Reviewer Comment 4

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.

This article “Experiments with large number of GNSS-RO observations through the ROMEX collaboration in the Met Office NWP system” by Bowler and Lewis discusses the impact by increasing GNSS-RO observations during ROMEX on numerical weather prediction. First, it was stated that forecast scores were degraded caused by stratospheric effects which required adjustments e.g. to the refractivity operator and data processing. After these corrections, assimilating more GNSS-RO data—particularly around 20,000 observations per day—greatly improved forecast accuracy, especially in the southern-hemisphere extra-tropics.

This manuscript highlights nicely the potential but also the challenges coming from assimilating a high number of GNSS-RO data, which wasn’t done before. The authors looked into many different aspect, e.g. bias correction and changes to the forward operator to address the challenges. I recommend publishing these results after major revision.

Main points:

In general, I think the manuscript needs to be slightly trimmed and restructured perhaps. e.g. the explanation of RMSE/ bias is done after using this measure in various score cards and with using only observations as a reference. Also, many sensitivity studies have been done to tackle the bias in geopotential. Rather showing all the different scorecards and only have a little discussion about them, I would focus on the most important ones but keep mentioning the results obtained from the various experiments with e.g. bias correction. I truly believe this makes the manuscript more readable.

> Following the suggestions of other reviewers, we have removed Figure 9, and have added a table summarising the list of experiments which are presented. This table includes references to the section in which the experiment is discussed, which will hopefully allow the user to navigate the paper more easily.

Also, I am a bit reluctant to accept the main conclusion that there is a saturation in impact for the 20.000 daily occultations. As the authors discussed, an alternative flavour of this sample showed lower forecast impact, which could cause different fits as previously. This needs to be shown and discussed referring to a possible impact of data quality. However, here I would be careful not to only make the quality of FY3E responsible for that behaviour without making an additional analysis. Nevertheless, differences in timing of the observation, geographical coverage and quality of the observations are playing a key role in their impact.

> We feel that this study shows that the quality of the observations is important, and one cannot treat all observations as equal. We did not intend to suggest that one satellite (or constellation) is responsible for the apparent saturation of benefit and have reworded that discussion appropriately.

> However, our experiments have shown that we see limited benefit from the additional observations above the control in certain (well-observed) regions. We have avoided using the word "saturation" in either the abstract or conclusions, as this word can carry certain unwanted implications.

Other points

Missing brackets for citing other literature throughout the paper

> This error derives from the history of the manuscript (it was originally written in a format which only allowed the use of the \cite command). This has now been fixed.

P1, abstract: I find it confusing to read first about negative impacts and then later substantial improvement in forecast quality. Please make it clearer what changed to get better forecast with more RO data.

> We have added an additional sentence near the start of the abstract to highlight be improvements before going on to address the challenges. The paragraph in question now begins:
>
> After making various changes to the observation operator used to assimilate the ROMEX observations it was found that the additional observations made substantial improvements to the forecast quality. Without these changes the additional data was seen to degrade the forecast quality, highlighting the importance of understanding the biases within the NWP system. The negative impacts were largely due…

P1, l21/22: This sounds like a hypothesis which has not been analysed in this study - hence, I would avoid stating that.

> This is a reference to the experiments which constructed an alternative version of the 20,000 occultations per day dataset, and demonstrated smaller benefits with this. We have reworded this to be more explicitly linked to our results:
>
> Overall the largest forecast improvements were seen when assimilating 20,000 occultations per day. An alternative dataset was also created with 20,000 occultations per day, but with a different choice of satellites. This alternative dataset gave smaller benefits than the official one, indicating that the quality of the data from each satellite is also important.

P3, l78: typo in observation

> Thank you, corrected.

P4, l91 Would add "horizontal" before "resolution"

> Done.

P 4, Section 2.1: Please state which metric is being looked at. RMSE?

> We've updated this first sentence to read:
>
> 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.

P4, l.105: How big is "large". Please quantify.

We have brought forward the discussion of the forecast biases (promoting Figure 6 to be Figure 2). The discussion of biases is now split between Section 2.1 (where the initial experiments are discussed) and Section 2.2 (which applies a bias correction to the observations).

P5, l120-123: What do you mean with adjusting the observations? Modifying the bending angles which are assimilated or the corresponding impact? parameter?

Yes, we have modified the observed bending angles. We have modified the wording to make this clear:

If we modify the observed bending angles in this region to be larger, then the bias in O-B in this region will be reduced. An experiment was run where the observed bending angles were adjusted by a factor, linearly increasing from 1 (no adjustment) at 7 km impact height, to 1.025 at 0 km impact height (noting that this is below the earth's surface).

P8, l137 – same as in previous comment. What do you increase here?

As before, we have changed “observations” to “observed bending angles”.

P12, Figure 12: What is pert_all? A perturbation of bending angles in all height levels or an average of pert_5km and so on. If the latter, how can pert_all be positive but the individual impact heights are mostly negative?

The perturbation is to the bending angle at all heights. The pressure increment in Figure 7 is indeed positive for pert_all, whereas the pressure increment for the pert_{height} experiments is mostly negative. This is easiest understood from Figure 8, which looks at the change in the increment as a result of the perturbations. This shows that the effect of the perturbation is to increase the pressure in all the tests, with the biggest increase for pert_all. Since the unperturbed increment to the pressure is negative, these increases are being added to a negative baseline.

P12, L178: Which change? The change due to adding ROMEX compared to the control?

We have updated this sentence to be clearer:

The above results indicate that the reduction in the short-range geopotential height forecasts when assimilating the additional GNSS-RO observations is likely due to a systematic reduction in the atmospheric pressure.

P 13, l190: “may be preferred” sounds a bit out of place here. Maybe typo?

This is a bit of loose language. We have improved the sentence to read:

The constants used in this equation are derived from rather old experiments, and more recent formulations of the refractivity may may lead to more accurate simulations of the atmosphere (Healy, 2011; Aparicio and Laroche, 2011).

P15, l 200-204: Here or even better earlier it would be good to discuss if this change in the mean/bias is solely or partly responsible for degradations seen in RMSE (scorecards). What about std deviation? Does this also contribute to an increase in RMSE?

> We have included a forward reference to the changes in the bias, so that the reader understands that these topics will be discussed.

P15, Eq (3): This would be good to define earlier in the manuscript.

> It is hard to see an alternative location where this would fit, so we have left the equation in its current location.

P17, Eq(4): I would rename oi,t as truth or reference, which can be observations but in other instances analysis.

> We have adjusted this term to “verifying reference value”.

P18: This part jumps a bit from verifying against observations and against analysis - hence I'd define standard deviation and rmse with using reference or truth rather than observations.

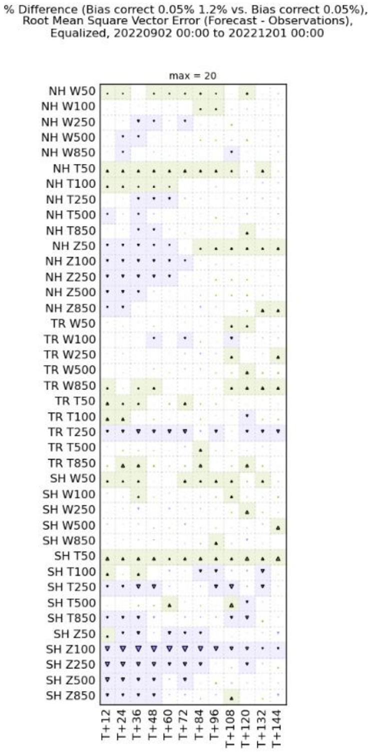
> Yes, the use of observation was an oversight.

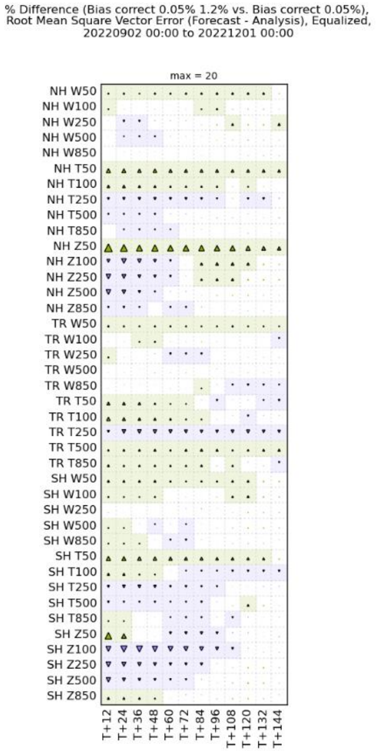
P25, l 314: remove one of the “at”s

> Deleted.

P25, paragraph l 318-324: It would be good to see the impact in forecast scores for that experiment.

> Experience tells us that changes to observations at high altitudes tend to have relatively small impacts on the NWP forecast skill. This is partly because the main NWP forecast quantities are in the troposphere and partly because the DA system gives low weight to high-altitude RO observations. One experiment that we have run which is not included in the paper is to bias correct observations below 34 km impact height by -0.05%, and observations above 47km by +1.2%, with a linear variation between. Compared to the experiment only perturbing observations by -0.05%, this experiment had relatively limited and somewhat mixed impacts (see figures below).





P26; paragraph l.342-349: This suggests that FY3E caused the degradation. Can this be said or are the results only suggestions it could be due to different "quality"? Looking at Fig 20 I'd said the latter. Could Fig 19 be edited including the alternative 20k sample?

Figure 19 has been updated to include the alternative 20k experiment.

It was not our intention to suggest that FY-3E was the principal culprit in the behaviour that we observed. Therefore, we have reworded the paragraph to put this mention closer to the end:

The apparent degradation when assimilating all the observations caused some confusion. However, we noted that the 20,000 occultations per day dataset is not a random sampling of the full dataset, but excludes certain satellite constellations. It excludes observations from the following satellites and satellite constellations: FY-3, Tianmu, Yunyao, KOMPSAT-5 and GeoOptics. Therefore, if observations from these satellites are less beneficial, or even harming the forecast quality then we would expect the 20,000 occultations per day dataset to be the best performing. Separate experiments testing the assimilation of FY-3E into our operational environment showed a degradation in performance (Lewis, 2025). Therefore, we speculate that different observations are of different quality and that the 20,000 occultations per day dataset kept those which are most beneficial to the Met Office's NWP system. However, it should be noted that during January 2025 the Chinese Meteorological Administration introduced an update to their processing of Fengyun observations, which appears to have led to substantial improvements (Yan Liu, personal communication).

P28, l370-371: Is this still valid to say after showing that the alternative sample in the 20k experiment did show a different behaviour? Hence, using the alternative 20k experiment might give you a different answer.

> We have rephrased this sentence to:
>
> Since we are noting that the Met Office system gains little additional benefit from the full set of observations due to the final set of observations being less beneficial in the Met Office system, it is interesting to consider which variables are driving the behaviour.

P33, l408: Please, mention in which forecast score metric this degradation can be seen and why it was unexpected.

> We have rephrased this sentence to be more explicit:
>
> The initial experiments with the additional observations showed an increase in the RMSE of forecast error. This was unexpected as one would expect the additional information provided by the observations to improve forecast quality.

P34, l424: I would rephrase this part of the sentence "standard deviation of forecast error was improved". I'd rather say that compared to the reference the standard deviation decreased, which can be interpreted as an improvement or something similar.

> This has been rephrased to "the standard deviation of forecast error was reduced when assimilating the additional observations".