upper altitude range lies within the strong IFS sponge layer (Sec. 2.2). The three altitude ranges are evaluated for each month by plotting their histograms with a bin size of 1 K.

We also analyse monthly root-mean-square-error (RMSE) profiles

$$\text{RMSE}(z) = \sqrt{\frac{\sum [T_{\text{ECMWF}}(z,t) - T_{\text{lidar}}(z,t)]^2}{\#\text{total}}} \tag{4}$$

where, in contrast to $\overline{T_{\text{diff}}}(z)$, temperature differences in the individual profiles do not cancel out in the monthly means. It is investigated whether wintertime GW amplitude and/or phase deviations give rise to enhanced RMSE between IFS and lidar data. Only for the following part of the analysis, where phase differences are quantified, lidar temperature profiles were smoothed with a 2-km running mean in order to neglect the smallest scales hardly resolved in the IFS due to increasing vertical grid spacing with altitude.

GW perturbations in terms of temperature fluctuations (T') are determined by applying a fifth-order Butterworth high-pass filter with a cut-off wavelength of 15 km to individual vertical profiles (Ehard et al., 2015, 2018). Therefore, the GW spectrum in our analysis is limited to GWs with vertical wavelengths smaller than approximately 15 km (note that our Butterworth filter does not have a sharp cut off). Afterwards, the perturbation amplitude $\sqrt{\langle T'^2 \rangle}$ is computed with a running mean over 15 km (angle brackets). Only profiles with an average amplitude > 3 K are considered. We derive the dominant vertical wavelengths and the respective phases as a function of altitude with wavelet analysis. The procedure consists of the following steps: the temperature perturbations are normalized with $\sqrt{\langle T'^2 \rangle}$ to ensure unbiased wavelet spectral power with altitude, and, between the lidar and the IFS. The wavelet analysis is performed with the code provided by Torrence and Compo (1998) and a Morlet wavelet with a normalized frequency $\omega_0 = 2$ is used in order to get high resolution in vertical space. The wavelet power spectrum is given by the square of the absolute value of the complex wavelet transform. The phase is defined via the arc tangent of the ratio between the imaginary and real part of the wavelet transform. A profile of the approximated dominant vertical wavelength is determined by finding the maximum in the wavelet power spectrum at each altitude. Taking the phase at these maxima results in a phase profile. The comparison of the phases determined for lidar and the IFS allows us to identify and quantify phase differences ($\Delta\phi$). The comparison of the vertical wavelengths in the lidar and the IFS data allows us to assess, whether phase differences are due to the misrepresentation of the vertical wavelenghts of the dominant GW in the IFS.

Last but not least, GW activity measured as GW potential energy per unit mass is compared between the lidar and the IFS data

$$E_p(z,t) = \frac{1}{2}\frac{g^2}{N^2(z,t)}\frac{\left\langle T'^2(z,t)\right\rangle_{15\mathrm{km}}}{T_0^2(z,t)} \tag{5}$$

$$\text{with} \quad N^2(z,t) = \frac{g}{T_0(z,t)}\left(\frac{dT_0(z,t)}{dz} + \frac{g}{c_p}\right), \tag{6}$$

where $T_0 = T - T'$ is the background temperature, N is the Brunt-Väisäla frequency, $g = 9.81$ m s^{-2} is the acceleration due to gravity, and c_p is the heat capacity of dry air at constant pressure (Ehard et al., 2015, 2018). For a monochromatic wave, E_p is based on T'^2 that is either integrated along height for one wavelength or along time for one wave period (Tsuda et al., 2004). For our individual profiles irregularily distributed in time, we use vertical averaging with a sliding window (Baumgaertner & McDonald, 2007) with a width of 15 km, i.e. the maximum wavelength in the T'-data, which is marked by the angle brackets in Eq. (5) (i.e., similar to the previous calculation of perturbation amplitudes for wavelet analysis). To avoid edge effects, the uppermost and lowermost 5 km of the E_p-profiles are discarded (Ehard et al., 2015). We limit our comparison to E_p and do not consider the vertical flux of horizontal momentum because the horizontal wavenumber needed in the computation (Ern et al., 2004; N. Kaifler et al., 2020) is not available from ground-based lidar measurements and corresponding vertical IFS profiles.

The annual cycle of E_p is analyzed in the middle atmosphere for 45 km $< z <$ 55 km. The distributions of E_p are determined for the altitude ranges 35 km $< z <$ 45 km and 45 km $< z <$ 55 km for May and August 2018. It was previously found that stratospheric E_p and GW momentum fluxes show a log-normal distribution rather than a normal distribution (Baumgaertner & McDonald, 2007; Hertzog et al., 2012). The probability density function for the log-normal distribution is given by

$$y = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln x - \mu)^2/2\sigma^2}, \tag{7}$$

where μ is the expected value and σ is the geometric standard deviation (Baumgaertner & McDonald, 2007). Taking this into account, monthly mean $\overline{E_p}$ are given based on the logarithmic mean (or geometric mean of the log-normal distribution) of E_p

$$\overline{E_p} = e^{\hat{\mu}}, \tag{8}$$

$$\hat{\mu} = \frac{\sum \ln[E_p(z,t)]}{n} \tag{9}$$

$$\text{and} \quad \hat{\sigma}^2 = \frac{\sum (\ln[E_p(z,t)] - \hat{\mu})^2}{n} \tag{10}$$

(Baumgaertner & McDonald, 2007) where $E_p(z,t)$ represents either all (n) values used in the monthly mean calculation in an particular altitude range ($\overline{E_p}$) or all values at each altitude ($n = \#$total) to calculate monthly mean $\overline{E_p}$-profiles.

However, distributions of GW activity over mountainous regions may have even larger tails that are not adequately described by a log-normal distribution (Plougonven et al., 2013). This enhanced intermittency of GW activity is caused by more frequent extreme GW events over mountainous regions compared to flat landscapes and ocean surfaces. The intermittency of GWs is important because the vertical profiles of GW momentum flux convergence determine the wave forcing of the mean wind, which is different for sporadic GWs with large amplitudes versus GWs with same mean momentum but smaller amplitudes (Minamihara et al., 2020). GW intermittency can be well quantified by the Gini coefficient (popular in economics) as in Plougonven et al. (2013) for GW momentum flux

$$I_g = \frac{\sum_{n=1}^{N-1}(n\bar{f} - F_n)}{\sum_{n=1}^{N-1} n\bar{f}}, \tag{11}$$

where in our case, F_n is the cumulative sum of $E_p(z,t)$ sorted in ascending order having an average $\bar{f} = F_N/N$. I_g is zero for a constant time series and one for a very intermittent data series. Near orography (e.g., the Antarctic Peninsula) enhanced values of 0.6-0.7 were found in the lower stratosphere in mesoscale simulations for austral spring 2005 (Plougonven et al., 2013).

3 Results

3.1 Temperature differences and seasonal variability

First, we quantify the temperature differences between CORAL and IFS (Eq. (2) and 15th/85th percentiles), including their altitudinal structure and seasonal variability, i.e. how they compare between the extended summer (October to March) and the extended winter (April to September, i.e. the GW-active season) periods. Monthly mean temperature differences for 2018 are overall < 2 K in the mid-stratosphere below $z = 40$ km (Fig. 2). Although a reduced number of data profiles is available at these altitudes (Sec. 2.1), the figure shows a small cold bias in the IFS with respect to the lidar below $z = 30$ km for Río Grande for March-September 2018, with the largest difference in August. While most of the months show a cold bias in the IFS up to $z = 45$ km, there is a 2 K warm bias at $z = 40$ km in August 2018. Around the stratopause at 45 km $< z <$ 55 km, the sign of the IFS temperature bias is changing throughout the year, with the largest warm bias (4 K) occuring in May 2018 and the largest cold bias (-10 K) in August 2018. There is a cold bias in the IFS (up to -4 K) for the extended summer period. Overall, lidar and IFS temperatures above $z = 45$ km show a good agreement in the extended summer period (quantified by a linear Pearson correlation cofficient > 0.7 for around 95 % of the profiles). In the extended winter period, the agreement is worse (linear Pearson correlation cofficient > 0.7 only for around 60 % of the profiles). The results are most reliable at altitudes above 30 km, because the uncertainty of the lidar measurements is < 1 K at 30 km $< z <$ 60 km (Sec. 2.1).

The comparisons for May and August 2018 are also repeated for forecast lead times of 25 to 35 hours and the warm IFS bias at $z = 50$ km for May and at $z = 40$ km for August is found to be 1-3 K larger (not shown). This indicates that a warm mid-stratosphere bias in IFS grows for longer lead times.

The 15th/85th percentile, that describe how much the temperature differences between the IFS and lidar for individual temperature profiles vary within the month, is significantly larger and increases with altitude in the extended winter period (April to September) compared to the other months (Fig. 2). In the upper stratosphere, the percentiles deviate from the mean by up to ~ 10 K in August 2018.

When other years are considered, the mean temperature differences in the upper stratosphere for 40 km $< z <$ 50 km are smaller in May 2019 and 2020 in comparison to May 2018 (Fig. 2). For July 2019 and 2020, a cold bias of -10 K is present around the stratopause (45 km $< z <$ 50 km) in the IFS. This is not the case for July 2018, but a similar bias is found for August 2018 (Fig. 2). These changing biases are likely due to variability in the overall atmospheric conditions. Monthly mean stratopause temperatures (not shown) are higher (approx. 268 K) in August 2018, July 2019 and 2020 in comparison to July 2018 (approx. 258 K). The IFS does not capture these enhanced stratopause temperatures which explains the larger monthly mean temperature differences at 45 km $< z <$ 55 km for these three months, independent of the IFS cycle. Further, the spread between the 15th/85th percentile in May and July is similar or slightly smaller for 2019 and 2020 compared to 2018. The spread increases with altitude also for 2019 and 2020, i.e. in the updated IFS cycles.

The temperature differences and their variability in the course of the year are investigated in more detail for the three middle-atmospheric altitude ranges (Eq. (3)) by computing histograms. The distribution of the temperature differences is narrowest for the summer months (exemplarily shown for January and October 2018) for all three altitude ranges and differences between individual profiles are rarely found outside the range of ± 5 K (Fig. 3). The largest differences, exceeding ± 5 K, are found in the winter months mainly above $z = 45$ km. There, the IFS experiences a warm bias of up to 15 K (May, July 2018) and a cold bias of more than -15 K (August 2018). The distributions are very similar for May and July 2019 and 2020 (gray shaded panels in Fig. 3) and for 2018. However, the distributions are better centered at zero for May 2019 and 2020 around the stratopause (45 km $< z <$ 55 km), which results in smaller differences in the mean profiles in Figure 2. In contrast, the distributions for July 2019 and 2020 are clearly shifted to negative values in comparison to July 2018, i.e. temperatures are more frequently underestimated by < -5 K in the IFS, as is found for August 2018 (Fig. 3).

The corresponding RMSE profiles are shown for all months in Figure 4. Again, the results are most reliable at altitudes above 30 km because the uncertainty of the lidar is smallest and the total number of profiles larger for 30 km $< z <$ 60 km (Sec. 2.1). Overall, the RMSE is mostly smaller than 5 K up to $z = 45$ km but clearly increases with altitude and can exceed 10 K in the extended winter period (April to September). In the stratosphere (i.e. below 55 km altitude), the RMSE is found to be largest in August 2018 and June 2019 and 2020. Our hypothesis is that the presence of GWs in the middle atmosphere can cause large differences for individual temperature profiles during this time of the year due to amplitude and phase errors (analyzed in the following section).

The annual cycle for 2018 of the absolute monthly mean temperature differences ($|\overline{T_{\mathrm{diff}}}|$) and the RMSE averaged for 45 km $< z <$ 55 km is shown in Figure 5. There is no winter maximum or robust annual cycle detected for $\langle|\overline{T_{\mathrm{diff}}}|\rangle$. Minima are found for May 2019 and 2020 (symbols in Fig. 5) because the monthly mean profiles agree well up to $z = 55$ km (Fig. 2). However, the RMSE shows maximum values in the extended winter period continuously larger than 7 K. This illustrates the seasonal variability discussed above for the individual months. The annual cycle is later correlated to $\overline{E_p}$ in the middle atmosphere over Río Grande to relate the seasonal variability of middle atmosphere temperature differences to GW activity.

3.2 Amplitude and phase deviations

As the largest temperature differences between IFS and lidar occur in winter, at the time of enhanced GW activity over Río Grande (next section and Figure 8), we now investigate whether GW amplitude and/or phase deviations in the IFS are causative. Figure 6 shows an example of such amplitude and phase deviations for two individual profiles in May 2018. The profiles for both days show qualitative agreement in phase and amplitude up to $z = 45$ km (Fig. 6a,c). Higher up, there is an amplitude error of more than 20 K on 31 May 2018 (Fig. 6a) and a clear phase error on 21 May 2018 (Fig. 6c). It was already mentioned that the sponge damps GW amplitudes in the IFS in the middle atmosphere. Reducing the sponge strength may also reduce temperature differences caused by GWs. This is illustrated by the purple profile in Figure 6a where the sponge was removed in the experimental IFS simulations leading to a reduction of the amplitude error at 60 km. However, the removal of the sponge can lead to even larger temperature differences at certain altitudes for cases that show a phase error even though the GW amplitude itself is closer to the observations (purple profile in Figure 6c).

Phase deviations between lidar and IFS are quantified based on wavelet analysis (see Section 2.3). Up to $z = 45$ km, phase shifts are less than 90 degrees for both cases in May 2018 (Fig. 6b,d) and the vertically averaged values for 35 km $< z <$ 45 km are 45 degrees and 33 degrees for 21 May and 31 May 2018, respectively. Above $z = 45$ km, phase shifts increase beyond 90 degrees for 21 May 2018 (Fig. 6d) and the vertically averaged value for 45 km $< z <$ 60 km is 59 degrees. The phase shift at these altitudes is related to longer vertical wavelengths in the IFS compared to lidar (Fig. 6d). To determine the role of phase deviations, we separate the profiles into those with good phase agreement ($\Delta\phi < 50$ deg) between lidar and IFS and those with poor phase agreement ($\Delta\phi \geq 50$ deg). The number of profiles that have poor phase agreement at 45 km $< z <$ 60 km is larger for May 2018 (66 % of the profiles) compared to August 2018 (39 % of the profiles).

In Figure 7, mean vertical wavelength and phase differences for May and August 2018 are shown. In general, the mean vertical wavelength of the dominant GWs in the lidar data in May 2018 increases from around 7 km to 12 km between $z = 20$ km and $z = 45$ km and then drops down to less than 10 km aloft. This drop is not found in the IFS up to $z = 60$ km. This was already seen for 21 May 2018 (Fig. 6c, d) and appears to also be a dominant feature in the monthly mean (Fig. 7a). In contrast, the vertical wavelength is fairly constant and larger than 10 km above $z = 30$ km in August 2018 (Fig. 7b). The vertical wavelengths in the IFS and lidar agree better at $z = 50$ km than in May 2018. The mean phase difference at this altitude is almost 90 degrees in May 2018 while it is close to 45 degrees in August 2018 (Fig. 7).

3.3 Gravity wave activity, intermittency, and effect of the model sponge

The GW potential energy E_p (Eq. (5)) is independent of the wave phase, and thus can be used to quantify GW amplitude deviations between IFS and lidar. Figure 8 shows the annual cycle of $\overline{E_p}$ for lidar and IFS for the altitude range 45 km $< z <$ 55 km. The annual cycle with maximum (minimum) GW activity in the winter (summer), that is characteristic for the Southern Andes region (Schroeder et al., 2009), is well reproduced by the IFS also above $z = 45$ km i.e., within the sponge layer.

Monthly mean $\overline{E_p}$ in the IFS is generally underestimated due to GW amplitude errors (and therefore underestimated T'). However, the reduction of $\overline{E_p}$ for May and July 2020 compared to 2018 is reproduced by the IFS (see markers in Fig. 8). E_p of all individual profiles, vertically averaged for the same altitude range, are also shown in Figure 8. This shows that even though E_p is calculated following Ehard et al. (2015) with T'^2 averaged in the vertical (Tsuda et al., 2004), our E_p values are qualitatively similar to the E_p values in Reichert et al. (2021) (see their Fig. 6). Moreover, E_p uncertainties due to lidar temperature uncertainties are insignificant at altitudes between 30 km and 60 km (Reichert et al., 2021). E_p for the individual profiles also reveals that IFS indeed captures high E_p values of some strong GW events like the one in June 2018 (crosses in Fig. 8), which was analyzed in detail by N. Kaifler et al. (2020).

Coming back to the seasonal variability of the temperature differences between the IFS and lidar, one finds that GW activity (Fig. 8) and the RMSE (Fig. 5) show a similar annual cycle. The correlation coefficent between lidar $\overline{E_p}$ and the RMSE is 0.96 for 2018. The correlation is smaller (0.42) for lidar $\overline{E_p}$ and $\langle|\overline{T_{\text{diff}}}|\rangle$. This suggests that the monthly mean temperature differences are not dominated by the misrepresentation of GWs.

The distributions of E_p for altitudes weakly affected by the model sponge (35 km $< z <$ 45 km) and strongly affected by the sponge (45 km $< z <$ 55 km) are shown in Figure 9 for May and August 2018. The distributions are in general log-normal with partly larger tails, as can be seen by comparing to the probability density function computed from Eq. (7) using $\hat{\mu}$ and $\hat{\sigma}$. The expected or mean value $\hat{\mu}$ and the geometric standard deviation $\hat{\sigma}$ are better suited to describe the distributions than the arithmetic mean and standard deviation. $\hat{\sigma}$ of the lidar and IFS distributions for the two months is close to unity and clear differences are found for $\hat{\mu}$. Overall, GW activity is larger in August compared to May. $\hat{\mu}$ for the IFS is 59 to 67 % of $\hat{\mu}$ for the lidar measurements in the lower altitude range, leading to $\overline{E_p}$ in the IFS only reaching around 35 % of the $\overline{E_p}$ in the lidar (Fig. 9a,c; Fig. 11). Nevertheless, the IFS captures some events of enhanced E_p as can be seen for example for May (E_p of 80 J/kg in Figure 9a).

In the upper altitude range, the comparison of the E_p distribution and the corresponding probability density function reveals that the IFS is missing the highest E_p values in the tail of the log-normal distribution, especially in August (Fig. 9b,d). $\overline{E_p}$ for the IFS is only 10-17 % of the $\overline{E_p}$ for the lidar (Fig. 9b,d; Fig. 11). The 'no-sponge' IFS simulations show that the missing high E_p values and fairly low $\overline{E_p}$ are partly due to the sponge (Fig. 10b,d). The removal of the sponge leads to an increase of $\hat{\mu}$ and corresponding $\overline{E_p}$ to 26 % and 42 % of $\overline{E_p}$ for the lidar for May and August 2018, respectively (Fig. 10b,d; Fig. 11). Longer lead times of 25 to 35 hours further increase $\overline{E_p}$ in the 'no-sponge' simulations to 31 % for May 2018, while $\overline{E_p}$ stays almost the same (45 %) for August 2018 (not shown). At altitudes 35 km $< z <$ 45 km, E_p remains similar in the 'no-sponge' simulations with values generally smaller than 120 J/kg (Fig. 10a,c).

In addition to the effect of the sponge layer, small scale GWs that are not resolved in the vertical in the IFS contribute to the underestimation of E_p in the IFS when compared to lidar. Regridding lidar temperature data to the 137 IFS vertical levels prior to the E_p calculation on the 100-m-grid eliminates GW structures from the lidar data that cannot be represented by the IFS solely due to the limited vertical resolution. The high E_p values and averaged $\overline{E_p}$ of the lidar measurements are reduced by a similar amount as E_p values increase in the IFS when the sponge is removed (Fig. 10; Fig. 11). Clear differences between the E_p distributions of the original lidar data and the regridded lidar data can be seen for E_p values larger than 200 J kg^{-1} (240 J kg^{-1}) for May (August) for 45 km $< z <$ 55 km (Fig. 10b,d; Fig. 10b,d). The contribution of unresolved scales in the IFS is likely even larger because this estimate does not consider the effective vertical resolution or scales not resolved horizontally. The lidar data does not provide any information on horizontal scales. Given that the effective horizontal resolution of the model is approximately 6-10 times the grid spacing due to explicit and implicit model diffusion, the IFS is unlikely to resolve horizontal wavelengths smaller than $\sim 50 - 90$ km outside the sponge layer. In the sponge layer, the effective resolution is much coarser than that due to a hyperviscosity type sponge that acts on the horizontal wavenumber.

To quantify the importance of extreme GW events (i.e., large E_p values and intermittent GW activity), the Gini coefficient (Eq. (11)) is calculated for the two altitude regions for May and August 2018 (Tab. 2). Weaker extreme GW events in combination with smaller mean GW activity for May results in a similar Gini coefficient as for August, when extreme GW events are stronger and the mean GW activity is larger. The lidar and the IFS agree in terms of GW intermittency for 35 km $< z <$ 45 km. Above, the intermittency slightly decreases for the lidar while it is almost constant for the IFS for August 2018. The inter-

Table 2. Gini coefficient (Eq. (11)) for May and August 2018

data	month	35-45 km	45-55 km
CORAL	May 2018	0.50	0.46
IFS	May 2018	0.53	0.50
IFS no sponge	May 2018	0.55	0.56
CORAL	Aug 2018	0.51	0.43
IFS	Aug 2018	0.53	0.52
IFS no sponge	Aug 2018	0.50	0.45

mittency in the IFS slightly decreases (increases) for August (May) at 45 km $< z <$ 55 km when the sponge is removed. The latter finding can be reproduced by repeating the analysis with better statistics for the full hourly IFS data set for May and August 2018, i.e. not limited to times where lidar observations are available.

4 Discussion and Summary

Similar to previous studies for Europe (Le Pichon et al., 2015; Ehard et al., 2018; Marlton et al., 2021), we found a generally good agreement between the IFS and lidar temperature data up to 45 km altitude at higher mid-latitudes in the Southern Hemisphere, in the lee of the Southern Andes. Monthly mean temperature differences between the IFS and lidar are $<$ 2 K for altitudes 20 km $< z <$ 40 km for all months, and, apart from August 2018, usually IFS exhibits a cold bias with respect to lidar. Near the stratopause at 45 km $< z <$ 55 km, which is above the peak altitude of assimilated radiances (1-2 hPa) in the IFS and influenced by the strong sponge, there is more time variability and the sign of the monthly mean temperature differences changes throughout the year. The largest monthly mean warm bias in the IFS with respect to lidar (4 K) occurs in May 2018 and the largest cold bias (-10 K) occurs in August 2018, July 2019, and July 2020 and is related to the warm stratopause (approx. 268 K). This suggests that the IFS cold bias in the upper stratosphere at Río Grande in winter lies within the range found for the older IFS cycle 41r1 (-8 K) and cycle 41r2 (-20 K) in the Northern Hemisphere for December 2015 (Ehard et al., 2018). For the extended summer period (October to March 2018), the monthly mean cold bias in the IFS is at most -4 K for 45 km $< z <$ 55 km and the differences for individual profiles are rarely found outside the range of ± 5 K. The spread of the difference distribution (15th/85th percentiles), the RMSE, and maximum differences for individual profiles are significantly larger and increase with altitude in winter ($> \pm 10$ K). The lidar and the IFS temperatures show better correlation in the extended summer period than in the extended winter period. The better agreement between the IFS and lidar in the summer months previously found for the Northern Hemisphere (Le Pichon et al., 2015) also manifests for the Southern Hemisphere and a more recent IFS cycle. The high correlation between the annual cycle of the RMSE and of the GW activity supports the hypothesis that the seasonal variability of the temperature differences over Río Grande is related to the middle atmosphere GW activity.

The wavelet analysis of individual profiles for May and August 2018, revealed that the GWs in the lidar measurements and IFS have similar vertical wavelenghts and are largely in phase ($\Delta\phi < 50$ deg) below $z = 45$ km. This means that the temperature differences at these altitudes are mainly due to deviations in amplitudes. Enhanced phase deviations ($\Delta\phi \geq 50$ deg) are found to be a feature of the upper stratosphere and lower mesosphere and are therefore likely a result of the propagation and representation of GWs in the middle atmosphere in the IFS. The vertical wavelength is clearly overestimated in the IFS compared to the lidar in the monthly mean for May 2018, though better agreement was found for August 2018. Resulting temperature differences at these altitudes are as such a combination of amplitude and phase deviations that are related to differences in the vertical wavelengths. Differences in the vertical wavelengths could be caused by errors in the horizontal wind (strength and/or direction) and/or inadequate vertical resolution in the IFS at these altitudes. The larger number of profiles that show poor phase agreement for May 2018 (66 %) compared to August (39 %) could be the reason why satellite observations in the upper stratosphere are rejected by the 4D-Var in the IFS more frequently in May. To the best of our knowledge, a quantitative evaluation of phase deviations in the wintertime temperature perturbation profiles that are shaped by GWs has not been published for the IFS before. For an eight-day period with strong GW activity in June 2018, N. Kaifler et al. (2020) found good agreement between lidar and IFS in amplitude and phase of the mountain waves over Río Grande. Such information can only be extracted when instantaneous temperature profiles are available instead of nightly means (e.g., Le Pichon et al., 2015) and when the analysis is not only restricted to monthly mean statistics (e.g., Ehard et al., 2018).

The analysis of the annual cycle of GW activity in the middle and upper stratosphere complements the findings by Schroeder et al. (2009) for the Andes and reveals that the IFS captures the winter maximum and summer minimum well also at altitudes above 30 km. In general, the IFS underestimates E_p in the middle atmosphere over Río Grande and the discrepancy is increasing with altitude. $\overline{E_p}$ of the IFS above $z = 45$ km is only around 10 % of $\overline{E_p}$ derived from the lidar observations. Similar results are found for ERA5 in Strelnikova et al. (2021) who show that GW potential energy densities of ERA5 at $z = 55$ km are on average one order of magnitude smaller (i.e., reaching only 10 %) when compared to two European lidar stations. However, there can be a good agreement below $z = 45$ km for individual events like the one at Río Grande in June 2018 analyzed in detail by N. Kaifler et al. (2020). While the removal of the sponge in the IFS can lead to increasing temperature differences at certain altitudes for profiles with phase deviations, it has a positive effect on E_p (i.e., an increase) above $z = 45$ km because E_p is independent of the GW phase. $\overline{E_p}$ increases from only ≈10 % of the $\overline{E_p}$ of the lidar measurements to 26 % and 42 % for May and August 2018, respectively, when the sponge is removed. This shows that the sponge is an important but not the only cause for a reduced $\overline{E_p}$ in the IFS. Given this, the plan at ECMWF is to reduce the depth of the sponge layer in the upcoming IFS upgrade as well as to remove the weak damping on the zonal-mean by the sponge. In addition to the sponge, a too low model resolution is likely important as some of the GWs are unresolved in the IFS. In particular, the coarse vertical resolution in the upper stratosphere and lower mesosphere likely plays a role.

GW intermittency has been previously quantified by the Gini coefficient for GW momentum fluxes determined from e.g., balloon (Plougonven et al., 2013), satellite (Wright et al., 2013; Hindley et al., 2019) or radar (Minamihara et al., 2020) measurements. These different observations are sensitive to different parts of the GW spectrum and focus on different time periods and locations than discussed in this study. Therefore, it is not reasonable to directly compare GW intermittency for GW momentum fluxes in the aforementioned studies to the E_p-intermittency here. Hence, the discussion here is limited to the relative changes in the Gini coefficient with altitude over Río Grande. GW intermittency slightly decreases for the lidar measurements from 35-45 km to 45-55 km altitude. It is almost constant for the operational IFS data for August 2018 but slightly decreases with altitude when the sponge is removed. In regions where orographic GWs dominate, the intermittency decreases with height when GWs with large momentum flux are removed at altitudes where the background wind matches the ground-based phase velocity of the GWs (Minamihara et al., 2020). However, this mechanism cannot explain the steep decline of GW intermittency found around the tropopause in the PANSY MST radar data at Syowa station, Antarctica. Instead, partial reflection due to discontinuities in static stability at the tropopause, is mentioned as one possible mechanism (Minamihara et al., 2020). Changing static stability in the vicinity of the stratopause at around 50 km (Fig. 1) can have a similar effect on the GW intermittency in the middle atmosphere over Río Grande. In addition, large-amplitude orographic GWs can break or dissipate well below their critical level at the mesopause in winter or propagate horizontally out of the observational volume of the ground-based lidar (Ehard et al., 2017). All these processes are potentially important and could lead to decreasing intermittency with altitude at the location of Río Grande. However, the differences and changes we found in the Gini coefficient lie below the differences between orography (0.8) and ocean (0.5) found in the lower stratosphere (Plougonven et al., 2013). A stronger decrease in intermittency is found over Río Grande above 60 km altitude in winter (0.22) and can be related to the saturation of the GW spectrum (Reichert et al., 2021). Overall, the GW intermittency in the IFS is close to the intermittency in lidar measurements, even though the E_p distributions of the IFS are shifted to smaller E_p values compared to the lidar measurements.

In summary, this study presents the first detailed analysis of local differences between middle atmosphere lidar temperature measurements and IFS temperatures for the GW hot spot region of the Southern Andes. It was found that the ability of the IFS to accurately represent temperatures over Río Grande depends on the altitude range and season. In particular, conditions in summer are better captured by the IFS than the more complex wintertime conditions with large-amplitude GWs. The shortcomings in the representation of middle atmosphere GWs in the IFS are characterized by amplitude and phase differences that contribute to the site-specific temperature differences. While amplitude deviations in the IFS are due to the sponge and unresolved GWs, the origin of the GW phase shift often observed in the upper stratosphere and lower mesosphere between the IFS and the lidar data, is related to differences in the vertical wavelength. In the mid-stratosphere, the IFS has a good representation of the GW vertical wavelengths and phases. Investigating this topic in more detail could help to understand why phase deviations are happening frequently in fall, i.e. May, and improving the vertical wavelength and phase representation could help preventing the rejection of satellite observations in the IFS data assimilation system. Misrepresentation of the middle atmosphere winds over Río Grande in early winter, when the polar vortex is not yet fully formed, or wind variations by tides or planetary waves could be parts of the issue. Moreover, improving GW amplitudes in the upper stratosphere and lower mesosphere by e.g., a weaker sponge, will help only if GW phases are represented correctly.

Acknowledgments

Part of this research was conducted within programs of the German Science foundation under Grants GW-TP/DO1020/9-1 and PACOG/RA1400/6-1 in the frame of the DFG research group MS-GWAVES (FOR1898) and of the German Federal Ministry for Education and Research under Grant 01LG1907 (project WASCLIM) in the framework of ROMIC (Role of the Middle Atmosphere in Climate). BK and NK thank Jose-Luis Hormaechea and Alejandro de la Torre for support with logistics and operation of the lidar at Estacion Astronomica Río Grande (EARG) station in Río Grande. SG and IP thank the ISSI Team "New Quantitative Constraints on OGW Stress and Drag" at the ISSI Bern for fruitfull discussions of related results. A 2-months visit of SG at ECMWF in collaboration with IP and IS was funded by ECMWF. Access to ECMWF operational forecast data was granted through the special project "Deep Vertical Propagation of Internal Gravity Waves".

Data availability The ECMWF IFS and CORAL temperature profile data used in the study are available at the HALO-database (DLR/ECMWF, 2022, license CC BY 4.0 and ECMWF's Terms of Use apply). Dataset numbers are: 7905-7925. Wavelet software was provided by C. Torrence and G. Compo (Torrence & Compo, 1998, 2022).

References

Alexander, M. J. (1998). Interpretations of observed climatological patterns in stratospheric gravity wave variance. *J. Geophys. Res. Atmos.*, *103*(D8), 8627–8640. doi: 10.1029/97JD03325

Baldwin, M. P., & Dunkerton, T. J. (2001). Stratospheric harbingers of anomalous weather regimes. *Science*, *294*(5542), 581–584. doi: 10.1126/science.1063315

Baumgaertner, A. J. G., & McDonald, A. J. (2007). A gravity wave climatology for Antarctica compiled from Challenging Minisatellite Payload/Global Positioning System (CHAMP/GPS) radio occultations. *Journal of Geophysical Research: Atmospheres*, *112*(D5). doi: 10.1029/2006JD007504

Byrne, N. J., & Shepherd, T. G. (2018). Seasonal persistence of circulation anomalies in the southern hemisphere stratosphere and its implications for the troposphere. *Journal of Climate*, *31*(9), 3467 - 3483. doi: 10.1175/JCLI-D-17-0557.1

DLR/ECMWF. (2022). SOUTHWAVE [dataset]. *HALO-DB*, https://halo-db.pa.op.dlr.de/mission/111.

Eckermann, S. D., Ma, J., Hoppel, K. W., Kuhl, D. D., Allen, D. R., Doyle, J. A., ... Love, P. T. (2018). High-altitude (0–100 km) global atmospheric reanalysis system: Description and application to the 2014 austral winter of the deep propagating gravity wave experiment (deepwave). *Monthly Weather Review*, *146*(8), 2639 - 2666. doi: 10.1175/MWR-D-17-0386.1

Ehard, B., Kaifler, B., Dörnbrack, A., Preusse, P., Eckermann, S. D., Bramberger, M., ... Rapp, M. (2017). Horizontal propagation of large-amplitude mountain waves into the polar night jet. *J. Geophys. Res. Atmos.*, *122*(3), 1423–1436. doi: 10.1002/2016JD025621

Ehard, B., Kaifler, B., Kaifler, N., & Rapp, M. (2015). Evaluation of methods for gravity wave extraction from middle-atmospheric lidar temperature measurements. *Atmos. Meas. Tech.*, *8*, 4645-4655. doi: 10.5194/amt-8-4645-2015

Ehard, B., Malardel, S., Dörnbrack, A., Kaifler, B., Kaifler, N., & Wedi, N. (2018). Comparing ECMWF high-resolution analyses with lidar temperature measurements in the middle atmosphere. *Q.J.R. Meteorol. Soc.*, *144*(712), 633-640. doi: 10.1002/qj.3206

Ern, M., Preusse, P., Alexander, M. J., & Warner, C. D. (2004). Absolute values of gravity wave momentum flux derived from satellite data. *Journal of Geophysical Research: Atmospheres*, *109*(D20). doi: 10.1029/2004JD004752

Fritts, D. C., & Alexander, J. M. (2003). Gravity wave dynamics and effects in the middle atmosphere. *Rev. Geophys.*, *41*, 1003. doi: 10.1029/2001RG000106

Garfinkel, C. I., & Hartmann, D. L. (2011). The Influence of the Quasi-Biennial Oscillation on the Troposphere in Winter in a Hierarchy of Models. Part I: Simplified Dry GCMs. *Journal of the Atmospheric Sciences*, *68*(6), 1273 - 1289. doi: 10.1175/2011JAS3665.1

Gisinger, S., Dörnbrack, A., Matthias, V., Doyle, J. D., Eckermann, S. D., Ehard, B., ... Rapp, M. (2017). Atmospheric Conditions during the Deep Propagating Gravity Wave Experiment (DEEPWAVE). *Mon. Weath. Rev.*, *145*(10), 4249-4275. doi: 10.1175/MWR-D-16-0435.1

Hertzog, A., Alexander, M. J., & Plougonven, R. (2012, 11). On the Intermittency of Gravity Wave Momentum Flux in the Stratosphere. *Journal of the Atmospheric Sciences*, *69*(11), 3433-3448. doi: 10.1175/JAS-D-12-09.1

Hindley, N. P., Wright, C. J., Smith, N. D., Hoffmann, L., Holt, L. A., Alexander, M. J., ... Mitchell, N. J. (2019). Gravity waves in the winter stratosphere over the Southern Ocean: high-resolution satellite observations and 3-D spectral analysis. *Atmospheric Chemistry and Physics*, *19*(24), 15377–15414. doi: 10.5194/acp-19-15377-2019

Hoffmann, L., Xue, X., & Alexander, M. (2013). A global view of stratospheric gravity wave hotspots located with atmospheric infrared sounder observations. *J. Geophys. Res. Atmos.*, *118*(2), 416–434.

Kaifler, B., & Kaifler, N. (2021). A Compact Rayleigh Autonomous Lidar (CORAL) for the middle atmosphere. *Atmospheric Measurement Techniques*, *14*(2), 1715–1732. doi: 10.5194/amt-14-1715-2021

Kaifler, N., Kaifler, B., Dörnbrack, A., Rapp, M., Hormaechea, J., & Torre, A. (2020, 09). Lidar observations of large-amplitude mountain waves in the stratosphere above Tierra del Fuego, Argentina. *Scientific Reports*, *10*, 14529. doi: 10.1038/s41598-020-71443-7

Le Pichon, A., Assink, J. D., Heinrich, P., Blanc, E., Charlton-Perez, A., Lee, C. F., ... Claud, C. (2015). Comparison of co-located independent ground-based middle atmospheric wind and temperature measurements with numerical weather prediction models. *Journal of Geophysical Research: Atmospheres*, *120*(16), 8318-8331. doi: https://doi.org/10.1002/2015JD023273

Marlton, G., Charlton-Perez, A., Harrison, G., Polichtchouk, I., Hauchecorne, A., Keckhut, P., ... Steinbrecht, W. (2021). Using a network of temperature lidars to identify temperature biases in the upper stratosphere in ECMWF reanalyses. *Atmospheric Chemistry and Physics*, *21*(8), 6079–6092. doi: 10.5194/acp-21-6079-2021

Minamihara, Y., Sato, K., & Tsutsumi, M. (2020). Intermittency of Gravity Waves in the Antarctic Troposphere and Lower Stratosphere Revealed by the PANSY Radar Observation. *Journal of Geophysical Research: Atmospheres*, *125*(15), e2020JD032543. doi: https://doi.org/10.1029/2020JD032543

Plougonven, R., Hertzog, A., & Guez, L. (2013). Gravity waves over Antarctica and the Southern Ocean: consistent momentum fluxes in mesoscale simulations and stratospheric balloon observations. *Quarterly Journal of the Royal Meteorological Society*, *139*(670), 101-118. doi: 10.1002/qj.1965

Polichtchouk, I., Hogan, R., Shepherd, T., Bechtold, P., Stockdale, T., Malardel, S., ... Magnusson, L. (2017). What influences the middle atmosphere circulation in the IFS? *ECMWF Technical Memoranda*(809). doi: 10.21957/mfsnfv15o

Preusse, P., Eckermann, S. D., & Ern, M. (2008). Transparency of the atmosphere to short horizontal wavelength gravity waves. *Journal of Geophysical Research: Atmospheres*, *113*(D24). doi: https://doi.org/10.1029/2007JD009682

Preusse, P., Ern, M., Bechtold, P., Eckermann, S. D., Kalisch, S., Trinh, Q. T., & Riese, M. (2014). Characteristics of gravity waves resolved by ECMWF. *Atmospheric Chemistry and Physics*, *14*(19), 10483–10508. doi: 10.5194/acp-14-10483-2014

Reichert, R., Kaifler, B., Kaifler, N., Dörnbrack, A., Rapp, M., & Hormaechea, J. L. (2021). High-Cadence Lidar Observations of Middle Atmospheric Temperature and Gravity Waves at the Southern Andes Hot Spot. *Journal of Geophysical Research: Atmospheres*, *126*(22), e2021JD034683. doi: https://doi.org/10.1029/2021JD034683

Schroeder, S., Preusse, P., Ern, M., & Riese, M. (2009). Gravity waves resolved in ECMWF and measured by SABER. *Geophysical Research Letters*, *36*(10). doi: https://doi.org/10.1029/2008GL037054

Shepherd, T. G., Semeniuk, K., & Koshyk, J. N. (1996). Sponge layer feedbacks in middle-atmosphere models. *Journal of Geophysical Research: Atmospheres*, *101*(D18), 23447-23464. doi: https://doi.org/10.1029/96JD01994

Simmons, A., Soci, C., Nicolas, J., Bell, B., Berrisford, P., Dragani, R., ... Schepers, D. (2020). Global stratospheric temperature bias and other stratospheric aspects of ERA5 and ERA5.1. (859). doi: 10.21957/rcxqfmg0

Strelnikova, I., Almowafy, M., Baumgarten, G., Baumgarten, K., Ern, M., Gerding, M., & Lübken, F.-J. (2021). Seasonal Cycle of Gravity Wave Potential Energy Densities from Lidar and Satellite Observations at 54° and 69°N. *Journal of the Atmospheric Sciences*, *78*(4), 1359 - 1386. doi: 10.1175/JAS-D-20-0247.1

Torrence, C., & Compo, G. P. (1998). A practical guide to wavelet analysis. *Bull. Amer. Meteor. Soc.*, *79*, 61-78. doi: 10.1175/1520-0477(1998)079⟨0061:APGTWA⟩2.0.CO;2

Torrence, C., & Compo, G. P. (2022). Wavelet software [software]. *website*, http://atoc.colorado.edu/research/wavelets/.

Tsuda, T., Ratnam, M. V., May, P. T., Alexander, M. J., Vincent, R. A., & MacKinnon, A. (2004). Characteristics of gravity waves with short vertical wavelengths observed with radiosonde and GPS occultation during DAWEX (Darwin Area Wave Experiment).

J. Geophys. Res. Atmos., *109*(D20).

Wright, C. J., Osprey, S. M., & Gille, J. C. (2013). Global observations of gravity wave intermittency and its impact on the observed momentum flux morphology. *Journal of Geophysical Research: Atmospheres*, *118*(19), 10,980-10,993. doi: https://doi.org/10.1002/jgrd.50869

Yamashita, C., Liu, H.-L., & Chu, X. (2010). Gravity wave variations during the 2009 stratospheric sudden warming as revealed by ECMWF-T799 and observations. *Geophysical Research Letters*, *37*(22). doi: https://doi.org/10.1029/2010GL045437

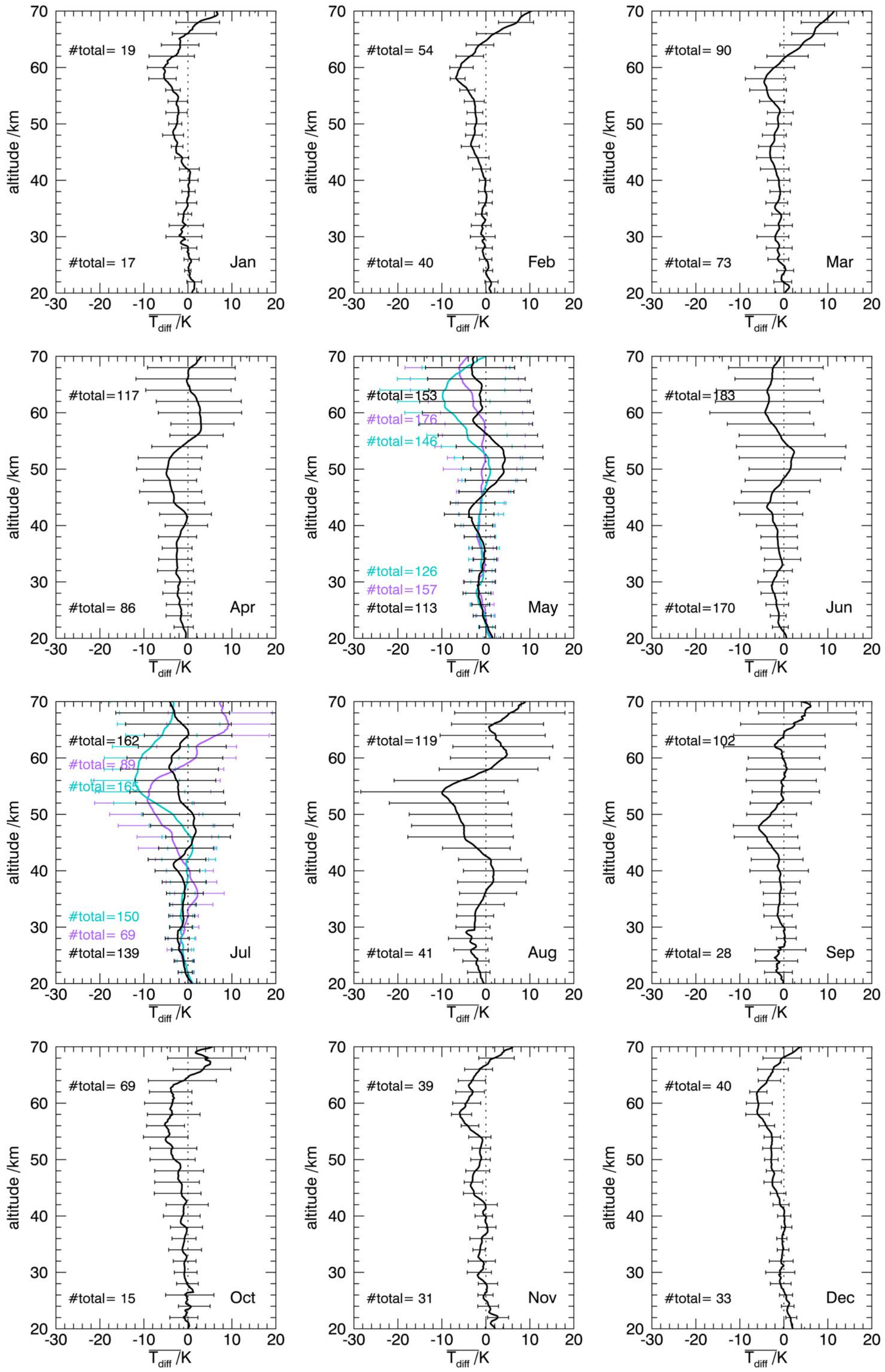

Figure 2. Monthly mean temperature differences (profiles) and 15th/85th percentiles (horizontal bars) between lidar and IFS for 2018 (black), for May and July 2019 (purple), and for May and July 2020 (turquois). The number of profiles at 20 km (50 km) altitude is given at the bottom (top) part of the panels and gives of the amount of profiles that determines the monthly means below and above 30 km altitude (Tab. 1). Negative (positive) values mean that temperatures in the IFS are underestimated (overestimated).

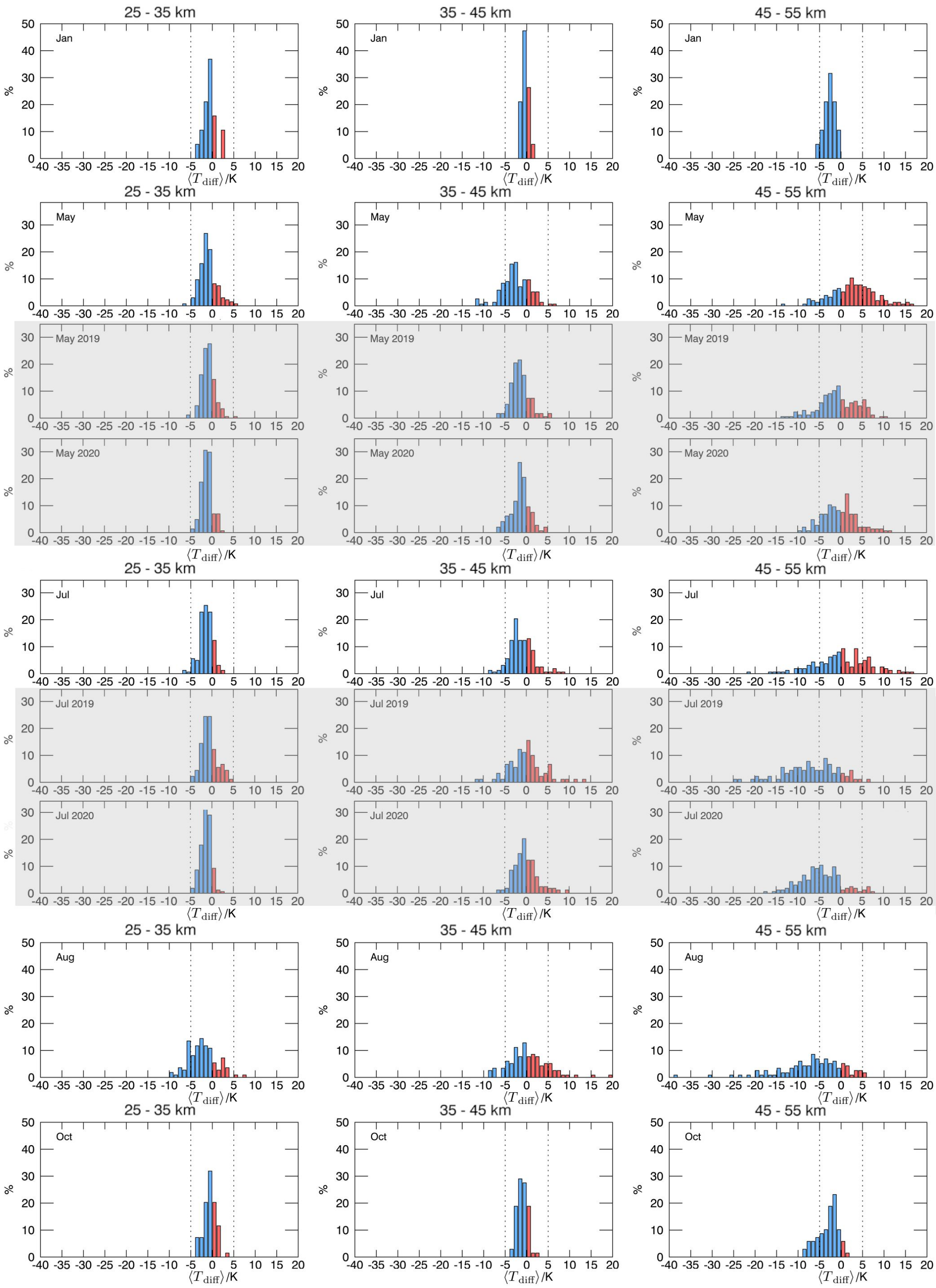

Figure 3. Distribution of temperature differences between lidar and IFS for January, May, July, August, and October 2018 (gray shaded panels: May, July 2019, 2020) averaged for 25-35 km altitude (left), 35-45 km altitude (middle), and 45-55 km altitude (right). Negative (positive) temperature differences are blue (red). Vertical dashed lines mark the ±5 K range.

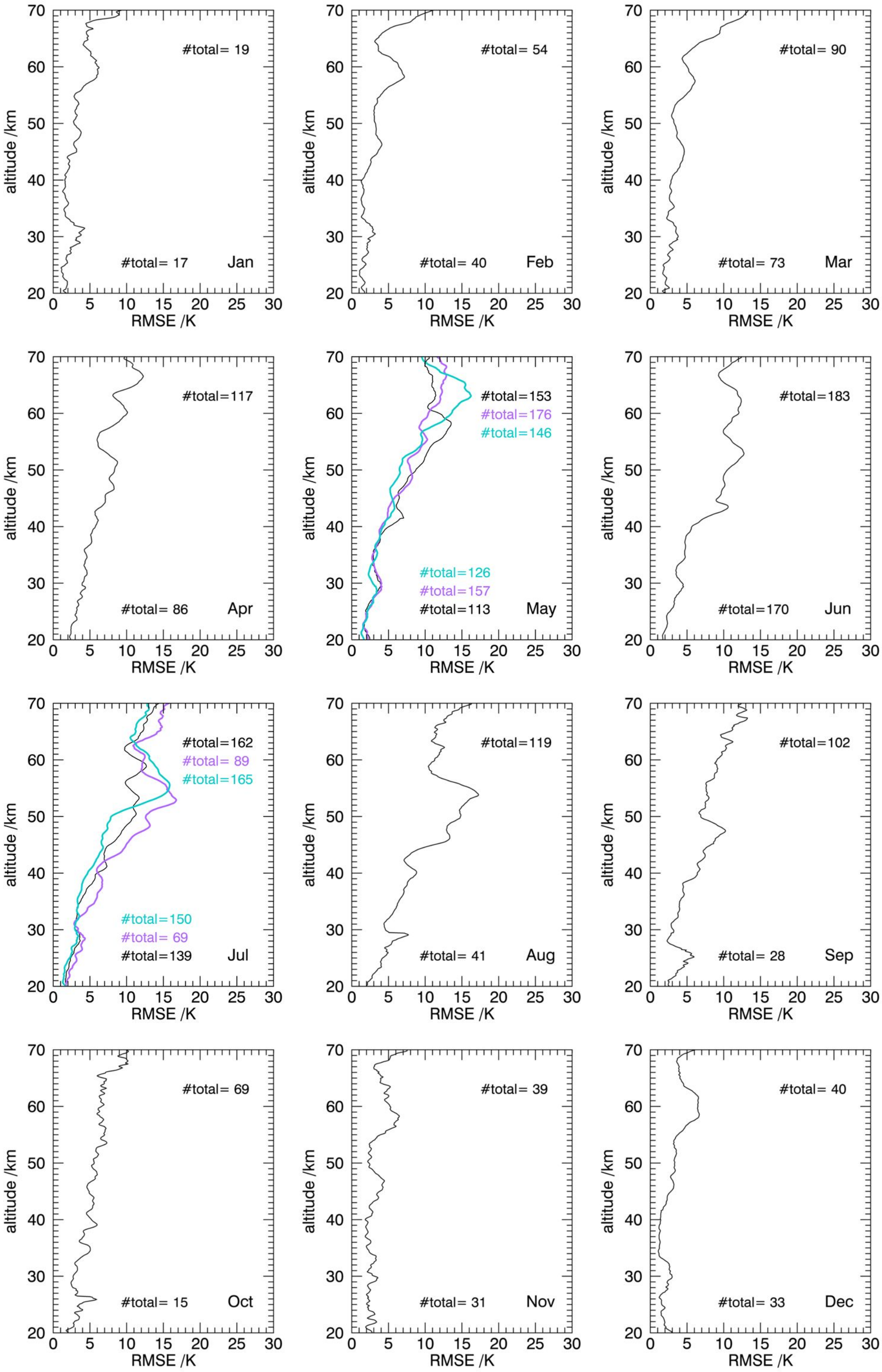

Figure 4. Temperature RMSE for IFS, verified against lidar for 2018 (black), for May and July 2019 (purple), and for May and July 2020 (turquois). The number of profiles at 20 km (50 km) altitude is given at the bottom (top) part of the panels and gives of the amount of profiles that contribute to the RMSE below and above 30 km altitude (Tab. 1)).

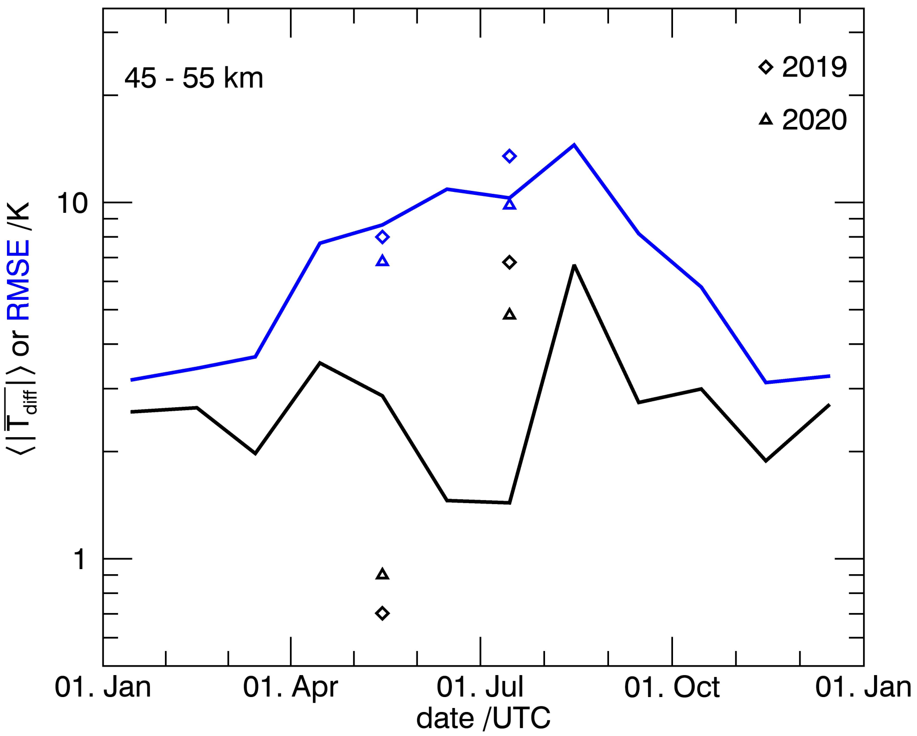

Figure 5. Vertically averaged (45 km < z < 55 km) absolute monthly mean temperature differences (black) between lidar and IFS and the RMSE (blue) for 2018. Diamonds and triangles are for May and July 2019 and 2020, respectively.

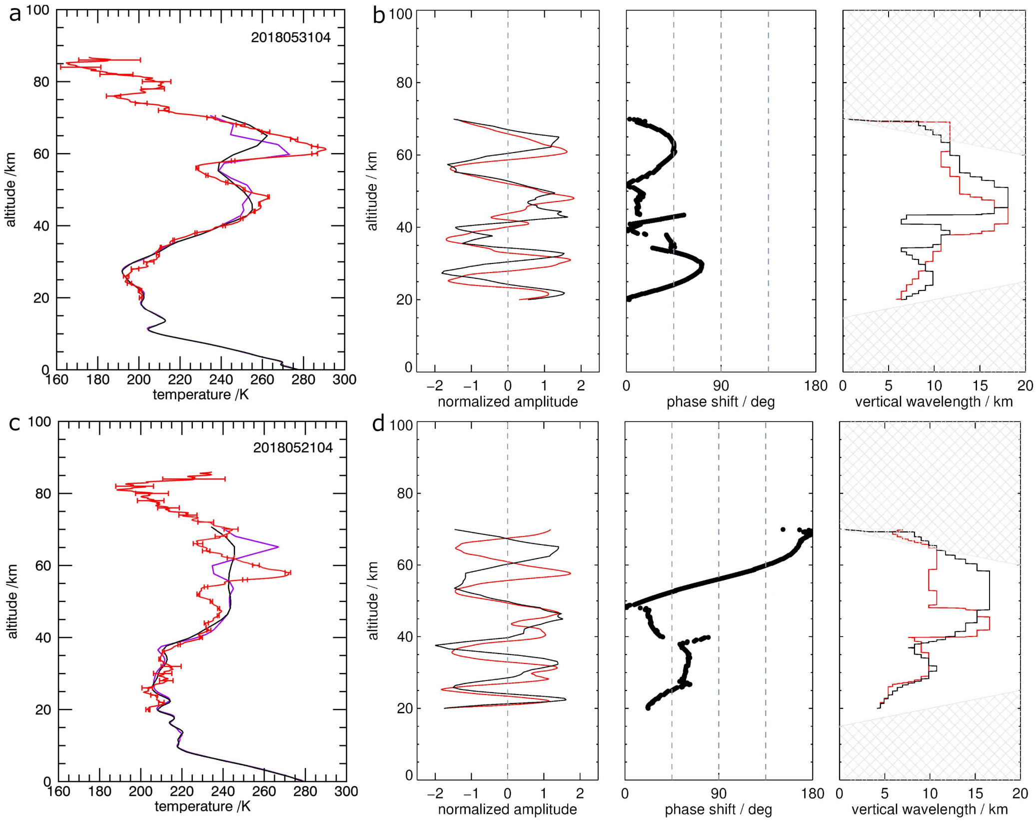

Figure 6. Example profiles for (a) 31 May 2018 04 UTC and (c) 21 May 2018 04 UTC of IFS temperature for the operational forecasts (black) and the experimental forecasts without the sponge (purple) and lidar temperature (red) with horizontal bars marking the uncertainty of the measurements. (b, d) corresponding perturbation profiles (T') as normalized amplitudes and results from wavelet analysis, i.e. phase difference between lidar and IFS (dotted) and vertical wavelengths. Hatched areas mark the cone of influence of the wavelet analysis.

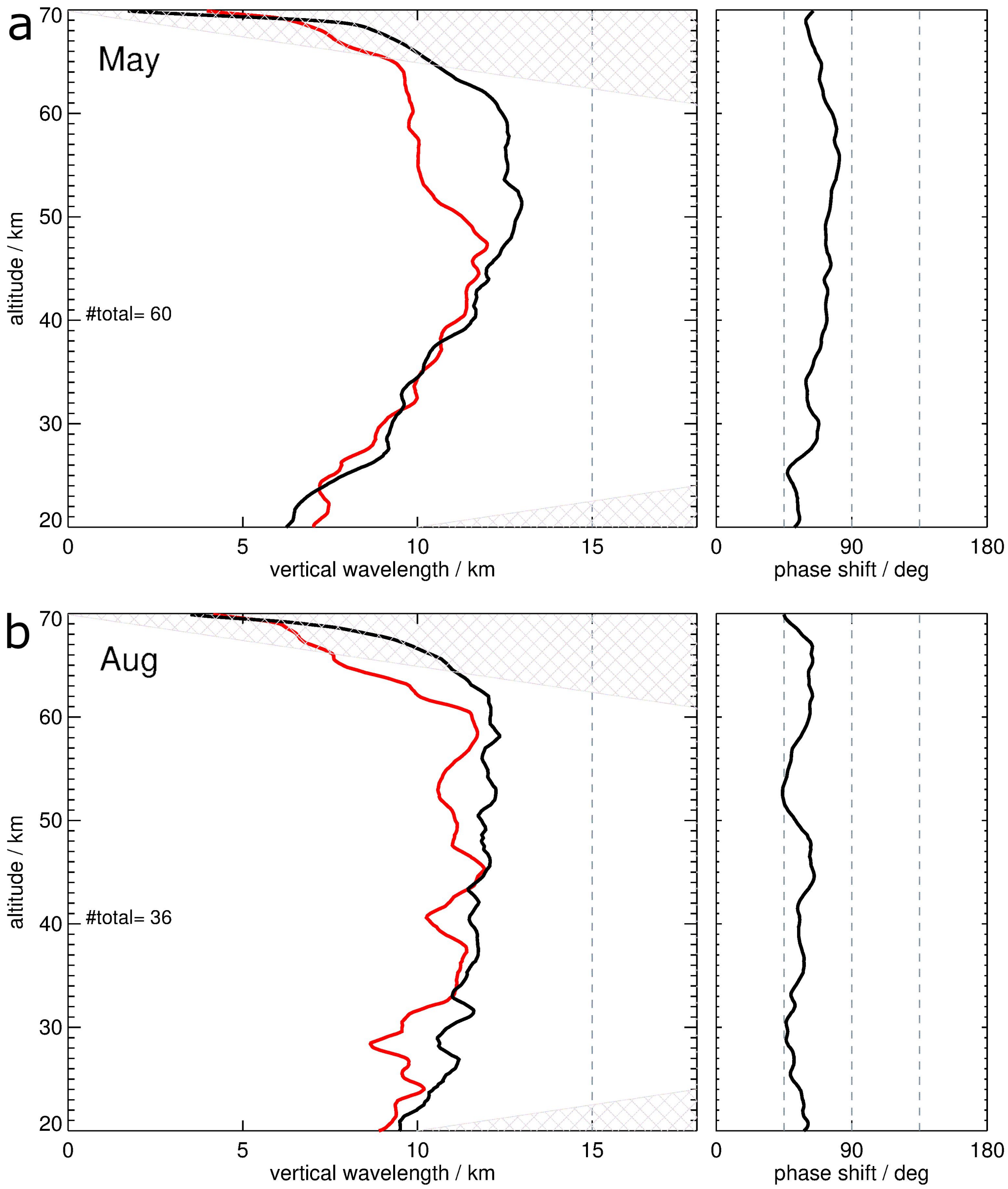

Figure 7. Mean vertical wavelengths (lidar: red, IFS: black) and phase difference for (a) May 2018 and (b) August 2018 determined from wavelet analysis of continuous profiles with mean $T' \geq 3$ K in the middle atmosphere. Hatched areas mark the cone of influence of the wavelet analysis.

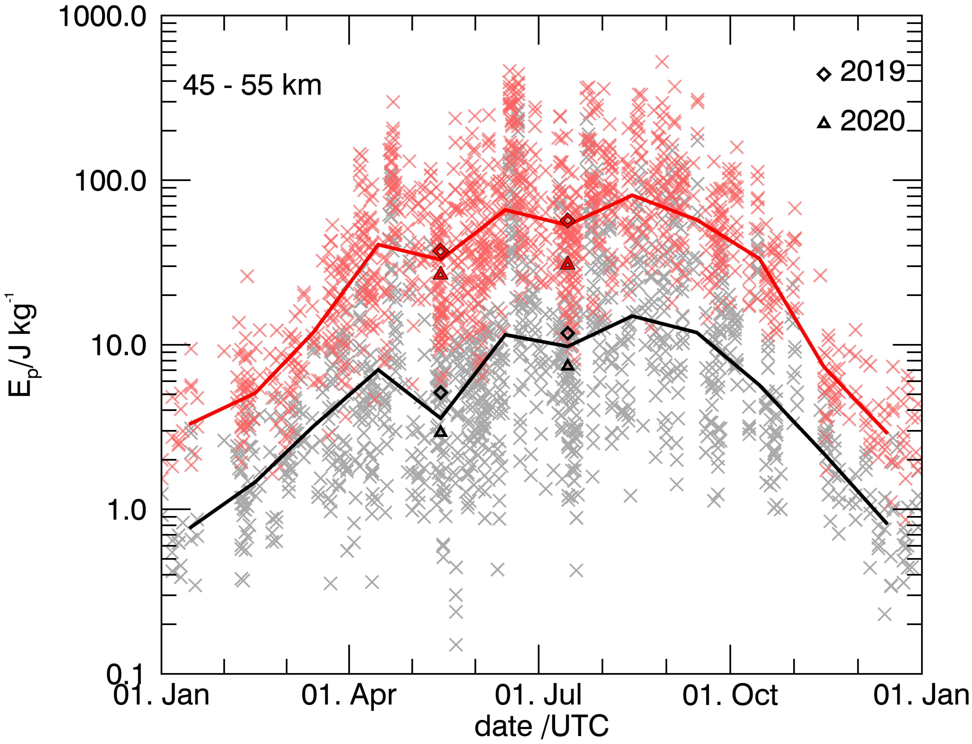

Figure 8. Annual cycle of $\overline{E_p}$ for the IFS (black) and for the lidar measurements (red) in the altitude range of 45 to 55 km for 2018. Diamonds and triangles show $\overline{E_p}$ for May and July 2019 and 2020, respectively. Crosses in the background show E_p of all the individual profiles in 2018 vertically averaged for the same altitude range.

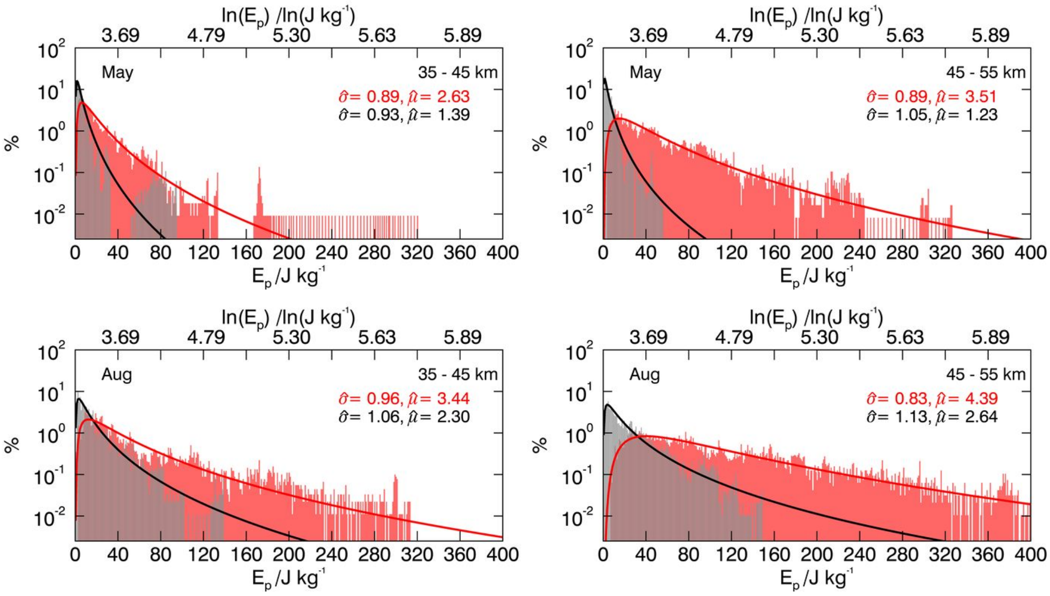

Figure 9. Distribution of E_p for the IFS operational forecasts (gray) and for the lidar measurements (light red) at an altitude range of 35-45 km (left) and 45-55 km (right) for May 2018 (top) and August 2018 (bottom). $\hat{\sigma}$ and $\hat{\mu}$ are the geometric standard deviation and expected value of the data distribution, respectively. Solid black and red lines show the probability density function of the log-normal distribution (Eq. (7)) computed with $\hat{\mu}$ and $\hat{\sigma}$.

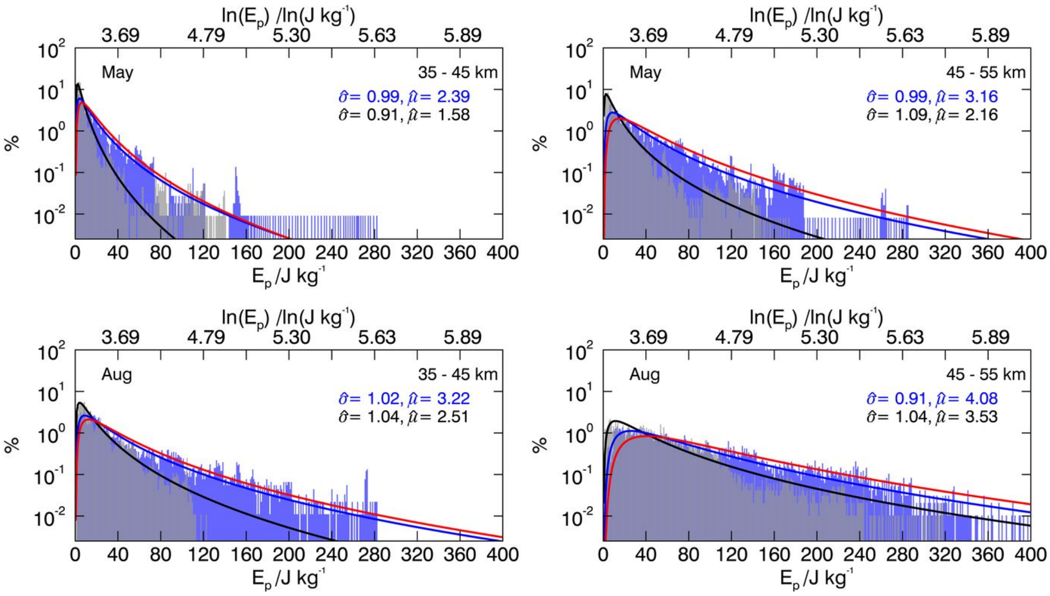

Figure 10. Same as Figure 9 but for the experimental IFS forecasts without the sponge (gray) and lidar data regridded to 137 vertical IFS levels prior to the analysis (light blue). Red line is from the original lidar data for direct comaprison (taken from Fig. 9).

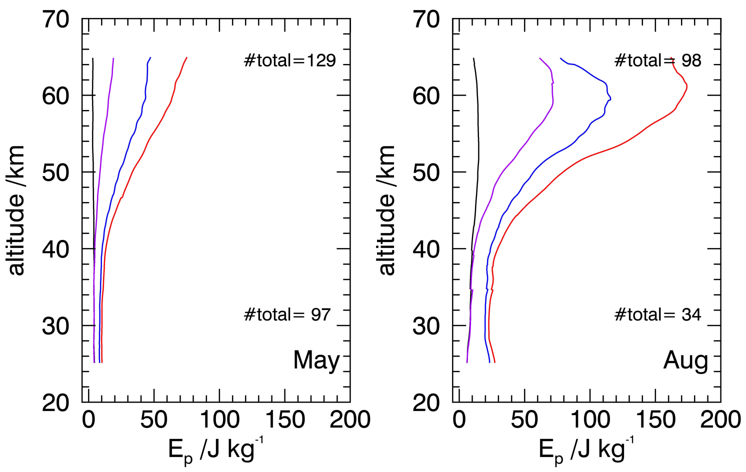

Figure 11. Monthly mean profiles of E_p for the operational forecasts (black), the experimential forecasts without the sponge (purple), the original lidar data (red), and the lidar data regridded to 137 vertical IFS levels prior to the analysis (blue) for May 2018 (left) and August 2018 (right). The number of profiles used for the statistics below (above) 30 km altitude is given at the bottom (top) part of the panels.