Referee Report
Reviewer's Comments:
This paper is a valuable first step towards a cross-correlation of 21 cm emission with other spectral lines during the epoch of reionization. It uses both 21 cm and IR data to perform an initial cross-correlation study, while also using simple models of foregrounds for both observations to explain the effects seen and extrapolate to future observations. The paper is generally quite well written. As it stands, the observed correlation and simulations are interesting, although they are not always rigorous enough to completely justify some of the conclusions of the authors (more on this below). However, I believe the ultimate cross spectrum result (Figure 14 and section 5.4) is highly problematic for two reasons.

(1) The redshifts of the two observations are only slightly overlapping. The radio observations cover 21 cm redshifts of 6 - 7.3, whereas the IR observations cover Lyman alpha redshifts of 5.1 - 6.3. A simple correlation of these two data sets cannot be interpreted as an upper limit on the cross spectrum; for this to be a valid limit, the authors would need to simulate the amount of correlation that is actually present in such a small overlap of redshift space. The authors seem to recognize this and state in section 3.2 that "it overlaps sufficiently for our purpose of characterizing the noise and foregrounds." I agree with this sentiment, since the foregrounds should not dramatically evolve over these wavelengths/frequencies. The data can still serve the purpose of a "Foreground and Sensitivity Analysis" (as stated in the title), but the cross spectrum result needs to be modified/abandoned.

This is an important point regarding interpretation of our ultimate cross spectrum results. We have added clarification throughout the paper that the purpose of the paper is not a formal scientific limit on the 21cm/Lya cross spectrum from the EOR for the reason the reviewer points out. Our purpose is instead to study the foregrounds in 21cm/Lya correlation measurements from the EOR and their implications for experimental design. Along these lines, we demonstrate how quadratic estimator methods can be used in cross spectrum analyses of real radio and IR images. We have corrected the text wherever we referred to an upper limit on on the EOR cross spectrum, describing the result instead as a limit on foregrounds in a preliminary 21cm/Lya correlation measurement.

(2) Independent of the above fact, the current cross spectrum analysis as conducted is also potentially problematic for the purposes of a foreground study. The details of the analysis are not clear enough for me to be sure this is an issue, so if nothing else, the analysis needs to be better explained. Fundamentally, the question is about the nature of the error bars in Figure 14. In the analyses cited for the OQE formalism (e.g. Dillon et al. 2014), the error bars include the contribution from the foreground covariance (which is often estimated empirically, in, e.g., Dillon et al. 2015). Thus the errors are uncertainties on any 21 cm signal - a detection of foreground power well above the thermal noise levels will still result in error bars consistent with zero if the foreground covariance is

well-enough modeled. That may be useful for a signal upper limit, but for a study of foregrounds and sensitivities, those inflated errors hide much of the useful information. A better explanation of what goes into these errors and the associated covariance matrices is required. The authors use the residual 21 cm map power spectrum to estimate the 21 cm covariance matrix (section 5.3), so my suspicion is that the errors do include foreground (since the residual power spectra from Beardsley are still dominated by foreground power). What fraction of the errors in Figure 14 come from the thermal noise in the 21 cm map, and what fraction from the foregrounds? The IR covariance matrix is better explained, but the relative contributions of the different covariance terms should be estimated. The fact that the points in Figure 14 are both positive and negative is indeed indicative of uncorrelated signals, but the exact breakdown of the uncertainties in these measurements are needed to justify many of the conclusions of the paper, e.g., that the 1% geometric flux correlation is below the sensitivity of this study.

We thank the referee for raising this interesting point which gives us the opportunity to clarify our method. Indeed in the standard case of measuring the 21cm auto power spectrum, if one models the foreground covariance very well, then the estimator will accurately select and downweight the foreground-prone regions of 3D k space. This weighting limits the 1D bandpowers to the foreground-free regions, as the referee alludes to, thus depriving them of foreground information.

The analogous question in our 2D cross spectrum case is: are we preferentially downweighting the region of 2D k space where the radio-IR cross spectrum is large? As we discuss in Appendix C, we don't downweight by any model of the radio-IR cross spectrum at all. Indeed even the radio-IR correlation observed in Sec. 4.2 is only at the few percent level, meaning its cross spectrum is 100x smaller than the geometric product of the radio and IR auto spectra. Neglecting these terms gives our simplification of the quadratic estimator of the cross spectrum in Eqns. C5 and C6.

Of course, one could still ask whether even by downweighting the 21cm and IR foregrounds separately, we still end up preferentially downweighting the part of fourier space where the foreground cross spectrum lives. To check whether this is the case, we plot an FFT-based cross spectrum of the same residual MWA and ATLAS images (grey) alongside the optimal quadratic estimator (OQE) cross spectrum (red) in Fig. 14. In fact the two spectra are quite similar. In most bins the OQE 2sigma limits are lower, but are not significantly closer to the fiducial 1% radio/IR foreground correlation. To understand why the FFT results are only slightly worse, consider that after our IR foreground masking, only ~5% of coarse pixels end up masked. With different masking parameters more or fewer coarse pixels ended up masked, in some cases necessitating the OQE to deal with large masked regions. But with the best masking parameters we settled on, the OQE turns out to only give a modest improvement over the FFT.

The major critiques need to be addressed before I can recommend this paper for publication. I also have several other questions/critiques, as well as a number of other minor points that I list at the end of this report.

1) It is stated that FHD outputs "odd" and "even" data cubes to eliminate a noise bias in the power spectrum estimation. Is this necessary for a cross spectrum? Or are the authors throwing away SNR by keeping this separation?

In our study we work with band-averaged data cubes, i.e., 2D images. In this case, the foreground residuals are much larger than the thermal noise, and so averaging the odd and even cubes together to further reduce thermal noise is unnecessary. For simplicity we use the odd cube for auto and cross spectrum calculations, and only use the even cube when we form the difference cube for the purpose of estimating the thermal noise level.

2) The authors project the wide field MWA maps onto an orthographic coordinate grid. Is this lossless? Or are the effects simply unimportant given the smaller 4 degree ATLAS fields?

In fact when we project the MWA maps onto an orthographic grid, we are merely undoing the HEALPix gridding performed by the Fast Holographic Convolution software used by Beardsley+ 16. The natural fourier dual space of radio interferometer visibilities is in fact an orthographic projection of the sky. And over our small fields of view, we are justified in treating this orthographic projection as the true sky, neglecting a complete wide field treatment. This is the same approach used by Dillon+ 13, 14, 15 and Beardsley+16.s

3) Just after equation 12, the authors state the units are "easily seen," but don't actually state the units of I, just that V is in Jy. It's not obvious what units I has.

We state that V(u) has units of Jy and that I_nat(theta) = \sum_j V(u_j)e^(-2pi i u_j*theta) du^2. Observe that I_nat(theta) must have units of Jy [du]^2, and observe further that argument in the exponential must be dimensionless, so [u] = 1/rad. The units of I_nat(theta) are thus Jy/rad^2.

4) The authors use APASS cross matching to better determine the flux scale of their ATLAS observations, but only 20% of sources are found to have a match. Is this low fraction expected? Why are the matches relatively poor?

As we have added to this section, APASS is complete down to ~3mJy, while ATLAS is complete down to 0.5mJy, explaining why so many ATLAS sources don’t appear in APASS.

5) The authors speculate that some of the fall in coherence vs. ell (right panel of Figure 6) is due to the MWA resolution, which they say corresponds to a maximum ell of ~4000. The coherence already seems to have asymptoted at ell ~ 1000, though. Is this consistent?

As we explain in the text, we believe the fall is “...due in part to the MWA's 3' resolution at 185\,MHz, corresponding to a maximum \ell of roughly 4000, and in part to the similar fall in the 4.5\,\mu m catalog power spectrum.”

6) Equations 15 and 16 are not substantiated well enough. I was able to work out equation 15 with moderate algebra, but equation 16 needs to be derived more thoroughly.

We have added an intermediate step, and in fact have simplified our final form of beta slightly.

7) Last paragraph in section 4: why are the luminosity bins for the mock catalogs logarithmically spaced and not linear? It's not surprising this yields a better result, but should be explained.

We use logarithmic bins to best sample the large dynamic range of the luminosity funcitons.

8) The first paragraph in section 5.1 is misleading, particularly the second half. It says 3D 21 cm experiments need long integrations to detect the signal, but that a broadband experiment quickly detects foreground residuals. However, a 3D experiment detects strong foregrounds with similar low integration times. Is the claim simply that the uv plane is filled more quickly using multifrequency synthesis?

Yes, this paragraph was a bit muddled. We have simplified it to read:

> "We begin by quantifying the 185\,MHz foreground residuals in angular power spectrum measurements. In broad band (i.e., multifrequency synthesis) images, thermal noise quickly integrates below foreground residuals over across of fourier space. Reaching the cosmological signal should therefore be thus a matter of foreground mitigation and not the long (e.g., thousand hour) averages needed to measure the 3D power spectrum \citep[e.g.][]{beardsley13,PoberNextGen}. We check this hypothesis, asking how how much observation time is required to achieve the best foreground subtraction."

9) The authors perform a "cross-check" of their analysis by comparing to the k_parallel = 0 bin of the Beardsley result. It looks like this doesn't really agree with any of the "raw" power spectra for the data. Why is this?

The slight ~10-20% discrepancy is perhaps distracting from the main point of the figure that Beardsley's k_parallel=0 bin very closely agrees our band averaged power spectrum. To drive home this point, we have omitted Beardsley's "raw" power spectrum, whose slight discrepancy away from our band averaged raw spectrum could be due to an any number of artifacts due to bright foregrounds. These would need to be investigated in future work.

10) The authors claim ionosphere-related errors cause data from a single night to integrate down slower than data spanning multiple nights. This is one possible cause, but hardly conclusive. It's fine to offer this as a possible explanation, but unwarranted to conclude a change of observation strategy is necessary (which is suggested in section 6) without a more thorough investigation.

We have clarified in Sec. 5.1 that future work would be needed to understand this effect in more detail, and clarified in Sec. 6 that the change of observation strategy is contingent on ionosphere errors being shown to be the culprit.

11) The 5 and 12 sigma cuts in IR source masking are confusing. Why are they applied as separate steps? Why aren't sources just masked with the larger masking radius to begin with?

Of course we could have applied all four foreground masking steps at once, but our intention was to motivate our parameter choices by showing the steps separately in both image space in power spectrum space. Naively, 12 sigma might seem to be overkill, and so we felt it necessary to show that 5 sigma is just not sufficient.

12) The discussion of normalization prior to equation 19 is confusing. The authors discuss using the peak value of the appropriate row of W for normalization instead of the sum to avoid bias. This seems like a fundamentally different normalization, in the sense that only one can be correct. This should be explained.

We feel that this is already a bit too much detail for the paper, but in any case our simulations showed that using the peak is the correct normalization if the true power spectrum is very steep. In the case where we are sample variance noise dominated, the quadratic estimator downweights by the true power spectrum. This results in window functions with asymmetric tails. The tail in the direction that the true power spectrum increases is very small, but the tail in the other direction (direction of decreasing true power spectrum) is very large. Taking the sum of this window function wrongly includes this large tail which is actually irrelevant to the normalization because it the true power spectrum is very low in that region.

13) The authors state that they estimate their ATLAS power spectrum error bars "conservatively" by taking the standard deviation of the bandpowers across the four fields. This is incorrect. If one has a Gaussian distribution and draws samples from it, then measures the standard deviation of those samples, those standard deviations will be drawn from a Chi^2 distribution. For a Gaussian distribution with standard deviation 1.0, drawing 4 samples and measuring their standard deviation has an expectation value of 0.8. So it's a small bias, but it does mean the error bars are being incorrectly estimated, and not conservatively-so.

We have clarified our thinking as: “Instead of predicting the bandpower errors from the input covariances, we conservatively bootstrap the error bars by computing the standard deviation of each bandpower over the four fields.”

14) In the third to last paragraph of section 5.2, the authors state that increasing any of the masking parameters doesn't change the results. This is confusing, since increasing the masking radius from 5 to 12 sigma clearly changes the answer. Do they mean going beyond 12 sigma has no additional effect?

Again, we have broken up the foreground masking process into four steps purely for the reader's benefit. The final foreground masking radius is 12 sigma, and yes, increasing that number does not reduce the foreground power.

15) Section 5.3 needs more detail in several places:

a. What is the redshift range over which the Gong et al. results are valid? Does it match well with the z = 7 range in which the mocks are being simulated?

We use the Gong+ results as our model Ly-alpha power spectrum because they present an error band including errors in the escape fraction of ionizing photons, fraction of radiation emitted at Ly-alpha-, star-forming rate, and IGM clumping factor. They don't give a specific redshift range beyond "z~7" for their simulated power spectrum.

b. The Pober et al. paper is using 21cmFAST from Mesinger et al. for its simulations. That paper should be properly cited.

Done

c. Why are the authors using a z = 8 power spectrum for their pessimistic z = 7 mock? What redshift are they using for the optimistic mock? And in both cases, what is the actual neutral fraction at z = 7 in these models?

Again, the text states "Combining simulations from all these sources allows us to better estimate the modeling uncertainty. Future work is needed to more self-consistently model these fields and their correlation over a range of possible reionization scenarios".

d. The Heneka coherence functions are not discussed in any depth. Where do they come from? How are the mock 21 cm and LyA cubes actually generated to have the given coherence functions? Are the coherence functions expected to be redshift dependent?

The z=7 coherence functions come from the EOR simulations (n-body with radiative transfer) of Heneka+. They should certainly evolve somewhat over redshift, though future work will be needed to model this evolution.

e. The authors use Gaussian statistics in their mock cubes, but they note that reionization is expected to become "somewhat less Gaussian" as it proceeds. Reionization, especially near the tail end, is highly non-Gaussian. The authors should comment on what effect this would have on their simulations.

We have clarified that more work is needed to understand the effect of non-gaussianity in this context.

16) The authors say that the formulae in Appendix C are valid if the correlation between the two images is small. How small is small? Are these equations valid if no foreground removal or masking is performed?

By small by mean that the cross spectrum must be small relative to the geometric product of the auto spectra. Without any foreground removal/masking, as in Fig. 6, the radio/IR

foreground correlation is larger due to geometric effects, so the approximation of zero correlation would be less accurate. It is important to note that this approximation will not result in wrongly stringent limits on the power spectrum, instead it will give limits that are not as stringent as they could be due to not completely optimal weighting.

17) How realistic is it for the Hubble 2' field to be mosaicked into a 4 degree field? How many orbits of Hubble would that require?

OENTUHANSTEOHUSNATOEHUSNTAOEHUSNTAOEHUS

18) Figure 15 and associated text: how significant are the "detections" corresponding to the edges of the orange and gray squares? The SKA+Hubble/DES detections are described as "convincing", but if those are 1 sigma errors, those are anything but convincing.

OENTUHANSTEOHUSNATOEHUSNTAOEHUSNTAOEHUS

19) At the every end of section 5, the authors mention the expected anti-correlation of LyA/21 cm as a clear way to distinguish between geometric foreground correlation and the signal. This is not discussed previously in any detail, and if it's really so clear, why spend so much time calculating the geometric foreground correlation?

We have studied the geometric foreground correlation to understand the impedidents that first generation cross correlation experiments will face. That the expected EOR correlation negative is merely a signature we are looking for, it does not give us a shortcut around the foreground effects.

20) In the third to last paragraph of the conclusions, the authors say they "simulate the signal loss"; this is technically true, but signal loss is a loaded term in cosmology. The authors, through simulated mock cubes, calculate the amplitude of the 21 cm and LyA power spectra compared with 3D cubes (for Gaussian fields). There are many other sources of signal loss left unexplored.

Our exact phrasing is “we simulate the signal loss in 2D versus 3D EOR intensity mapping experiments”, which specifies the effect we referring to.

21) The final concluding paragraph suggests that the predicted EOR anticorrelation is "close to being within reach." Figure 15, frankly, suggests otherwise.

OENTUHANSTEOHUSNATOEHUSNTAOEHUSNTAOEHUS

Minor points:

- First sentence in second paragraph in section 1: ...the only way to extract the EOR... is hyperbole, or at least unsubstantiated.

We have clarified the sentence to read “...cross correlation with 21cm maps may be the cleanest way to extract the EOR component of the near infrared background.”

- Proper names of mathematicians are repeatedly left uncapitalized when referring to functions they are credited with (e.g. Fourier, Poisson, Cartesian, Gaussian, etc.).

Fixed.

- The authors frequently use Delta(vector{k}) and C(vector{ell}). These functions do not make sense when the arguments are vectors. Delta, in particular, must be a spherically averaged power spectrum; only P(vector{k}) is well defined.

We have corrected this oversight.

- Equations 6 & 7: a and b are not defined.

We have clarified that a,b are integers with $-N/2 \leq a,b \leq N/2$.

- The color scale in Figure 5 is not defined. Is it simply number counts?

Yes, we have clarified in the caption that the colorbar indicates the counts in each cell.

- 5th paragraph in 4.3: fiducial survey parameters have a zmax and a dmin. Is there a reason for this inconsistency?

We’ve clarified the text to read “We pick fiducial survey parameters of d_min=20 Mpc and d_max=3000 Mpc (z_max=0.75)...” The minimum distance corresponds to a redshift of 0.005, which we don’t find as helpful as simply stating the distance in Mpc.

- Just after equation 17, the phrase "difference cube" is used. I think this corresponds to the difference between the "even" and "odd" cubes mentioned earlier, but it should be defined.

We have clarified this sentence to read “We estimate the thermal noise power spectrum by computing the power spectrum of the difference between the interleaved odd and even cubes discussed earlier, which contains only thermal noise.”

- The very last sentence of section 5.1 is confusingly worded. I understand what it's trying to say, but it could be put much more straightforwardly.

We have broken up this sentence into “Note that as expected, the thermal noise of the deep integration is 10 times lower in power than the 3\,hour integrations. Because these are band averaged images, even the 3\,hour thermal noise is at least 100 times lower than the foreground residuals.”

- C_ft in equation 20 is undefined.

We have clarified that C_ft is the “diagonal matrix with a guess of the true power spectrum on the diagonal” described in the previous sentence.

- In the list of radio surveys, SKA-LOG is mentioned. I presume this means SKA-Low.

Fixed.