## Reviewer Comment 3

Thank you for your review. The original text is maintained in black, whereas our responses are marked in red. Any changes to the revised manuscript are also indicated in red text.

The manuscript presents a nice study on the impact of large numbers of real GNSS-RO data on the MetOffice NWP system in the context of the ROMEX project and shows some interesting lessons that can be learned with such a large dataset. It is well readable and mostly very clear, nevertheless it might need explanations in some places to address readers outside of the radio-occultation community or who are not familiar with NWP systems or data assimilation for NWP.

Futhermore the manuscript could be slightly improved by addressing the minor issues and needed clarifications as discussed below.

- Page 2, line 40: CGMS is not a subgroup of WMO, but a "multi-lateral coordination and cooperation across all meteorological satellite operators in close coordination with the user community such as WMO, IOC-UNESCO, and other user entities" (text taken from https://cgms-info.org/about-cgms/)

> Thank you for pointing this out. This paragraph has been updated to:
>
> The Coordination Group for Meteorological Satellites (CGMS) provides recommendations on the number of observations that should be made each day by the various observation platforms. Whilst the CGMS is not able to mandate the number of observations to be made, it does provide guidance to national meteorological centres and world meteorological organisation (WMO) members on the number of observations that should be made.

- Page 4, line 90ff: the description of the NWP system could be slightly more specific. Analyses every 6 hours? Which forward model used for GNSS-RO? There are several implementations, and technical details like refractivity expression etc. are of interest. (It seems clear to me that bending angle is used, but some centres use refractivity.) A reference, ideally also outlining the modeling of observation error would be great.

(Section 2.4 does discuss refractivity in passing.)

> This paragraph has been greatly expanded to include more details of the modelling system. It now reads as follows:
>
> All the experiments documented in this report were run using the low-resolution version of the Met Office's global NWP trial workflow. The resolution of the NWP forecast is described as N320, meaning that it has 640 by 480 grid points, corresponding to a resolution of around 40 km in the mid-latitudes. The model also uses 70 levels in the vertical, stretching from 20 metres above the surface to 80 km altitude. The forecasts are for the atmosphere only, and take a prescribed sea-surface temperature from the OSTIA data assimilation system (Fiedler et al., 2019). The data assimilation system is a hybrid 4D-Var system (Rawlins et al., 2007; Clayton et al., 2013), meaning that a portion of the background-error covariances are derived from the operational ensemble. The system is run in

"uncoupled" mode, meaning that an ensemble forecast is not run as part of the experiments, but the ensemble information is taken from the archive of the operational system. The data assimilation system runs on a 6h cycle, ingesting observations from a wide variety of sources, including GNSS-RO, microwave and infrared radiances, atmospheric motion vectors, conventional observations and many more. The Met Office's forward operator for GNSS-RO bending angles (Burrows et al., 2014; Burrows, 2014) is a one-dimensional operator. As discussed later, the formulation of Smith and Weintraub (1953) is used in the calculation of the atmospheric refractivity. The observation uncertainties are assigned according to Bowler (2020b), which also gives a brief description of the quality control procedures used for GNSS-RO. For satellites not available at that time an appropriate observation uncertainty estimate is used (Metop-C uncertainties are copied from Metop-A, FY-3E from FY-3C and so on). All other GNSS-RO uncertainties are taken from the COSMIC-1 estimates, although commercial GNSS-RO observations are assumed to have a 6μ rad minimum uncertainty, rather than 3μ rad as is used for most satellites.

Did the authors make initial assumptions about data quality from different missions? Some NWP centres choose not to assimilate all data equally, and the authors themselves noted (line 342) that using data from a particular mission showed a degradation of the scores.

Each satellite uses different observation uncertainties, but otherwise they are treated equally.

Line 91: Not really important, but does the lowest model layer have a full-level height of 20m, or is it the actual depth of that layer?

The Met Office's model uses a staggered vertical grid. The heights of the "theta" levels (which hold the mass variables) start at 20 m above ground and smoothly proceed to 80 km above the geoid. The wind and pressure levels start at 10 m above ground and stop at approximately 76 km above the geoid (although there is also a fictious pressure level at approximately 84 km).

- Line 102ff: when presenting forecast scores, always specify against which "truth" the verification is performed. Neither the text nor the caption of figure 1 explain the observation type being used (radiosondes)? The caption also refers to surface observations (SYNOP), but I cannot find related entries in the figure.

The observations used are radiosondes and this has been added to the figure caption. The reference to 2m temperature and 10 wind was an error, and these have been deleted.

- Line 116 and figure 2:

  – does the figure show the mean error for all observations processed, or has a quality control been applied?

This is after quality control, so text has been added to note this.

– there are several spikes (or wiggles) of varying amplitude visible in the plot. What is the origin of these spikes? Interpolation of observations or of model equivalent? E.g. the region between 15 km and 20 km is so noisy that the reader can hardly guess the average bias. The spikyness also varies a lot with latitude band (fig.15). Can this have an effect on the results shown in fig.16?

  These spikes are caused by the interpolation of atmospheric variables from model levels to the observation impact heights. They were reduced by the work of Burrows et al., (2014) but were not eliminated. They may have an effect on the distribution of biases which are seen in Figure 16, but not on the cause of COSMIC-2 being an outlier. That was caused by the monitoring system (from which the plots were produced) failing to account for the drift of the tangent point of each observation within a profile. That issue is now described in the appropriate section.

– does this spiky behavior also occur in the forward operator used in the variational assimilation? If so, does it affect the modeling of error (background / observation)?

  Yes, this is an interpolation issue in the forward operator which is being used and therefore affects the variational assimilation. The observation uncertainties diagnosed from the Desroziers method (Bowler, 2020) contained these oscillations and so explicit smoothing was applied to the uncertainties. We have added the following text to the paper:

  Between 15 and 20 km there is a small-scale oscillations in the statistics. This is due to the interpolation of the atmospheric quantifies from model levels to the impact heights of the observations. These oscillations were reduced by the use of pseudo-levels (Burrows et al., 2014), but were not fully eliminated.

- Page 8, line 134: the "golden region" may be colloquially known to experts in the field, but either a reference or a less colloquial description would help the non-experts.

  We have added a reference to Anthes et al., 2025 which explains the term.

- Line 148: "This shows a large reduction in the forecasts ..."
The reduction is shown here for *bias*. It appears that geopotential bias is globally averaged (without stating so). Is this effect seen similarly in the extratropical hemispheres? Regarding the quantification of the reduction: this is against ECMWF analyses! While it is numerically fine, one could also argue that analysis biases are more consistent after the adjustment. On the positive side, the *global drift* of bias during forecast seems much reduced.

  The caption was missing stating that the verification is for the northern extra-tropics, this has been added. Similar results are seen for the southern extra-tropics and when verifying against sondes. The paragraph has now been updated to read:

As previously noted much of the degradation in the forecast quality was due to a change in the bias of geopotential height forecasts in the troposphere. As well as the original results, Figure 2 shows the forecast bias for 500 hPa geopotential height of the experiments which apply a bias correction to the observations. The experiment which adjusts the observations by 0.05% approximately halves the negative bias in the short-range forecasts of this quantity. The experiment adjusting by 0.1% entirely eliminates the negative bias replacing with a slight positive bias, similar to the control NWP system. With increasing lead time, the forecast tends towards a positive 500 hPa geopotential height bias. The change of bias in the short-range forecast seems to be the main reason that the adjusted experiments perform better than the initial experiment — they are able to remove the large negative bias in the geopotential height forecasts. Figure 2 shows verification against ECMWF analyses in the northern extra-tropics. Similar results are seen in the southern extra-tropics and for verification against sondes.

- Page 15, section 2.5: please specify a "typical" (or reference) lead time of background forecasts.

We have added the following:

Since a 6h data assimilation window is used, the background forecast lead time is between 3 and 9h.

- Line 21: "... fit to the independent observations ..."
What does "independent" refer to here? Does this express that these observations were not assimilated, and the set of observations in fig.12 is different from those in fig.11?

They are observations which are used in the assimilation (both the control and experiments), but are separate from GNSS-RO. However, we realise that the use of “independent” may have been misleading, so we have deleted it.

- Page 25, line 319: "... whereas the ECMWF system has little bias".
Everybody may argue that this is true to a good extent, but some would rather say "... presumably has a smaller bias", or similar.

This remark is based on plots such as this from the ROM SAF monitoring: https://rom-saf.eumetsat.int/monitoring/images/2025/2025-08-01/12_new/global_BA_GRAS-C_EUM_month.png
Hence, we’re attempting to suggest a measurable difference between the two systems. We’ve updated the wording in an attempt to signify this:

The measured bending angle bias at high altitudes is different between the ECMWF and Met Office systems. The Met Office system has a positive bias at high altitudes, whereas the ECMWF system's observed bias is closer to zero.

- Line 320ff: "This may be partly due to the model ..."
I find the description of the treatment of variables in the forward operator very confusing and not helpful. Could issues in the forward operator also explain the mean error patterns seen in fig.2/fig.15, or is there a relation?

The treatment in of variables in the forward operator affects the results that are seen in Figures 2 & 15. However, since we have not explored this in detail we don't wish to show any results. Rather, we have reworded the paragraph to hopefully be clearer:

The measured bending angle bias at high altitudes is different between the ECMWF and Met Office systems. The Met Office system has a positive bias at high altitudes, whereas the ECMWF system's observed bias is closer to zero. One feature of the Met Office system is that the forward operator doesn't ingest the forecast temperature directly. Instead it is provided with the air pressure and specific humidity, and the virtual temperature is derived from these. We conducted a very brief experiment which altered the operator to work directly from the temperature provided by the model indicated that this reduced the bias at high altitudes, bringing the statistics to be much closer to those of ECMWF (not shown). Therefore, applying this bias correction to the observations is unjustified since we are correcting the observations to look more like the model. However, applying a bias correction at high altitudes appears to be effective at improving the forecast quality, and therefore it is included in the experiments.

- Page 26, line 346ff: reference or personal communication?

Unfortunately, we don't have a reference for this – we are told that a paper is in preparation on the upgrade. We have added (Yan Liu, personal communication) to note this.

- Page 26, section 4.2: please specify the background lead time.

This has been added.

- Page 28, line 371ff: the text here and the caption of figs.22-25 are not consistent. The text refers to a change in RMSE, where the reader expects a decrease if the systems improves (similarly to fig.21 for the (O-B) statistics), while the figure captions refer to the "RMSE scorecard", where an increase denotes better results. Can the authors please resolve this?

As with the "overall" figure that is printed at the top of each scorecard, the values plotted in Figures 22 to 25 are positively-oriented (a reduction in the RMSE gives a positive overall figure). We accept that this is not clearly explained in the manuscript, so we have changed the presentation to be negatively-oriented, so that we are plotting the percentage reduction in RMSE. The figures have been changed and we feel this makes a lot more sense within the text.

- Page 30, line 388: "differences in locations of the verifying data points"
It is not the locations alone but data density and spatial sampling that skews the verification against observations where each profile is both used and counted, while verification against analyses is less affected. I recommend to reformulate slightly.

We have reworded the beginning of this paragraph to account for this point. It now reads as follows:

There are many possible reasons which could lead to the differences noted above in the verification against sondes and ECMWF analyses. It seems likely that the

location and density of the verifying data points is an important factor.  Sondes in the northern extra-tropics are concentrated over land, and particularly over Europe…