**We apologize to both the editor and the referee for the delay in replying. In the course of replying to the referee’s comments, we discovered several errors in the original analysis and f**

**One of these was a computer bug in calculating the cosmological k corresponding to the original figures 7, 8, and 10 (now XXX, YYY, and ZZZ). This changes the conclusions in Section 4.1 about the extent of the region in k-space free of foregrounds.**

**Section 4 has been extensively re-written to clarify the conclusions that can actually reliably be drawn from the study.**

**A very simple calibration achieves qualitative redundancy, reasonable amplitude accuracy**

**The instrument shows reasonable stability**

**The foregrounds are *isolated* in delay about as well as the simulations would predict (width in k-space)**

**Features of the direction-dependent leakage are captured by the beam model**

1) INTRODUCTION

---------------------------------------

a) Overall, the introduction section is poorly written. There are several missing words and badly phrased sentences. This section should be improved.

**We have reworked much of the introduction as indicated.**

2) LEAKAGE MODES

---------------------------------------

a) In section 2.1, simulations of direction dependent Mueller matrix at two different frequencies are shown in figure 1. The fractional leakage \sqrt(M\_01^2 + M\_02^2), should be shown. This makes it easier for the readers to compare the fractional leakage predicted by simulations with the one observed from the power spectra.

**We have added a plot of this "linear leakage beam" as Figure 2 and an explanation of how to interpret it in Section 2.1.**

b) Section 2.1 mentions that emission in Stokes I is extremely bright compared to Stokes Q, U and V, and that the leakage dominates over this emission. However, Jelic et al. (2015) and Lenc et al. (2016) observed linearly polarized structures of ~ 5-10 K. Since, HERA-19 has maximum baseline of ~54 m, and if the Mueller matrix terms (smaller fractional leakage) shown in figure 1 are to be believed, the diffuse polarized emission should dominate in Stokes Q and U. This should be explained.

**The referee's point is well taken that the relative contribution of polarized and unpolarized power to the pseudo-Stokes Q and U power spectra (Figures 7 and 8) and to the images in Figure 4 should be better quantified. Indeed, it turns out that the fractional polarization observed in the power spectra is consistent with that**

**However, the Lenc and Jelic data were taken far from the Galactic Plane where this data were taken, and it is not**

**The pseudo-Stokes I power spectrum shows that the simulations and data agree on the total about of power in the combined RA 10.5 – 23 dataset within a factor of 2, with the simulation slightly high on the shortest baselines, where it is reasonable to expect that the GSM without any point source contribution is likely to be most accurate. This would correspond to the absolute level of the calibration applied to the real data being wrong by about 40%, averaged over this LST range. It is possible there are other reasons for the discrepancy.**

**It is reasonable to suppose that the contribution of actual polarized emission to the pseudo-Stokes Q and U should be comparable and**

**Uniformly, the simulation fails to account for the observed power level in pseudo-Stokes V. This is very likely due to the lack of inclusion of direction-independent D-terms**

Referring to Figure 10, the measured amplitude of the Q and U power spectra are ~10^13.5 and I is ~10^16, implying a ratio of power spectra of about 2.5 orders of magnitude. The expected I -> Q,U leakage based on the simulations into Q and U is about 70% of the measured value. Removing this, and taking the remainder to be intrinsically polarized, this corresponds to an intrinsic polarized fraction of ~0.03. At 160 MHz (high band), we expect the sky temperature to be ~200 K, so the polarized fluctuations measured on the shortest baseline are order 6 K, indeed consistent with the polarized emission measured by Jelic and Lenc.

We have added this calculation to the text, in the Discussion section.

3) OBSERVATIONS AND DATA REDUCTION

---------------------------------------

a) In Section 3.1, calibration uses a clean model of the Galactic Center (as a 1 Jy point source) for calibration. However, the GC is extended in nature, what is the extent of GC ? Also, the visibilities are scaled with GC flux extrapolated from the GSM at 408 MHz, which has ~5% uncertainty at lower frequency. I think this is a poor choice of calibration strategy for the analysis shown in paper. This should be justified. Aside, previous PAPER analyses have used sources such as Pictor A and Fornax A for flux calibration, and since HERA and PAPER are at same location, why wasn't Pictor A and Fornax A used in the calibration ?

**We have revised the calibration strategy, and expanded our description of it. We note that Fornax A and Pictor A were beneath the horizon for the duration of these observations. An improved model of the GC was used to obtain the calibration, and we have not used the extrapolated flux as given by Equation 11 in the original draft, but rather the GSM at the observing frequencies directly. We note that Equation 11 appears to have simply had an error in it; this is the primary reason the calibration has shifted in the current draft.**

**The 5% uncertainty of the GSM is not markedly worse that the few percent uncertainty generally attached to the flux scale of other observations, and does not significantly affect the conclusions of the paper.**

**We acknowledge the limitations of this calibration strategy and point to ongoing work to improve it.**

b) The calibration step uses gaincal and bandpass functions of CASA package. What are the inputs for these functions? The time and frequency resolution of calibration solutions should be specified. Were the solutions obtained per snapshot or a single solution was applied per 10 minutes?

**These points have been clarified in the text.**

c) Figure 4 shows the multi-frequency synthesis images of the Galactic center. Are these produced using a 10 minute dataset or by combining a full night (after applying appropriate delay corrections) ?

**The images are only of the time interval used in the calibration dataset; this has been clarified in the text.**

d) Figure 5 shows the bandpass solutions. The bandpass for different antennas seem to vary from each other in terms of shape as well as amplitude, with maximum bandpass amplitude (solid green line) being ~5 times greater than the minimum (dashed orange line). This effect is not explained anywhere in the text. What is the origin of this antenna to antenna variation ? How does it affect the results ?

**We thank the referee for bringing up this point. It appears that part of the spread was due to calibration to a point source rather than extended model. The new solutions are much closer to one another in both shape and amplitude (Figure 5), with only one antenna showing significant deviation. The remaining variation is consistent with the manufacturing variability in the analog electronics.**

e) Figure 6 shows the phases of the visibilities post calibration. Although the global features in the phases for different redundant baselines seem to agree with each other, the phases seem to vary on smaller time and frequency scales. What is the reason for these variations ? A ratio or a difference plot between phases of two redundant baselines will highlight any small variations and should be shown.

**Various other studies (e.g. Carilli et al 2018) have shown that the HERA baselines are not in fact terribly redundant, and therefore we do expect variations between calibrated baselines.**

**The effect of small differences in phase is less crucial to this study, since we never combine baselines in phase, but only in power spectra. We do know that the calibration is not introducing significant spectral structure (Figures 5 and XXX)**

f) In section 3.2, Blackman-Harris window function is used to minimize the sidelobes. However, Thyagarajan et al. (2016) showed foreground isolation of >= 12 orders of magnitude using 24 hours of simulated data with HERA-19 array, compared to ~6-7 orders of magnitude isolation observed here. Reason for the choice of this window function should be explained.

**Indeed, the Blackman-Harris window was used to minimize sidelobes, and therefore increases the dynamic range of the Fourier transform to windowing artifacts. The statements made in the original paper’s Section 4.1 comparing this to other studies, rather confusingly conflated foreground *isolation,* or how well the signal stays within the wedge, with *dynamic range*, taken here as the ratio between the maximum and the minimum values in the power spectrum. We have now attempted to clarify the distinction in Section 4.2. Obviously, if a window has very poor dynamic range, it will move power from inside to outside the wedge. However, we point out that the real data are limited in dynamic range not by the window, but by the noise level, and the simulations indeed achieve > 12 orders of magnitude dynamic range with the Blackman Harris window (now standard in the HERA pipeline), and they also show that not much power is shifted from inside to outside the wedge above this floor. We note that the filter used in Thyagarajan et al. (2016) (a Blackman-Harris window squared in frequency domain) in does indeed achieve greater dynamic range, at the usual tradeoff in poorer spectral resolution. As our study is already limited in spectral resolution, but not in dynamic range, we have kept the Blackman-Harris window.**

g) In section 3.2, the delay spectra spaced by 10.7 seconds were cross multiplied to avoid noise bias. Was this done for every 10 minute chunk, or for the full night ? How were the power spectra for each night created : (i) by phasing all the data to a single point, or (ii) by creating a power spectrum for each 10 minute chunk and combining them later in Fourier domain ? This should be explained properly.

The text is already very explicit that this cross multiplication was performed at every 10.7s integration. See Equations 13 and 14. In the final paragraph of Section 3.2, we explain the averaging (over all times for baselines of identical length) used.

4) RESULTS AND DISCUSSION

---------------------------------------

a) The major theme of the paper is to discuss the polarization leakage effect due to HERA-19 dishes, but the total polarized intensity (P = Q + iU or |P| = \sqrt(Q^2 + U^2)) power spectrum is not shown anywhere in the paper. Total polarized intensity power spectrum is a better metric and makes much more sense compared to individual Stokes Q and U power spectra while discussing the polarization leakage effect due to direction-dependent beam. The polarized intensity power spectrum should be shown or at least there should be an explanation for not calculating and presenting it.

**It is not entirely clear if the referee is defining the total polarized intensity P in the image domain, but that appears to be the case. Unless we have misunderstood, then the power spectrum of P requested here will simply be the sum of the Stokes Q and U power spectra individually (PP\* = Q^2 + U^2) and these are the power spectra shown (and which can be easily added from the new Figure XXX.) Since we are not calculating the power spectrum from images, but directly from the visibilities, we will have only the pseudo-Stokes approximations to these of course. We also show the equivalent beam causing such total (linear) polarized intensity leakage in the new Figure 2 to give some intuition for its effects. However, the Q and U power spectra are affected differently both by I -> Q, U leakage (compare their different leakage beams in Figure 1) and by direction-independent calibration effects. Thus it is of some interest to consider them separately. Ultimately, of course, we agree with the referee that it is their sum which affects the result, but do not think it adds substantially to the paper.**

b) In figure 9, the power levels outside horizon (k\_parallel = 0.2 h/Mpc) are shown, and for some cases these are even lower than theoretical noise levels. It is difficult to trust this result, specially given the poor choice of calibration. The second paragraph of section 4 mentions theoretical noise levels for high and low band. However, it doesn't mention how these numbers were estimated except for a reference to Parsons et al. (2012). The noise calculations should be mentioned.

**We have now provided more detail on the noise calculations, including a reference to the underlying software calculating them. The concern than the power levels outside the horizon are lower than theoretical values**

c) The color bars in both figures 7 and 8 are the same for I, Q, U and V power spectra. Stokes Q, U and V power spectra appear to have a dynamic range of ~10,000 (this is clear from bottom rows in figure 7 and 8), but the color scale used is > 8 orders of magnitude in dynamic range. Changing to lower dynamic range might reveal low level features, if any. This should be fixed.

>>> <<<

d) All four Stokes power spectra shown in the bottom rows of figure 7 and 8 exhibit variations of ~2 orders of magnitude in power. These variations are not visible in the 2D power spectra probably due to large dynamic range of the color scale. What is the origin of these variations ?

**Presumably this refers to**

e) In section 4.1, the comment "In similar studies of 2D polarized power spectra, both PAPER and LOFAR measurements found 'filled' regions of Fourier space out to the edge of the EoR window .." doesn't make sense. This is not a valid comparison, as the longest baseline in HERA-19 is 54 m (which is smaller than shortest LOFAR baseline and most of the PAPER baselines) and doesn't resolve the power within/on horizon, which is mentioned in next paragraph.

**The comparison to LOFAR has been removed.**

f) Section 4.3 mentions that simulations using an unpolarized diffused GSM reproduced 75% and 35% of the Stokes I power seen within horizon for high and low band respectively. It's difficult to get these numbers from figure 10, specially because the y-axis is in log scale and cannot reflect small variations. A ratio between simulated power and observed power within horizon should be shown which reflects these numbers. Same applies for Stokes Q and U. Only Stokes V comparison is clear enough to guess the ratio.

>>> <<<

g) Section 4.3 mentions that a fraction of observed power ,< 25%, in pseudo-Stokes Q and U is maybe due to linearly polarized foregrounds. How was this number calculated ?

**This was the excess polarized power above predicted leakage levels: now clarified in the text, and expanded in discussion according to comment 2(b).**

h) Section 4.3 paragraph 2 is confusing. Jelic et al. (2015) and Lenc et al. (2016) observed bright diffuse polarized structures of ~5-10 K which is 3 orders of magnitude higher than the 21cm signal. However, the leakage from P to I is small enough such that its contribution is similar to EoR levels. This was also shown by Asad et al. series on 'Polarization leakage in EoR window' and should be cited here. This paragraph should be modified accordingly.

The purpose of that paragraph (the one beginning with the Lenc et al. (2016) citation) was to place the power spectrum of Galactic synchrotron \_outside\_ the wedge, i.e. the Faraday-rotated component, into perspective with EoR observations. We see now that it is misleading, and have removed it, since it did not contribute to the overall narrative of the section anyway.

Minor Corrections/Comments:

a) Notation used for neutral hydrogen (HI) has different font type/size at several locations in the text. E.g. see first sentence of the abstract and 3rd sentence in the first paragraph of Introduction. This should be fixed.

**We found only one instance of this inconsistency - the one you pointed out - and it has been fixed. If there were other locations in the text this was missed in, we would appreciate your input.**

b) Asad et al. 2018 also talks about wide-field nature of polarization leakage. This paper should be cited.

**Although Asad et al. 2018 is based on the same topic as this paper, their failure to quote noise power or equivalent uncertainty on their measurements make it difficult to draw meaningful comparisons. We find the work too qualitative to merit citing here.**

c) HERA memos are cited at several location in the text. Weblink to these memos should be mentioned.

**Every mention of HERA memos are already weblinks - we have underlined them for additional emphasis. We think this is appropriate as ApJ is now an electronic-only journal.**

d) The observational parameters, such as, number of nights, duration of observation, frequency range, resolution of the recorded data, primary beam size, SEFD of the dishes etc. should be mentioned in a table. This makes it easier for the readers to find the info at one place rather than searching through the text.

**This has been added as Table 1**.

e) In section 2.1, a separate subsection should be created for the RFI excision task.

**This is done.**

f) In equation 7, \tilde{b} should be changed to \vec{b}.

**The TeX has \vec, but for some reason it renders as \tilde! We will inform the editor about this issue.**

g) Figure 9 shows the Stokes I power (I presume) but it is not mentioned in the text or image caption. This should be mentioned.

**This is done.**

h) Title of section 5 should be 'CONCLUSIONS'.

**This change is made.**