Draft version January 15, 2016
Preprint typeset using LaTeX style emulateapj v. 01/23/15

THE MURCHISON WIDEFIELD ARRAY EPOCH OF REIONIZATION PROJECT: 21 CM POWER SPECTRUM ANALYSIS METHODOLOGY

Daniel C. Jacobs[1,2], B. J. Hazelton[3,4], C. M. Trott[5,6], Joshua S. Dillon[7], B. Pindor[5,8], I. S. Sullivan[3], J. C. Pober[3,9], N. Barry[3], A. P. Beardsley[1,3], G. Bernardi[10,11,12], Judd D. Bowman[1], F. Briggs[5,13], R. J. Cappallo[14], P. Carroll[3], B. E. Corey[14], A. de Oliveira-Costa[7], D. Emrich[6], A. Ewall-Wice[7], L. Feng[7], B. M. Gaensler[15,5,16], R. Goeke[7], L. J. Greenhill[12], J. N. Hewitt[7], N. Hurley-Walker[6], M. Johnston-Hollitt[17], D. L. Kaplan[18,3], J. C. Kasper[19,12], HS Kim[5,8], E. Kratzenberg[14], E. Lenc[5,16], J. Line[5,8], A. Loeb[12], C. J. Lonsdale[14], M. J. Lynch[6], B. McKinley[5,8], S. R. McWhirter[14], D. A. Mitchell[20,5], M. F. Morales[3], E. Morgan[7], A. R. Neben[7], N. Thyagarajan[1], D. Oberoi[21], A. R. Offringa[5,22], S. M. Ord[5,6], S. Paul[23], T. Prabu[23], P. Procopio[5,8], J. Riding[5,8], A. E. E. Rogers[14], A. Roshi[24], N. Udaya Shankar[23], Shiv K. Sethi[23], K. S. Srivani[23], R. Subrahmanyan[5,23], M. Tegmark[7], S. J. Tingay[5,6], M. Waterson[13,6], R. B. Wayth[5,6], R. L. Webster[5,8], A. R. Whitney[14], A. Williams[6], C. L. Williams[7], C. Wu[25,3], J. S. B. Wyithe[5,8]

Draft version January 15, 2016

ABSTRACT

We present the 21 cm power spectrum analysis approach of the Murchison Widefield Array Epoch of Reionization project. In this paper, we compare the outputs of multiple pipelines for the purpose of validating statistical detections of cosmological hydrogen at redshifts between 6 and 12. Multiple, independent, data calibration and reduction pipelines are used to make power spectrum limits on a fiducial night of data. Comparing the outputs of imaging and power spectrum stages highlights differences in calibration, foreground subtraction and power spectrum calculation. The power spectra found using these different methods span a space defined by the various tradeoffs between speed, accuracy, and systematic control. Lessons learned from comparing the pipelines range from the algorithmic to the prosaically mundane; all demonstrate the many pitfalls of neglecting reproducibility. The remaining contrast between results is due to variations in the hypothesis tested. We briefly discuss the way these different methods attempt to handle the question of evaluating a significant detection in the presence of foregrounds.

Subject headings: cosmology: dark ages, reionization, first stars — methods: data analysis — techniques: interferometric

[1] Arizona State University, School of Earth and Space Exploration, Tempe, AZ 85287, USA
[2] e-mail: daniel.c.jacobs@asu.edu
[3] University of Washington, Department of Physics, Seattle, WA 98195, USA
[4] University of Washington, eScience Institute, Seattle, WA 98195, USA
[5] ARC Centre of Excellence for All-sky Astrophysics (CAASTRO)
[6] International Centre for Radio Astronomy Research, Curtin University, Perth, WA 6845, Australia
[7] MIT Kavli Institute for Astrophysics and Space Research, Cambridge, MA 02139, USA
[8] The University of Melbourne, School of Physics, Parkville, VIC 3010, Australia
[9] Brown University, Department of Physics, Providence, RI 02912, USA
[10] Department of Physics and Electronics, Rhodes University, Grahamstown 6140, South Africa
[11] Square Kilometre Array South Africa (SKA SA), Park Road, Pinelands 7405, South Africa
[12] Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA
[13] Australian National University, Research School of Astronomy and Astrophysics, Canberra, ACT 2611, Australia
[14] MIT Haystack Observatory, Westford, MA 01886, USA
[15] Dunlap Institute for Astronomy and Astrophysics, University of Toronto, ON M5S 3H4, Canada
[16] The University of Sydney, Sydney Institute for Astronomy, School of Physics, NSW 2006, Australia
[17] Victoria University of Wellington, School of Chemical & Physical Sciences, Wellington 6140, New Zealand
[18] University of Wisconsin–Milwaukee, Department of Physics, Milwaukee, WI 53201, USA
[19] University of Michigan, Department of Atmospheric, Oceanic and Space Sciences, Ann Arbor, MI 48109, USA
[20] CSIRO Astronomy and Space Science (CASS), PO Box 76, Epping, NSW 1710, Australia
[21] National Centre for Radio Astrophysics, Tata Institute for Fundamental Research, Pune 411007, India
[22] Netherlands Institute for Radio Astronomy (ASTRON), PO Box 2, 7990 AA Dwingeloo, The Netherlands
[23] Raman Research Institute, Bangalore 560080, India
[24] National Radio Astronomy Observatory, Charlottesville and Greenbank, USA
[25] International Centre for Radio Astronomy Research, University of Western Australia, Crawley, WA 6009, Australia

1. INTRODUCTION

Study of primordial hydrogen in the early universe via 21 cm radiation has been forecast to provide a wealth of astrophysical and cosmological information. Hydrogen is the principal product of big bang nucleosynthesis and is neutral over cosmic time from recombination until reionized by the first batch of UV emitters (stars and accretion disks). While neutral it is visible in the 21 cm radio line, which is both optically thin and spectrally narrow, making possible full tomographic reconstruction of a very large fraction of the cosmological volume. Reviews of 21 cm cosmology, astrophysics and observing can be found in Morales & Wyithe (2010); Furlanetto et al. (2006); Pritchard & Loeb (2012); Zaroubi (2013).

Direct detection of HI during the Epoch of Reionization (cosmological redshifts $5 < z < 13$) is currently the goal of several new radio arrays. The LOw Frequency ARray (LOFAR; Yatawatta et al. 2013), the Donald C. Backer Precision Array for Probing the Epoch of Reionization (PAPER Parsons et al. 2014) and the Murchison Widefield Array (MWA; Tingay et al. (2013); Bowman et al. (2013)) are all currently conducting long observing campaigns.

The analysis of the resulting data presents several challenges. The signal is faint; initial detection is being sought in the power spectrum with thousands of hours of integration (accumulated over several years) required. This faint spectral line signal sits atop a continuum foreground four orders of magnitude brighter. At the same time, the instruments are fully correlated phased arrays with wide fields of view that strain the conventional mathematical approximations of radio astronomy practice. The methods used to arrive at a well calibrated, foreground-free estimation of the power spectrum are all under development in the sense of the algorithms as well as the implementation.

The path from observation to power spectrum can be roughly divided into two parts: removal of foregrounds and estimation of power spectrum. Methods for estimating the power spectrum, particularly those which minimize the effects of foregrounds, have been studied and implemented by Morales et al. (2006a, 2012); Dillon et al. (2013, 2014); Liu & Tegmark (2011); Trott et al. (2012); Liu et al. (2014a,b). Common elements include using knowledge about the instrument and foregrounds to minimize covariance, applying an optimal quadratic estimator to make a minimal error estimate, and studies of effects related to including the spectral dimension in the Fourier transform. One significant problem studied has been minimizing the impact of any residual foregrounds by down-weighting or minimizing correlation with contaminated band powers. In this paper we compare power spectra calculated using a range of methods.

Most power spectrum analyses require removal of bright foregrounds to some level. Recently, two sorts of foreground removal have been suggested: methods which exploit detailed knowledge of foregrounds and those which are relatively agnostic. Among the latter, several authors have described methods for fitting and removing smooth spectrum foregrounds from image cubes Morales et al. (2006b); Bowman et al. (2009); Liu et al. (2009); Liu & Tegmark (2011); Chapman et al. (2013); Dillon et al. (2013); Yatawatta et al. (2013). These methods have been demonstrated to robustly remove foregrounds near the field center but are less effective for sources far from the central lobe of the primary beam (Thyagarajan et al. 2015a,b, Pober et al, in press). A second class of agnostic methods is the delay/fringe-rate filtering approach (Parsons et al. 2012; Liu et al. 2014a,b), which has been applied to data from PAPER (Parsons et al. 2014; Ali et al. 2015; Jacobs et al. 2015). Applying time and frequency domain filters to the time ordered data, this technique uses a small amount of knowledge about the instrument to filter modes likely to be dominated by foregrounds. This method removes smooth spectrum foregrounds across the entire sky and is comparatively robust in the face of uncertainty about the instrument at the cost of losing some sensitivity. Meanwhile, full forward modeling and subtraction of a sky model such as that implemented for LOFAR (see e.g. Jelić et al. (2008); Yatawatta et al. (2013)), requires a much higher fidelity model of the instrument and the sky (Datta et al. 2010; Vedantham et al. 2012a).

The MWA foreground removal approach leverages the array's optimization for imaging to directly subtract known foregrounds in addition to the full range of treatments of residual foregrounds, including foreground avoidance and foreground suppression. If successful, direct subtraction opens the most sensitive power spectrum modes, substantially improving the ability of early measurements to distinguish between reionization models (Beardsley et al. 2013; Pober et al. 2014). Recent work towards the goal of foreground subtraction includes better algorithmic handling of wide field imaging effects (Tasse et al. 2012; Bhatnagar et al. 2013; Sullivan et al. 2012; Ord et al. 2010; Offringa et al. 2014), and continually improving catalogs of sky emission (de Oliveira-Costa et al. 2008; Jacobs et al. 2011; Jacobs et al. 2013; Hurley-Walker et al. 2014). Ongoing operation of the next generation low frequency arrays –LOFAR, PAPER and MWA are all in their third or fourth year of operation– continues to push the refinement of instrumental models (e.g. the work of Neben et al. (2015) in mapping the primary beam with satellites) and improve the accuracy of model subtraction. At the same time, more complete surveys of foregrounds are currently under way. These include the MWA GLEAM[1] survey (Wayth et al. 2015) and the LOFAR MSSS[2] (Heald et al. 2015)..

In turn, efforts with these currently operational experiments are having a major influence on how future, larger, EoR experiments will be designed and conducted. Primary among these future experiments will be programs using the low frequency Square Kilometre Array (Koopmans et al. (2014)) and the Hydrogen Epoch of Reionization Array (HERA Pober et al. 2014). Specifically, the MWA is one of three official precursor telescopes for the SKA and the only one of the three fully operational for science. The low frequency SKA will be located at the MWA site in Western Australia, giving the MWA special significance.

Given the challenges of using newly developed methods to reduce data from a novel instrument to make a low sensitivity detection, it is reasonable to consider the question of how one knows one is getting the "right"

[1] GLEAM: GaLactic and Extragalactic All-sky MWA
[2] MSSS: Multi-frequency Snapshot Sky Survey

answer. One option is to generate, as accurately as possible, a detailed simulation of the interferometer output and then input that to the pipeline under test. Such forward modeling is an essential tool for checking correct operation of portions of the pipeline, however the model will always be an imperfect reflection of reality, leaving open multiple interpretations of any differences between model and data. Forward modeling the instrument response is also difficult to divorce from the analysis pipeline being tested; often the same software doing the analysis is used to perform the simulations. A second option, and the focus of this paper, is comparison between multiple independent pipelines.

In section 2 we summarize the observing strategy used to collect our data, section 3 explains our multiple pipelines and comparison strategy. In section 4 we show comparisons of images, 2D diagnostic power spectra and 1D power spectrum limits, section 5 lists some lessons learned from the comparison process and section 6 offers some conclusions.

2. OBSERVING

2.1. *The MWA*

The MWA is an interferometric array of phased array tiles operating in the 80-300 MHz radio band. Each tile consists of a 4x4 grid of dual polarization bow-tie shaped dipoles that are used to form a beam on the sky with a full width of $26^\circ(\lambda/2)$ at the half power point. Signals from individual antennas are summed by an analog delay-line beamformer which can steer the beam in steps of $6.8^\circ cos(l)$. The signal is digitized over the entire bandwidth but only 30 MHz are available at any one time. This 30 MHz of bandwidth is broken into 1.28 MHz "coarse" bands by a polyphase filter-bank in the field and sent to the correlator (Ord et al. 2015) where it is further channelized to 40kHz, cross-multiplied and then averaged at 0.5 second intervals. More details on the design and operation of the MWA can be found in Lonsdale et al. (2009) and Tingay et al. (2013).

2.2. *The 21 cm Observing Program*

The MWA reionization observing scheme spans two 30 MHz tunings, 140-170 MHz ($9.2< z <7.5$) and 167-196 MHz ($7.5< z <6.25$) and two primary minimal foreground regions (RA 0h and 4h, Dec -27°); both transit the zenith at the MWA's latitude and are near the galactic pole. A third pointing towards Hydra A is also observed; see Figure 1 for an overview. Here we focus on the low redshift tuning, and the RA=0h pointing, where the band is chosen for its lower sky temperature and pointing is chosen for its ease of calibration –having fewer bright, resolved sources; see Table 1 for a listing of observing parameters.

The analysis presented here is on three hours of data, one of 400 nights which have been collected as part of the observing program; 150 nights are thought to be necessary for a detection of typical models (Beardsley et al. 2013).

2.3. *Data Included Here*

During observing, the beam-former was set such that the target region repeatedly drifted through the field of view. With an available beamformer step size of 6.8°; each drift was about 30 minutes long. This was done for a total of 6 pointings in a night, or about 3 hours. The data included here include the two pointings leading up to the target crossing zenith, the zenith pointing, and then three more pointings after the transit crossing. Data were recorded in 112 second units for a total of 96 snapshots. These snapshots are the basic unit of time on which many operations become independent - eg RFI flagging, FHD calibration and imaging.[1] Each snapshot is flagged for interference using the AOFlagger (Offringa et al. 2010)[2] algorithm and then averaged to 2 seconds and 80kHz. As described in Offringa et al. (2015), the interference environment at the Murchison Radio-astronomy Observatory is benign and generally requires flagging of about 1% of the data. Though the full set of linear polarization parameters are correlated, and Stokes I images and power spectra are the final product of interest, at this stage of the analysis the instrumental polarizations have been found to be more instructive; with one exception, only the linear east-west polarization is examined here. No significant differences are seen in the north-south data. The same set of snapshots is used in every pipeline run.

3. POWER SPECTRUM PIPELINES

In this section we introduce the basic pipeline components, define some terms common to all, and then in sections 3.1-3.5 give finer grain descriptions of the specific implementations.

The 21 cm brightness at high redshift is weak and detectable by first generation instruments only in statistical measures such as the power spectrum. The spectral line signal is a three dimensional probe, two spatial dimensions and a third from the mapping of the spectral axis to line-of-sight distance via the Hubble relation. 3D power spectra are computed at multiple redshift slices through the observed band and then, taking advantage of statistical rotational symmetry, averaged in shells of constant wavenumber k. The power spectrum is well matched to an interferometer, which natively measures spatial correlation; the baseline vector maps to the perpendicular wavemode $k_\perp$. An additional Fourier transform in the spectral dimension provides $k_\parallel$.

The principal challenge to detecting 21 cm at very high redshifts is foreground emission. At frequencies below 200 MHz the dominant sources are synchrotron emissions from the local and extragalactic sources. Synchrotron is generally characterized by a smooth spectrum which rises as a power law towards lower frequencies. The local Galactic neighborhood has a significant amount of spatially smooth power appearing at short $k_\perp$ modes, extragalactic point sources appear equally on all angular scales and dominate over the Galaxy on long $k_\perp$ modes.

Our analysis pipeline has two main components: one which removes foregrounds –leaving as small a residual as possible– and a second which computes an estimate of the power spectrum. Foreground subtraction is generally the domain of calibration and imaging software where the focus is on building an accurate forward model of the telescope and foregrounds. Challenges include: ionospheric

[1] Note that this is not true in the RTS which uses a time interval scaled by the baseline length.

[2] `sourceforge.net/projects/aoflagger`

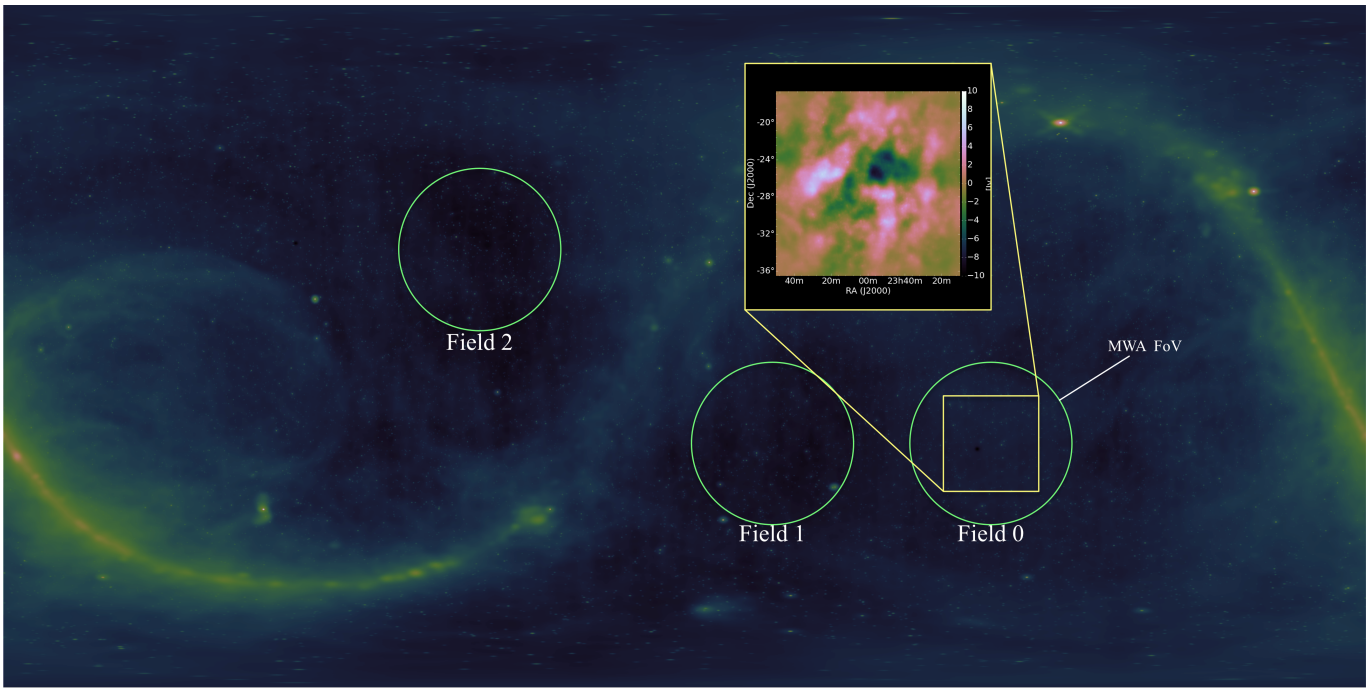

FIG. 1.— An overview of the MWA reionization observation strategy. The background image is a cartesian view of the sky at radio wavelengths and the circles indicate the deep fields observed by the MWA EoR project. Here we are focusing on field 0, centered on Dec -27° and RA 0h. Inset is a foreground subtracted image of the field made using the Real Time System (described more completely in 3.1). A model of smooth (Galactic and unresolved) emission has not been subtracted and dominates the residual map of this 22° wide image. Not visible in this average map is variation from channel to channel caused by sources far beyond the field of view which shows up as the wedge in 2d power spectra...

TABLE 1
MWA EOR OBSERVING PARAMETERS

Parameter	Value
Field of View	$26°\,\lambda$ FWHM
Tuning	166-196 MHz redshift range $7.56 < z < 6.25$
Target area	(RA,Dec) 0h00m, -27°00m
Primary beam pointing grid	6.8°
Snapshot length	112 seconds
Time and frequency resolution	0.5 s, 40 kHz
Post-flagging resolution	2s, 80 kHz
Time	3 hours on August 23, 2013, six 30 minute pointings or 96 snapshots[a]

[a] The same data set is used in every pipeline run

distortion, a very wide field of view, primary beam uncertainty, polarization leakage, and catalog inaccuracy. Though a number of calibration and imaging software packages –such as CASA and Miriad– are available, these challenges have necessitated the creation of custom software. As an added benefit, having developmental control of the imager enables the export of the export of additional information describing the observation which are necessary for the calculation of power spectrum uncertainties as described below.

As we mentioned in the introduction, a horizon-to-horizon model of the sky must be subtracted at high precision from each two minute snapshot across thousands of hours of data. At this scale, deconvolution and self-calibration of each snapshot image is not computationally tractable. In both FHD and RTS the sky model is mainly kept static, rather than peeling a large number of sources the focus has been on refining the instrument model used to subtract catalogs. This instrument model also provides information on the instrumental covariance which is used by the power spectrum estimators.

Detailed knowledge of instrumental covariance is essential to overcoming the two main challenges in estimating the power spectrum: 1) minimizing the effects of residual foregrounds and 2) faithfully recovering the underlying 21 cm power. As discussed in the introduction, simulations and early observations have shown that foregrounds tend to "contaminate" only specific k modes; using a model of instrumental covariance the power can be isolated to fewer modes. Accurate recovery of the 21 cm background will, to first order, depend on the ability to correctly calculate uncertainties. Initial power spectra are expected to be of low signal to noise, an accurate estimate of error is essential to estimating the significance of any putative detection (Pober et al. 2014; Beardsley et al. 2013).

Within the MWA collaboration, efforts have centered around multiple independent paths from raw data to a power spectrum. As described in Figure 2, these pipelines are generally divided into a component which performs calibration, foreground subtraction and imaging, and one which computes the power spectrum. During development, each power spectrum code was paired with a "primary" foreground subtraction method, FHD with *εppsilon* and RTS with CHIPS. The main results come from these primary paths (as depicted by the thin lines in Figure 2).

The primary difference between these pipelines is the division of responsibilities between foreground subtraction and power spectrum calculation. Some power spectrum methods take as input spectral image cubes output by the calibration and foreground subtraction system. The imager also provides a model of the telescope window function and measurement errors in the form of cubes of weights and variances. The weights are formed by gridding down ?1?s in the same way as the data while the variances are constructed by gridding ?1?s with the square of the beam. Together these encode the full covariance of the telescope's window function. Each set of cubes is generated with both even and odd sample cadences; the cross multiplication provides a power spectrum free of noise bias and the difference an estimate of noise.

Methods which take time-ordered data as input generate their own instrument model internally. The pipeline submodules names and citations are listed in Table 2 and described individually in sections 3.1 - 3.6.

3.1. *Calibration and Imager #1: RTS*

The MWA Real Time System (RTS; Mitchell et al. (2008); Ord et al. (2010)) was initially designed to make wide-field images in real time from the MWA 512-tile system (Mitchell et al. 2008). On the de-scoped 128 element array, it has been implemented as an offline system, where it has been adjusted to compensate for the lower filling factor (Ord et al. 2010). The RTS incorporates algorithms intended to address a number of known challenges inherent to processing MWA data, including; wide-field imaging effects, direction-dependent (DD) antenna gains and polarization response, and ionospheric refraction of low-frequency radio waves. Each MWA observation (112s) is processed through a separate instance of the RTS. The RTS is also parallelized over frequency so that each coarse channel (1.28 MHz broken into 40 kHz channels) is processed largely independently of the other coarse channels, with only information about peeled source offsets communicated between processing nodes.

The RTS calibration strategy is based upon the 'peeling' technique proposed by Noordam (2004) and a foreground model using a cross-matching of heritage southern sky catalogs[1] with the MWA Commissioning Survey. The cross-matching is done using the PUMA code which uses Bayesian inference to build a self-consistent set of SEDs for sources using data from catalogs with varying frequency and resolution. The brightest apparent calibrators in the field of view are sequentially and iteratively processed through a Calibrator Measurement Loop (CML). During each pass through the CML; i) the expected (model) visibilities of known catalog sources are subtracted from the observed visibilities. For the data processed in this work, 1000 sources are subtracted for each observation. ii) The model visibilites for the targeted source are added back in and phased to the catalog source location. Any ionospheric offset of the source can now be measured by fitting a phase ramp to the phased visibilities. iii) The strongest sources are now used to update the direction-dependent antenna gain terms, while weaker sources are only corrected for ionospheric offsets. For this work, 5 sources are used as full DD calibrators and 1000 sources are set as ionospheric calibrators. The CML is repeated until the gain and ionospheric fits converge to stable values. A single bandpass for each tile is found by fitting a 2nd order polynomial to each 1.28 MHz-wide coarse channel. The ∼1000 strongest sources are then subtracted from the calibrated visibilities. Calibration and model subtraction parameters are summarized in Table 3. Model subtracted visibilities are passed to the RTS imager and to the CHIPS power spectrum estimator.

The RTS imager uses a snapshot imaging approach to mitigate wide-field and direction-dependent polarization effects. Following calibration, the residual visibilities are first gridded to form instrumental polarization images which are co-planar with the array. These images are then regridded into the HEALpix (Górski et al. 2005) frame with wide-field corrections. Weighted instrument polarization images are stored, along with weight images containing the Mueller matrix terms, so that further integration can be done outside of the RTS. It is also possible to use the fitted ionospheric calibrator offsets to apply a correction for ionospheric effects across the field during the regridding step or subtraction of catalog sources, but in this work this correction has not been applied. These snapshot data and weight cubes are then integrated in time to produce a single HEALpix cube. This cube, averaged over the spectrum, is shown, with and without foregrounds, in Figure 3.

3.2. *Calibration and Imager #2: FHD*

Fast Holographic Deconvolution (FHD, Sullivan et al. (2012)) is a calibration and imaging algorithm designed for very wide field of view interferometers with direction- and antenna-dependent beam patterns. Variable beam patterns are used to grid visibilities to the *uv* plane and its inverted operation for de-gridding simulations to form model visibilities, this careful accounting of weights provides a necessary accounting of information loss caused by the inherent size of the dipole element.

The FHD calibration pipeline generates a model data set, computes a calibration solution which minimizes the difference with the data, smooths the calibration solution to minimize the number of free parameters, and outputs the residual. The calibration model is formed from sources found by deconvolving, in broadband images, about 75 of the 96 snapshots included here and retains those which are common to all snapshots and pass other consistency checks (Carroll et. al. in prep). In each snapshot sources are included in the model if they are at or above 1% of the peak primary beam, this amounts to about 7000 sources and a flux limit of about 80mJy with slight variations snapshot to snapshot. Most sources in

[1] See Table 3

TABLE 2
MWA EoR Pipeline components

Short Name	Name	Citations
Cotter	AOFlagger + Averaging	Offringa et al. (2010)
RTS	Real Time System	Mitchell et al. (2008); Ord et al. (2010)
FHD	Fast Holographic Deconvolution	Sullivan et al. (2012)[1]
εppsilon	Error Propagated Power Spectrum with InterLeaved Observed Noise	Hazelton et al 2015, in prep[2]
CHIPS	Cosmological HI Power Spectrum	Trott et al. (2016)
EmpCov	Empirical Covariance Estimator	Dillon et al. (2015)

[1] `github.com/miguelfmorales/FHD`
[2] `github.com/miguelfmorales/eppsilon`

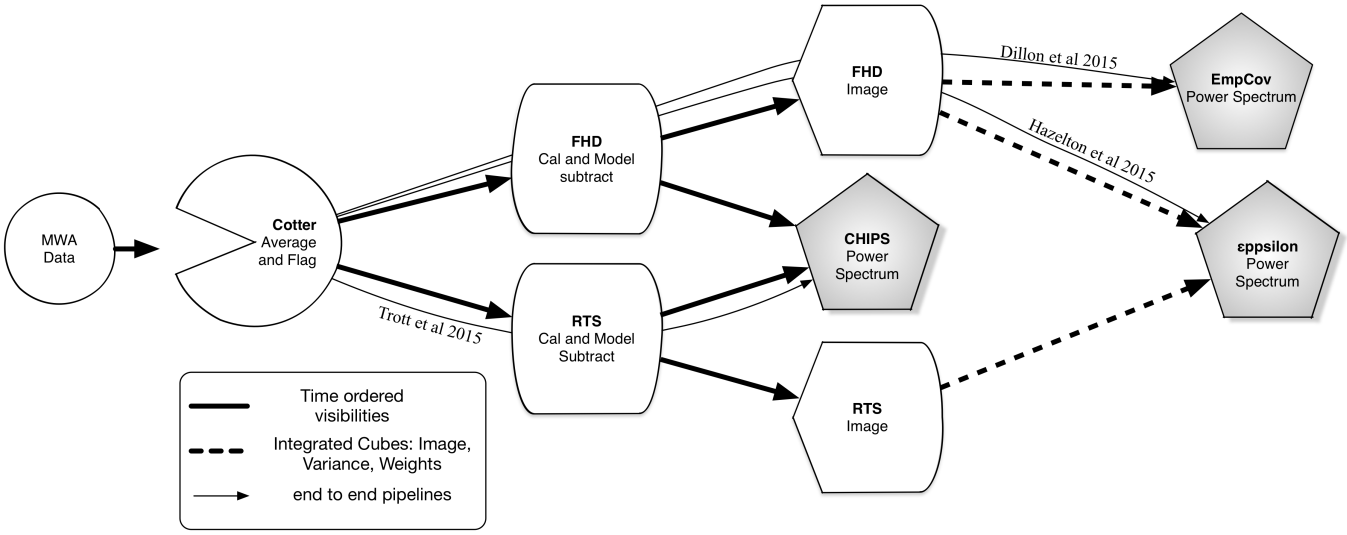

Fig. 2.— Parallel pipelines with cross-connections after foreground subtraction and imaging are compared against each other. Pipelines used to reach the cited power spectrum results are indicated with thin lines; citations for each block are listed in Table 2. Cotter uses AOFlagger to flag RFI and averages by a factor of 8. The averaged data are passed to either FHD or RTS for calibration, foreground subtraction and imaging. Both of these packages generate integrated residual spectral image cubes as well as matching cubes of weights and variances. *εppsilon* and EmpCov use these cubes to estimate the power spectrum. Meanwhile, CHIPS taps into the RTS and FHD data stream to get calibrated and foreground-subtracted time-ordered visibilities which it then grids with its own instrument model to estimate the power spectrum.

TABLE 3
MWA Reionization Calibration and Model subtraction Parameters. Counts are per-snapshot unless otherwise noted

Parameter	RTS	FHD
per cable passband	NA	384 channels[a]
per antenna passband	48 per tile[b]	3[c]
per antenna gain	2[d]	2[d]
peeling parameters	4[e]	None
peeled sources	5	None
subtraction catalog	Line[f]	Carroll[g]
number subtracted	1000	6932
Total free parameters	**6,420**	**880**

[a] for ea. of 6 cable types, averaged over 96 snapshots
[b] 2nd order polynomial per coarse channel
[c] poly fit over full band, 2nd order for amp, 1st for phase
[d] amplitude and phase
[e] Direction Dependent (DD) gain fits
[f] MWA Commissioning SurveyHurley-Walker et al. (2014),(Lane et al. 2014, VLSSr),(Large et al. 1991, MRC),(Mauch et al. 2003, SUMSS),(Condon et al. 1998, NVSS),cross matched using PUMA (Line et al, in prep) `github.com/JLBLine/PUMA`
[g] Combination catalog of legacy catalogs and sources deconvolved from this data (Carroll et al in prep)

this catalog have spectral indexes between 0 and -2 with the majority near a mean of −0.8, which corresponds to a 13% difference across the 30 MHz. Spectral index is not directly modeled, spectra are simulated to be flat, however, during catalog subtraction, the data are multiplied by a positive spectral index of 0.8 such that most sources will appear to be flat. Full spectral modeling of catalog sources will be included in future analyses.

The goal of FHD's calibration is to minimize the number of free parameters with the twin goals of minimizing potential signal loss and in building a deep understanding of instrumental systematics. Here we give a brief description of the instrument model, a listing of all the parameters mentioned here is also given in Table 3 along with a rough accounting for the total number of fitting parameters. Initial complex gain solutions computed using the Alternating Direction Implicit technique described in Salvini & Wijnholds (2014) for each antenna, channel and polarization. This generates a gain and phase for every channel on every tile, for each 112s snapshot. Most antennas have similar solutions with the main features corresponding to the exact type and length of analog cable feed of which there are 6 different types; per-tile solutions are further averaged into per-cable-type

and averaged over the entire 3 hour observation. After these solutions are divided out, the residual per-tile solutions are further fit for a second order amplitude spectral polynomial and a 1st order phase slope. This is done on every snapshot to account for temperature-driven amplifier gain changes. One systematic easily visible in the power spectra is a small reflection corresponding to the 150m cables. This is fit and removed as a phase delay with a ∼0.1dB amplitude in the time averaged per-tile bandpass solutions.

The residual time-ordered visibilites are then passed to CHIPS and to FHD imaging for formation of spectral cubes. The FHD imager produces snapshot cubes using the MWA beam model described by Sutinjo et al. (2015) and averaged in time. This image, averaged over the spectral dimension, is shown, with and without foregrounds, in Figure 3.

3.3. *Comparing Calibration and Imaging Steps*

Through the parallel-but-convergent development of these imagers have emerged two very similar systems, however some differences remain in the analysis captured here. The two primary differences are in the treatment of calibration and in the subtracted catalogs.

In both pipelines the calibration is a two step process. First, calibration solutions for each channel, and antenna are computed by solving for the least-squares difference with a model data set. Next, those solutions are fit to a model of the array; for example fitting a polynomial to the bandpass. FHD and RTS take different approaches to this step, a fact reflected in the the number of free parameters in this fit. A smaller number of parameters minimizes the possibility of cosmological signal loss; more free parameters can absorb physics missing from the instrument model. As tabulated in Table 3 the RTS fits for 6,420 free parameters while FHD fits for ∼880.

In practice, some parameters will be averaged over more than a single night which will further reduce the number of free parameters per observation, though hundreds to thousands of free parameters is still typical. This is a large number but it is considerably smaller than the 180 million data points typically recorded in a two minute observation.

As has been noted, there is nearly an order of magnitude difference in the number of free parameters, which is worth considering. The primary difference is in the treatment of the passband. There are a number effects which show up in the passband calibration. The edges of the 1.28 MHz bands are known to be subject to aliasing from adjacent coarse channels as well as under-sampling when cast to 4bit integers by the correlator (van Vleck corrections) and so are flagged. This flagging creates a regular sampling function which shows up as the characteristic horizontal lines in a 2D power spectrum. Added to this is a small amount of interference flagging. Additionally, reflections at analog cable junctions show up as additional spectral ripple corresponding to the length of the cables.

The RTS fits for a low order polynomial on every 1.28 MHz chunk on every antenna, while FHD averages each channel over all antennas to get a common passband for all and then fits a low order polynomial to get any tile to tile variation. This significantly reduces the number of free parameters and the likelihood of signal loss, though leaving open the possibility of additional un-modeled instrumental effects.

The construction of the foreground subtraction model is also a point of difference between the two pipelines. As noted in Table 3, foreground/calibration models contain different numbers of sources which have been derived by different means. The RTS catalog cross-matches multiple heritage southern sky catalogs with the MWA Commissioning Survey using the Bayesian cross-matcher PUMA (Line et al in Prep). The FHD subtraction model contains sources found in a deep deconvolution of this same data set. Both catalogs have the goal of producing a reliable set of sources that minimizes false positives and accurately reflects resolved components, though they go about it in different ways. The FHD catalog focuses on the reliability aspect by performing a deconvolution on every snapshot used in the observation and selecting sources which appear in most observations (Carroll et al, in prep). The RTS catalog has used the somewhat less precise MWA commissioning catalog but by cross-matching these sources against many other catalogs of known sources and fitting improved positions and fluxes, the accuracy is seen to increase.

3.4. *Power Spectrum #1: εppsilon*

εppsilon calculates a power spectrum estimate from image cubes and directly propagates errors through the full analysis, see Hazelton et al 2015, in prep for a full description. The design criteria for this method is to make a relatively quick and uncomplicated estimate of the power spectrum to provide a quick turnaround diagnostic. The input to *εppsilon* is gridded image cubes for each 112s snapshot, such as are produced by FHD or RTS imaging, in which the data has been split into interleaved time samples (referred to as even and odd cubes) along with matched cubes containing the modeled instrumental weighting and variance. These snapshot HEALpix cubes are integrated in time keeping pixels with a beam weight of 1% or more, a cut which effectively limits the field of view to ∼20°. The accumulated data, weight and variance cubes are Fourier transformed along the two spatial dimensions into uvf space, where the spatial covariance matrix is assumed to be diagonal. This is approximately true if the uv pixel size is well matched to the primary beam size, so the *εppsilon* DFT grid size is restricted to being equal to the width of the FT primary beam; ie 1/(field of view). The data (variance) cubes are then divided by the weight cubes (weight cubes squared) to arrive at the best estimates of the sky and variances. Next the sum and difference of the even and odd cubes are computed with variances given by adding the reciprocal of the even and odd variances in quadrature. The difference cube then contains only noise (as long as the time interleaving is fine enough) and the sum cube contains both sky signal and noise.

The next step is to Fourier transform in the frequency direction. Here we choose to use the full 30 MHz spectral window, weighted by a Blackman-Harris window function, which heavily down-weights the outer half of the band to effectively sample 15 MHz; a cosmological redshift range of 0.86. This weighting scheme minimizes the covariance of bright foreground modes between power spectrum modes as described in Thyagarajan et al. (2013); Parsons et al. (2012); Vedantham et al. (2012b),

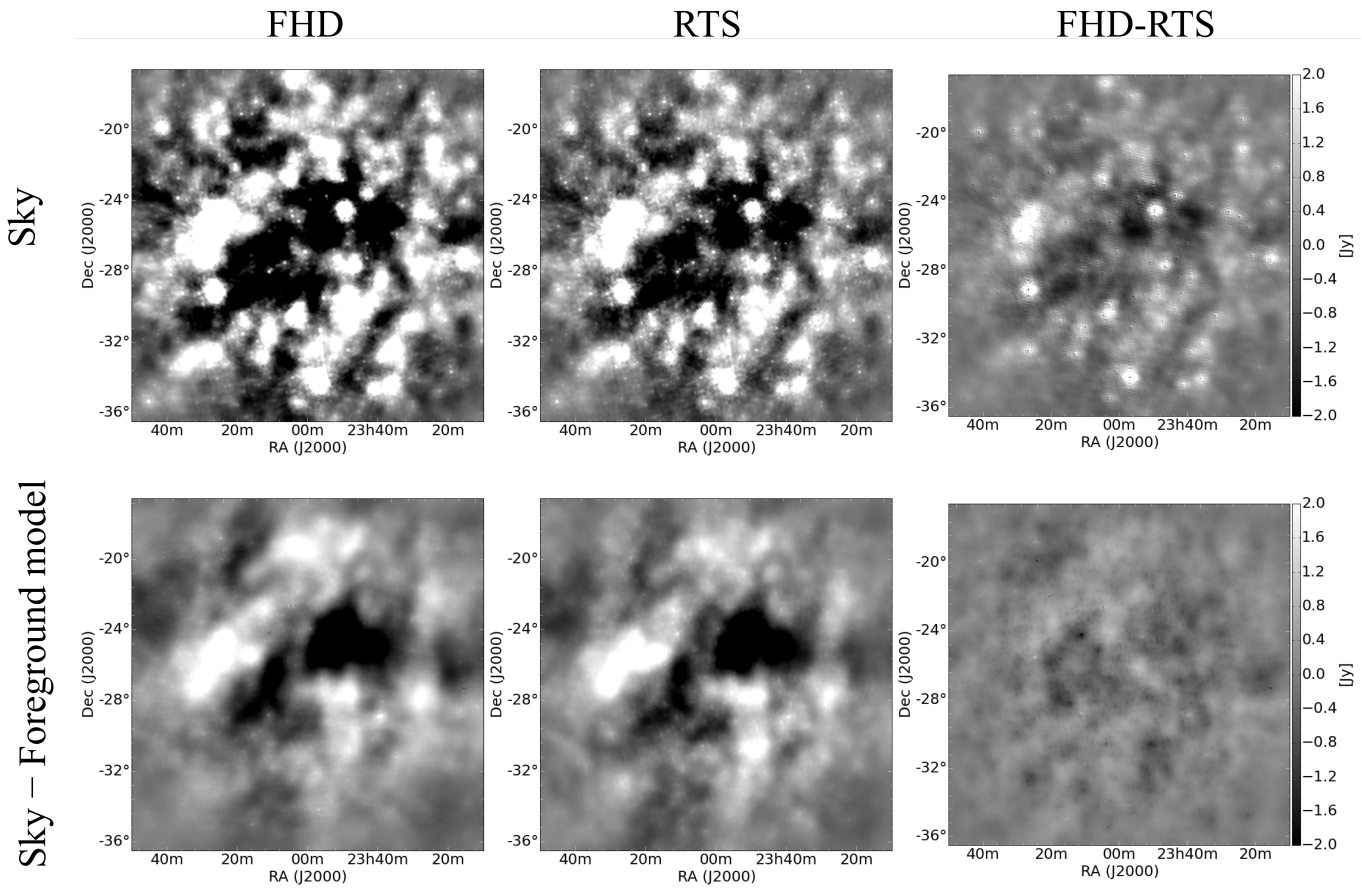

FIG. 3.— A comparison between the image outputs of the FHD (left), RTS (center) and their difference (right) averaged in the spectral dimension and projected from native HEALpix to flat sky. The images have been left in the natural weighting used by image-based power spectrum schemes and no deconvolution has been applied. In the top row, no foreground model has been subtracted; the residual shown represents a 15% difference. On the bottom both have subtracted their best model of the sky containing similar sets of thousands of sources, most pixels the difference is 30% or lower. The difference between foreground subtracted images reveals a good agreement on large scale structure and small differences in the fluxes of a few sources.

among others. The transform of the spectral dimension is done using the Lomb & Scargle periodogram to minimize the effects of regular gaps in the spectrum which occur every 1.28 MHz. The sky signal power is estimated by the square of the sum cube minus the square of the difference cube, which is mathematically identical to the even/odd cross power if the even and odd variances are identical, while the square of the difference cube provides a realization of the noise power spectrum. Diagnostic power spectra are generated by averaging cylindrically to a two dimensional $k_\parallel, k_\perp$ power spectrum. These are shown in the left column of Figure 4. One dimensional power spectra (shown in Figure 6), are calculated by a similar process but only including the inner third of the spectrum (for a redshift range of 0.3) and then averaging along shells of constant k, masking points within the foreground wedge[1] and weighting by variance.

3.5. *Power Spectrum #2: CHIPS*

The CHIPS power spectrum estimation method computes the maximum likelihood estimate of the 21 cm power spectrum using an optimal quadratic estimator formalism and is more completely described in Trott et al. (2016). The design criteria for this method were to fully account for instrumental and foreground induced covariance in the estimation of the power spectrum. The approach is similar to that used by Liu & Tegmark (2011), but with the key difference of being performed entirely in uvw-space, where the data covariance matrix is simpler (block diagonal), and feasible to invert. This approach also allows straightforward estimation of the variances and covariances between sky modes by direct propagation of errors. CHIPS takes as input calibrated and foreground subtracted time-ordered visibilities. Tapping into the pipeline post-calibration but before imaging, CHIPS uses its own internal instrument model to estimate and propagate uncertainty.

The method involves four major steps: (1) Grid and weight time-ordered visibility channels onto a uvw-cube using the primary beam model, (2) compute the least squares spectral (LSS) transform along the frequency dimension to obtain the best estimate of the line-of-sight spatial sky modes (this technique is comparable to that used by *εppsilon*), (3) compute the maximum-likelihood estimate of the power spectrum, incorporating foregrounds and radiometric noise, averaging k_x and k_y modes into annular modes on the sky, $k_\perp$; (4) com-

[1]Here defined, conservatively, as the light travel time across the baseline plus the delay associated with the pointing furthest from zenith. It is indicated as a solid line on Figure 4.

pute the uncertainties and covariances between power estimates. The first step is the most computationally-intensive, requiring processing of all the measured data. The main departure point for CHIPS from *εppsilon* is in the much finer resolution of the *uv* grid. Using an instrument model, CHIPS calculates the covariance between *uv* samples as a function of frequency. Since the beam and *uv* sampling function are both highly chromatic, extra precision in this inversion is thought to be highly beneficial. After a line of sight transform similar to that used by *εppsilon*, this covariance information is inverted to find the Fisher Information, the maximum likelihood power spectrum, and covariances between measurements. The maximum likelihood estimate of the power in each $k_\perp, k_\parallel$ mode is shown in the right column of Figure 4 and averaged in spherical bins in Figure 6. This last averaging step includes an additional weighting by the known power spectrum of a confused foreground in a process described in more detail for these data by Trott et al. (2016). The power spectra shown in Figure 6 show an excess of power in excess of the expected noise. This excess is notably similar between both calibration/foreground subtraction pipelines. The amount of power in the excess, as compared with the error bars, also depends rather dramatically on the range of k bins included in the final averaging to the 1D. These are discussed in more detail in section 4.1.

3.6. *Power Spectrum #3: Empirical Covariance*

The EmpCov power-spectrum estimation method computes a 1D power spectrum using a quadratic estimator formalism. The method and its application to this data is described in more detail by Dillon et al. (2015) .

The quadratic estimator method of Liu & Tegmark (2011) treats foreground residuals in maps as a form of correlated noise and simultaneously downweights both noisy and foreground-dominated modes, keeping track of the extra variance they introduce into power spectrum estimates. This technique can be computationally demanding but using acceleration techniques described by Dillon et al. (2013), has been applied to the previous MWA 32T results of Dillon et al. (2014) while a very similar technique, working on visibilities rather than maps, was used for the recent PAPER 64 results of Ali et al. (2015). Dillon et al. (2015) build on these methods to mitigate errors introduced by imperfect mapmaking and instrument modeling through empirical covariance estimation, assuming all data covariance is sourced by foregrounds.

EmpCov takes as input FHD calibrated images with foregrounds subtracted as well as possible, split into even and odd time-slices and averaged over many observations. From these cubes, it estimates the frequency-frequency foreground residual covariance in annuli in uvf space, assuming that different uv cells have uncorrelated foreground residuals. This assumption, similar to that made by CHIPS, allows the combined foreground and noise covariance to be inverted directly. The resulting power spectrum is shown in Figure 6 for direct comparison with the CHIPS result.

3.7. *Benefits of Comparison*

One benefit from having multiple pipelines is the freedom to investigate different optimization axes. The design of the *εppsilon* power spectrum estimator emphasizes speed and relative simplicity, choices motivated by the need to understand the effect, on the power spectrum, of processing decisions such as observation protocol, flagging, and calibration. Using *εppsilon* we have discovered and corrected multiple systematic effects, primarily those of a spectral nature which were not obvious in imaging but quite apparent in the 2D power spectrum. With the ability to quickly form power spectra on different sets of data, *εppsilon* has been an important tool for selecting sets of high quality data.

In contrast, CHIPS starts from time-ordered data and in its calculations emphasizes a more full accounting of instrumental and residual foreground covariance. Not only does this higher resolution covariance calculation provide a more accurate accounting of the instrumental window function on the power spectrum, but it also allows for more precise weighting schemes based on knowledge of the statistical properties of the residual foregrounds. This is useful when making 1D power spectra where foreground-like modes can be down-weighted in the average.

4. COMPARISON DISCUSSION

Inspecting a comparison of the images and power spectra reveals several common features. Images before and after foreground subtraction are shown in Figure 3, presented in the natural weighting used by the power spectrum estimators without application of any further cleaning. Putting the same 3 hours of MWA data into each pipeline, we inspect output images before and after foreground subtraction. The pre foreground-subtracted (sometimes called the "dirty" image) have residuals at about the 15% level, after foreground subtraction the differences are somewhat larger at 30%. Residuals in the dirty maps are largest around bright sources. This is most likely due to slight differences in the calculation of image plane weights which are dramatically emphasized by the broad psf from the natural weighting. As evidenced by the clean residual maps, the point source subtraction is well modeled when subtracted in the visibilities. The foreground subtracted images (sometimes called "residual" images) show a much closer agreement both around the subtracted sources and in the large scale structure. Large scale structure is more difficult to distinguish. Inspection of the snapshot images before averaging in time and frequency revealed that the structure is constant across both time and frequency, which suggests real Galactic emission rather than sidelobes or aliasing.

4.1. *Power Spectra*

Application of our two independent power spectrum estimators to our two calibration and foreground subtraction pipes gives us a total of four different power spectra (Figures 4 and 5). Each power spectrum estimator has been developed to target the output from a "primary" calibration and foreground subtraction process – the diagonal elements of Figure 4– and have been highly optimized to that up-stream source of data. The off-diagonal power spectra were created using auxiliary links which import the data and the metadata produced by the foreground subtraction step. Since they are less highly optimized, lacking as they do the advantage of a close working relationship, these pathways represent an upper

limit on the variance to be expected from small analysis differences but allow us to look for effects common to foreground subtraction or to power spectrum method.

Properties shared by all are the large amount of power at low $k_\parallel$ roughly at an amplitude of 10^{15} $\mathrm{mK}^2/(\mathrm{Mpc/h})^3$. This emission is approximately flat over most of $k_\perp$ but rises steeply rise at low $k_\perp$. The amplitude agreement is particularly apparent in Figure 5. A model of smooth galactic emission has not been subtracted which likely contributes to this steep rise. The "wedge" shaped linear dependance on baseline length in the 2D power spectra is due to the inherently chromatic response of a wide field instrument to smooth spectrum foregrounds; sources entering far from the phase center appear as bright pixels at higher $k_\parallel$ with sources on the horizon at the edge indicated by Figure 4's solid black line. The solid and dotted lines in the figure indicate the upper boundaries of power from sources at the horizon and at the beam half power point, respectively. With the exception of some instrumental features foreground power is well isolated within this expected boundary. This emission is also visible in the image cubes as sidelobes extending from outside of the imaged area which move as function of frequency. Observations recorded when the Galactic plane is near the horizon have a much larger wedge component and have been excluded from this analysis. See Thyagarajan et al. (2015a) and Thyagarajan et al. (2015b) for a detailed discussion of the foreground contributions to the power spectra in this data.

The two main instrumental systematics are horizontal striping due to missing or poorly calibrated data at the edges of regular coarse passbands and vertical striping due to spectral variation near uneven uvf sampling. The former can be minimized by careful calibration of the passband, the latter by uv rotation synthesis and by accounting for covariance between uvf samples.

The most noticeable difference between the different pipeline paths is in the power level in the window above the horizon and below the first coarse passband line (between 0.1 and 0.3 $k_\parallel$ and 0.01 and 0.05 $k_\perp$). FHD to *εppsilon* displays a noise-like window in the 2D space, with a number of points dipping below zero while the other methods are noise like only at at much higher ks. One commonality between all power spectra with this positive bias is a relatively higher amplitude of the coarse passband lines.

The final analysis step is to average into 1D power spectra along shells of constant k. These are shown in Figure 6 for three of the four analysis tracks[1] shown in Figure 4 with the addition of the Dillon et al. (2015) points and a theoretical sensitivity curve calculated using the 21CMSENSE sensitivity code[2] by Pober et al. (2014).

The positive biases visible in the 2D power spectra are also apparent here. Only a few a few points are consistent with zero at 2σ, however most some are very close to the theoretical sensitivity curve and have errors matching those predicted for noise. The power spectra fall into two groups, those calculated from input image cubes (*εppsilon* and EmpCov) and those calculated directly from visibilities using CHIPS. The image-based points are somewhat deeper at low k, as noted in the 2D plots. Points from CHIPS are biased more strongly at low k but the slope is flatter and converges with the other pipelines at higher k.

[1]The RTS→*εppsilon* spectrum is excluded here because image-plane uncertainties are not yet available as outputs from the RTS.

[2]`github.com/jpober/21cmsense/`

4.2. *CHIPS bias and the interpretation of error bars*

Part of the CHIPS bias is due to the calculation of weightings. Default CHIPS analysis uses a statistical model of confused foregrounds to down-weight biased modes, particularly those correlated with the wedge power. For this reason it is desirable to include the wedge modes in that 1D average. However it significantly changes the interpretation of error bars; points in which a significant amount of power have been down-weighted will have error bars much larger than thermal. In the interest of comparing with the other methods, these power spectra have been calculated using points only lying outside the wedge horizon. This limits the amount of wedge-to-window covariance CHIPS can remove and contributes to the larger bias.

Including the full wedge in the CHIPS covariance calculation offers foreground suppression, but also introduces a foreground component into the error bars. Compare in Figure 7 the RTS→CHIPS power spectrum in Figure 6 with that given by Trott et al. (2016) which used the same data shown here, though only 1/3 of the 30MHz band. In both, CHIPS has down-weighted by a model of foreground covariance formed by propagating a statistical model of confused sources. The only difference is that black excludes the wedge but red does not. When the wedge is included the modeled foregrounds in those voxels dominate the covariance weights. Applying these weights essentially moves the foreground bias into the error bars and asserts that, given our best model of foregrounds, the power spectra are completely consistent with noise and foregrounds.

The power spectra in Figure 6 show the range of results possible given the same input data. Though they do not all agree, they do paint a consistent picture. Differences partly come from the definition of error bars but also indicate the relative difficulty of methods. Methods which rely on an imager seem to perform somewhat better. This is perhaps unsurprising. CHIPS computes the instrumental correlation matrix in visibility space using beam, bandpass, etc. As the CHIPS analysis exists entirely in the visibility space, errors in modeling the instrument model elements are perhaps more difficult to detect than they are in the image space. However, we do not suggest that visibility-based calculations like CHIPS are doomed to failure; rather the opposite. The instrument models will continue to improve, and this improvement will be easily validated by comparison with the other pipelines.

5. LESSONS FROM COMPARING INDEPENDENT PIPELINES

A data analysis pipeline is necessarily built on a complex software framework which is only imperfectly described in prose and is susceptible to human error. Comparison between independently developed analysis paths, each with their own strengths and limitations, is essential to placing believable constraints on the Epoch of Reionization. The ongoing comparison between independent MWA pipelines has revealed a number of issues both sys-

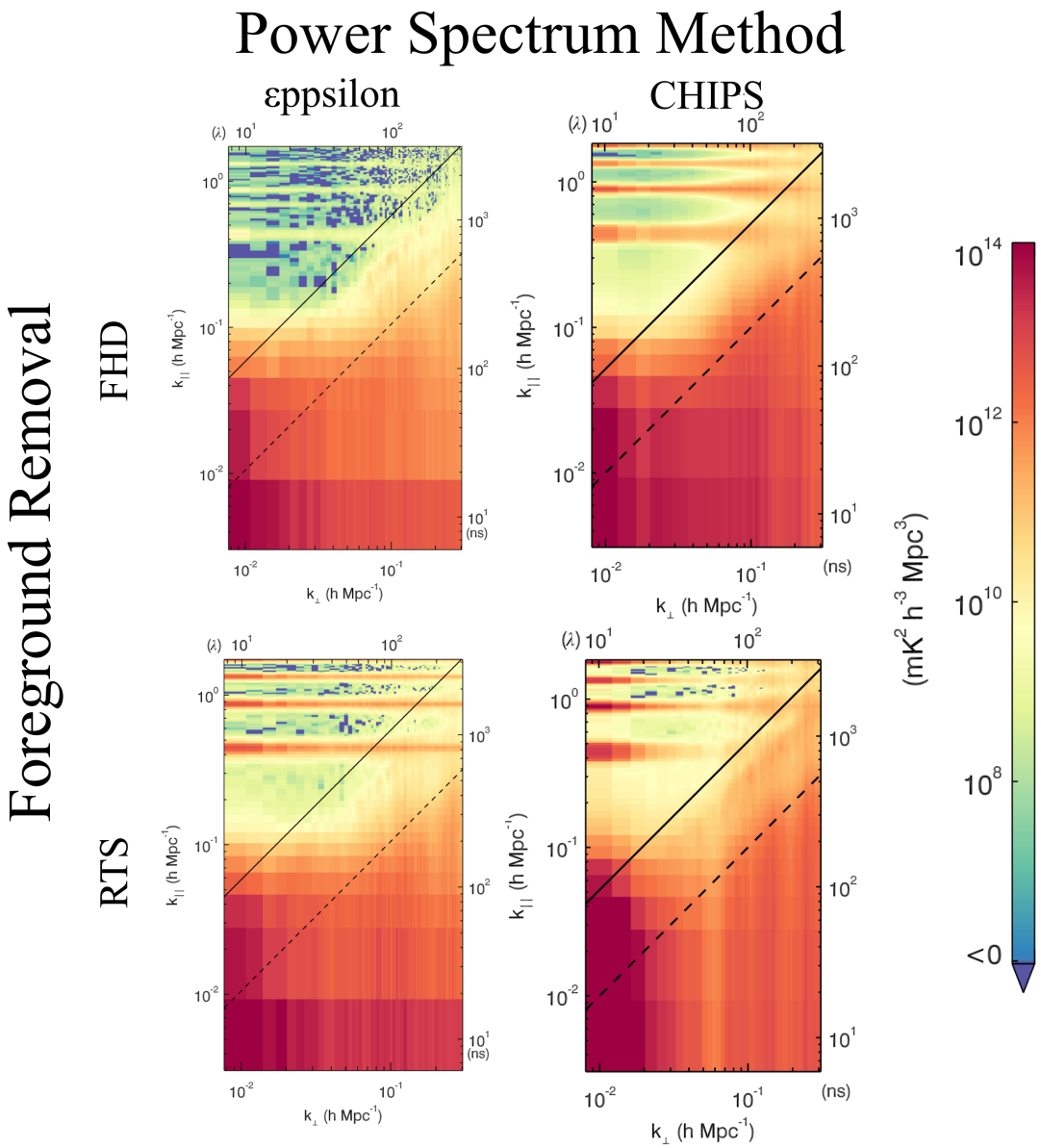

FIG. 4.— Power spectra computed using two foreground subtraction methods and two power spectrum estimation methods on the data shown in Figure 3; the power spectrum has been computed in 3D spectral line cubes and then averaged cylindrically. In the top row data have been calibrated and foreground subtracted using the Fast Holographic Deconvolution method, in the bottom row by the MWA Real Time System. In the left column, power spectra have been estimated with *εppsilon*, which emphasizes speed and full error propagation, in the right column, CHIPS corrects more correlation between k modes. All spectra display the now well-understood "wedge"-shaped foreground residual and horizontal stripes caused by evenly spaced gaps in the instrument pass-band. Because all the power spectra are calculated by cross-multiplying independent data samples, measurement noise remains zero mean; negative regions are therefore indicative of noise-dominated regions.

tematic (related to our understanding of the instrument or foregrounds) and algorithmic (optimizing our use of this knowledge) which we will briefly mention here.

- *Systematic example: cable reflections*

  As discussed above, one significant difference between the two pipelines is the number of free parameters fit in the calibration step, particularly in the spectral dimension. Both calibration pipelines begin by calibrating each channel and then averaging over a number of axes. The RTS fits a low order polynomial, piecewise, to each of the 24 1.28MHz sub-band solutions, while FHD fits a similar order polynomial to the entire band's calibration solution. Inspection of power spectra calibrated using the FHD scheme revealed previously unknown spectral features corresponding to reflections on the analog cables at the -20dB level (1.5%). FHD calibration now includes a fit for these reflections and the feature is substantially reduced. These features are fully covered by the RTS fit (which uses of order 10 times as many free parameters as FHD).

- *Calibration example: number of sources*

  In early comparisons between RTS and FHD images one immediately apparent difference was the

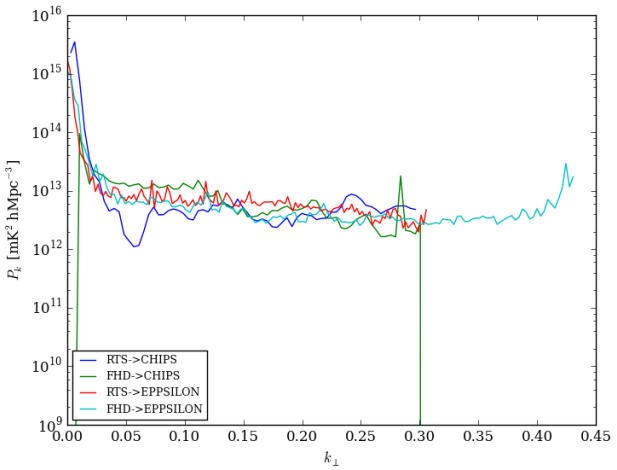

FIG. 5.— Horizontal cut sampling the $k_\parallel = 0$ mode of the 2D power spectra shown in Figure 4 indicating good agreement on flux scale and foreground power shape over most k modes. The foreground subtraction model only includes point sources. The steep rise is likely due to the bright, smooth galactic foreground emission visible in the residual images in Figure 3 and power spectra by Thyagarajan et al. (2015b).

somewhat lower dynamic range of the RTS images. This was traced to the largest (at the time) difference between the two approaches; RTS used the more traditional radio astronomy practice of calibrating to a pointing on a bright source at the beginning of each night and then transferring the calibration to the rest of the observations, whereas FHD was calibrating against the foreground model using the cataloged sources within the field of view (a few thousand). This dramatically highlighted the breakdown of approaches designed for traditional dish telescopes with a narrow field of view. The MWA field of view is so wide, that even the calibration pointing included many sources of brightness comparable to the calibrator. These sources were not included in the calibration model and thus limited the accuracy of the calibration. Also, owing to the phased array beam steering, the primary beam for the calibrator pointing is very different from the beams used for the primary reionization observations, particularly in polarization response. So, though the instrument itself is highly stable in time over many hours, calibrations must be carefully matched up with the observing parameters or experience a dramatic loss of imaging dynamic range, both spatially and spectrally. The addition of "in field" calibration, where the foreground subtraction model is also the calibration model, significantly improved the RTS images and brought the two imagers into substantial agreement.

- *Algorithmic example: full forward modeling for absolute calibration and signal loss*

  During the comparison process, one way in which all pipeline results differed from each other is in the overall amplitude of the power spectrum scale. Flux calibration, weightings, Fourier conventions and signal loss must all be well understood for good agreement to be reached. Signal loss, in particular, must be examined closely. Unintentional or unavoidable down-weighting or subtraction of reionization signal could occur at multiple stages such as bandpass calibration, uvf gridding, or inverse covariance weighting. These effects are best calibrated via forward modeling of simulated sky inputs. and in the process provides verification of the overall power spectrum scale. For example, detailed simulations of reionization signals through FHD and *εppsilon* found that in areas of dense uv sampling, simulated power spectra experienced a 50% reduction of detected power Hazelton et al 2015, in prep. These simulated reionization data sets have been calibrated internally by comparing outputs at every step of the imaging and power spectrum process, and so are well understood at a detailed level, and suitable for use calibration standards for new pipelines.

- *Algorithmic example: w-planes in power spectrum calculation*

  Many of the differences found between power spectra during the comparison were traced to the post-foreground-subtraction steps, particularly the implementation of new imaging and power spectrum estimation codes. One example was an anomalous loss of power in CHIPS power spectra which particularly effected longer baselines. CHIPS grids in a coordinate space defined by the baseline vector $\vec{b}$ and spectral mode η and then uses an instrument model to diagonalize and sum in this sparse power spectrum space. Unlike FHD which uses snapshots to avoid directly handling the third or 'w' term of the baseline vector, CHIPS accumulates the entire observation into a full uvwη hyper-cube. The number and size of the voxels in this space, particularly in the w direction is a somewhat free parameter and relates to the precision of the instrument model, the amount of time included and other factors. Subsequent, more detailed foreground simulations suggested a factor of 4 w resolution increase which eliminated the signal loss and dramatically improved agreement.

- *Interchange standards*

  Finally, in the interest of transparency, we offer a somewhat prosaic but perhaps vital lesson regarding nomenclature. For fixed dipole arrays there are (at least) two popular and mutually exclusive traditions. Tradition A: In keeping with the customary abscissa of latitude longitude plots, the east-west oriented dipole is labeled X. Tradition B: Astronomically, the X polarization is measured as the amplitude of a dipole aligned with lines of constant Right Ascension; which for a source at zenith maps to north-south. We humbly suggest that those pursuing a cross comparison effort select one standard at the outset.

We must stress that without the ability to compare between independent pipelines, most of these effects would have gone un-detected or mis-diagnosed as algorithmic

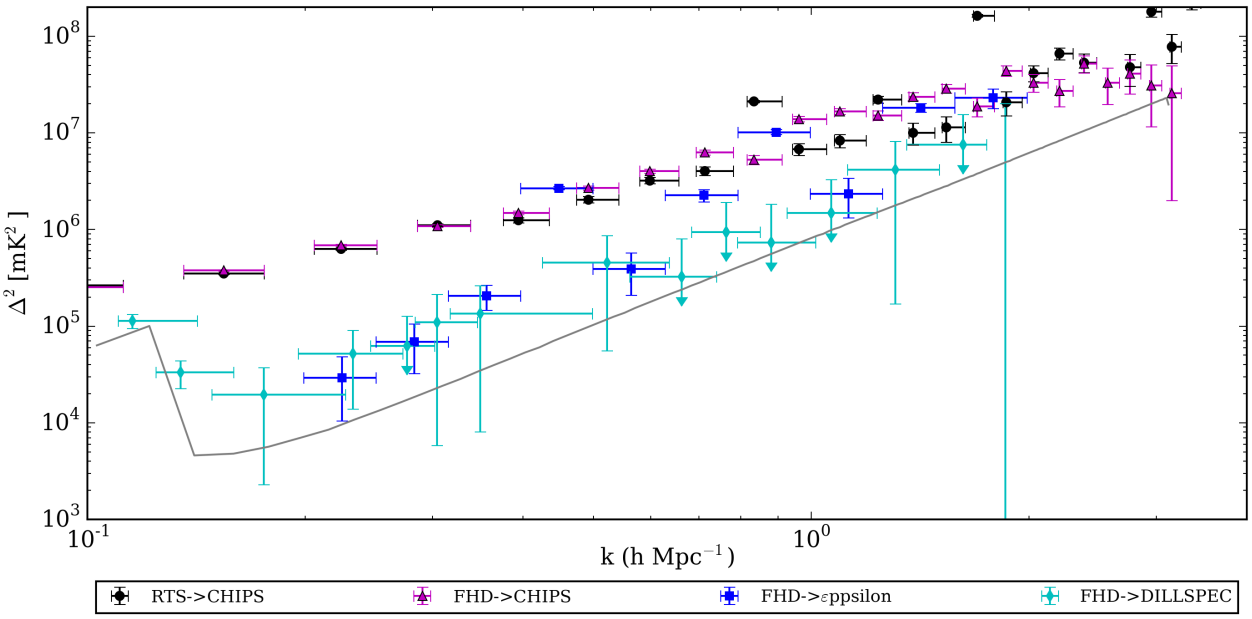

FIG. 6.— Power spectra averaged along shells of constant $|k|$ with 2σ errors. In three hours of data, four different methods demonstrate different kinds of limits on the power spectrum. Note that of the four pathways shown in Figure 4, only three are included here, but we have now included the power spectrum from Dillon et al. (2015) . The *εppsilon*power spectra of RTS outputs are excluded because error bars are not available. Many of the features visible in the 2D plots are also visible here: the excess in the CHIPS spectra is clearly visible as is a smaller excess in the *εppsilon* spectrum. The black line indicates 2σ bounds for points dominated by noise. Power levels for typical theoretical models are typically in the 5 to 10 mK^2 range across these k modes.

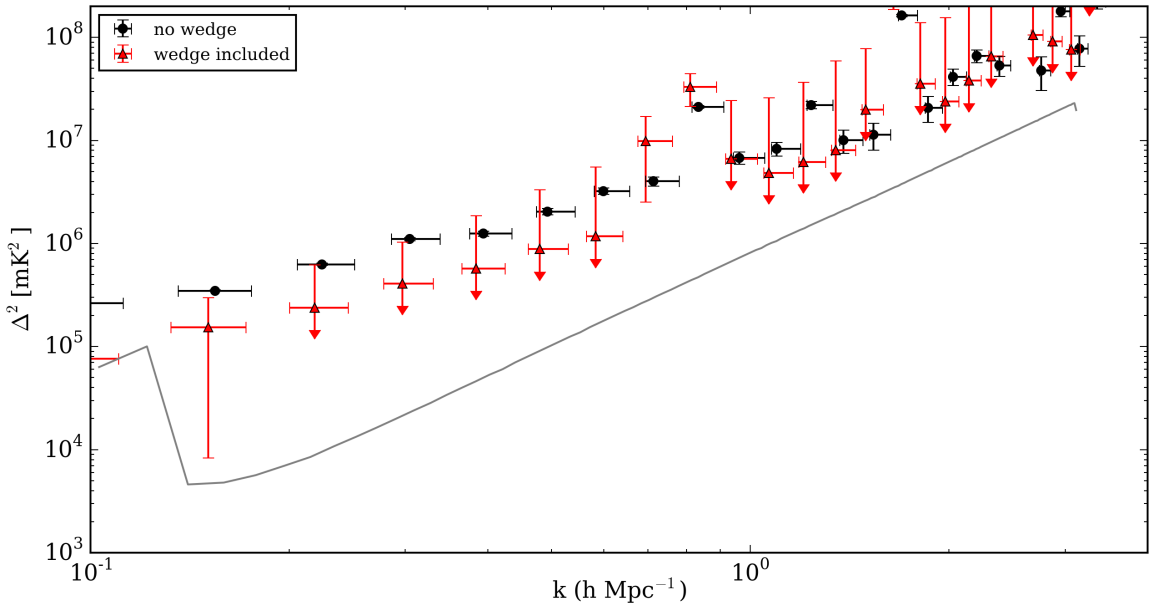

FIG. 7.— An example of the dramatic impact that weighting and covariance minimization has on the interpretation of error bars. Here we compare the RTS→CHIPS power spectrum from Figure 6 with that given by Trott et al. (2016). The latter was made with the same data but only 1/3 of the 30MHz band, and so slightly higher error bars. In both, CHIPS downweights by a model of foreground covariance formed by propagating a statistical model of confused sources. The only difference is that black excludes the wedge but red does not. When the wedge is included the modeled foregrounds in those k-space voxels dominate the covariance weights. Applying these weights essentially moves the foreground bias into the error bars and asserts that, given our best model of foregrounds, the power spectra are completely consistent with noise and foregrounds and do not provide evidence for a significant cosmological 21 cm signal.

deficiencies and have persisted into the final result or motivated additional fitting parameters resulting in higher signal loss as well as a vague disquiet. In addition to pipeline redundancy, forward modeling can provide some important checks, for example the absolute calibration of FHD and *εppsilon* described in Hazelton et al 2015, in prep, however the result is only as good as the model itself.

6. CONCLUSIONS

In this overview paper we have provided a top level view of foreground subtraction and power spectrum estimation methods of the MWA Epoch of Reionization project, described more completely in companion papers Hazelton et al 2015, in prep, Trott et al. (2016), and Dillon et al. (2015) . In this comparison we see that both foreground subtraction methods are able to reliably remove similar amounts of power. Differences between the images are smaller than the remaining residual foregrounds by a factor of 3.2, suggesting an overall ∼30% error on the aggregate calibration, foreground subtraction and imaging between the two pipelines. The power spectra of these foreground subtracted outputs agree on the scale and distribution of power, though with some differences in the leakage of power into the window. These differences are partly due to definition of error bars and whether they include just noise or also foreground terms.

Including foregrounds in the error calculation is a key exercise because it lets us answer more nuanced questions. Rather than simply: is the data inconsistent with a 21cm detection in the presence of noise? With CHIPS we can ask: Is the data inconsistent with a 21cm detection in the presence of noise and an a-priori foreground model? With EmpCov we can ask: is the data inconsistent with a 21 cm detection in the presence of noise and a foreground model fit to the data? These are all good questions.

21cm cosmology experiments have very wide fields of view, dense samplings, drift scanning observing and the reionization science levies a requirement for very high, 10,000:1, spectral dynamic range. All of this has necessitated the development of new algorithms for calibration and imaging, as well as the surrounding scaffolding to process thousands of hours of data to achieve this precision. This paper is the first step towards validating these pipelines and providing robust repeatable results.

This work was supported by the U. S. National Science Foundation (NSF) through award AST–1109257. DCJ is supported by an NSF Astronomy and Astrophysics Postdoctoral Fellowship under award AST–1401708. JCP is supported by an NSF Astronomy and Astrophysics Fellowship under award AST-1302774. CMT is supported by an Australian Research Council DECRA Award, DE140100316. This scientific work makes use of the Murchison Radio-astronomy Observatory, operated by CSIRO. We acknowledge the Wajarri Yamatji people as the traditional owners of the Observatory site. Support for the MWA comes from the U.S. National Science Foundation (grants AST-1410484, AST-0821321, AST-0457585, PHY-0835713, CAREER-0847753, and AST-0908884), the Australian Research Council (LIEF grants LE0775621 and LE0882938), the U.S. Air Force Office of Scientific Research (grant FA9550-0510247), MIT School of Science, the Marble Astrophysics Fund, and the Centre for All-sky Astrophysics (an Australian Research Council Centre of Excellence funded by grant CE110001020). Support is also provided by the Smithsonian Astrophysical Observatory, the MIT School of Science, the Raman Research Institute, the Australian National University, and the Victoria University of Wellington (via grant MED-E1799 from the New Zealand Ministry of Economic Development and an IBM Shared University Research Grant). The Australian Federal government provides additional support via the Commonwealth Scientific and Industrial Research Organisation (CSIRO), National Collaborative Research Infrastructure Strategy, Education Investment Fund, and the Australia India Strategic Research Fund, and Astronomy Australia Limited, under contract to Curtin University. We acknowledge the iVEC Petabyte Data Store, the Initiative in Innovative Computing and the CUDA Center for Excellence sponsored by NVIDIA at Harvard University, and the International Centre for Radio Astronomy Research (ICRAR), a Joint Venture of Curtin University and The University of Western Australia, funded by the Western Australian State government.

REFERENCES

Ali, Z. S. et al. 2015, ApJ, 809, 61
Beardsley, A. et al. 2013, Monthly Notices of the Royal Astronomical Society, 429, L5
Bhatnagar, S., Rau, U., & Golap, K. 2013, ApJ, 770, 91
Bowman, J. et al. 2013, Publications of the Astronomical Society of Australia, 30, 31
Bowman, J., Morales, M., & Hewitt, J. 2009, The Astrophysical Journal, 695, 183
Chapman, E. et al. 2013, Monthly Notices of the Royal Astronomical Society, 429, 165
Condon, J. J., Cotton, W. D., Greisen, E. W., Yin, Q. F., Perley, R. A., Taylor, G. B., & Broderick, J. J. 1998, The Astronomical Journal, 115, 1693
Datta, A., Bowman, J. D., & Carilli, C. L. 2010, The Astrophysical Journal, 724, 526
de Oliveira-Costa, A., Tegmark, M., Gaensler, B. M., Jonas, J., Landecker, T. L., & Reich, P. 2008, Monthly Notices of the Royal Astronomical Society, 388, 247, (c) Journal compilation © 2008 RAS
Dillon, J., Liu, A., & Tegmark, M. 2013, Physical Review D, 87, 43005
Dillon, J. et al. 2014, Physical Review D, 89, 23002
Dillon, J. S. et al. 2015, Phys. Rev. D, 91, 123011
Furlanetto, S. R., Oh, S. P., & Briggs, F. H. 2006, Physics Reports, 433, 181, elsevier B.V.
Górski, K., Hivon, E., Banday, A., Wandelt, B., Hansen, F., Reinecke, M., & Bartelmann, M. 2005, The Astrophysical Journal, 622, 759
Heald, G. H. et al. 2015, A&A, 582, A123
Hurley-Walker, N. et al. 2014, PASA, 31, 45
Jacobs, D. C. et al. 2011, The Astrophysical Journal, 734, L34
Jacobs, D. C. et al. 2013, ApJ, 776, 108
—. 2015, ApJ, 801, 51
Jelić, V. et al. 2008, Monthly Notices of the Royal Astronomical Society, 389, 1319, (c) Journal compilation © 2008 RAS
Koopmans, L. et al. 2014, in Proceedings of Advancing Astrophysics with the Square Kilometre Array (AASKA14). 9 -13 June, 2014. Giardini Naxos, Italy. (http://pos.sissa.it/cgi-bin/reader/conf.cgi?confid=215), 1

Lane, W. M., Cotton, W. D., van Velzen, S., Clarke, T. E., Kassim, N. E., Helmboldt, J. F., Lazio, T. J. W., & Cohen, A. S. 2014, MNRAS, 440, 327
Large, M., Cram, L., & Burgess, A. 1991, The Observatory, 111, 72
Liu, A., Parsons, A. R., & Trott, C. M. 2014a, Phys. Rev. D, 90, 023018
—. 2014b, Phys. Rev. D, 90, 023019
Liu, A. & Tegmark, M. 2011, Physical Review D, 83, 103006
Liu, A., Tegmark, M., & Zaldarriaga, M. 2009, MNRAS, 394, 1575, (c) Journal compilation © 2009 RAS
Lonsdale, C. J. et al. 2009, Proceedings of the IEEE, 97, 1497
Mauch, T., Murphy, T., Buttery, H. J., Curran, J., Hunstead, R. W., Piestrzynski, B., Robertson, J. G., & Sadler, E. M. 2003, Monthly Notice of the Royal Astronomical Society, 342, 1117
Mitchell, D. A., Greenhill, L. J., Wayth, R. B., Sault, R. J., Lonsdale, C. J., Cappallo, R. J., Morales, M. F., & Ord, S. M. 2008, IEEE Journal of Selected Topics in Signal Processing, 2, 707
Morales, M., Bowman, J., Cappallo, R., & Hewitt, J. 2006a, New Astronomy Reviews
Morales, M., Bowman, J., & Hewitt, J. 2006b, The Astrophysical Journal
Morales, M. F., Hazelton, B., Sullivan, I., & Beardsley, A. 2012, ApJ, 752, 137
Morales, M. F. & Wyithe, J. S. B. 2010, Annual review of astronomy and astrophysics, 48, 127, oise
Neben, A. R. et al. 2015, Radio Science, 50, 614
Noordam, J. E. 2004, Ground-based Telescopes. Edited by Oschmann, 5489, 817
Offringa, A. R., de Bruyn, A. G., Zaroubi, S., & Biehl, M. 2010, in RFI Mitigation Workshop, 36
Offringa, A. R. et al. 2014, MNRAS, 444, 606
—. 2015, PASA, 32, 8
Ord, S. M. et al. 2015, PASA, 32, 6
Ord, S. M. et al. 2010, Publications of the Astronomical Society of the Pacific, 122, 1353
Parsons, A. R. et al. 2014, ApJ, 788, 106
Parsons, A. R., Pober, J. C., Aguirre, J. E., Carilli, C. L., Jacobs, D. C., & Moore, D. F. 2012, The Astrophysical Journal, 756, 165
Pober, J. C. et al. 2014, The Astrophysical Journal, 782, 66
Pritchard, J. R. & Loeb, A. 2012, Reports on Progress in Physics, 75, 6901
Salvini, S. & Wijnholds, S. J. 2014, A&A, 571, A97
Sullivan, I. S. et al. 2012, The Astrophysical Journal, 759, 17
Sutinjo, A., O'Sullivan, J., Lenc, E., Wayth, R. B., Padhi, S., Hall, P., & Tingay, S. J. 2015, Radio Science, 50, 52
Tasse, C. et al. 2012, Comptes Rendus Physique, 13, 28, académie des sciences
Thyagarajan, N. et al. 2015a, ApJ, 804, 14
—. 2015b, ApJ, 807, L28
Thyagarajan, N. et al. 2013, The Astrophysical Journal, 776, 6
Tingay, S. et al. 2013, Publications of the Astronomical Society of Australia, 30, 7
Trott, C., Wayth, R., & Tingay, S. 2012, The Astrophysical Journal, 757, 101
Trott, C. M. et al. 2016, ArXiv e-prints
Vedantham, H., Shankar, N. U., & Subrahmanyan, R. 2012a, The Astrophysical Journal, 745, 176
—. 2012b, The Astrophysical Journal, 745, 176
Wayth, R. B. et al. 2015, PASA, 32, 25
Yatawatta, S. et al. 2013, Astronomy & Astrophysics, 550, 136
Zaroubi, S. 2013, in Astrophysics and Space Science Library, Vol. 396, Astrophysics and Space Science Library, ed. T. Wiklind, B. Mobasher, & V. Bromm, 45