Response to Reviewer 1

The authors thank Anonymous Reviewer #1 for his or her time in reviewing our manuscript, and for their helpful comments. Addressing them has enabled us to improve this manuscript. Our point-by-point responses are indicated in blue below.

The manuscript "How Does Assimilating a Large Commercial GNSS RO Dataset Impact HAFS Hurricane Forecasts? An Evaluation in Support of the ROMEX Experiment" by William Miller et al. presents a thorough study on the impact of large amounts of real GNSS-RO data on the prediction of four 2022 Atlantic hurricanes. The manuscript is well readable and accessible to readers unfamiliar with hurricane modeling. It is mostly very clear and almost ready for publication.

The manuscript could still be slightly improved by addressing the minor issues discussed below:

- Page 5-6, line 143: "...assimilated in the lower troposphere can positively impact HAFS forecasts, given the tendency for these observations to have larger forward operator errors and/or likelihood of rejection there, compared to RO data from the middle or upper troposphere." While the reader may understand the gist of this statement, it is slightly inaccurate and ambiguous. In the lower troposphere, observations may have larger errors due to the complex path that GNSS signals propagate and the resulting processing to bending angles. On the other hand, forward modeling may use an overly simple operator (e.g. 1d Abel integral instead of ray-tracing), have a large representativity error, and model background error is larger. I recommend writing "...to have larger forward modeling errors..." or something similar to summarize this.

We agree, and we re-phrased this passage to "…given the tendency for these observations to have larger retrieval or forward modeling errors leading to greater likelihood of rejection…" (lines 159-161).

- Page 7, line 181: "a 15-foot peak storm surge". Can the authors improve this so that a reader used to SI units does not have to calculate?

Replaced the phrase with "a 4.6-meter peak storm surge" (line 197).

- Page 8, lines 199ff: The description of the model configuration lacks a specification of the model top and number of model levels. What is the type of nesting? 1-way, or 2-way with feedback to the coarser model? (Presumably the former.)

Our HAFS-A model has 81 vertical levels with a 2-hPa model top. Yes, the nesting uses two-way feedback.

We incorporated this information into Section 2b (lines 217-221):

"HAFS-A v2.0 is dual nested with two-way feedback between a fixed ~ 75° × 75° sized outer domain with 5.4-km horizontal resolution and a TC-following ~ 12° × 12° sized nest with 1.8-km horizontal resolution. The 81-level sigma pressure hybrid vertical coordinate system has a 2-hPa model top, and its lower boundary is coupled to the Modular Ocean Model version 6 (MOM6)."

It appears that the outer domain is not part of the ROMEX experiment in the sense that the additional RO data are not assimilated there. The authors might wish to clarify this already here. Only on page 15, lines 357-359 state that the outer domain stays the same for all experiments, and in the conclusions on page 31, lines 706ff.

Our manuscript mentions earlier, in Section 2a, that only the HAFS inner moving nest assimilates observations (lines 221-226):

"Initial conditions and lateral boundary conditions for the outer domain are provided by NOAA's operational Global Forecasting System (GFS). The Gridpoint Statistical Interpolation (GSI; Wang et al. 2013) software assimilates observations into HAFS-A's moving nest every six hours using a Four-dimensional Ensemble Variational (4DEnVar) algorithm that extracts flow-dependent background error covariances from the 80-member operational GFS ensemble."

But you are right to point out that this point should be re-emphasized because it is important for interpreting the track forecast evaluations, and we added these sentences to Section 2d when describing the HAFS experiment setup (lines 295-298):

"The operational GFS provides the initial and boundary conditions for HAFS-A's outer domain during each DA cycle in all four experiments. Therefore, since HAFS-A only assimilates observations in the moving nest, this study measures assimilated GNSS RO observation impacts on a TC vortex and its near-storm environment within a few hundred kilometers."

- Page 9, lines 236-237: "... background super-refractivity (SR) layer where the vertical refractivity gradient is large." Note that "large" could be specified more clearly as, e.g., above the critical value, if this threshold is chosen.

We edited this passage to mention the value of the critical threshold, and it now reads as (lines 256-258):

"…by the RO ray's tangent point being located within or below a background super-refractivity (SR) layer where the vertical refractivity gradient exceeds 75 percent of the -157 N-units km^{-1} critical threshold."

Using 75 percent of the critical N-gradient value identified in prior research allows for the fact that the model's vertical grid resolution may not be able to fully capture some background SR layers, thus providing additional margin of safety.

Our HAFS-GSI code has an additional "observation-based" SR QC criterion that flags observations within a background super-refractivity layer has a vertical refractivity gradient exceeding a lower threshold of 0.5 × (-157 N-units km^{-1}) if the additional condition of the observed bending angle exceeding 0.03 radians is met. In this case, the entire profile is checked to find its largest bending angle, and all observations below the height of maximum bending angle are discarded. This observation-based SR QC check, developed by Cucurull (2015; also cited in the References), has also been used in the operational GFS's RO data assimilation algorithm. For brevity, we think that it's sufficient to mention only the 0.75 × (-157 N-units km^{-1}) criterion, given that the focus of this paper on forecast impact evaluation rather than DA algorithm optimization and direct the reader to Cucurull (2015) for additional details.

Cucurull, L., 2015: Implementation of quality control for radio occultation observations in the presence of large gradients of atmospheric refractivity. *Atmos. Meas. Tech.,* **8,** 1275-1285.

- Page 4, line 97, and page 9, line 245, 255: The official spelling of "MetOp" is now "Metop", e.g., https://www.eumetsat.int/our-satellites/metop-series

Corrected everywhere in the text (lines 110, 265, 268-270, 283). Thank you for catching this.

- Page 12, lines 312ff: When discussing possibly extreme O-B outliers in the lower troposphere, it is important to note that Spire's processing is known to involve a screening of profiles before sending them out, unlike UCAR's processing of COSMIC-2. Therefore, a fair comparison seems difficult. However, the comparison of rejection rates above 25 hPa could be adjusted to a top at about 1 hPa (~ 45 km), if possible.

Thank you for pointing out this difference between the RO data processing methodologies of Spire and COSMIC-2. Spire's profile screening may not affect our study, since EUMETSAT processed the ROMEX profiles from lower-level Spire occultation data.

We re-plotted Figure 3 (reproduced below), using the same format except for (1) lowering the top of the highest pressure level bin to 15 hPa, (2) moving all three missions onto the same panels to make comparison easier, and (3) adjusting the scheme for coloring the bars according to mission to be consistent with Figures 2 and 4. The 15-hPa level is below the 30-km high-altitude cutoff impact height applied to all commercial RO bending angles in HAFS-A. Non-commercial RO data, including COSMIC-2, can be assimilated up to a 50-km high altitude cutoff in HAFS-A. We verified these settings after checking log files from our HAFS experiments, and we updated the last sentence of Section 2c with this information (lines 274-277). In the manuscript's first submitted version, we had erroneously reported the commercial and government mission high altitude cutoff heights to be 45 and 55 km respectively, which are the settings used in the operational GFS. Limiting the QC rejection rate comparison shown in Figure 3 to the sub-15 hPa layer enables a cleaner comparison of these missions' data characteristics that is not influenced by the tunable high-altitude cutoff parameter applied to data in the stratosphere.

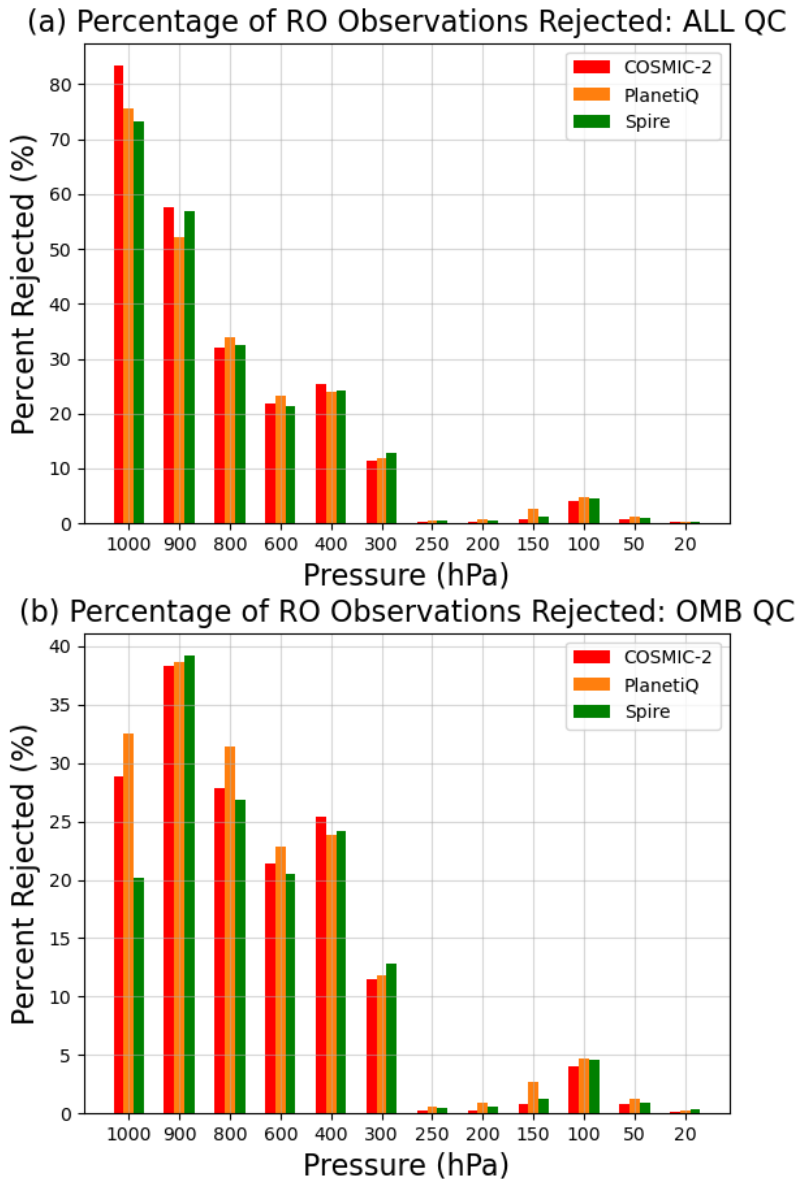

Figure 3. (a) Percentage of RO bending angle observations binned by pressure level that are rejected by HAFS-GSI QC screens. Pressure level height (hPa) bins labeled on the abscissa as 1000, 900, 800, 600, 400, 300, 250, 200, 150, 100, 50, and 20 are bounded as > 950, 950-850, 850-700, 700-450, 450-350, 350-275, 275-225, 225-175, 175-125, 125-75, 75-25 and 25-15 hPa, respectively. Red, orange, and green bars show COSMIC-2, PlanetiQ, and Spire rejection percentages respectively. (b) As in (a), except showing only the percentage of observations rejected by the statistical check for O-B outliers.

- Page 15, Fig.5a: It is very difficult to understand the statistical significance of the results from this plot. Would it be possible to present the results in a different way, perhaps by splitting?

We acknowledge that the lines showing the mean absolute TC position errors in Figure 5a lie nearly atop one another, which results from 1) relatively small impacts of the RO observations

since they are only assimilated in HAFS's inner nest and 2) the large *y*-axis scale needed to capture the TC position error growth over the 126-h forecast period. To more clearly show TC position error differences among the experiments, we compare the Control-minus-EXP difference in the HAFS experiments' mean position errors normalized by the Control mean position error at each verification time in the Figure 5b panel. The relative skill lines shown in Fig. 5b are consistent with the statistical significance times marked in Fig. 5a. For example, at *t*=12 h, when the ROMEX forecasts have statistically significant larger mean position error compared to Control (Fig. 5a), Fig. 5b shows a sharp negative peak in relative skill (indicating degradation), whereas at *t*=30 h, the only time when ROMEX has statistically significant smaller mean position errors than Control, it has its largest positive relative skill (improvement) during the forecast period.

- Page 17, lines 400ff and Fig.8/9(b,d,f): In the mean differences shown here, ERA5 should cancel out, right? Or is it mean absolute differences to ERA5?

Yes, ERA5 cancels out in the bias differences shown as shaded colors in Fig. 8/9(b,d,f). Thus, for example, negative bias differences overlaying positive Control biases (shown as solid black contours) with larger magnitudes than the differences indicate that the experiment's fields are still positively biased relative to the ERA5, but less positively biased than Control. To clarify this, we wrote the equation used to compute the bias differences into the Figure 8 caption, which now reads as "(b) Time-height plot showing the difference between the ROMEX and Control temperature biases [(ROMEX − ERA5) − (Control − ERA5)] in shading (K)..." (lines 538-539).