The authors would like to thank the reviewer for their suggestions, which have made the manuscript clearer and more reproducible. Responses to the individual points are included below. Line numbers refer to the new diffed manuscript.

Major comments:

1) Figure 4: Please overlay in another color or plot symbol, the ionosondes that were included in the assimilation.

The figure has been updated to show ionosondes that were ingested but not used for validation, ingested and used for validation, and used only for validation. Since different time periods are used for this study, different ionosondes are used for ingest and validation for each run. The validation only sites are not ingested in any of the model runs but are not necessarily used to validate every time period since data may not be available. The ingested and used for validation sites were ingested for at least one of the runs and also used to validate at least one run but for any given time period they may not have been used for both ingestion and validation.

Text has been updated in the Figure 4 caption and Lines 175–179 in the diffed PDF.

2) Overfitting/hmF2 Anomalies in NIMO hmF2: While I commend the authors for selecting cases that include such outliers and their honesty in this analysis, its presence raises the concern that the model is overfitting the ionosonde data. This returns as a concern when looking at Figure 11 later, where there are a number of issues in the NIMO hmF2 that are likely attributable to autoscaling errors in the assimilated dataset. If the paper is not seeking to be diagnostic and does not want to understand what is causing the issue, that's fine (this can be a pure performance validation); however, the issue still needs to be acknowledged and it would be valuable to include at least some mention of potential causes. At present the authors have just smoothed over the problem (in the literal sense) as a mitigation without explaining why it should be necessary in the first place. There is virtually no discussion in the manuscript of the issues highlighted so clearly in Figure 11. I'm not suggesting that the authors do considerably more work, but I feel it essential that there at least be some discussion of the issue and whether resolutions have been implemented to mitigate it in the future (beyond the smoothing).

It is out of scope to discuss the work that has been and will be done in later versions of the model, but we added some text to this effect on lines 432–433. Text has also been added to the conclusions at lines 526–528 and 533.

3) Metrics Consistency: I still believe that the authors should be

consistent in their use of metrics throughout and if not, at least give an explanation or justification in the manuscript for why specific metrics were chosen for some parameters and not others. While I don't believe that anything nefarious is going on and it's likely that the other metrics are not of substantive additional value in such cases (and likely would just take up more space presenting for no reason), justification or explanation is necessary to avoid the appearance, even in the very slightest, that the authors have cherry picked complementary metrics. This is not the opinion of the reviewer, but it is something that can be easily mitigated and would avoid misunderstanding. For example, why is MAE a relevant metric for foF2 but not for TEC?

Added a paragraph addressing the variety of metrics on lines 302-308.

4) Figure 10 and line 417: the IRI time series doesn't agree with my expectations. The hmF2 is missing a third peak that one typically would expect in IRI output at that location and season and the diurnal phase is off. Jicamarca should be at 76W, not 76E, as listed in the plot title. Your IRI curve is consistent with what I would expect at 76E (and have been able to reproduce with my copy of IRI-2016) but not at the Jicamarca location as line 417 implies. Also, not relevant here but just in case, just a general quirk of the IRI, when running the Fortran code, it doesn't actually wrap longitudes into negatives (requiring input between 0 and 360 degrees), but also it doesn't have a check in place to error out or warn the user when a negative longitude is supplied and just gives bad output. Maybe it's finally been fixed but can't hurt to be safe.

I had plotted the wrong location, probably in a rush and didn't double check. It turns out I plotted a spot in southern India. The new version of the plot is the correct location. We also verified against the CCMC version of IRI to make sure different mistakes weren't introduced.

This in depth look at IRI also led us to realize another mistake in the manuscript. Although the default hmF2 for the version of IRI we were using was AMTB, it turns out that the flag had actually been set to change it to Shubin. This has been corrected in the text (Lines 127-130) and our IRI output matches the Shubin hmF2 on CCMC.

5) Figure 11: I can't reproduce these IRI curves either with my local compiled copy of IRI-2016 or with the CCMC one, while my copy agrees with the CCMC IRI-2016 output with default selections. This brings me back a bit to your response to my previous comment #6. There is no iri.dat file included with the Fortran. Am I over-interpreting here and in your response to reviewers are you just referring to all of the .dat files that you would need to run the model? If you're using a version of the IRI that was in fact "frozen" from before 2021, that would imply that it also has the read limit problem still, in which

case the model wouldn't be able to run with updated apF107.dat and ig_rz.dat files, at least not the latest ones. In November 2024, the files became larger than allocated variable size and so would result in a read fault. Fixing the issue to be able to use newer index files with an older copy of IRI2016 requires modifying the Fortran in a few places. The other alternative is trimming the back of the file off, but the IRI assumes that its index files start at a specified date, so doing so would cause the entire index database to index incorrectly. Since none of your periods are after 2020, however, you could use one of the older working snapshot files, the last viable one being the November 2023 file. In which case my comments here are moot; however, it would put me at a loss for why your IRI-2016 output in Figure 11 doesn't agree with what I get running my compiled version of the IRI-2016 Fortran or with the CCMC's output with the same version. This is may be a separate issue from that in Figure 10.

We verified things and found the issue was with the starting time for the figure. This has been corrected in the figure caption.

6) Line 440-444: My concern in my previous review regarding the ISRs using the ionosondes for calibration was not about the calibration itself, although I think it is very good to include the explanation of the calibration as you have. My concern was that the ISRs are collocated with ionosondes, which were assimilated, and the parameter that the validation is limited to is that which is actually measured by ionosondes and assimilated. While the assimilated data is autoscaled, it remains challenging to consider the validation with the ISRs as independent or an indication of general performance of the systems under such circumstances (just because two different measurement techniques are being used, they are still measuring the same thing). Depending on the uncertainty that you attribute to the ionosonde data, and thereby how heavily it's weighted in the assimilation, these validations could ultimately just end up as effectively an assessment of autoscaling performance, mind you that would be only in the extreme case.

We find it essential to determine the quality of the model performance near and far from the assimilation locations.

Minor Comments:

1) Lines 112-115: Just to confirm, are you using the SAO files or the .EDP files in the assimilation? Not critically important, but the EDP files include additional processing, where data were inverted with both POLAN and ARTIST's NHPC software and run through QualScan as a quality assurance step. The EDP data would thereby not be explicitly the same as that available from other sources, like GIRO, albeit coming from the same autoscaled virtual height trace. The QualScan performance metrics are provided in the top line of the .EDP files and may be useful to the authors in future studies, if they were using

the .EDP files here.

The EDP files are used, not the SAO files. This has been added to text on Line 112.

2) Lines 184-186: If you know which stations are in or out of the validation and ingestion datasets, why have you not conducted a validation with them in separate groupings, one as a "residuals" assessment and the other as an independent validation?

Not all ingested stations are included in the assimilation. Additionally, the stations chosen change throughout a validation run period. This is discussed on lines 175-179.

3) Lines 246-247: It may be more appropriate to say that it has "already been receiver bias calibrated".

No, as stated in the text on line 109, we assimilate the relative TEC. This variation in the TEC is calculated and removes any contribution from the receiver bias. We added a reminder to the reader on lines 236-239, in case this was overlooked.

4) Lines 247-248: "Distinct" is doing a lot of work here. While the processing is slightly different, the fundamental measurement is the same. To an extent it really depends on whether you see this as a validation of the system as a whole or as a validation to provide diagnostic understanding of the assimilation system. If the former, this is largely fine, albeit requiring caveating, as validation of TEC in that circumstance is largely a validation of the data processing front end and other upstream processing and error weighting rather than an assessment of the entire system.

We disagree that "distinct" is doing a lot of work here, as calibrated, vertical TEC is a completely different beast from the relative TEC that was assimilated. Line 238-239 now highlights that the bias, especially, is a statistic that is not affected by the same receiver-satellite pairs being used in the assimilation and validation.

7) Figure 10: Purely curiosity but has there been any smoothing or other processing applied to NIMO here? I'm surprised that the variation is so clean given the smaller timescale variations we see in your other time series examples. Is NIMO just having an easier time with this period than your other ones?

The new Figure 10 is not as smooth, but no smoothing was applied to the older figure. The smoothness is just a matter of scale and location. For example, plotting the NIMO data in bottom panel of Figure 7 over several days would also look smooth.

8) Lines 416-418: It is worth being careful with the wording here. "Better capture" is not appropriate, as the figure does not include a reference truth from which to make that assessment. There is a more pronounced uplifting in the NIMO output, but without observations, we don't know whether it is "better" capturing uplifting-related impacts. For example, observations suggest that the hmF2 peaked only at 450km during that interval.

Reworded, now on lines 411-412.