Response to editor's comments

Thank you for your comments. All line numbers refer to the revised manuscript (not the tracked-changes version).

In addition to the requested changes, we have applied a small number of changes to update and clarify the text. One of the changes is to alter the article's title.

L118: A summary of the changes in the RMSE scores for various variables and forecast lead times for the initial experiments are shown in Figure 1 suggest to clarify what are the initial experiments -> the initial experiments (all ROMEX observations, compared to the control)

> We've updated this sentence to now reads (L117):
>
> "A summary of the changes in the RMSE scores for various variables and forecast lead times for the initial experiments assimilating all ROMEX observations, compared to the control, are shown in Figure 1."

L145: If we modify the observed bending angles in this region to be larger: the previous sentences referred to 15-20km. Therefore, "this region" in this sentence should be clarified, referring to 0-7km.

> Thank you, we have updated this line to read (L145):
>
> "If we modify the observed bending angles between 0 and 7 km impact height to be larger, then the bias in O-B in this region will be reduced."

L417: For verification against sonde observations (Figure 22 Missing the right parenthesis

> Fixed.

Response to reviewer 3

Version 3 of the manuscript, "Experiments with large number of GNSS-RO observations through the ROMEX collaboration in the Met Office NWP system", by Neill Bowler and Owen Lewis, satisfactorily addresses all of my questions and comments on the previous version. I therefore consider it now acceptable for publication.

Thank you for your continued review. All line numbers refer to the revised manuscript (not the tracked-changes version).

However, I would like to suggest a few minor improvements to the text that was added or changed in this revision. I consider these improvements to be purely editorial and do not believe they require further review. They may be approved by the editor.

- Page 1, abstract, lines 6-8: I find the transition from the first paragraph to the second very difficult and confusing. Since the intent of the experiments is to assess the potential impact of the dataset, so this should be described first. The means and attempts to achieve a positive impact should come afterward.

> The current formulation of this part of the abstract was taken in response to comments from another reviewer, concerned that a negative statement near the beginning of the abstract might lead the reader to think that ROMEX had been a failure. However, we agree that starting with the final experiment is confusing. Therefore, to try and balance these two concerns we've revised these sentences to read (L6):
>
> Whilst the final conclusion is that the ROMEX observations have a beneficial impact, the initial tests with the expanded dataset demonstrated a degradation in forecast performance, largely due to forecast biases in geopotential height.

- Page 7, line 143: "atmospheric quantifies" -> "atmospheric quantities"

> Corrected.

- Page 15, line 235: "... fit of other observations ..." -> remove "other"

> Deleted

- Page 22, line 305: "biase" -> "bias"

> Corrected.

- Page 26, lines 359-361:
"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)."

Am I right in assuming the authors wanted to write:

"We conducted a very brief experiment in which we altered the operator to work directly from the temperature provided by the model. This reduced the bias at high altitudes, bringing the statistics much closer to those of ECMWF (not shown)."
(or similar)?

Thank you, we have used the suggested text – this is much better (L358).

Response to reviewer 4

Review 2nd round: “Experiments with large number of GNSS-RO observations through the ROMEX collaboration in the Met Office NWP system” by Bowler and Lewis.

Thank you for considering all the reviewers’ comments and amending the document accordingly. I have the following remaining remarks, which would need to be addressed.

Thank you for your continued review. All line numbers refer to the revised manuscript (not the tracked-changes version).

- Thanks to addressing my comment to “P1, l21/22: This sounds like a hypothesis which has not been analysed in this study - hence, I would avoid stating that.” In my original statement I meant to say that the research if data quality defines the change in impact seen in the two 20,000 samples cannot be concluded with the research solely done in this study. Of course, it could be an important factor which determines the magnitude in impact on forecast scores, but it could also be other factors, like e.g. timing, geographical distribution. In Section 4 (l.403ff) the quality of FY-3 is discussed again, and albeit the improved provided FY-3 data since 2025 to NWP centres, I believe in ROMEX EUMETSAT has done the processing of FY-3 data. Hence, I’m not sure if the quality of that particular dataset is the only reason for the differences in forecast scores seen with the two different 20,000 datasets. Furthermore, as stated in this manuscript Anthes et al. 2025 showed that the differences between observation dataset is small (l.133), which would contradict that the quality between sensors is the leading cause for alternative 20k to perform so differently than the original version tested. Please, rephrase your reworded statements, especially l.25-27 in the abstract: “This alternative dataset gave smaller benefits than the official one, indicating that the quality of the data from each satellite is also important.”

Although the FY-3 observations for ROMEX are labelled as being processed by EUMETSAT, they were actually processed by CMA. EUMETSAT received the data in a format other than BUFR and therefore had to generate the BUFR files. Hence, they were labelled as EUMETSAT despite being processed by CMA.

The paper by Anthes et al. (2025) focused on the performance of COSMIC-2, Spire and Yunyao as these are the most numerous satellites in ROMEX. Only Figures 9 and 10 compare the quality of the FY-3 satellites, and Figure 9 shows that the FY-3 satellites have the largest standard deviations between 20 and 30 km, which is consistent with our evaluations.

When using the term "the quality of the data" we were intending to mean quality in the broadest sense, as opposed to the number of the observations. To make this clearer we reworded to the following:

"This alternative dataset gave smaller benefits than the official one, indicating that differences in the properties of the observation data source are important and not simply the number of observations."

We have also added a brief additional note to the discussion in the text (L387).

- I appreciate discussing the change in standard deviation for forecast scores next to discussing RMSE in section 2. I simply miss a sentence or two why this seems to be a good thing to do.

We cannot work out where the extra sentences would be needed.

- Table 1 Which RMSE values are computed here. Please define.

We've added a note to state that it's the average change in RMSE over the Met Office scorecard (L114).

- Figure 2, I understand that this is a consistent change in bias, which needs to be addressed but it is also rather small: 2.5m bias at 500hPa is about 0.05% reduction for a standard atmosphere. I think it would be fair to mention the magnitude here.

We feel that the existing text is probably sufficient in this case (L125):

"Whilst it is hard to be certain about the true value of the average geopotential height of a given pressure level, the shift in the bias is large enough to be confident that it is detrimental change."

- Again, I would drop one of the scorecards discussed in section 2.2, preferably Fig 5 as it is not discussed in detail – I believe the (summarised) RMSE values in Table 1 give already a good picture of what happens here and could be references in the concerning paragraph.

That seems sensible – we have deleted Figure 5.

- L 217, remove one “may”

Done

- Fig. 15 Caption, isn’t this normalised by B and not average as described

That’s correct, we have updated the caption for figures 15 and 16.

- Tracked Manuscript: P.27, l.409. Maybe good to cite this very recent publication, if it fits: https://amt.copernicus.org/articles/19/659/2026/

Unfortunately, this article addresses the quality of the occultations before the processing change, therefore we don’t feel that it fits.

- P33, l408 (originally): I think it would be unexpected to see an improvement out of the box by almost tripling (or even 5 x) the number of assimilated GNSS-RO observations without any adjustments. This statement also applies to l.474/475. Please amend.

We have deleted the sentence in the conclusions which speaks about expectations – it was not necessary.