We thank the Referees and RJ Sica for their helpful comments (italic). We have incorporated them into the revised manuscript.

Please find our point-by-point answers below.

**Referee 1**

General comment:

*The authors simulate the effects of different GW extraction methods using a monochromatic wave with, of course, distinct properties and discrete resolution in altitude and time. Altitude range and length of the data are set to typical, but fixed values. Depending on simulation either the vertical wavelength or period of the wave are changed. Unavoidable, part of the results depend on the chosen set of parameters and the limited resolution. Unfortunately, in the description of the results often it remains unclear which effect can be generally expected by the particular method, and which is only valid for this special combination of parameters. I suggest making this more clear for every case study. Please compare specific comments below.*

We address these issues at the specific points. Please compare to the following comments.

Specific comments:

*p. 9054, l. 3: The authors should make clear that the overestimation is a result of the selected wave period and running mean width. As, e.g., visible in Fig. 3, other periods result in underestimation or even stronger overestimation of the "true" perturbation. Obviously, depending on period, different fractions of the wave variance are attributed to the background.*

We changed the sentence in question to: "However, the running mean slightly overestimates the extracted temperature perturbations which is due to the choice of a specific period of $\tau = 1.9\,\mathrm{h}$ (cf. Fig. 3e)."

*Figure 2 a, b: I assume that the patterns visible in the figures is mainly due to numerical effects and not a general feature of the filter. Please comment.*

The pattern is due to an aliasing effect and arises because the nightly mean extracts the temperature perturbations completely only at those altitudes where the temperature perturbations average to zero during the entire observational period. Otherwise the residual is attributed to the background temperature. Since this feature depends on the vertical phase speed $c_z = \lambda_z/\tau$ and thus on the vertical wavelength of the added perturbations, slight variations in the spectral response appear at varying altitudes. By extending the timeseries these effects are reduced since the formerly mentioned residual decreases in magnitude.

*p. 9054, l. 6: It should be explained whether the 9 km feature is a general result of the polynomial filter method with this particular parameters.*

This feature arises due to the specific parameter choice. To state this more clearly we added the following sentence to Section 5.2: "However, vertical wavelengths of $\approx 9\,\mathrm{km}$ are still slightly underestimated with this parameter set."

*p. 9055, l. 1–3 (Fig. 3b, e, f): I do not understand the reason for the oscillations at very small periods. While the slow decrease of T' (yellow line) at large periods is reasonable and obvious, it is unclear why short period oscillations should hide in a 3 h averaged background. Please explain!*

We investigated this issue further and found that the oscillatory behavior is due to the coarse time resolution of 0.5 h. Simulations with the same setup as for Figure 3 but with a higher time resolution showed a reduced oscillatory behavior (see Fig. 1 at the end of this document). The remaining oscillations are due to the fact that only perturbations with a period of $\tau = 3\mathrm{h}/n$

($n =$ integer) average out completely in a 3 h runningmean.
To clarify this in the text we added the following sentences to Section 5.1: "The strongly oscillating spectral response of the running mean method for short periods (Fig. 3e) arises due to the coarse temporal resolution of 0.5 h used in the simulations, which is a typical temporal resolution of lidar measurements. If the temporal resolution of the simulations is increased these sharp peaks for periods shorter than 1 h vanish (not shown)."

*Figure 3 a, b: As in Figure 2, I assume that the fast change in response with altitude is a numerical effect and not a general feature of the filter. Please comment.*

You are right, it is the same effect as explained for Figure 2. Please see the corresponding comment above.

*Figure 3 c, d: Please comment on the strong change in response below 30 km and above 80 km.*

This arises due to the fact that in the first (last) few kilometers less information is available for the polynomial fit or the Butterworth filter. Thus, as a consequence perturbations on a larger scale cannot be detected by these methods. This is somewhat comparable to the cone of influence used in wavelet transformations (cf. Torrence and Compo, 1998, A practical guide to wavelet analysis. Bull. Amer. Meteor. Soc., 79, 61−78).
To address this issue in the discussion we added the following sentence to Section 5.2: "One disadvantage of both spatial filtering methods is the dampening of vertical wavelengths larger than 5 km at the upper and lower edge of the analyzed altitude window due to edge effects."

*p. 9055, l. 5−7: I think it is not only a potential phase delay being responsible for the over- and underestimation of the wave. Both vertical filtering methods react similarly with periodic changes in response by up to 20% for long-period waves. Potentially this is because spatial and temporal scales of the wave are related via Eq. 6. The authors should explain their filter characteristics for long periods in more detail, especially since they propose their application for interrupted data sets, or generally speaking for data sets where the continuous time series is shorter than the GW period. It should also be noted that the flat response of the Butterworth filter (as well as the polynomial fit) (Fig. 3e, f) is partly a result of the 10 km vertical averages of T'.*

The periodic changes in the spectral response for large periods are mostly caused by a phase delay. To illustrate this we show the mean absolute temperature perturbations between 25 and 50 km altitude in Figure 2 (end of this document). From these the spectral response is calculated. It can be clearly seen that a phase delay arises above 30 km altitude. Below this altitude the short vertical wavelength is not extracted correctly and thus the large deviations arise. This phase delay also appears at shorter periods. However, it is not visible for the shorter periods since these waves have a larger vertical phase speed ($c_z = \lambda_z/\tau$). As a result they propagate over a larger altitude range during the 8 h of simulation and hence the phase delay averages out in Figure 3 and 5.
We stress the averaging over 10 km which was applied to produce Fig. 3e and f by adding the following sentence to Section 3.1: "This oscillation is not seen in Figure 3e and f due to the vertical averaging over 10 km."

*p. 9055, l. 22: This overestimation by the nightly mean method is a very interesting result. Do you can explain why the overestimation happens at this specific altitude and whether this result can be generalized to other situations?*

This overestimation at this altitude arises purely due to the added time dependent variation of the background temperature which have their maximum amplitude at these altitudes (cf. Eq. 8)

*p. 9058, l. 10−11: The increase of GWPED with altitude is an important topic. It would be interesting to learn why the running mean method results in a stronger increase of GWPED with altitude, compared to other methods?*

At this moment we are not sure why this is the case. We speculate that it might be due to one of the following two reasons:
1) The runningmean method extracts only a small part of the gravity wave spectrum with short periods. It might be that these short period gravity waves propagate with less damping into the middle atmosphere compared to gravity waves with longer periods. Which would be in accordance to the findings of Preusse et al. (2008) [Transparency of the atmosphere to short horizontal wavelength gravity waves, J. Geophys. Res., 113, D24104].
2) Somewhat similar might be the effect observed by Kaifler et al. (2015a) who found that large amplitude quasi-stationary mountain waves are less likely to propagate deep into the mesosphere but will dissipate most of the time in the stratosphere. This could lead to the lower growth rate of the two spatial filters compared to the runningmean method since the latter does not capture mountain waves.
However, the exact cause remains unknown and will be investigated further in the future.

*p. 9059, l. 25: As stated correctly by the authors, tidal signals are hard to retrieve from lidar observations of limited duration. Seemingly, in this case they discern GW and tides by their vertical wavelengths only. Unfortunately this is partly ambiguous because similar signatures might occur also from GW of very large vertical wavelength or a superposition of waves at very different scales.*

The confirmation that the signal shown in Figure 6c is caused by the semidiurnal tide was not inferred due to the large vertical wavelength but from a composite analysis. We changed the sentence in question to state this more clearly: " The broad descending maximum in temperature perturbations is caused by the semidiurnal tide, which was confirmed by a composite analysis over several days (not shown)."

*p. 9063, l. 14: I think this statement is too strong, even if the nightly mean method has indeed its limitations. Nevertheless, a correct application of this method takes the different measurement durations into account.*

We relaxed the sentence in question to: "This makes the nightly mean method an improper choice for compiling gravity wave statistics if a dataset with a varying length of observational periods is analyzed."

*p. 9064, l. 7−8: The results presented here are mainly true for (nighttime) RMR lidar data only - or generally speaking for observations covering a large vertical range but limited time. Resonance lidar soundings are the main tool in the mesopause region, but cover only a limited vertical range even if they are partly not restricted to nighttime. This suggests other methods for GW analysis. I recommend making this clear and not writing about "lidar" in general.*

We changed the sentence in question to: "Based on the results presented here, two methods are recommended for gravity wave extraction from lidar temperature measurements covering a large altitude range:"

Technical corrections:
*p. 9058, l. 18: I would suggest calling this "increase" a "decrease with (increasing) altitude".*
We changed the phrase in question to: "Another striking feature in Fig. 7 is the enhanced

GWPED below 35 km altitude (...)"

*p. 9059, l. 14/15: Rauthe et al., (2006) write about resolved wave periods of 1.5−12 h in winter (1.5−3 h in summer), not lengths of observation.*

We corrected the sentence in question, which now reads: "They resolved gravity waves with periods of 1.5–12 h during winter and 1.5–3.5 h during summer."

*p. 9062, l. 22: "may by" should read "may be"*

Your suggestion was added to the text.

*Figures 3 a and 5 a: Insert "z" behind "Altitude"*

Your suggestion was added to the Figures.

**Referee 2**

*(...) The case study showing real data shows that the sliding polynomial and the Butterworth filter yield similar fluctuations (Figure 6 e and f). However, the 3 h running mean clearly captures waves with different vertical phase speeds, and high- lights the importance of understanding what wave periods and scales are contributing to the resolved wave-field. Analysis of the two-dimensional gravity wave spectrum would help address this issue. Future 2-d Spectral analysis of Rayleigh lidar data sets (without gaps) could provide valuable insights.*

You are right, in principle a 2-d spectral analysis would be the best possible solution. For such an analysis, one would ideally like to have measurements over a time longer than typical gravity wave periods in order to resolve the complete wave spectrum. Unfortunately such long measurements are often interrupted due to the presence of clouds. These measurement gaps provide a huge problem for spectral analysis since most spectral filters apply a Fourier transformation which relies on an uninterrupted dataset. Hence, a large part of lidar measurements cannot be analyzed by such methods.

*The results in Figure 7 would be enhanced by showing errors or uncertainties in the potential energy per unit mass profiles. Are the profiles significantly different above 60 km? The addition of error bars would allow the reader assess these differences.*

We estimated the uncertainties in the GWPED by means of a Monte Carlo simulation. The uncertainties in the GWPED are too small to be clearly visible in Figure 7. Thus, we added the following sentence to section 4.2 to state the uncertainties: "Relative uncertainties of the GWPED for all methods are on the order of 0.5% in the stratosphere and increase to approximately 5% at 80 km altitude which is considerably smaller than the variations of the GWPED due to the geophysical variability."

*The estimate of relative temperature fluctuations is sensitive to two distinct issues; the choice of background profile, and the wave scales included in the fluctuations. The former may yield systematic changes of 10−20%, however the latter may yield changes on order of over 100%. In the presentation of the temperature fluctuations in Figure 6 , it may be useful to add a presentation of the relative temperature fluctuations. Furthermore, subtle filtering effects can yield differences in estimates of gravity wave activity by a factor of 2−3.*

We also plotted relative temperature perturbations instead of the absolute temperature perturbations (see Figure 3 at the end of this document). Only very little difference can be seen between the relative and the absolute temperature perturbations. Thus, we chose to show the absolute temperature perturbations since these are the quantity directly inferred from the different methods.

*The authors note the use of a reflection technique to minimize the discontinuities in the temperature profiles that would yield artifacts in the spectral estimates. Have the authors considered the use of established windowing techniques to address these issues.*

We also applied different windowing techniques. However, these resulted in a strongly reduced spectral response at the upper and lower edge of the analyzed altitude window. We found that this effect is stronger compared to the reduced spectral response by the reflection technique.

*The authors discuss the limited bandwidth of temperature fluctuations due to the fact that the Nyquist frequency is limited by the temporal resolution of the lidar measure- ments. Temperature measurements can be limited by the need to have sufficient sig- nal at the upper altitudes to allow reliable temperature profiles. Use of lidar density measurements can allow higher resolution measurements at lower altitudes. Can the authors cite recent density-based retrievals of gravity wave activity.*

You are right that in principle density measurements allow for higher resolution measurements at lower altitudes. However, temperatures with a higher resolution at lower altitudes can also be retrieved if subsequent integrations with increasing temporal resolution are performed. A general problem of an altitude dependent resolution is that one resolves a larger part of the gravity wave spectrum at lower altitudes compared to higher altitudes which complicates the interpretation of the results. To avoid this effect we decided to use the same temporal and spatial resolution at all altitudes.
Addressing density-based retrievals, we added the following sentence to section 2: "For different methods of extracting gravity waves from density measurements see e.g. Sica and Russell (1999), Thurairajah et al. (2010) and Mze et al. (2014)."

**Referee 3**

*I do believe the statement criticizing the use of the profile mean to extract the gravity wave parameters is too strong. Most studies utilize data collected over many hours of observations. Futhermore, most workers understand that to obtain an accurate sample mean profile requires many hours of observations because the gravity wave perturbations are large and the wave-spectrum correlation time is long. However, I do agree that the filtering method advocated by these authors has the advantage that it can be applied to short datasets.*

We relaxed the statement in the conclusions to: "This makes the nightly mean method an improper choice for compiling gravity wave statistics if a dataset with a varying length of observational periods is analyzed."

**RJ Sica**

*1. FYI there are now more advanced methods to extract temperature from Rayleigh- lidar, ones that in principle do not require the assumption of hydrostatic equilibrium (e.g. Sica, R. J., and A. Haefele (2015), Retrieval of temperature from a multiple-channel Rayleigh-scatter lidar using an optimal estimation method, Appl. Opt., 54(8), 1872–1889, doi:10.1364/AO.54.001872.)*

We added the following to the paragraph in question: "Recently Sica and Haefele (2015) proposed a temperature retrieval using optimal estimation methods."

*2. Careful using temperature for GW analysis! When you assume hydrostatic equilibrium the temperatures are correlated with height (see the lidar averaging kernels using the assumption of hydrostatic equilibrium in the work above to get an idea of the magnitude of this effect). How do you account for these correlations? Consider in the future using the "purer" quantity, density fluctuations, which are uncorrelated.*

As stated in section 2 we use a temperature based analysis since in connection with resonance and/or rotational Raman measurements a larger altitude range can be covered than by pure Rayleigh measurements. Additionally, due to varying signal strength density measurements have the disadvantage that the raw signal has to be normalized to a reference density if temporal filtering is to be applied. This normalization is likely to introduce uncertainties to the analysis as well. Due to these reasons we decided to use temperature measurements instead of density measurements.

*3. You might be interested in several novel techniques we have used to investigate gravity waves from lidar measurements which involve parametric modelling (which in effect builds filters of different orders chosen using the noise of the measurement), e.g. Sica, R., and A. Russell (1999), Measurements of the effects of gravity waves in the middle atmosphere using parametric models of density fluctuations. Part I: Vertical wavenumber and temporal spectra, J Atmos Sci, 56(10), 1308–1329, Sica, R. (1999), Measurements of the effects of gravity waves is the middle atmosphere using parametric models of density fluctuations. Part II: Energy dissipation and eddy diffusion, J Atmos Sci, 56(10), 1330–1343 and Sica, R., and A. Russell (1999), How many waves are in the gravity wave spectrum? Geophys Res Lett, 26(24), 3617–3620.*

We added the following sentence to section 2: "For different methods of extracting gravity waves from density measurements see e.g. Sica and Russell (1999), Thurairajah et al. (2010) and Mze et al. (2014)."

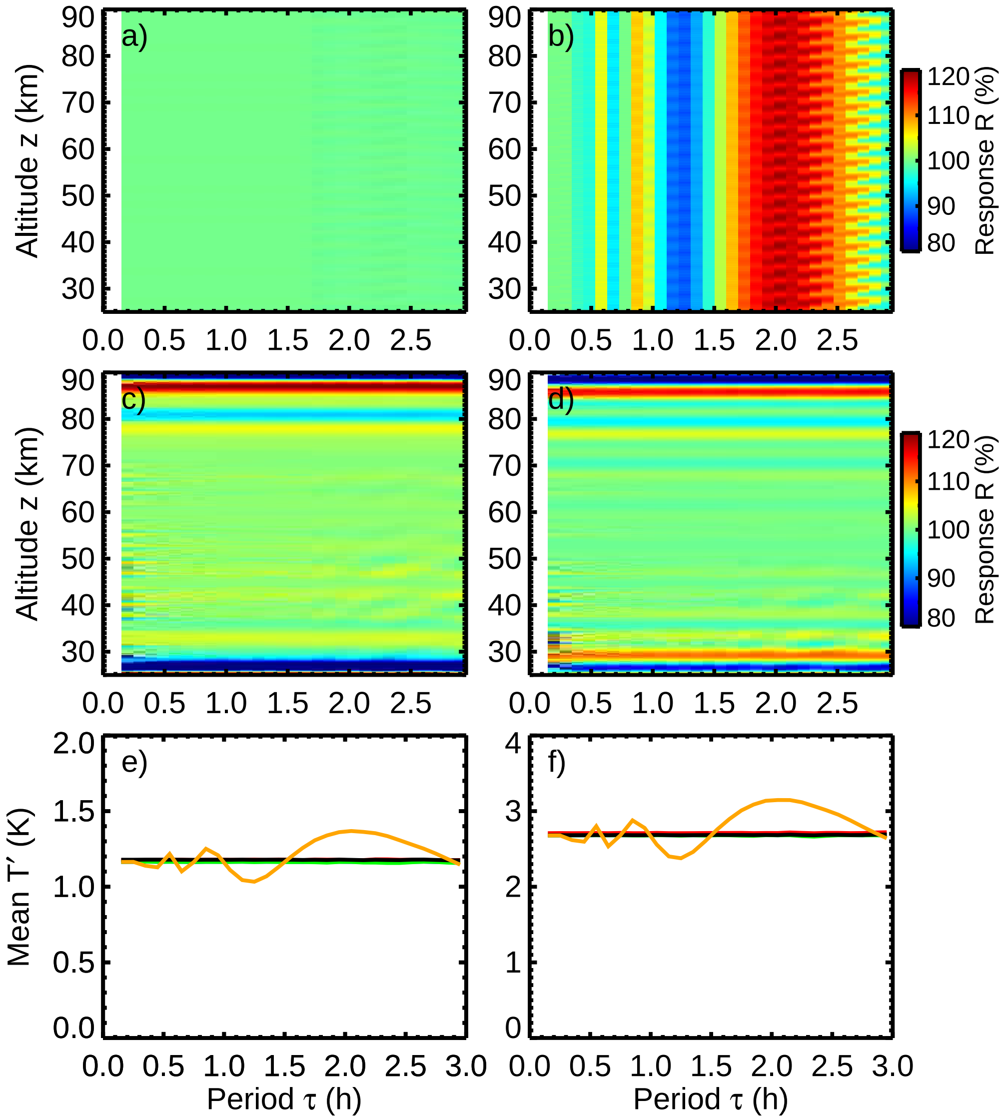


Figure 1: Spectral response as a function of period (cf. Figure 3). The simulations were carried out with a higher temporal resolution of 6 min. The other parameters were kept as described in Section 3.

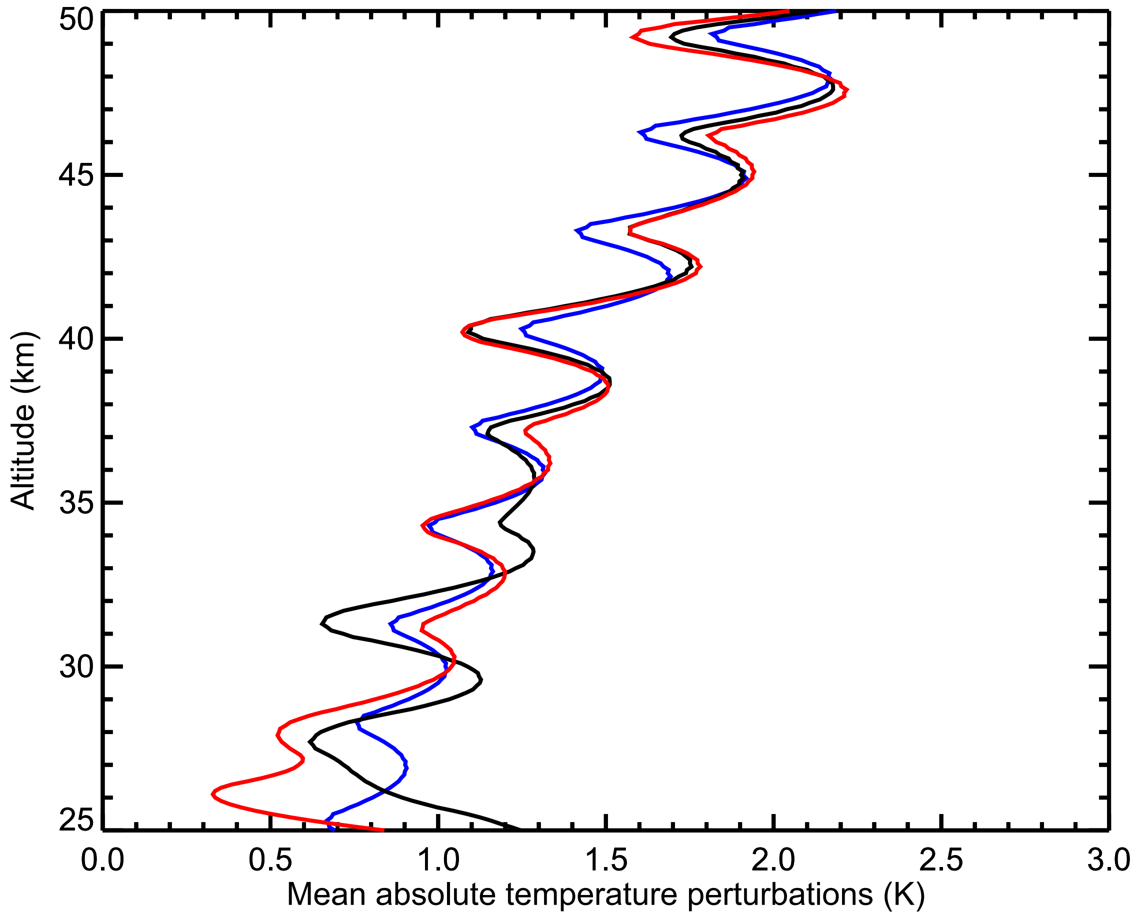


Figure 2: Mean absolute temperature perturbations calculated from Eq. (6) with $\lambda_z$=6 km, $\tau$=14 h. All other parameters were kept the same as described in Section 3. Blue line – original, black line – extracted with the Butterworth filter, red line – extracted with the sliding polynomial fit method.

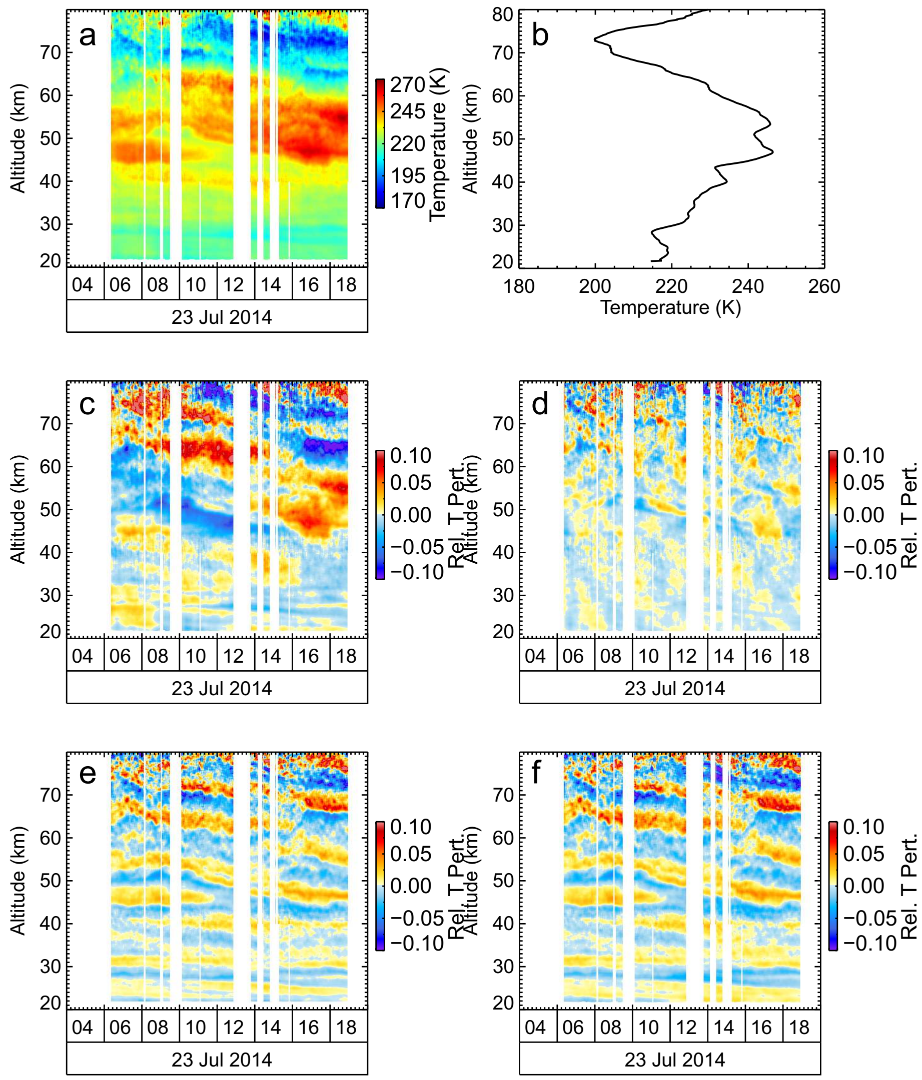


Figure 3: Same as Figure 6 but with relative temperature perturbations in panels c-f instead of absolute temperature perturbations.