Stratospheric gravity-waves over the mountainous island of SouthGeorgia: testing a high-resolution dynamical model with 3-D satellite observations and radiosondes, Hindley et al

This paper presents an extensive comparison of wintertime gravity wave activity and gravity wave properties in mesoscale model simulations and AIRS satellite observations above and around South Georgia. The authors have analyzed numerous aspects of the gravity wave field using state of the art methods and visualization. In the end, the comparison is summarized in 7 key findings addressing among other things the amount and distribution of momentum flux in the model. Overall, the paper presents interesting results that can help gain a better understanding of gravity waves in mesoscale model simulations without gravity wave parametrization. However, methods, results, and discussion are not strictly separated in the paper and, due to the length of the paper, this makes it hard to stay focused during reading. Results are often directly followed by 1-2 sentences of discussion (sometimes speculation). I recommend to address the three major remarks mentioned below to improve the paper (mainly readability) before final publication.

Major remarks

I) Why don't you use a similar 4^{th}-order polynomial fit as for the AIRS data to determine temperature perturbations that contain orographic and non-orographic GWs? Using SG and smoothed nSG to determine perturbations seems nice way to get the contribution of the orographic GWs in the model but doesn't immediately sound like the best choice for a comparison AIRS (unless you can show that 4^{th}-order polynomial leads to similar perturbations as the procedure described here and/or your results are not very sensitive to the background removal). [Moreover, I realized that contributions of non-orographic waves seem to be important in several sub-sections later in the manuscript. SG and nSG could be used to separate and quantify some of the non-orographic contributions in the model simulations (not all the analyses need to be done but for some quantities it could strengthen the findings and conclusions with respect to the non-orographic GWs).]

II) Structure of the paper:

- strictly separate "data and methods" and "results"
  This would also mean help the reader to already know by the end of Sec. 2 what to expect in the result section of the paper.
  - First part of Sec 3 and 3.1 describing the data processing is better moved to Sec 2 (which could then be called Data and Methods)
  - First part of Sec 3.3 Should be moved to Sec 2
  - First part of 3.3.2 should be moved to Sec 2.
  - Sec 4.3: Gini coefficient can be introduced in Sec. 2
  - First part of Sec 4.4. should be moved to Sec 2.
  - Sec. 2.4 already presents results and could be moved to Sec. 3 (or create new Sec 3 with only content of Sec. 2.4)
- separate results and discussion
  - L397: This sentence can be left for later discussion.
  - L403-407: This sentence can be left for later discussion.
  - …and so on

III) Subjective and expletive words like "overwhelmingly" or "very" (>20 occurrences) can be reduced without losing information. There is also a large amount of speculation (some contradictions, some repetitions) in the paper (>25 occurrences of likely and >30 could explanations) that lack quantification.
Just some examples: "…should result in simulated conditions over South Georgia that are very close to reality for the given time periods." "…time separation is very small and the local wind conditions

can be expected to be very close to reality." "A small fraction of this distribution is likely to be measurement error, but the results may still be significant.", "Since these are clear mountain wave structures, it suggests that this could be due to errors in the speed and direction of the background wind in the model.")
I recommend looking through the paper and deciding if such expressions/sentences are essential for the main content/message of the paper and if they can be justified or quantified. If not, they could be removed. Instead of listing every possible explanation for some of the observed differences between model and observations, the explanations could be limited to the one or two most relevant ones.

Minor comments

L9: "high" instead of „very high"; you may want to add "without gravity wave parametrization" over South Georgia.

L23: please specify which scales are meant by short and long

L40: not all but "a large amount of these short vertical and horizontal scales are too small to be resolved even in recent GCMs"; pls add more recent citation (e.g., Plougonven et al 2019, How does knowledge of atmospheric gravity waves guide their parametrizations?)

L43: In some cases? Isn't it rather the norm than the exception?

L84: GWs can propagate large horizontal distances, and from this point of view the Andes are not too far at all. (compare L705: The island lies only 2000 km east of the southern tip of South America, a region associated with the largest stratospheric mountain wave activity observed anywhere in the world)

L88: range of scale sizes? Please clarify.

Fig. 1: Why are the soundings of January 2015 shown here? They are not relevant for the content manuscript. (see also comment on L190)

L137: "…to study gravity waves." Not the whole spectrum of gravity waves is small scale.

L140: Is the fit applied horizontally or vertically?

L176: …much finer than the 3 km vertical grid of the AIRS retrieval: this is kind of a change in the objective of the paper. "when a model is allowed to run at very high spatial resolution over South Georgia, how realistic are the simulated gravity waves compared to observations?" vs how realistic are simulated gravity waves in the observational window of AIRS?
Moreover, can the vertical grid spacing of the model be directly compared to the vertical grid spacing of the retrieval? At least in the horizontal effective resolution is more like 5-10 times the grid spacing.

L183: Does "no gravity wave parametrization" also mean no non-orographic parametrization?

L190: I would expect that there wasn't much mountain wave activity at all in the stratosphere in summer, so I think January 2015 can be omitted.

L194: Can you revise this sentence being more specific and naming the simulated conditions you are interested in, i.e. gravity waves. Then a large part of the wave spectrum can be expected to be close to reality but not the small scales.

L196: This sentence can be omitted.

L213: How can this have an effect at all on the data above 20 km? Is this due to the analysis performed later on?

L235: Was the radiosonde data assimilated in the operational analyses? This should be mentioned here.

L244: Do you mean a wind reversal in the meridional wind? Meridional wind direction is also changing at 30 km on 21st of July and end of July 2013.

L271: Measurement errors and artifacts should be removed from the measurement data before doing the comparison. They are not physically meaningful and are too obvious in the profiles (especially in Fig. 3b, d but also in Fig. 3g above 15 km). Moreover, it would probably help to filter the small scale fluctuations in the sounding data that are well below the vertical resolution of the model data. Fig 3b, e would then look smoother and easier to compare to 3c, d.

L286-290: “slight southward directional bias”, “more northward”: please revise this paragraph. The wording is very circuitous. It’s easier to just say that the model tend to slightly overestimate (underestimate) the southward (northward) winds in the mid-stratosphere. Because the mean difference is zero for the zonal component, this then not only tends in a small directional bias but also in a bias in the horizontal wind speed.

L287: the initial and boundary conditions

L293-L299: In my view, this paragraph is too speculative and can be omitted. Moreover, real time forecast of one to multiple days is different from short-term forecasts of up to 6h used here. Positional errors larger than then the horizontal grid-spacing of the model (everything smaller than that does not really an influence) are hopefully not contributing to the spread because they do not occur (or rarely occur and should then be removed from the sounding data before doing the comparison).

L301: Comparison is concluded and then starts again with discussing the surface winds. They are already included in L281 and local topographic effects are mentioned as possible reason. So L301-L306 can be removed. Detailed discussion of topographic differences between model and reality would include a comparison of the model topography to high-quality elevation data of the island. I don’t think it’s relevant for the rest of the paper.

L227: Can you specify what scales are meant? Vertically it’s clear to me from Sec. 2 (8-9km) but not horizontally (3 times footprint size, e.g. > approx. 80km?).

L330: I cannot follow the reasoning of this sentence. Is this because the model runs without GW parametrization or why/how does the generation of long scale waves depend on the smallest scales?

Figure 4: It is probably better not to show the model data above 58 km where the damping layer is located. With the saturated amplitudes and vertical phaselines, it distorts the visual perception.

L464: applied to the

L470ff: Can you provide some values for a more quantitative comparison? For example, max. amplitude (and later on horizontal and vertical wavelength) at 20 and 40 km above the island and the downstream values you are referring to for both AIRS and model.

L523: Really “measurement error of AIRS” or rather an uncertainty in the analysis and determination of the sign of m?

Sec 4 “results”: Section 3 contains already plenty of results. 4.1 could be just labelled 3.4 and so on

L643: I cannot follow. Isn’t this a conclusion resulting from comparing model to model as AIRS? There is clearly more MF in the model outside the observational window of AIRS.

L979: It would be interesting to repeat the analysis with the output of the UKMO global configuration in the near future. Or was something similar already done in the past? If yes, you could add the reference here.