Dear Editor and Referee,

We would like to thank the referee for providing a helpful review of our manuscript “First Season MWA EoR Power Spectrum Results at Redshift 7” by myself and others (Submission AAS01756). Below we provide responses to all comments from the referee. The most significant changes are:

- A new figure showing the 2D noise levels to correspond to the 2D power plots we show (Fig. 3).
- A correction to the residual power improvement when using a diffuse model (discussed below).
- Updated references.
- Many expanded explanations on items pointed out by the referee.

All changes are bolded in the revised manuscript. We feel the changes have strengthened the article and made it more accessible to the intended audience.

On behalf of the coauthors,

Adam Beardsley
Arizona State University

==========
Specific responses
==========

Major comments

---The authors emphasize the diagnostic importance of the kperp-kpara plot, and I appreciate the care with which such plots were discussed. However, I found the absence of a kpara-kperp error plot to be rather conspicuous. I would imagine that an error plot would also be enormously informative, and I'd like to see one included, along with a short discussion of its features.

We have added an error plot in Section 3.3 (now Figure 3), along with a paragraph at the end of the section describing the structure.

==========

---11th paragraph of Section 3.1. "At this stage we also average in frequency by a factor of two..." Is there a specific motivation for this? Or is it just a convenient way to reduce the data volume?

Yes, this is done to reduce data volume. We have added a sentence explaining this, and pointing out that it does not affect our

sensitivity. (It does however, effect the kpar modes we access, which is the cause of the different k modes shown for FHD+eppsilon and RTS +CHIPS that you point out below.)

==========

---After Eq. (5): "After fitting for all other parameters, we fit for the reflection mode". Why do this instead of a global fit for all the parameters, which would presumably give a better overall fit? I've noticed that with the form of the g-hat, if A_i,1 and A_i,2 are zero, the reflection mode parameters are degenerate with the "normal" calibration parameters. Is the separate fit a desire to avoid the near-degeneracy that results when A_i,1 and A_i,2 are small?

That's essentially it. We found that if we don't first subtract the polynomial fit, our reflection fits were dominated by the low order terms, and the results did not make physical sense. Even though the degeneracy should be low (the reflection ripple is quite fine compared to other parameters), the relatively huge amount of power in the polynomial terms was completely obscuring the reflection fits. We have added a sentence to this effect in the text.

==========

---Penultimate paragraph of Section 4.2. This reminds me of the Hazelton et al. (2013) paper about multi-baseline wedge effects. It would seem like the current paper is suggesting that the Hazelton et al. effect can be reduced with a fine kernel resolution. However, Hazelton et al. seem to imply that their effect is fundamental and unavoidable. Can the authors explicitly clarify this?

This effect is distinct from other wedge features discussed in the literature, and we have added a paragraph (3rd paragraph of Sec 4.2) pointing this out. The main difference is that this effect is due to the computational limits of the analysis itself, whereas other effects are inherent in the data. In the specific case of Hazelton et al. (2013), the effect is due to multiple baselines migrating in and out of uv cells, and the effective weighting of a given cell is frequency dependent. While the exact mixing would be dependent on the gridding kernel used, it would still exist in the limit of infinitely precise models because the frequency dependent weighting is due to the inherent chromaticity of the instrument.

==========

---2nd paragraph of Section 4.3. "All other sources were given a spectral index of -0.8". What motivated this choice?

This is the median spectral index of sources based on studies below 1.4 GHz. We now mention this and cite the appropriate references.

==========

---Penultimate paragraph of Section 4.3. The authors find that using the two different foreground models gave different flux scales. Is there a sense in which the flux scale has converged?

This is a bit subtle, but the flux scale difference is in the calibrated data, not in the model itself. The difference is quite small (part in 10^3 in mK^2 units), but gets amplified when we show straight difference plots like we use for the diffuse subtraction demonstration. KGS did not find a significant difference in flux scale of the catalog itself, but using a more complete sky model requires lower amplitude (and more accurate) gains to match the visibilities to the model. We have clarified this point in the text.

==========

---Figures 9 and 10. Figure 10 seems to indicate that there is a "+5" pointing. Why is it not included in Figure 9?

This observing campaign did not actually contain data from a +5 pointing, because that is the point that the EoR1 field has risen higher than EoR0 so we switch fields. We have removed the +5 pointing from the cartoon and explain in the caption.

==========

---Figure 11, bottom panel. Since the "+3" portion seems so tightly clustered and is only excluded from the analysis due to some consistent offset, might this be fixable by adjusting calibration parameters?

While it may be possible to recover more pointings in future analysis, it is likely to require both improved calibration and foreground/ instrument modeling. The window power statistic is robust against calibration, and the excess power we see in the N-S polarization at this pointing is likely due to bright large scale structure in our sidelobes (which are different in the two polarizations). We would need to properly model and remove this power in order to use this pointing. The potential to recover pointings is mentioned in Sec. 5.1 (end of 2nd paragraph), and discussed more thoroughly in Sec. 6.1.

==========

---8th paragraph of Section 5.2. I think the vertical lines caused by incomplete uv coverage needs more explanation/discussion. A thinning uv coverage alone should not result in the vertical streaks. It needs to be coupled with increased leakage from within the wedge. The fact that reduced uv coverage causes vertical streaks is therefore also a

result of how the power spectrum is estimated (since that + the instrument is what determines the leakage). As an example, one could go with a "scorched earth" style analysis, sacrificing sensitivity to preserve the EoR window by doing single-baseline delay spectrum analyses for the longer baselines. The foregrounds would be contained within the wedge under such a procedure, independent of uv-coverage (since we're already at the extreme of analyzing one baseline at a time). The form of one's power spectrum estimator must therefore play a role here.

This is a good point, but we feel the explanation is better suited for the pipeline section. In the penultimate paragraph of Section 3.3 we now go into more detail of the streaks seen in Fig 2.

==========

---Along the same lines as the previous point, might an accounting of covariances between different power spectrum cells also help here, perhaps allowing for decreased leakage?

Indeed, and this is under development in the eppsilon pipeline. We now mention this at the same place as the previous point.

==========

---5th from last paragraph of Section 5.2. Why an increase in slope of 14% of the horizon line? I'm curious as to the thinking here, particularly versus the alternative I've seen in the literature where a constant delay buffer is added to the horizon line.

The cut in kpar and the increased wedge slope effectively achieve the same thing as the constant delay buffer. Our wedge slope is actually a bit more aggressive than one would get from the delay buffer because we saw leakage consistently above the “horizon” (previous 3 hour integrations were on the cusp of reaching the noise level to see this).

==========

---Figure 15. Why does the FHD + eppsilon pipeline not go to as high k as RTS + CHIPS? Is there some limitation that prevents the highest k regions from being explored?

This is simply due to the frequency averaging FHD+eppsilon does at the gridding step, which RTS+CHIPS does not do. This is now explained in the caption of Fig. 15.

==========

---It would be helpful in the discussion section to give a sense for

what data and results we can look forward to. As an outsider, for example, I am curious as to whether there is more data "in the can" that is just waiting to be analyzed, or whether any new results will come from data taken once the MWA expansion is past its construction/ commissioning phase.

We have injected a paragraph (2nd to last of the paper) providing this information. There is plenty of data available, but real progress will be made through understanding the systematics discussed in the paper.

==========

Minor comments

---Throughout: the authors seem to use \citealt{} or \citep{} for all citations, but when the citation is part of the sentence, \citet{} to be used (so that publication year but not the author name is enclosed in parentheses).

This has been fixed in the updated draft.

==========

---4th paragraph of introduction: Greig et al. also placed constraints on physical models, I think.

Yes, thank you for pointing that out. This citation has been added.

==========

---Final sentence of 4th paragraph of introduction: The two halves of the sentence don't really have a logical connection to each other. I was confused for a moment because I thought there was a missing part to sentence. In particular, when I read "...and several analysis pipelines are under active development" I thought the authors meant that there were specific efforts to preserve the EoR window, but I think it's just generic pipeline development.

We have rephrased to “… and several analysis pipelines are under active development to exploit this foreground isolation” to emphasize that the pipelines cited are following the EoR window framework.

==========

---Similarly, in the last paragraph of Section 2, when the authors say "(e.g., Jacobs et al. 2016...)", I'm not sure what the "for example" is referring to. I'm not sure how those other works demonstrated that "all techniques demonstrated here will use this golden data set"?

Those citations were misplaced and meant for the previous sentence.

This has been fixed.

==========

---Equation 1: V_ij is never defined as the "true" visibility.

It is now defined in the sentence after the equation.

==========

---11th paragraph of Section 3.1 and first paragraph of Section 3.3: what does "frame" mean? Is this a technical term? I'm not familiar with it.

“Holographic frame” is now introduced as a term referring to the specific weighting scheme that yields optimal maps. We also removed another use of “frame” in Sec 4.4 when referring to weighting.

==========

---Section 3.3: even though it can be inferred from the description in Section 3.2, it would be helpful to readers to mention the foreground subtraction step in Section 3.3 to make it clear where it fits into the power spectrum pipeline.

We have injected a paragraph to Section 3.3 to mention the foreground subtraction. This is now the third paragraph of the section.

==========

---Section 3.3, 3rd paragraph: "The three dimensional power spectrum cube can next be averaged in annuli to form..." For readers who may not be accustomed to P(kperp, kpara) spectra, it would be good to be more specific here. (E.g., annuli oriented in which direction?)

This is now clarified as the annuli of constant (kperp, kpar), (i.e. orthogonal to the kpar axis).

==========

---In all kperp-kpara plots: The units "ns" and "lambda" are given, but it's never stated what quantities (i.e., "delay" and "baseline length", respectively) are plotted.

We have added a description of these axes in the penultimate paragraph of Section 3.3 (and in the caption of Figure 2).

==========

---I don't think "KGS" is ever defined.

It is now spelled out in a footnote when we introduce the catalog.

==========

---Just before Equation 6. "Pixel" is a bit of a confusing term here. Perhaps something like "kperp-kpara cell" might be more appropriate?

This has been changed.

==========

---An inconsistency in citation style. Why is "J. L. B. Line" always given his initials while everyone else is cited by last name only?

Our understanding is that unpublished references should include the initials, while published references do not (http://journals.aas.org/authors/references.html#Unpublished). However, when double checking we noticed a couple unpublished references that were missing the initials, and have now added them.

==========

---Penultimate paragraph of Section 4.4. What does "total residual power" mean? Can we be more precise? Is it the power integrated over the whole kperp-kpara plane? If so, the relevant k scales of integration need to be specified.

First, the 90% number turned out to be not quite right. When we returned to our calculation to answer this question, we realized we had mistakenly included the number from another test we did after “locking in” the analysis for this paper, where we included a Stokes Q model. The correct number should have been 70%, and has been updated in the paper (in all places it is mentioned).

To answer the question at hand, the total residual power is calculated by squaring and summing the residual image cube. This is a naturally weighted image, so the weighting is exactly the inverse of our error figure (now Fig. 3). While we include all scales measured by the instrument, it is dominated by the large perpendicular scales. This is now described in the paper.

==========

---Last paragraph of Section 4.4. I am having trouble understanding the sentence "The top panel shows the model power with the diffuse subtracted from the model power without". Can we rephrase? Also, I'm a little confused by what "model power with the diffuse subtracted" means. Isn't the diffuse power "added" to the model on top of the point sources?

This was worded in a very confusing way. The sentence now reads: “The top panel shows our point source foreground model power spectrum minus the model power spectrum when including diffuse.”

==========

---Figure 8 caption: I don't understand the meaning of the parenthetical remark "(positive difference)".

This sentence had a lot of words to orient the reader between the color scheme corresponding to positive/negative, and which power spectrum was subtracted from which. We have broken it up into two sentences to be more clear.

==========

---Figure 8 caption (but also in the text): is it not a bit optimistic to say that the diffuse modeling "shows promise" just because 90% of power is removed? 90% sounds impressive if we think in linear terms, but on a logarithmic scale (which is what counts), it's just one order of magnitude out of many.

We have removed the subjective language.

==========

---Final paragraph of the paper. It's not quite true that imaging is necessary for cross-correlation. One could also take the images from other surveys and use them to predict visibilities, allowing for a cross-correlation in visibility space.

While this is technically true (and interesting), it does not avoid the challenges inherent in imaging. In order to achieve the visibility correlation, one would need to simulate the measured visibilities with their phase to high accuracy. In the literature we often lump all instrumental effects into the "A" matrix operating on a sky vector resulting in a set of measurements, assuming the calibration will fix all the small instrumental effects and make the real A match the model A. For a cross-correlation one would need to simulate with the real A, including phase corrections at the level needed for imaging. This can be sidestepped in power spectrum measurements because localization is irrelevant, but it is essential for cross-correlation, and the calibration uncertainties leaks back in in the difference between simulating with a model A and the real A. For all intents and purposes, cross-correlating in visibility space would still be an imaging analysis. There is actually a paper in the works (M.F. Morales et al) which will very carefully explore the definition and trade-offs of imaging analysis.