A high reliability survey of discrete Epoch of Reionization foreground sources in the MWA EoR0 field

P. A. Carroll,[1★] J. Line,[2,3★] M. F. Morales,[1★] N. Barry,[1] A. P. Beardsley,[1,4]
B. J. Hazelton,[1,5] D. C. Jacobs,[4] J. C. Pober,[1,6] I. S. Sullivan,[1] R. L. Webster,[2,3]
G. Bernardi,[7,8] J. D. Bowman,[4] F. Briggs,[3,9] R. J. Cappallo,[10] B. E. Corey,[10]
A. de Oliveira-Costa,[11] J. S. Dillon,[11,12] D. Emrich,[13] A. Ewall-Wice,[11]
L. Feng,[11] B. M. Gaensler,[3,14,15] R. Goeke,[11] L. J. Greenhill,[16] J. N. Hewitt,[11]
N. Hurley-Walker,[13] M. Johnston-Hollitt,[17] D. L. Kaplan,[18] J. C. Kasper,[16,19]
HS. Kim,[2,3] E. Kratzenberg,[10] E. Lenc,[3,15] A. Loeb,[16] C. J. Lonsdale,[10]
M. J. Lynch,[13] B. McKinley,[2,3] S. R. McWhirter,[10] D. A. Mitchell,[3,20] E. Morgan,[11]
A. R. Neben,[11] D. Oberoi,[21] A. R. Offringa,[3,22] S. M. Ord,[3,13] S. Paul,[23] B. Pindor,[2,3]
T. Prabu,[23] P. Procopio,[2,3] J. Riding,[2,3] A. E. E. Rogers,[10] A. Roshi,[24]
N. Udaya Shankar,[23] S. K. Sethi,[23] K. S. Srivani,[23] R. Subrahmanyan,[3,23]
M. Tegmark,[11] Nithyanandan Thyagarajan,[4] S. J. Tingay,[3,13] C. M. Trott,[3,13]
M. Waterson,[25,26] R. B. Wayth,[3,13] A. R. Whitney,[10] A. Williams,[13]
C. L. Williams,[11] C. Wu[26] and J. S. B. Wyithe[2,3]

Affiliations are listed at the end of the paper

ABSTRACT

Detection of the epoch of reionization H I signal requires a precise understanding of the intervening galaxies and AGN, both for instrumental calibration and foreground removal. We present a catalogue of 7394 extragalactic sources at 182 MHz detected in the RA = 0 field of the Murchison Widefield Array Epoch of Reionization observation programme. Motivated by unprecedented requirements for precision and reliability we develop new methods for source finding and selection. We apply machine learning methods to self-consistently classify the relative reliability of 9490 source candidates. A subset of 7466 are selected based on reliability class and signal-to-noise ratio criteria. These are statistically cross-matched to four other radio surveys using both position and flux density information. We find 7369 sources to have confident matches, including 90 partially resolved sources that split into a total of 192 sub-components. An additional 25 unmatched sources are included as new radio detections. The catalogue sources have a median spectral index of -0.85. Spectral flattening is seen towards lower frequencies with a median of -0.71 predicted at 182 MHz. The astrometric error is 7 arcsec compared to a 2.3 arcmin beam FWHM. The resulting catalogue covers $\sim$1400 deg^2 and is complete to approximately 80 mJy within half beam power. This provides the most reliable discrete source sky model available to date in the MWA EoR0 field for precision foreground subtraction.

Key words: catalogues.

★ E-mail: pcarroll@uw.edu (PAC); jline@student.unimelb.au.edu (JL); mmorales@phys.washington.edu (MFM)

1 INTRODUCTION

Observations of the epoch of reionization (EoR) 21 cm neutral Hydrogen (H I) signal are one of the key science goals of the Murchison Widefield Array (MWA; Lonsdale et al. 2009; Bowman et al. 2013; Tingay et al. 2013). Astrophysical foreground sources are estimated to be 4–5 orders of magnitude brighter than this signal, presenting a major obstacle and motivating a careful dedicated survey to be used for both calibration and foreground power removal within the EoR analysis.

During the commissioning phase of the MWA and early development of the EoR analysis pipeline, the Molonglo Reference Catalogue (MRC; Large, Cram & Burgess 1991) was used to model the foregrounds. The MRC is complete to 1 Jy at 408 MHz or about 2 Jy at 182 MHz with an assumed average spectral index of −0.8. This is not only much shallower than desired, but large errors are introduced by this naive flux density extrapolation.

The MRC catalogue was later replaced by the MWA Commissioning Survey (Hurley-Walker et al. 2014), giving us a deeper and frequency-specific sky model that greatly improved calibration and foreground power subtraction. Unfortunately the MWACS does not cover the northernmost $5°$ of the EoR fields, and as an early product of a partially built instrument it contains large flux density and astrometric uncertainties. While the GaLactic and Extragalactic All-sky MWA survey (GLEAM; Wayth et al. 2015, Hurley-Walker, in preparation) was still underway, an extragalactic survey tuned to the requirements of EoR foreground subtraction in the MWA EoR0 (RA = 0, Dec = −27) field was initiated.

Processing MWA data is different in a number of ways from traditional interferometric radio data processing. Traditional radio analysis recipes are well tuned for arrays with narrow fields of view, stable beams, sparse instantaneous uv coverage, and limited source confusion. MWA data break these assumptions. The MWA primary beam is ~ 1400 deg^2 and changes with time as the field drifts through, the instantaneous uv coverage is excellent, and the continuum confusion limit is reached very quickly. Background sky regions reach a signal-to-noise ratio (SNR) ~ 10 in a 112 s integration.

For these reasons most MWA analyses have deconvolved sources on times scales of a few minutes or less to minimize time-dependent beam effects and leverage the snapshot uv coverage. Many different deconvolution algorithms may be chosen, though typically a pixel based algorithm such as CLEAN is used (i.e. source components are located at pixel centres and may take either positive or negative values).

Going from radio deconvolution products to a source catalogue is often performed by fitting sources in a restored image. For the MWA this has meant restoring the individual snapshot observations, mapping to celestial coordinates to remove widefield distortions, and then co-adding or mosaicking (e.g. Wayth et al. 2015). Combining snapshots through co-adding increases the SNR of real sources and drives down the amplitude of false side lobe sources as the array beam (PSF) rotates. While this reduces contamination, it removes information that is valuable for determining reliability, that is whether or not a source is in fact true and real. True sources should be detected consistently across all observations, while noise and side lobes will vary in time. This information is lost if source finding is not performed prior to image stacking. Reliability is a primary concern for the EoR foreground model as it will be used for both calibration and subtraction (Barry et al. 2016).

Various source extractors have been developed that isolate flux density peaks in an image, fit an assumed PSF or morphological

shape, and measure the integrated flux density and background rms (e.g. Hancock et al. 2012). This approach does well but is not without its limitations. Hopkins et al. (2015) demonstrate clear variations in the performance of 11 different source finders in terms of completeness, reliability, and measurement accuracy. Blended or extended sources were particularly troublesome despite the fact that all sources in the test images were artificially positioned at pixel-centres. In this work we take a novel approach to source finding that does not require the use of restored images or an assumed source shape. Individual snapshots are deconvolved using the Fast Holographic Deconvolution (Sullivan et al. 2012, Sullivan et al., in preparation) software package. FHD deconvolution is similar to CLEAN in that sources are iteratively removed using point-like source components with a CLEAN gain. However, FHD differs in using a full direction-dependent PSF, centroiding each source component (not fixed at pixel centres), and using only positive components. For a bright unresolved point source, FHD will create a set of positive source components, each with a floating-point precision position, tightly scattered about the actual source position ($\ll$ PSF FWHM). Diffuse sources will be modelled by a cloud of source components approximating the extended flux density distribution.

The FHD source components trace the sub-resolution flux density distribution of sources remarkably well. We can therefore identify and extract sources simply by spatially clustering the components regardless of shape. This does not require a restored image, nor the assumption of a PSF or morphological model. Machine-learning classification methods are then used to self-consistently assess source reliability. This process for source finding, measurement, and classification has been termed KATALOGSS (KDD Astrometry, Trueness, and Apparent Luminosity of Galaxies in Snapshot Surveys; hereafter abbreviated to KGS). We use the KGS results in combination with a typical signal-to-noise detection threshold and cross-matching to maximize the overall completeness and reliability of the final catalogue.

This paper presents these methods and the resulting source survey of the MWA EoR0 field. The observations and pre-processing are described in Section 2; the process of source finding and association across snapshots in Section 3; and the reliability classification is detailed in Section 4. In Section 5 we introduce the Positional Update and Matching Algorithm (PUMA) used to cross-match the catalogue to other radio surveys and fit for the power-law spectral index. The final catalogue is described in Section 6.6 along with measures of the astrometric accuracy, spectral index distribution, and completeness. We further identify 25 new sources previously undetected at radio frequencies and discuss potential associations in Section 6.8.

2 DATA AND PROCESSING

The MWA EoR0 field is centred at RA = 0 h and Dec = −27°, and was chosen because it has no bright complex sources in the primary field of view. The FWHM of the antenna beam is approximately 20°, but sources in the edges of the beam and first few side lobes are clearly visible and should be subtracted (Thyagarajan et al. 2015a,b; Pober et al. 2016). For this catalogue we concentrate on identifying sources in the primary beam but go out to the 5 per cent power point (nearly first beam null, ~ 1400 deg ^{2}). The data for this catalogue include 75 2 min snapshot observations (112 s consecutive integrations with 8 s gaps) from the night of 2013 August 23. The observations were made at 182 MHz with 31 MHz bandwidth and cover 2.5 h in total.

The MWA antenna can only point in discrete locations in elevation and azimuth (Tingay et al. 2013), leading to a 'drift-and-shift' observation pattern where the sky drifts through the antenna beam for approximately half an hour before the antenna is re-pointed to follow the field. These 75 snapshots cover 5 antenna pointings centred on field transit. The visibility data were averaged in time and frequency to 2 s and 40 kHz, and flagged for RFI with Cotter (Offringa et al. 2010). Due to the remoteness of the Murchison Radio Observatory, less than 3 per cent of the data were flagged for RFI.

Fast Holographic Deconvolution (FHD; Sullivan et al. 2012) was used for calibration, imaging, and deconvolution.[1] FHD is an efficient implementation of A-projection (Myers et al. 2003; Bhatnagar et al. 2008; Morales & Matejek 2009), and enables direction dependent beam corrections by using the antenna holographic beam pattern in gridding. During each deconvolution iteration, the brightest pixels are identified and the sources are fit with a Gaussian-approximated synthesized beam shape to determine the flux density. Unlike standard CLEAN, the position is centroided to a floating point location and all components are strictly positive-valued. A gain of 0.2 is applied, and the source components are forward modelled through the direction dependent instrument model to update the residual sky image. The deconvolution loop is stopped if the rms of the residual image increases or if the source fitting fails and no valid components can be extracted. Otherwise, deconvolution halts after a maximum of 500 iterations or a maximum of 3×10^5 components are subtracted.

In four of the 75 snapshots, deconvolution stopped early due to lack of convergence on a single bright and extended source (starburst galaxy, NGC 253). These four snapshots were excluded from the remaining analysis. Among the remaining 71 snapshots, the deconvolution threshold ranged from 45–76 mJy with an average of 57 ± 6 mJy. The background noise level of the snapshot images is stable over time, averaging 9–11 mJy pixel^{-1} at beam-centre.

3 SOURCE FINDING

The FHD source components trace the sub-resolution flux density distribution of sources remarkably well. We can therefore identify and extract sources simply by identifying clusters of components. This does not require a restored image, nor the assumption of a PSF or morphological model, thus avoiding potential sources of error inherent to producing a restored image and using an image-based source finder (see Section 1).

A driving factor in distinguishing true sources from contamination is the consistency of detection. True sources should appear in most snapshots, while false sources (noise peaks and side lobe sources) vary in time and should appear in few. This is a valuable metric for determining source reliability and is lost when images are co-added. Source finding is therefore performed individually for each 112 s snapshot observation to preserve this information.

DBSCAN (Ester et al. 1996), a density-based hierarchical clustering algorithm, is used to identify spatially isolated clusters of source components produced by FHD. DBSCAN works by first identifying local maxima in the density distribution of components and then building clusters hierarchically outward from these cores. The input parameters for DBSCAN are the neighbourhood radius within which a point is considered to be a part of the same cluster, and the minimum number of points required within that radius for

a new core to be formed. For this application the minimum number of components was set to one so as not to exclude sources with only a single component extracted, and the neighbourhood radius was set to the half-width-half-max (HWHM) of the PSF.

Approximately 5000 clusters were identified in each snapshot, for a total of 3.58×10^5 across all snapshots. Each cluster is considered a single detection. For each detection, we calculated the position centroid, standard deviation, and rms width of the radial scatter of the component positions. The flux density, S, was estimated according to equation (1), where S_i and g_i are the flux density and clean gain factor for ith component and N_{comp} is the total number of components. This corrects for the fraction the flux density that was not deconvolved for a point source.

$$S = \frac{\sum_{i=1}^{N_{comp}} S_i}{1 - \prod_{i=1}^{N_{comp}} (1 - g_i)} \tag{1}$$

A second round of spatial clustering was then performed to match source detections across snapshots and create a single set of source candidates with detection rates. In choosing a neighbourhood radius, we found that using the PSF HWHM was maximized the number of sources detected in all snapshots, while a radius of one-quarter the beam width maximized the number of sources detected uniquely in all snapshots (i.e. not blended with another source in the same snapshot).

To understand this, consider the case of a radio galaxy with two lobes separated by an angular distance close to the PSF width. Due to variations in the beam response and positional uncertainty between snapshots, the lobes may be clustered into a single sources in some snapshots. On cross-association, a larger clustering radius will merge these into a single source detected in all snapshots but with multiple detections in those where the lobes were individually resolved. A smaller radius will find three sources, each detected in only a fraction of the snapshots, and the total flux density will be double counted.

We therefore use the PSF HWHM to maximize the number of sources detected in all snapshots for reliability determination. Although this means that some close pairs will be blended, we later use the quarter-width radius to split those indicated as having multiple components through cross-matching (Section 5.2).

A total of 9490 clusters, or spatially isolated source candidates, were identified with detections in at least two snapshots. A roughly equivalent number were detected in only a single snapshot and discarded. For each cluster, if multiple detections were found within any snapshot, the flux densities were summed and the position centroided into a single equivalent detection for that snapshot.

A 3σ clip was applied to the flux density distribution of all detections in a cluster, where σ is the standard deviation of the distribution, to exclude spatially coincident noise or side lobe sources. The mean position and amplitude gain of the primary beam (hereafter referred to simply as beam response or beam power) were then calculated.

Since sources drift through the beam over time, and the rms noise scales inversely with the beam response, we weight the mean and standard deviation of flux density by the inverse of the estimated variance in the residual image at the source position. A local estimate of the pixel rms in a region around the source position is difficult given the low resolution of the images (1 pixel = 2.24 arcmin at beam centre). We instead use the estimated thermal component of the variance by fitting pixel rms (Jy beam^{-1}) as a function of beam response for each snapshot and using the rms value predicted by source's position in the beam.

---

[1] FHD code and documentation is available on GitHub at https://github.com/miguelfmorales/FHD.

We use both the standard deviation σ_S and standard deviation of the mean $\sigma_{\overline{S}}$ as measures of uncertainty because σ_S is poorly constrained for small N_{det} while $\sigma_{\overline{S}}$ is based on an indirect estimate of the thermal rms in each snapshot. These differing limitations give complementary measures of uncertainty that, along with other observables, can be used to learn how true sources, noise fluctuations, and side lobes sources behave. Rather than apply a cut based on a strict SNR detection threshold, we classify each source in terms of its overall reliability to inform our selection.

4 RELIABILITY CLASSIFICATION

In determining the reliability of a source, the signal-to-noise ratio, or SNR, is a valuable statistic. For example, a 5σ confidence level is a common choice for robust statistical significance of a detection, however this operates on the assumption of a Gaussian noise distribution that can be sufficiently measured or predicted. Non-thermal sources of noise can lead to false confidence in spurious detections. Particularly problematic in radio images are the side lobes of bright sources that result from imperfect calibration and deconvolution. These side lobes can be significantly brighter than the background rms noise. They may also be spatially coincident between consecutive snapshots due to the short integrations (small uv rotation) and low resolution. This false confidence and seemingly non-spurious behaviour make it challenging to identify contamination in an automated way.

We expect a non-negligible number of false sources contaminating the sample and consider a 5σ selection to be insufficient on its own to exclude the brightest and most problematic contaminants. Cross-matching to overlapping surveys can help to identify contamination, but this is limited by the reliability of the comparison survey and the incidence of false matches. Requiring all sources to be previously detected in another survey can also result in a loss of completeness.

Here we summarize a classification scheme developed to assess source 'trueness' and measurement reliability in a self-consistent way. The resulting classifications are used in conjunction with cross-matching to inform the final catalogue selection. Details of the classification steps and intermediate results can be found in Appendix A for the interested reader.

Machine learning based classification algorithms require a set of input features that describe the population. Feature engineering is the process of scaling, combining, or otherwise manipulating fundamental parameters to increase the predictive potential of the model. The process of feature engineering and selection is somewhat of an art, driven by domain insight to the question at hand as much as the data available.

Features input to complex models are typically developed and honed through many iterations of trial and error. We ultimately define nine features based on observable measurements that result in a well-modelled distribution.

(i) *Log flux density*. The log of the weighted mean flux density of all source detections, $\log_{10}(S/\mathrm{Jy})$.

(ii) *Log signal to noise*. The log of the ratio of the mean flux density to the standard deviation, $\log_{10}(S/\sigma_S)$.

(iii) *Log signal to noise of the mean*. The log of the ratio of the mean flux density to the standard deviation of the mean, $\log_{10}(S/\sigma_{\overline{S}})$.

(iv) *Number of detections*. The number of snapshots in which a source was detected, N_{det}.

(v) *Expected number of detections*. The estimated cumulative probability that the mean source flux density lies above the deconvolution limit in each snapshot, N_{exp}.

(vi) *Reliability metric*. We define a normalized reliability metric that takes a value between 0 and 1, as $r_{det} = N_{det}/\sqrt{71\,N_{exp}}$. This is designed to down-weight the relative reliability of sources that drift out of the field (i.e. 71 of 71 expected is more reliable than 2 of 2 expected).

(vii) *Local density*. The number density of sources within a $1°$ radius of the source candidate $\rho_N\,(\pi\,\mathrm{deg}^2)^{-1}$.

(viii) *Distance to brightest neighbour*. Distance to the brightest source within a $1°$ radius of the source candidate, d_{bright} (deg).

(ix) *Flux density ratio to brightest neighbour*. The flux density ratio between the source candidate and the brightest neighbour within a $1°$ radius, S/S_{bright}.

The final three features offer additional information that help to differentiate likely side lobe sources which typically occupy regions of high number density in close proximity to a much brighter source. Principal component analysis was then used to reduce the parameter space from nine dimensions to three. Three components explain 83 per cent of the total feature variance and allow for simple visualization and model fitting. The 17 per cent information loss is later regained.

A model consisting of 10 independent three-dimension Gaussian components was fit to the data in the reduced parameter space. 10 components were found to model the distribution well without over-fitting. Each source candidate was labeled by the Gaussian component it was most probably associated with. The Gaussian labels were then used to train a more robust ensemble classifier. This allowed the classification boundaries to adjust according to the true feature distribution rather than the forced Gaussian approximation.

The classification labels were ordered according to the median r_{det}. Lower numbered classes therefore tend to be more reliable ($R0 - R2$) while mid-range classes tend to be fainter or further from field-centre ($R3 - R6$). The majority of false positives appear to be captured in classes $R8$ and $R9$ along with the faintest sources that are detected too few times to be considered reliable. A more detailed interpretation is given in Appendix A.

We make an initial cut on the set of 9490 source candidates to include only those detected with high confidence ($S > 5\sigma_S$ and $S > 5\sigma_{\overline{S}}$) or reliability ($R_{class} < 7$; all of which meet the $S > 5\sigma_{\overline{S}}$ criteria). This reduces the sample to 7466 source candidates. Looser criteria may be explored in the future in terms of the effectiveness of foreground power subtraction.

5 RADIO CROSS MATCHING

To check the reliability of the sources, a cross match was performed with the following catalogues: the 74 MHz Very Large Array Low Frequency Sky Survey redux (VLSSr; Lane et al. 2014); the 408 MHz Molonglo Reference Catalogue (MRC; Large et al. 1991); the 843 MHz Sydney University Molonglo Sky Survey (SUMSS; Mauch et al. 2003); and the 1.4 GHz NRAO VLA Sky Survey (NVSS; Condon et al. 1998). A summary of these catalogues is given in Table 1.

The EoR0 field is centred at RA 0h and Dec $-27°$ and the KGS sources reach a depth of $\sim$60 mJy at 182 MHz in the centre of the field. This coverage does not match well to any one of the comparison surveys. The VLSSr and NVSS catalogues cover the northern half of our field ($\gtrsim -30°$) while SUMSS covers the southern half

Table 1. A summary of the catalogues matched with PUMA to the KGS outputs including the reported or estimated completeness limit.

Survey	ν (MHz)	N_{sources}	Dec	PSF FWHM	S_{complete}
VLSSr	74	92 696	$\delta > -30°$	75″	~ 1 Jy beam^{-1}
MRC	408	12 141	$\delta_{1950} > -85°$	$\sim 3'$	~ 1 Jy beam^{-1}
SUMSS	843	211 050	$\delta < -30°$	45″	18 mJy beam^{-1}
NVSS	1400	1773 484	$\delta > -40°$	45″	2.5 mJy beam^{-1}

($\lesssim -30°$). The MRC catalogue covers the full sky area but is only complete to 1 Jy at 408 MHz or ~ 2 Jy at 182 MHz.

Cross matching was performed using the Positional Update and Matching Algorithm (Line et al. 2016). PUMA uses a combination of positional and spectral information to statistically test whether sources from multiple surveys in close proximity to one-another are true matches. We briefly describe the matching steps here, but for a full explanation of each step please refer to Appendix B and the code documentation.[2] Note this process is automated through Section 5.2 where we manually investigate outliers.

5.1 PUMA

Initially, PUMA attempts to match sources purely by position. In the first stage, a positional cross match is performed using STILTS (Taylor 2006). All sources within a radius of 2.3 arcsec from the base KGS source are selected. The choice of this radius is somewhat arbitrary (equal to the PSF FWHM) and intentionally liberal. For each cross match result, the probability P that all catalogues are describing the same source is calculated following Budavári & Szalay (2008), taking account of the positional errors. At this point, if a KGS source is matched to only one source from any catalogue and $P > 0.95$ it is accepted without further investigation. This is labelled as an `isolated` match.

The quoted uncertainties on position, particularly for blended and complex sources, are not necessarily accurate or directly comparable between surveys. If $0.8 < P < 0.95$, or all cross-matched sources lie within the resolution of the MWA (i.e. half of the original STILTS search radius), we investigate the spectral energy distribution (SED) by fitting a power-law spectral model, $S \propto \nu^\alpha$, using weighted least squares. If the fit is good, the source is accepted as an `isolated` match. Note that if a match is only found in one other catalogue, this fit always passes as there can be no residuals. Steps are taken later in Section 5.2 to account for any issues that could arise here.

Multiple matches to a single catalogue may occur due to confusion at the lower resolution or coincidental false source contamination. In the case where multiple sources from a comparison catalogue are matched to a single KGS source, PUMA first attempts to remove any false matches by fitting the spectral model to each possible combination of sources. If one match combination has smaller residuals than all others, as well as having $P > 0.95$, it is accepted as the `dominant` match.

If no `dominant` match is found, it is possible that a source is resolved into multiple components in the higher resolution catalogues. This is the common case for radio galaxies and star-forming galaxies with structure that is unresolved by the MWA. To test this, the spectral model is fit to the cumulative flux density of the matches at each frequency. If the fit is good, the source is accepted as a `multiple` match.

---

[2] PUMA code is open source. The code and documentation can be found here – https://github.com/JLBLine/PUMA.

Table 2. The total number and percent of catalogue sources in each PUMA decision category, and the number and percent of each category for which the match was flagged and manually modified in Section 5.2. The majority of sources (87 per cent) are matched to a single counterpart in other catalogues (`isolated` or `dominant`) with a 98.6 per cent automatic success rate. When confusion occurred (`mulitple` matches), PUMA chose the proper match combination in 84 per cent of cases. Most modifications were required for complex and extended sources.

Match result	Count	(per cent)	Modified	(per cent)
`isolated`	6119	(82.8)	75	(1.2)
`dominant`	310	(4.2)	11	(3.5)
`multiple`	940	(12.7)	153	(16.3)
`none`	25	(0.34)		
Total	7394	(100)	239	(3.2)

Table 2 details the number of each type of PUMA match decision. The flux density and uncertainty of matched sources are included in the KGS catalogue along with the measured broad-band spectral index (SI). The spectral index distribution is discussed in Section 6.4. The position of matches to the NVSS or SUMSS catalogues are also included for reference (a flux density weighted mean position is reported for `multiple` matches) and used to assess astrometric precision in Section 6.5.

5.2 Visual inspection

To check the robustness of the PUMA decisions we visually inspected any potentially suspect matches or atypical sources. These include 66 sources automatically flagged by PUMA when a confident match decision could not be made and 45 sources with a STILTS match that was automatically rejected by PUMA. For the sake of reliability, we also double check PUMA accepted matches that we manually flagged as outliers. These include all 900 sources accepted by PUMA as a `multiple` match and 205 defined as having either (a) spectral index in the 1 per cent tails of the distribution, $\alpha < -1.46$ or $\alpha > -0.17$; (b) positional offset from NVSS or SUMSS $> 3\sqrt{\sigma_{\mathrm{RA}}^2 + \sigma_{\mathrm{Dec}}^2}$.

In addition to visualizing the catalogue information and PUMA results, we looked at postage stamps of VLSSr, SUMMS, NVSS, and MWA images. Where appropriate, the PUMA decision or match information was modified. This could include removing matched sources that appeared spurious or ignoring a catalogue in a `multiple` match that appeared to be missing a source visible in its image. Fig. 1 shows an example of a source where both the catalogue information and images were inspected and used to modify the catalogue match.

By selecting outliers, we were able to discard 24 false `isolated` matches. These were typically sources with poor reliability, visually identified as bright side lobes, and coincidentally matched to either a true source or to apparent side lobe contamination in a comparison survey image. We erred on the side of reliability in these decisions.

Approximately 10 per cent of `multiple` matches were able to be deconstructed into two or more components. In Section 3, a radius of one-quarter the beam width maximized the number of sources detected uniquely in all snapshots. If multiple source candidates were found by using the tighter clustering radius, these were similarly cross-matched and substituted in manually if the overall result improved. A total of 90 sources were replaced with 192 counterparts. The reliability class of replacement sources was independently predicted using the original classifier.

By specifically targeting outliers, we not only discarded contamination but noted many interesting sources. For example, the unique
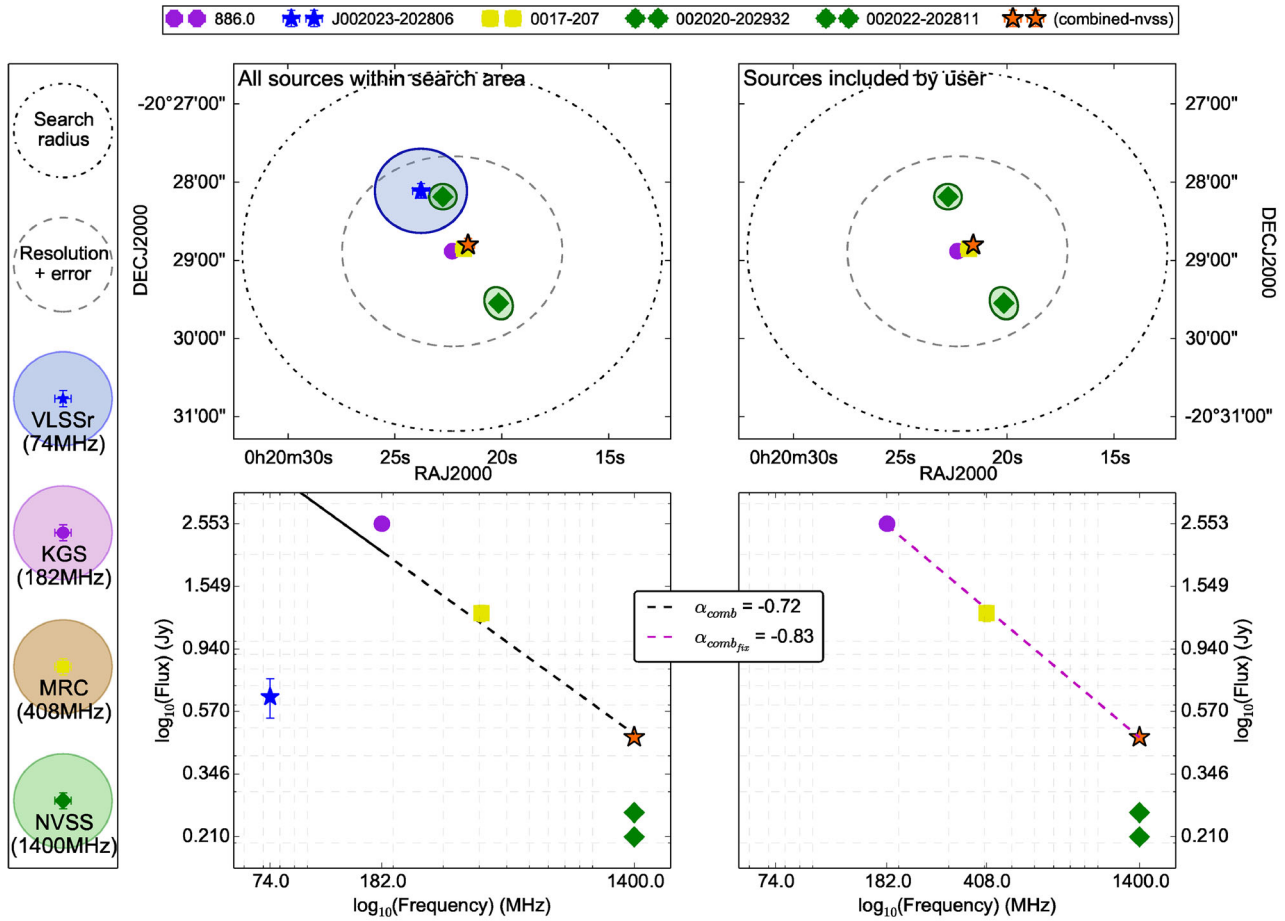
(a) Example of visualized position (top) and SED (bottom) information for a complicated match before (left) and after (right) manual modification. Ellipses indicate the reported major/minor axis and position angle.
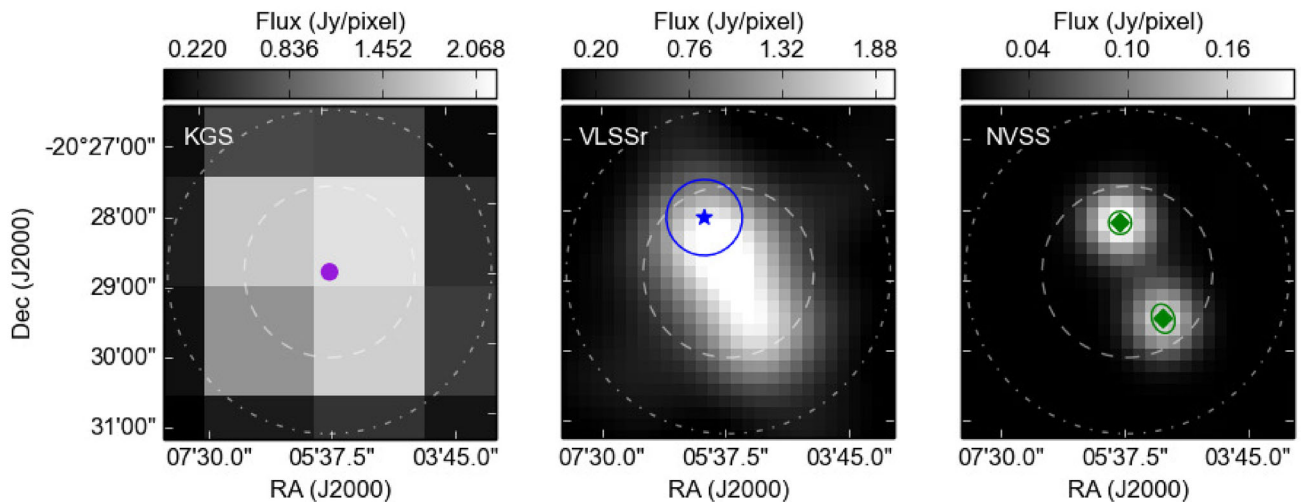
(b) Example postage stamp images inspected for complicated matches. The white dash/dotted circles correspond to the search radius and resolution+error as indicated in (a)

Figure 1. To investigate flagged matches, catalogue information on source position, shape, and flux density were plotted (a) and, for complicated sources, postage stamp images were obtained and compared (b). In this example of a `multiple` match, two sub-resolution sources appear present in both the VLSSr and NVSS images, but one is missing from the VLSSr catalogue. The centroid position of the NVSS sources agrees well with the KGS and MRC positions. The flux density of the NVSS sources are combined and the VLSSr source is excluded.

morphology of NGC 7793 resulted in its identification as a near face-on spiral galaxy in the Sculptor group. A subsequent search on the positions of other group members revealed detections of NGC 253 and NGC 55. NGC 253, the Sculptor galaxy, is the sixth brightest source in the catalogue and its extended morphology will require a more complex treatment in the foreground model. The low frequency emission from the Sculptor group galaxies will be further investigated by Kapinska et al. (in preparation). Other noteworthy examples of morphological and spectral outliers are presented in Section 6.6.

5.3 False sources and new detections

There are 167 source candidates not matched to another catalogue within the initial STILTS search radius, and another 12 STILTS matches that were rejected by PUMA. The majority of these are classified $R7 - R9$ and visual inspection supports their exclusion as noise or side lobe contaminants. 25 are included in the final catalogue. Of these, 20 are reliably classified $R0 - R6$. Five appear to be extended emission blended with a brighter source, but are reliably detected independently of that source.

We chose to include five faint sources classified $R8$ after careful consideration. These are interesting and illustrative. Detected in few snapshots, the mean and standard deviation are poorly constrained. The flux density is also likely to be over-estimated due to Eddington bias (see Section 6.2) near the detection threshold. In combination, these effects seem to have resulted in artificially high SNR measures, allowing their inclusion in the source candidate sample. We chose to keep these simply because they appear to be real in the images and were deemed deserving of followup. Many similar sources did not make the initial candidate selection. In terms of a foreground model, the level of contamination we risk through their inclusion is negligible.

The new source detections are explored further in Section 6.8 where we consider potential associations in other wavebands.

6 THE KGS EoR0 CATALOGUE

Since nearly all sources have accurate matches to other catalogues, we can explore the completeness, astrometric precision, spectral index distribution, and flux scale reliability of the catalogue.

6.1 Flux scale

To investigate how each matched catalogue contributed to the fitted spectral index, the flux density at each frequency was extrapolated using the fitted parameters. To be sure of a true matched SED, only `isolated` sources were used. A ratio between the reported catalogue flux density and extrapolated flux density was then calculated, as shown in Fig. 2.

On average there is no significant bias in the distributions, however it is interesting to note the width and skewness in these distributions. The VLSSr and KGS skew somewhat low and the MRC skews somewhat high. For sources detected in more than two catalogues, the spectral index fit used the quoted flux densities and uncertainties of the comparison catalogues. NVSS and SUMSS have lower uncertainties than the lower frequency catalogues and VLSSr has the largest. This is evident in the spread of the flux density ratio distributions; NVSS has a tight distribution centred at one, whereas VLSSr has a broader distribution. Clearly NVSS is being fit preferentially over the other catalogues. Although the median values are all consistent with unity, systematic effects are clearly present.
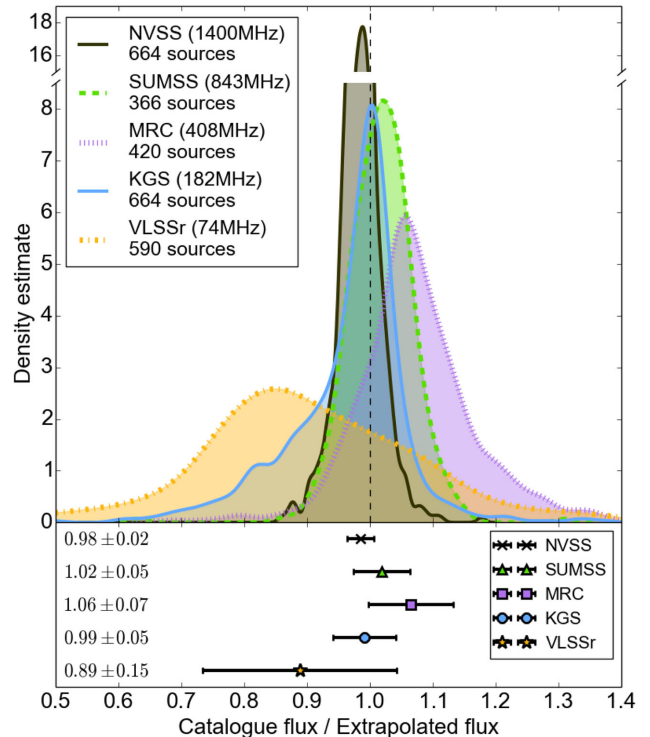
Figure 2. The ratio between observed flux density and extrapolated flux density from a fit to the SED is shown for every time a catalogue appeared in a match with at least two other catalogues for `isolated` sources. The upper panel shows a univariate kernal density estimation of each distribution (note broken y axis due to the sharp peak in the NVSS ratio distribution), while the lower panel shows the median and median absolute deviation of each distribution. The KGS spectral index agrees very well with no indication of flux bias on average.

It is difficult to distinguish catalogue flux biases from intrinsic spectral curvature effects, but it appears that only a sub-population of sources are affected rather than there being an overall shift in the distribution. A systematic under or over estimation due to the original flux scaling or calibration could account for this. No attempt was made to match flux scales across catalogues since all catalogues except for the VLSSr are tied to or derived from the Baars scale (Baars et al. 1977) based on measurements of Cassiopeia A. The VLSSr is tied to the RCB scale Roger, Costain & Bridle (1973), which is considered to be more accurate at low frequencies. Due to the low weighting of the VLSSr data points and the fact that all but three VLSSr matches are also matched to the NVSS, this scale difference negligibily impacts the SED fits and overall SI distribution. Nonetheless, in the following sections we divide all VLSSr flux densities by a factor of 1.1 to place them on the Baars scale and increase the flux density uncertainty by 5 per cent following Lane et al. (2014).

Alternatively, if a significant portion of SEDs display some degree of intrinsic positive curvature, the ratio distributions would be impacted predictably. The lowest frequency flux densities (VLSSr and KGS) would typically be overestimated by a power-law fit, while the central frequencies (MRC and SUMSS) would be underestimated. This is consistent with the flux density ratio distributions observed, however a more careful treatment is needed to conclude intrinsic curvature over systematic effects. We do not find evidence for flux scale bias relative to the comparison catalogues.

6.2 Eddington bias

Source flux densities near the detection threshold will be systematically high due to Eddington bias. At low apparent flux densities, statistical fluctuations below the sensitivity limit go undetected resulting in an overestimation of the true mean (Eddington 1913). To estimate a correction for this effect, we numerically solve for the true flux density that would most likely result in the observed average over all detections.

The probability density function of a source detection is assumed to be Gaussian, centred on the true flux density and with standard deviation equal to the background rms. The expectation value of the measurement is then the mean of the probability density function above the detection threshold. For each observation, the detection threshold is estimated as a function of position in the beam and the background rms is estimated from the residual images within a 20×20 pixel box around the source position.

The reported source flux density is the weighted mean of the snapshot detections. Using this as the initial guess for the true flux density of the source, we find the weighted mean of the expectation values for all detections as described above, and numerically solve for the true flux density that minimizes the difference between this expected mean and that observed.

We find the bias affects sources with apparent flux density ($S \cdot \text{beam}$) below about 100 mJy, but few sources are changed by more than 10 per cent. To gauge the accuracy of the true flux density estimates, we looked at the 182–1400 MHz spectral index distribution for `isolated` sources before and after correction. Above $S \cdot \text{beam} = 100$ mJy (Fig. 3, right), there is no significant difference in the distributions before and after correction. Below this threshold, Eddington bias is evident in the shift of the spectral index distribution towards more negative values (Fig. 3, left). After correction, the median value agrees well with that of the unbiased distribution.

Of the 2548 sources that would be corrected, the difference in flux density exceeds the standard error σ_S for only 177 sources (50 exceed $3\sigma_S$). For this reason, Eddington bias is not a major concern. The overall median SI change is small, from -0.850 to -0.843. However, the bias is significant for individual sources within the catalog, particularly at low apparent flux density. It is important to note that the validity of the correction factor for any source is contingent on there being a large enough number of detections that the mean is sufficiently constrained. It is therefore the least reliable for the most affected sources.

The catalog contains a column EB_corr that may be multiplied by the flux density to approximate a correction for Eddington Bias. The correction factor is 1 by default for a source if ($S \cdot \text{beam}$) > 100 mJy or if it is a `multiple` match (i.e. not a point source). We recommend adding the reported flux density uncertainty in quadrature with the absolute difference between the original and corrected flux density values. In the following sections we use the corrected flux density estimates.

6.3 Completeness

In order to assess completeness, we compare source counts to the predicted source counts of the NVSS and VLSSr surveys projected to 182 MHz (Fig. 4). Counts are considered only within the overlapping survey areas and each catalogue flux density is projected to 182 MHz using the median two-point spectral indexes $\alpha_{74}^{182} = -0.68$ and $\alpha_{182}^{1400} = -0.85$ for all matched `isolated` sources.

Because the sensitivity, and thus detection threshold, goes as 1/beam, the overall completeness falls off steadily below $\sim$1 Jy compared to the NVSS. Within the half-power point, we find the catalogue is complete to approximately 80 mJy. Source counts appear to be comparable to the VLSSr above 200 mJy, below which KGS sources are likely to go undetected at 74 MHz within the overlapping footprint.

6.4 Spectral index distribution

The spectral index (SI) distribution found by the match performed in Section 5 is shown in Fig. 5. We find an overall median of -0.85 and no bias is seen by including or excluding sources with poor spectral fits $\chi^2_{\text{red}} > 2$ in this estimate.

In Fig. 6, we show the SI distribution for all two-point SI measurements among matches to `isolated` KGS sources detected at an average beam power greater than 0.5 and flux density $S >$ 200 mJy. The two-point median spectral index is seen to range considerably, from -0.59 to -0.95, with a trend towards steeper
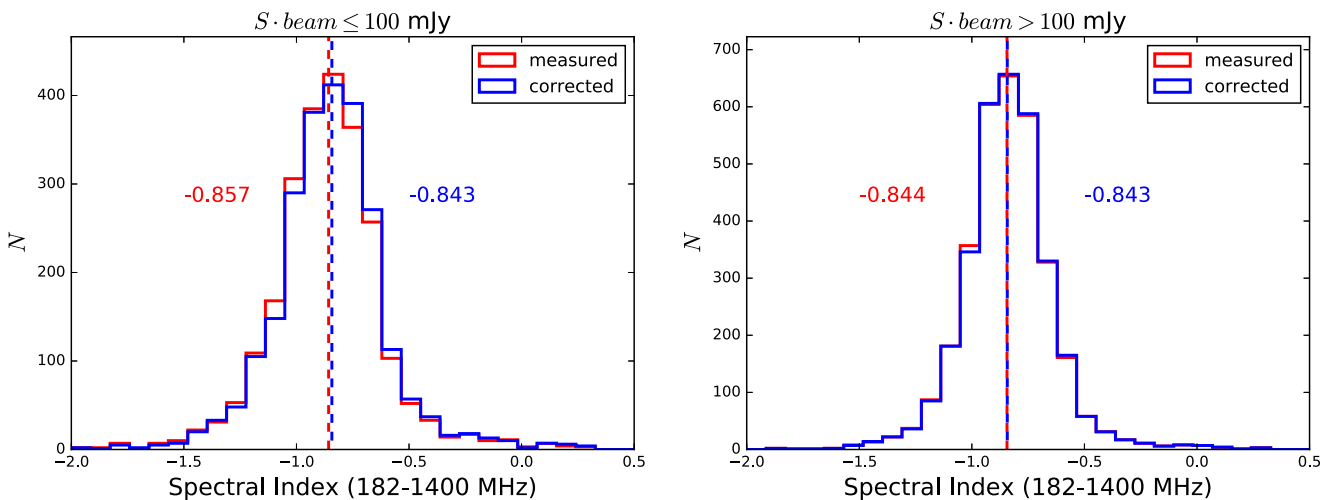

Figure 3. The two-point spectral index distributions of `isolated` sources with matches to NVSS at 1400 MHz before (red) and after (blue) correction for Eddington bias. Sources with apparent flux density $S \cdot \text{beam} < 100$ mJy (left) tend to be the most affected. After correction, the median spectral index agrees well with sources above this threshold (right) for which the estimated correction is negligible.
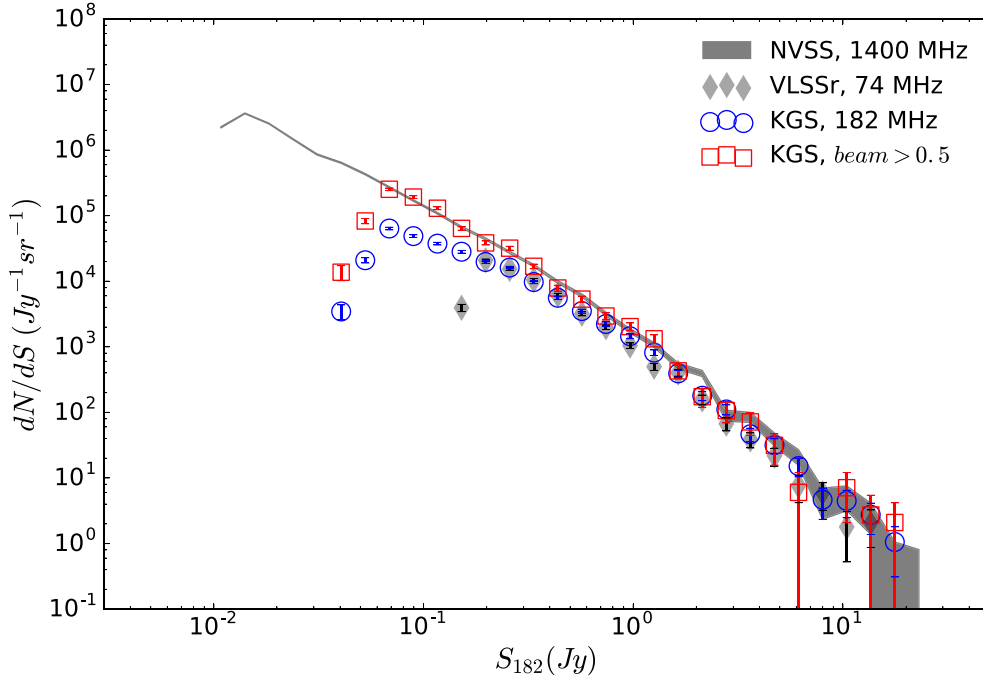
Figure 4. Differential source counts of the full catalogue (blue circles) compared to the NVSS (grey line) and VLSSr (grey diamonds). Counts are made within the overlapping footprint and the flux density values are projected to 182 MHz using the median two-point spectral index of all `isolated` matches. Completeness falls off as 1/beam below 1 Jy, however within half-beam (red squares) the catalogue is complete to approximately 80 mJy. The VLSSr shows comparable completeness to $S_{182} = 200$ mJy below which KGS sources are more likely to go undetected in the at 74 MHz.
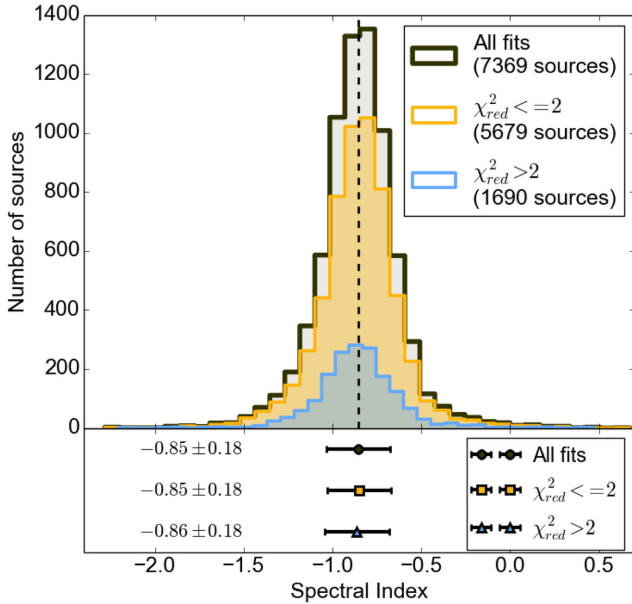
Figure 5. The SI distribution derived by matching to VLSSr, MRC, SUMSS and NVSS. The full sample is shown in black, with good spectral fits ($\chi^2_{red} \leq 2$) shown in gold and poor spectral fits ($\chi^2_{red} > 2$) shown in blue. The mean and standard deviation of SI distributions are shown in the lower panel.

spectra at higher frequencies. The lowest frequency measurements α^{182}_{74} and α^{408}_{182} give an average of -0.70 at 182 MHz.

By fitting a second order polynomial to the subset of 883 `isolated` sources detected in three or more catalogues, we predict a median 182 MHz spectral index of -0.71 with an interquartile range between -0.88 and -0.53 (the mean and standard devia-

tion are -0.71 ± 0.32). These results are consistent with Offringa et al. (2016), who directly measure the sub-band (132–198 MHz) spectral index for a highly comparable set of sources in the centre of the MWA EoR0 field. They find a median of -0.70 at 168 MHz (mean of -0.687 ± 0.275). For comparison, the Low Frequency Array (LOFAR) MSSS MVF survey (Heald et al. 2015) finds median values of $\alpha^{158}_{30} = -0.66$ and $\alpha^{158}_{119} = -0.77$ (mean values of -0.60 and -0.70, respectively) among 628 sources with $S_{150} > 200$ mJy.

6.5 Astrometry

Cross matching allows us to approximate the best astrometric position of a source based on higher frequency, higher resolution counterparts. Of the 7369 matched sources, all but three include a match to either the NVSS or SUMSS catalogues.

Fig. 7(a) shows the distribution of offset distances in RA and Dec from the NVSS or SUMSS match to `isolated` sources. The median offset is $\sim$10 arcsec in either dimension. While this is less than the median errors $\sigma_{RA} = 19$ arcsec and $\sigma_{Dec} = 15$ arcsec, a north-eastward systematic bias is clearly apparent. This is illustrated by a vector field in Fig. 7(b).

The source of this offset can be traced to errors in the MWACS catalogue used for calibration. Considering MWACS `isolated` matches on position, the median offsets for MWACS within half-beam are ΔRA $= -8$ arcsec and ΔDec $= -10$ arcsec. KGS offsets outside of half-beam are much larger in RA, but we suspect the root of the problem is in the calibration.

The bias is found to be well modelled by a second order polynomial as a function of (RA, Dec) position. The modelled bias is used to approximate a positional correction. The distribution of offsets after correction is shown in Figs 7(c)–(d). The median
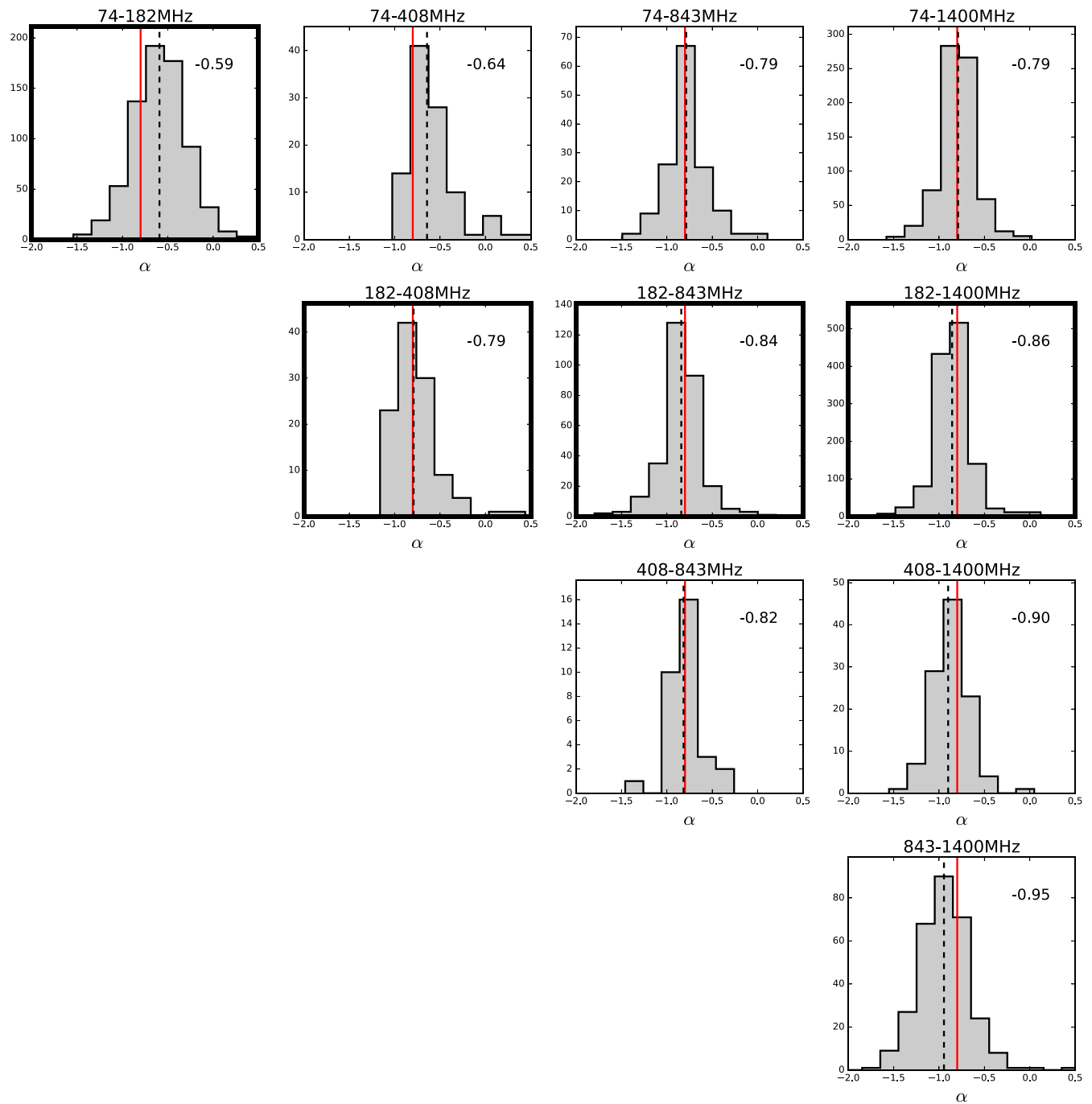
Figure 6. The two-point SI distributions for all catalogue matches to `isolated` KGS sources with $S > 200$ mJy and beam > 0.5. Bold axes indicate distributions using the KGS 182 MHz flux density. The median values are marked by dashed black lines and the red lines mark -0.8 for reference. The median becomes increasingly negative towards higher frequencies.

offset is reduced to <1 arcsec compared to ~ 10 arcsec in either dimension. The median absolute offsets are $|\Delta \text{RA}| = 4$ arcsec and $|\Delta \text{Dec}| = 3$ arcsec. In the catalogue, we report the bias-corrected KGS position as well as the matched catalogue (NVSS or SUMSS) position for comparison.

6.6 The catalogue

Table 3 lists a subset of the catalogue selected to represent a diverse sample. The PUMA position and spectral cross-match in-

formation for these are illustrated in Fig. 8. The complete catalogue of 7394 sources is included in the electronic supplement. The columns are:

(i) *Name*. Source name.
(ii) *RAJ2000*. Corrected mean J2000 Right Ascension in degrees of the snapshot detections.
(iii) *DECJ2000*. Corrected mean J2000 Declination in degrees of the snapshot detections.
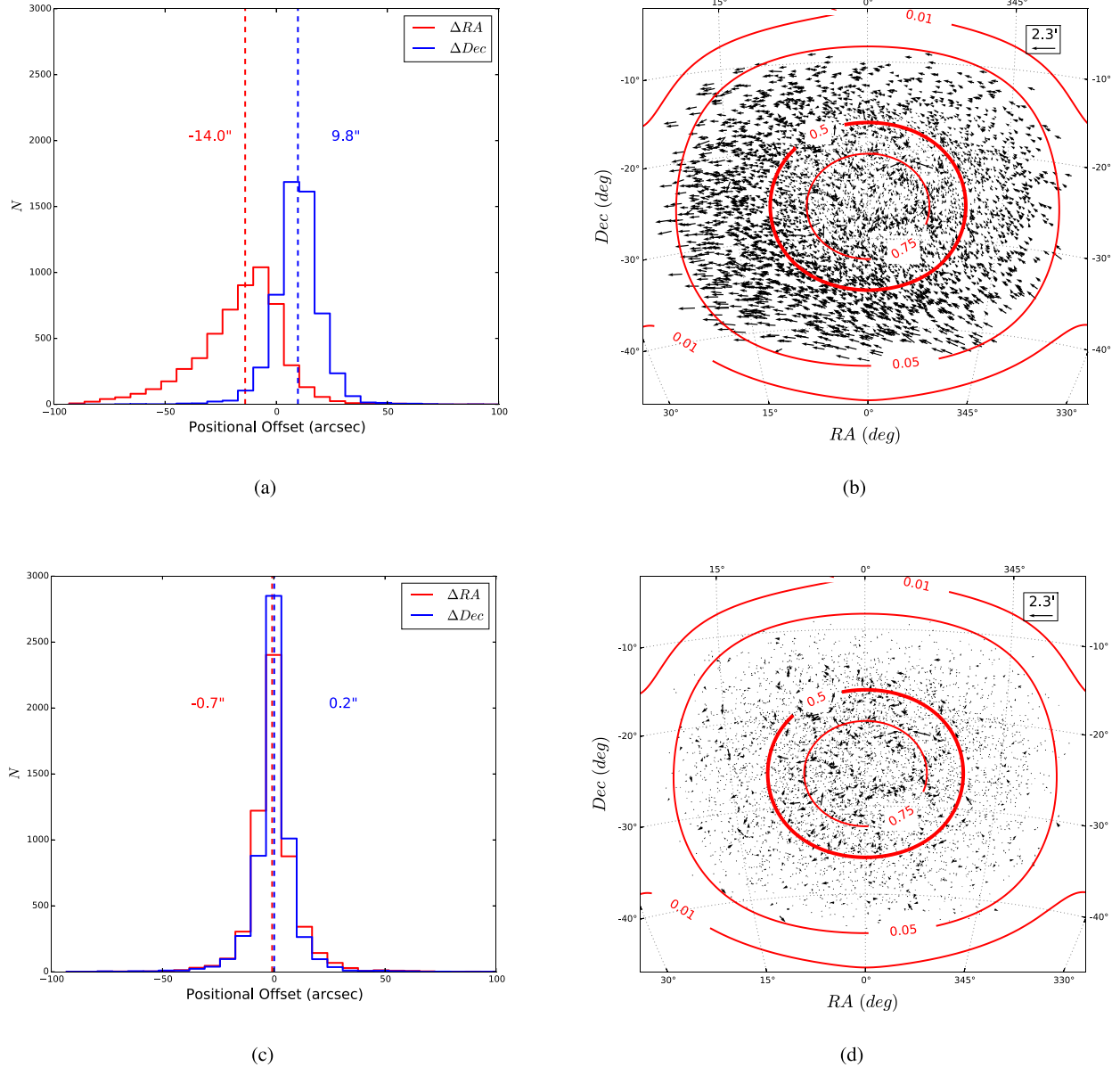
Figure 7. Left: the distribution of positional offsets compared to either NVSS or SUMSS counterparts for `isolated` sources before (a) and after (c) correcting for the north-eastward systematic bias. Right: vector fields illustrating the offset magnitude and direction before (b) and after (d) correction. The average analytic MWA beam power across all snapshots is contour plotted for reference.

(iv) *e_RAJ2000*. Standard deviation of the measured Right Ascension for all snapshot detections in arcseconds.

(v) *e_DECJ2000*. Standard deviation of the measured Declination for all snapshot detections in arcseconds.

(vi) *S_182*. Weighted mean integrated 182 MHz flux density measured in Jy.

(vii) *e_S_182*. Standard deviation of the measured flux density for all snapshot detections in Jy.

(viii) *EB_corr*. Estimated flux density correction factor for Eddington Bias.

(ix) *R_class*. The reliability classification (0-9).

(x) *Beam*. The mean relative beam response (0–1) at the source location.

(xi) *N_det*. Number of snapshots the source was detected in.

(xii) *Match_Type*. Type of match: `isolated`, `dominant`, `multiple`, `combine`, or `none`.

(xiii) *Inspected*. 0 if not visually inspected; 1 if the catalogue data and images were inspected; 2 if the match was modified by the authors.

(xiv) *Match_RAJ2000*. J2000 Right Ascension in degrees of the catalogue match to NVSS or SUMSS.

(xv) *Match._DECJ2000*. J2000 Declination in degrees of the catalogue match to NVSS or SUMSS.

(xvi) *e_Match_RAJ2000*. Uncertainty in Right Ascension of the NVSS or SUMSS catalogue match in arcseconds.

(xvii) *e_Match_DECJ2000*. Uncertainty in Declination of the NVSS or SUMSS catalogue match in arcseconds.

(xviii) *SI*. Spectral index α from a power-law spectral index fit $S \propto \nu^{\alpha}$ to all catalogue matches.

(xix) *e_SI*. Error on the spectral index parameter.

(xx) *S_74*. Flux density in Jy of the VLSSr catalogue match.

(xxi) *e_S_74*. VLSSr flux density error in Jy.

Table 3. A sample subset of the catalogue data. 10 sources were chosen to represent a diversity of characteristics. Ordering is by increasing distance from field centre ($0^h, -27°$). The corresponding cross-match results are shown in Fig. 8.

Name	KGS_RAJ2000	KGS_DECJ2000	e_RAJ2000	e_DECJ2000	S_182	e_S_182	EB_corr	R_class	Beam
KGS J001436-262208	3.65004	−26.36889	19.4	15.8	0.063	4e-3	0.45	8	0.94
KGS J002549-260214	6.45528	−26.03734	4.0	0.7	22.04	0.47	1.00	0	0.77
KGS J235701-344535	359.25422	−34.75985	2.9	1.4	22.21	0.36	1.00	0	0.57
KGS J234802-163113	357.01100	−16.52027	4.3	1.8	2.93	0.06	1.00	0	0.45
KGS J231310-315751	348.29558	−31.96420	10.8	1.8	3.79	0.22	1.00	1	0.38
KGS J225710-362746	344.29165	−36.46284	32.4	10.1	0.37	0.07	1.00	8	0.20
KGS J011535-314621	18.89879	−31.77273	14.0	7.2	1.92	0.19	1.00	4	0.15
KGS J004541-420550	11.42311	−42.09745	28.1	23.0	1.67	0.26	0.98	4	0.08
KGS J004617-420739	11.57300	−42.12757	11.2	6.5	28.33	2.59	1.00	4	0.08
KGS J013028-260952	22.61847	−26.16470	10.1	4.7	9.30	0.84	1.00	6	0.09

N_det	Match_Type	Inspected	Match_RAJ2000	Match_DECJ2000	e_Match_RAJ2000	e_Match_DECJ2000	SI	e_SI
5	isolated	1	3.65380	−26.37120	6.1	11.2	−1.46	
71	isolated	0	6.45490	−26.03690	0.4	0.7	−0.42	0.10
71	isolated	0	359.25280	−34.75880	0.7	0.7	−1.35	0.05
71	isolated	1	357.01090	−16.52020	0.4	0.7	−0.07	0.10
71	multiple	1	348.29430	−31.96400	0.7	0.7	−0.74	0.03
7	multiple	1	344.29450	−36.46190	0.7	0.7	0.47	0.550
41	multiple	1	18.89830	−31.77370	0.7	0.7	−0.88	
31	isolated	1	11.46000	−42.08360	1.8	2.2	−2.23	0.07
43	isolated	2	11.58250	−42.13760	1.4	1.8	−1.05	0.14
13	isolated	0	22.61700	−26.16680	0.4	0.7	−0.99	0.06

S_74	e_S_74	S_408	e_S_408	S_843	e_S_843	S_1400	e_S_1400	VLSSr	MRC	SUMSS	NVSS
						3.2E-3	6E-4			001436...	
13.55	1.65	17.00	0.51			8.75	0.26	J002549...	0023-263		002549...
55.20	6.72	8.70	0.35	3.02	0.09	1.28	0.04	J235700...	2354-350	J235700...	235700...
4.36	0.54	2.09	0.07			2.64	0.08	J234803...	2345-167		234802...
6.71	0.86	1.97	0.10	1.12	0.03	0.82	0.02	J231311...	2310-322	J231308...	231306...
		0.70	0.06	1.11	0.03	1.28	0.05		2254-367	J225710...	225710...
4.30	0.55	1.18	0.07	0.49	0.01	0.34	7E-3	J011535...	0113-320	J011533...	011533...
				5.5E-2	3E-3					J004550...	
		15.65	0.39	6.83	0.237				0043-424	J004613...	
26.68	3.26	5.36	0.17			1.46	0.05	J013027...	0128-264		013028...

(xxii) *S_408*. Flux density in Jy of the MRC catalogue match.

(xxiii) *e_S_408*. MRC flux density error in Jy.

(xxiv) *S_843*. Flux density in Jy of the SUMSS catalogue match.

(xxv) *e_S_843*. SUMSS flux density error in Jy.

(xxvi) *S_1400*. Flux density in Jy of the NVSS catalogue match.

(xxvii) *e_S_1400*. NVSS flux density error in Jy.

(xxviii) *VLSSr*. VLSSr source name.

(xxix) *MRC*. MRC source name.

(xxx) *SUMSS*. SUMSS source name.

(xxxi) *NVSS*. NVSS source name.

6.7 Caveats

There are two important caveats and potential sources of error we wish to emphasize to users of the catalogue.
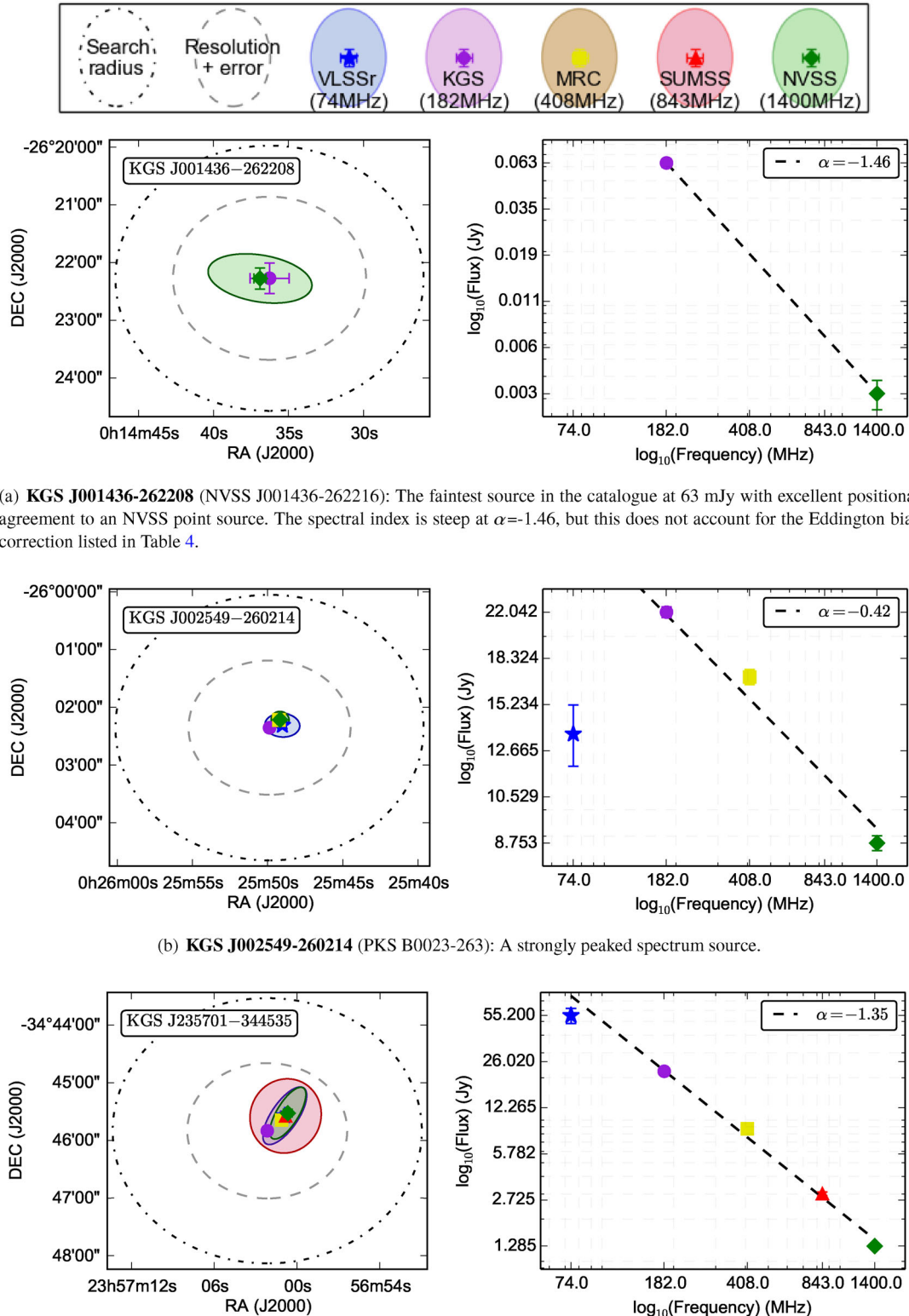
6.7.1 Primary beam model

The purpose of this survey was primarily to build a foreground model for the EoR analysis. As such, we've elected to include sources covering the full field, out to 5 per cent of the peak beam response. The accuracy of the source flux density measure-ments relies on the accuracy of the model of the primary beam response. *In situ* measurements for beam sensitivity characteri-zation are in progress but at the time of this analysis an ana-lytic MWA beam shape was assumed. As sources move through the beam, trends in the light curves near the edge of the field (beam < 0.2) suggest a 10–20 per cent flux density uncertainty that may not be sufficiently captured by the standard error reported in the catalogue.
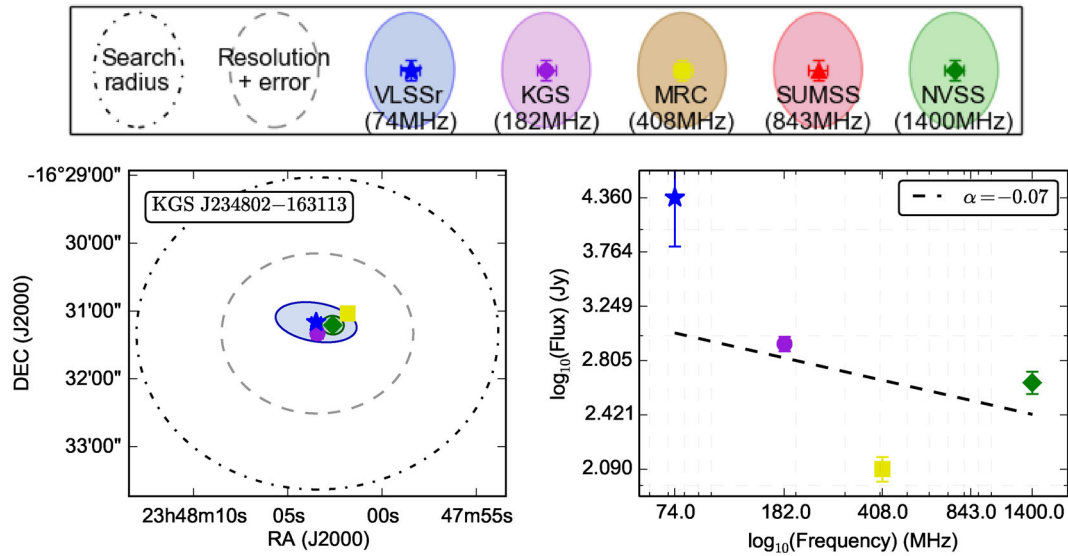
6.7.2 Extended sources

Among the 13 per cent of the sources flagged as `multiple` and visually inspected, many exhibit extended morphologies that are not well represented by the sub-components indicated in the higher res-olution catalogues. The MWA has many short baselines and much higher surface-brightness sensitivity than most radio telescopes. This has already led to the discovery of a number of large sources that were resolved out in previous surveys (e.g. a dying giant radio galaxy presented in Hurley-Walker et al. 2015). Diffuse emission picked up by the MWA will make interpreting the flux densities between surveys problematic for extended sources and care should be taken in this regard.

(a) **KGS J001436-262208** (NVSS J001436-262216): The faintest source in the catalogue at 63 mJy with excellent positional agreement to an NVSS point source. The spectral index is steep at $\alpha=-1.46$, but this does not account for the Eddington bias correction listed in Table 4.

(b) **KGS J002549-260214** (PKS B0023-263): A strongly peaked spectrum source.

(c) **KGS J235701-344535** (PKS B2354-350): This source demonstrates excellent positional agreement with a counterpart observed in all five comparison catalogues.

Figure 8. The PUMA match results for the 10 sources listed in Table 3, selected to demonstrate a variety of possible match and source types. The left plot shows the uncorrected catalogue positions, errors, and reported shape when available. The grey dashed line indicates the approximate PSF FWHM about the KGS source position. The dot–dashed black line marks the 2.3 arcsec initial search radius. The right plots shows the SED information. The black dashed line indicates the chosen power-law fit. Other lines indicate the fit to various possible source combinations when these were considered.

(d) **KGS J234802-163113** (PKS B2345-167): A good positional match with a poorly fit spectrum. This was subsequently identified as a variable QSO (Valtaoja & Valtonen 1992).
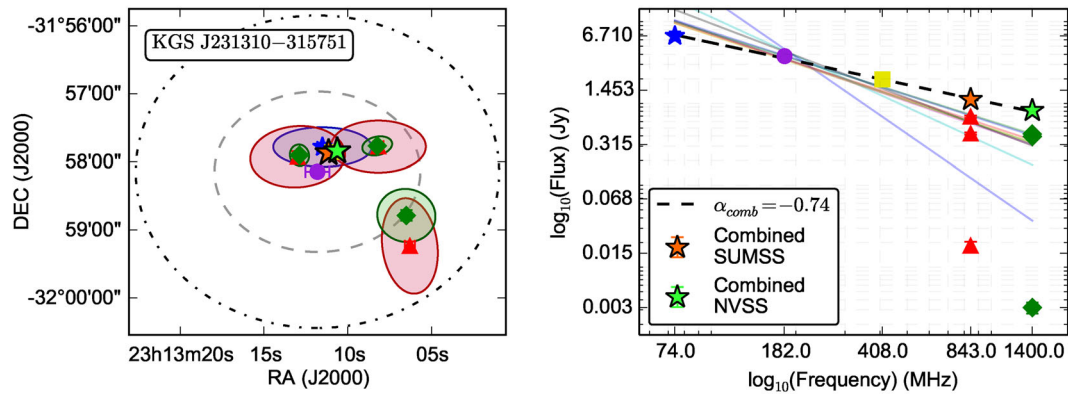
(e) **KGS J231310-315751** (PKS B2310-322): An extended source well-matched to a double at higher frequencies. The combined spectrum is well-fit by a power law. A confusing source (lower-right) within the initial search radius is ignored.
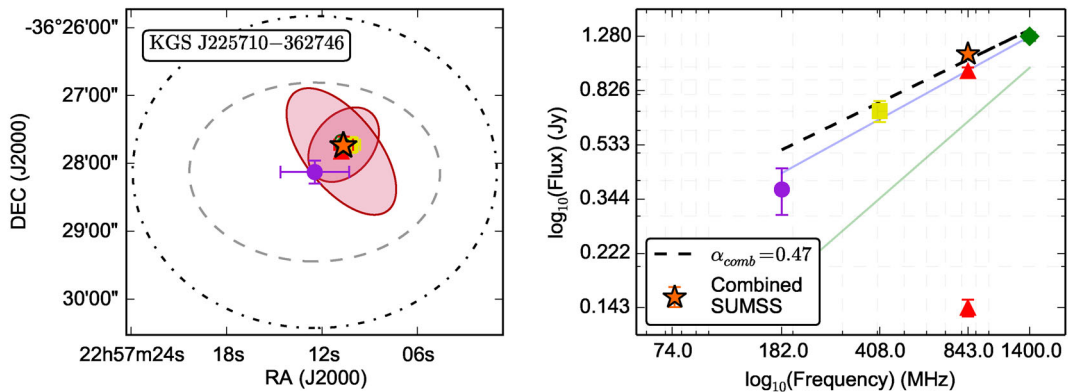
(f) **KGS J225710-362746** (PKS B2254-367): A source with a positive spectral index $\alpha \sim 0.5$.

Figure 8 – *continued*
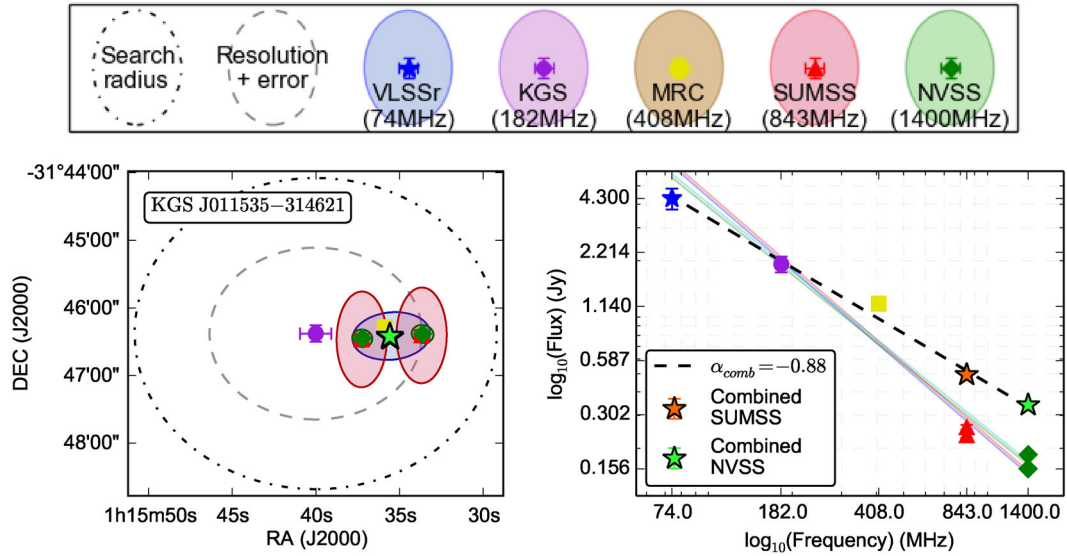
(g) **KGS J011535-314621** (PKS B0113-320): An example of a `multiple` match detected in all catalogues. When the components are combined, there is good spectral agreement with a power law fit, confirming the match despite the positional offset.
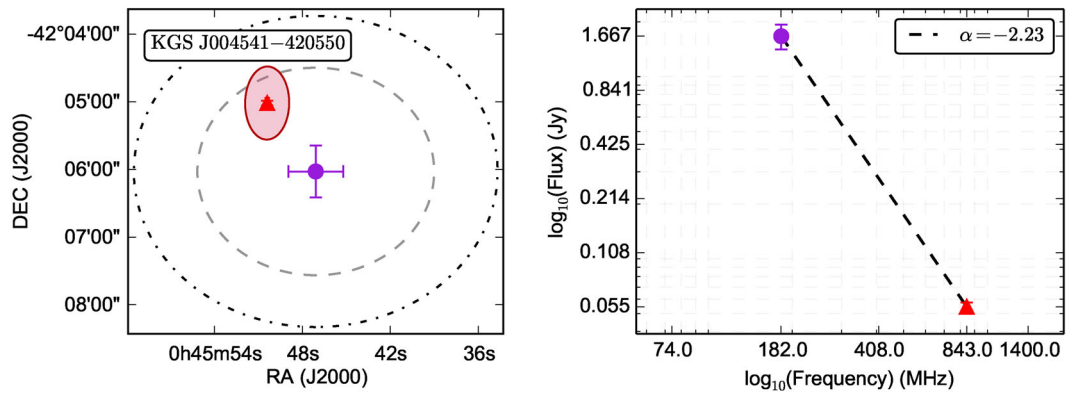
(h) **KGS J004541-420550** (SUMSS J004550-420501): This is an extremely steep spectrum source detected near the edge of the field with $\alpha = -2.23$. The flux density may be subject to error in the beam model.
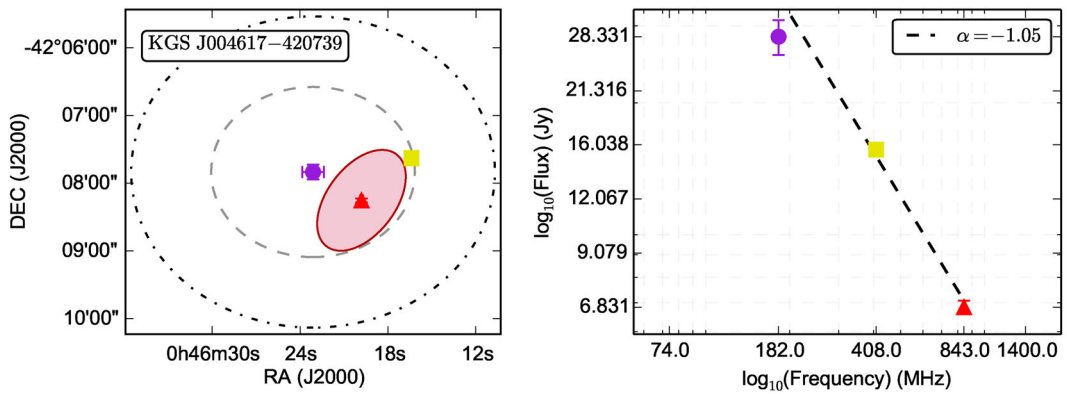
(i) **KGS J004617-420739** (PKS B0043-424): This is the brightest source in the catalogue at 28.3 Jy. It is near the edge of the field and exhibits a large positional bias.

Figure 8 – *continued*
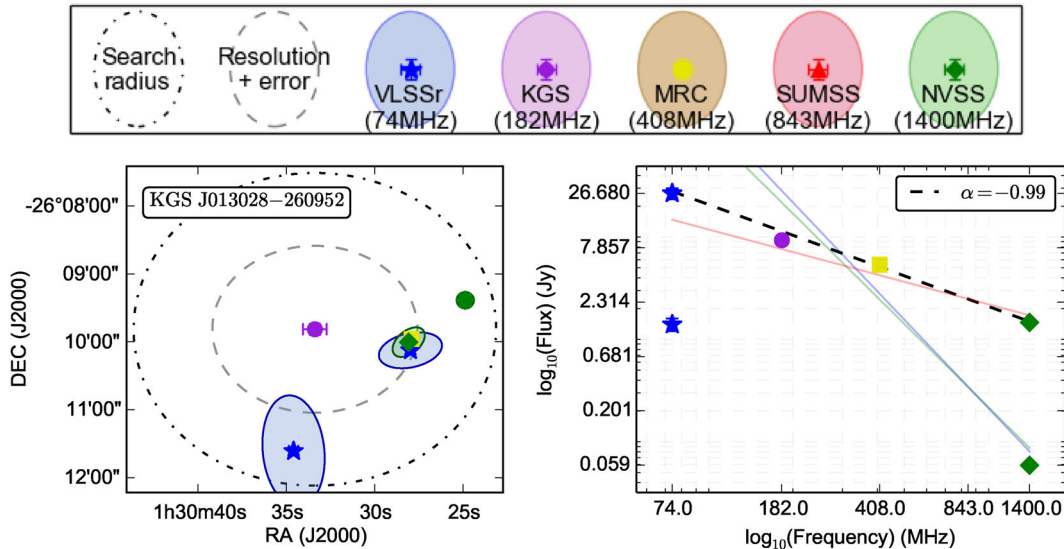
(j) **KGS J013028-260952** (PKS B0128-264): A good demonstration of the Bayesian positional match selection for a source with strong positional bias. The selected match has a posterior probability of 0.99 and is confirmed by the spectral fit.

Figure 8 – *continued*

6.8 New sources

We expect point sources with power-law spectra above $-30°$ declination to be detected in the NVSS or VLSSr surveys. Below $-30°$, we can expect new detections of ultra-steep spectrum sources (USS; $\alpha \lesssim -1.5$ with $S \propto \nu^\alpha$) that fall below the SUMSS completeness limit (18 mJy at 843 MHz) and are unobserved or undetected in the VLSSr. Transient, variable, or peaked spectrum sources may additionally lead to new detections depending on the spectral behaviour. Further, the low surface-brightness sensitivity of the MWA allows for the detection of faint extended sources that may be resolved and fall below the sensitivity limit of other surveys.

There are 25 sources with no previous radio detection identified. The properties of these sources are listed in Table 4 and postage stamp images are shown in Fig. 9. A search was made around the positions of these sources in the NASA Extragalactic Database[3] to find potential associations within a 2.3 arcmin search radius. 11 sources are unmatched at any wavelength within 30 arcsec. Eight possible associations with galaxy clusters are identified as well as three with galaxy groups. Others seem to trace galaxy over-densities, or may be high redshift (z) radio galaxy (HzRG) candidates. As these are likely to be some of the most interesting objects in the field, a proposal will be submitted to make follow up observations with the Australia Telescope Compact Array (ATCA). A summary for each source is described below (note the quoted flux densities are corrected for Eddington bias).

KGS J233620-313606: A 424 mJy source most likely associated with the galaxy cluster Abell S1136. The elongated shape suggests it may be a blended double. The cluster centre is located at a distance of 0.92 arcmin with a radius $R_c \equiv 1.72/z = 27.5$ arcmin. The source is most closely matched to GALEXASC J233618.66-313604.6 at

---

17 arcsec and the X-ray source SW J233617-313626 at 0.7 arcmin. Several other cluster members and sources at all wavelengths are also found within the 2.3 arcmin search radius.

KGS J232803-145208: A 249 mJy source most closely matched to the galaxy APMUKS(BJ) B232526.08-150914.0 at 0.6 arcmin from the source position. Five APMUKS(BJ) galaxies are found within 2.3 arcmin including one extended IR source 2MASX J23280750-1452221 at 1 arcmin separation.

KGS J000958-353932: A 164 mJy source 19 arcsec from the galaxy cluster member EDCC 408:[CGN95] 000726.8-3556 and extended IR source 2MASX J00095865-3539515. EDCC 408 is cross-identified with Abell 2730, centred 1.8 arcmin from the source position with a 14 arcmin estimated cluster radius. Many other sources are found within 2.3 arcmin including 12 other cluster members.

KGS J231311-230716: A 147 mJy blended double most closely matched to the UV source GALEXASC J231312.42-230715.1 with 20 arcsec separation. The centre of the galaxy cluster ABELL S1099 is found at a distance of 1.8 arcmin with a cluster radius $R_c = 15.6$ arcmin. Two cluster members CMW2004 388 and 339 are cross-identified with 2MASX extended IR sources and located at 1.35 arcmin and 1.82 arcmin, respectively from the source position.

KGS J235156-165850: A 133 mJy source unmatched within 30 arcsec. The extended IR source 2MASX J23515772-1657374 is 1.3 arcmin from the source position and many MRSS galaxy and GALEXASC UV sources are located with the search radius.

KGS J235021-194846: A 123 mJy source unmatched within 30 arcsec. 11 MRSS and APMUKS identified galaxies are found within 2.3 arcmin.

KGS J233116-192443: A 113 mJy source matched most closely to the UV source GALEXASC J233115.57-192441.2 at 7 arcsec. Several other UV sources and three MRSS galaxies are located within 2.3 arcmin.

KGS J231928-302751: A 114 mJy source most closely matched to the quasar 2QZ J231927.7-302845 at a 0.9 arcmin. Several other UV sources and one MRSS galaxy are found within 2.3 arcmin.

---

[3] The NASA/IPAC Extragalactic Database (NED) is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. https://ned.ipac.caltech.edu/help/intro.html.

Table 4. Properties of the new radio detections. There are 19 isolated sources (top) and 6 sources of apparently extended or diffuse emission (bottom). Postage stamp images are shown in Fig. 9.

Name	RA (deg)	Dec (deg)	S (mJy)	σ_S (mJy)	f_{EB}	N_{det}	Beam$_{mean}$	R_{class}
KGS J233620-313606	354.08570	−31.60165	424	57	1.00	71	0.67	0
KGS J232803-145208	352.01385	−14.86894	270	59	0.93	30	0.25	4
KGS J000958-353932	2.49438	−35.65899	164	27	1.00	47	0.50	4
KGS J231311-230716	348.29806	−23.12123	147	47	1.00	22	0.54	6
KGS J235156-165850	357.98357	−16.98081	136	14	0.98	14	0.51	6
KGS J235021-194846	357.58806	−19.81283	123	19	1.00	57	0.69	2
KGS J233116-192443	352.81680	−19.41199	118	20	0.95	11	0.59	6
KGS J231928-302751	349.86850	−30.46438	117	20	0.98	19	0.61	6
KGS J001054-341312	2.72537	−34.22013	116	18	0.97	14	0.60	6
KGS J233617-244958	354.07316	−24.83280	98	16	1.00	47	0.82	4
KGS J232926-255814	352.36171	−25.97067	93	8	0.99	17	0.80	6
KGS J234703-305612	356.76343	−30.93677	90	16	0.98	32	0.81	4
KGS J001640-215455	4.16737	−21.91550	82	12	0.97	7	0.84	8
KGS J000215-275242	0.56359	−27.87857	79	14	0.99	44	0.94	4
KGS J002837-261426	7.15760	−26.24063	76	13	0.87	5	0.85	8
KGS J234344-263049	355.93651	−26.51384	73	17	0.94	14	0.87	6
KGS J235556-224242	358.98621	−22.71193	70	9	0.85	5	0.88	8
KGS J234709-281746	356.78764	−28.29634	69	7	0.84	6	0.94	8
KGS J234851-232934	357.21345	−23.49292	68	6	0.66	6	0.91	8
KGS J002451-204048	6.21621	−20.68013	240	23	1.00	49	0.61	4
KGS J000821-193833	2.09041	−19.64273	217	65	0.99	34	0.68	4
KGS J000412-151811	1.05055	−15.30305	211	40	1.00	13	0.39	6
KGS J001702-312239	4.25956	−31.37749	155	31	0.99	63	0.74	3
KGS J001007-282942	2.53171	−28.49514	132	28	1.00	37	0.91	4
KGS J235139-255937	357.91289	−25.99377	81	20	0.98	11	0.92	6

Table 5. The resulting 10 relative reliability classifications. Columns from left to right are the reliability class, number of sources in that class, and the median values of detection rate, number of detections, expected number of detections, flux density, SNR, and SNR of the mean. Lower classes are more reliable while higher classes tend to be fainter or far from field centre. Classes *R*8 and *R*9 appear to capture sporadic and side lobe contaminants, but also faint sources near the detection threshold. The reliability classification is used to inform the final catalogue selection.

R_{class}	N	r_{det}	N_{det}	N_{exp}	S(Jy)	S/σ_S	$S/\sigma_{\overline{S}}$
0	2353	0.99	70	71	0.35	10.74	57.71
1	416	0.99	70	71	0.93	19.45	131.68
2	172	0.97	69	71	0.27	7.86	39.11
3	794	0.80	56	71	0.20	6.68	26.44
4	1162	0.60	38	60	0.20	6.54	18.35
5	204	0.38	21	38	0.31	6.61	11.21
6	1435	0.32	18	50	0.12	6.29	10.83
7	21	0.11	3	6	0.21	4.49	3.80
8	2630	0.09	4	33	0.14	6.88	4.63
9	303	0.05	2	32	0.11	36.68	3.29

KGS J001054-341312: A 112 mJy source most closely matched to 2MASX J00105255-3413132 at 19 arcsec and 2dF-GRS S495Z294 at 28 arcsec. The latter is one of 4 members of the 14-member galaxy group 2PIGG SGPGAL 5843 within 1.8 arcmin. The galaxy cluster and X-ray source APMCC 014 is 1.8 arcmin from the source position. A GALEXASC UV source and 1WGA X-ray source are also found at 20 arcsec and 26 arcsec, respectively.

KGS J233617-244958: A 98 mJy source. A total of 24 galaxies and UV sources are found between 11 arcsec and 2.3 arcmin.

KGS J232926-255814: A 92 mJy source most closely matched to the galaxy 2dFGRS S128Z262 at a distance of 40 arcsec. Five other galaxies and four UV sources are also located within the search radius.

KGS J234703-305612: A 88 mJy source unmatched within 30 arcsec. Seven galaxies are found within the search radius including two un-grouped 2dFGRS identifications at 1 arcmin and 2 arcmin from the source position. Two additional UV sources are also located within the search radius.

KGS J001640-215455: An 80 mJy source most closely matched to the UV source GALEXASC J001640.19-215516.7 at 21 arcsec. Two optical galaxies and 26 UV sources are located within the search radius.

KGS J000215-275242: A 79 mJy source unmatched within 30 arcsec. The galaxy 2dFGRS S198Z035 2 arcmin from the source position is part of the 32 member galaxy group 2PIGG SGP 2684. Five optical galaxies and four UV sources are found within the search radius.

KGS J002837-261426: A 66 mJy source unmatched within 30 arcsec. Four optical galaxies and many UV sources are found within the search radius.

KGS J234344-263049: A 68 mJy source most closely matched to the galaxy 2dFGRS S194Z277 at 0.84 arcmin. There are five galaxies and four UV sources within the search radius.

KGS J235556-224242: A 60 mJy source with four galaxies and eight UV sources within the search radius.

KGS J234709-281746: A 69 mJy source most closely matched to the X-ray source 2XMM J234709.2-281814 at a distance of 27 arcsec. The radio source ABELL 4038:[SPS89] 06iii is 860 mJy at 1.5 GHz lcoated 1.735 arcmin from the source position. Abell 4037 is centred 1.67 arcmin from the source position with a radius
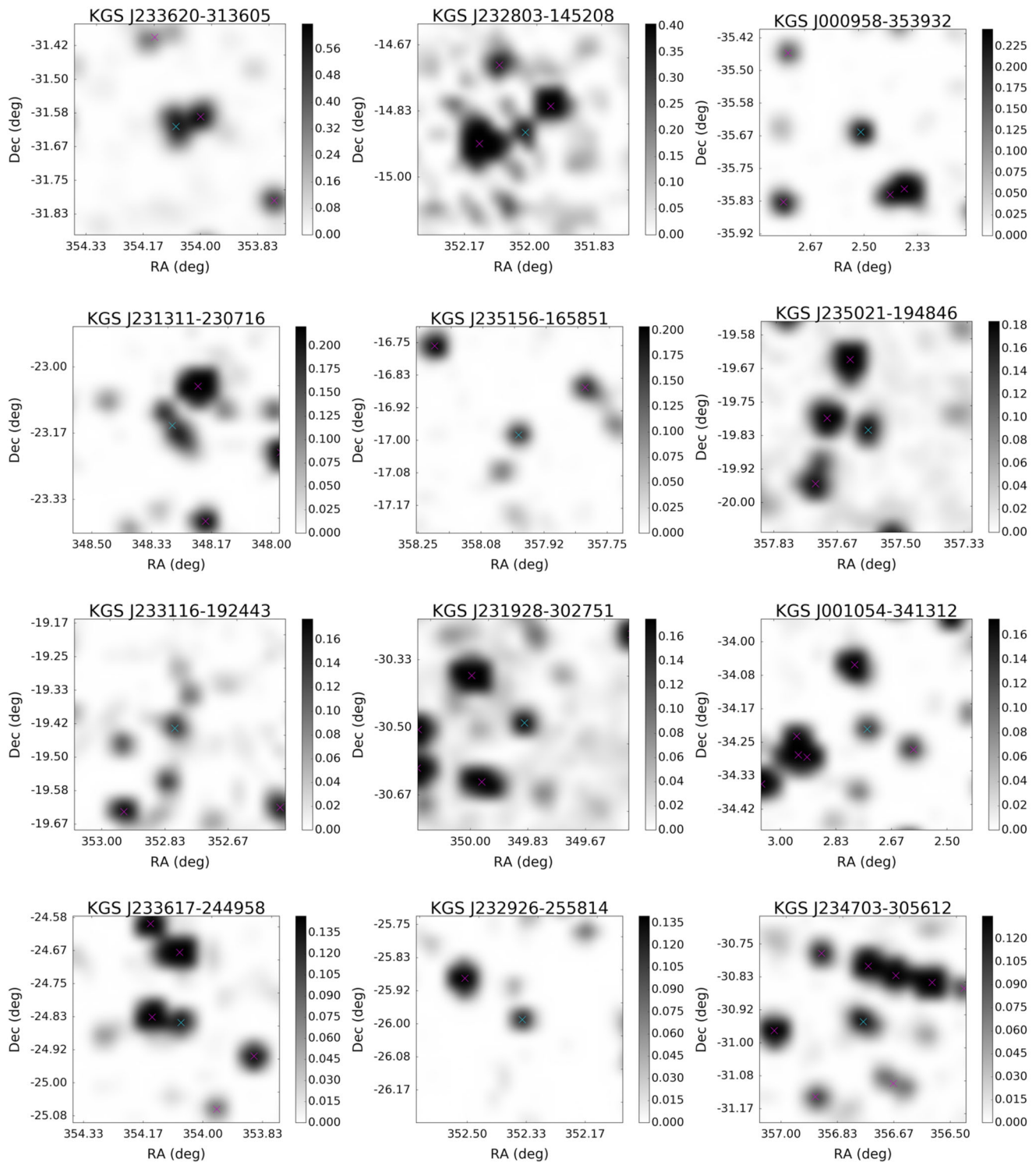
Figure 9. Postage stamp images of sources undetected in any of the comparison surveys. The first 20 are isolated sources ordered by flux density. Markers indicate the KGS mean positions of the unmatched source (cyan) and other catalogue sources (magenta). Images are 20×20 pixels, smoothed with a cubic interpolation. Units are approximately Jy beam^{-1}. The final five images are new detections that appear to be associated with a brighter counterpart. The low surface-brightness sensitivity of the MWA allows for the detection of diffuse and extended emission that may be be resolved out of the comparison surveys.
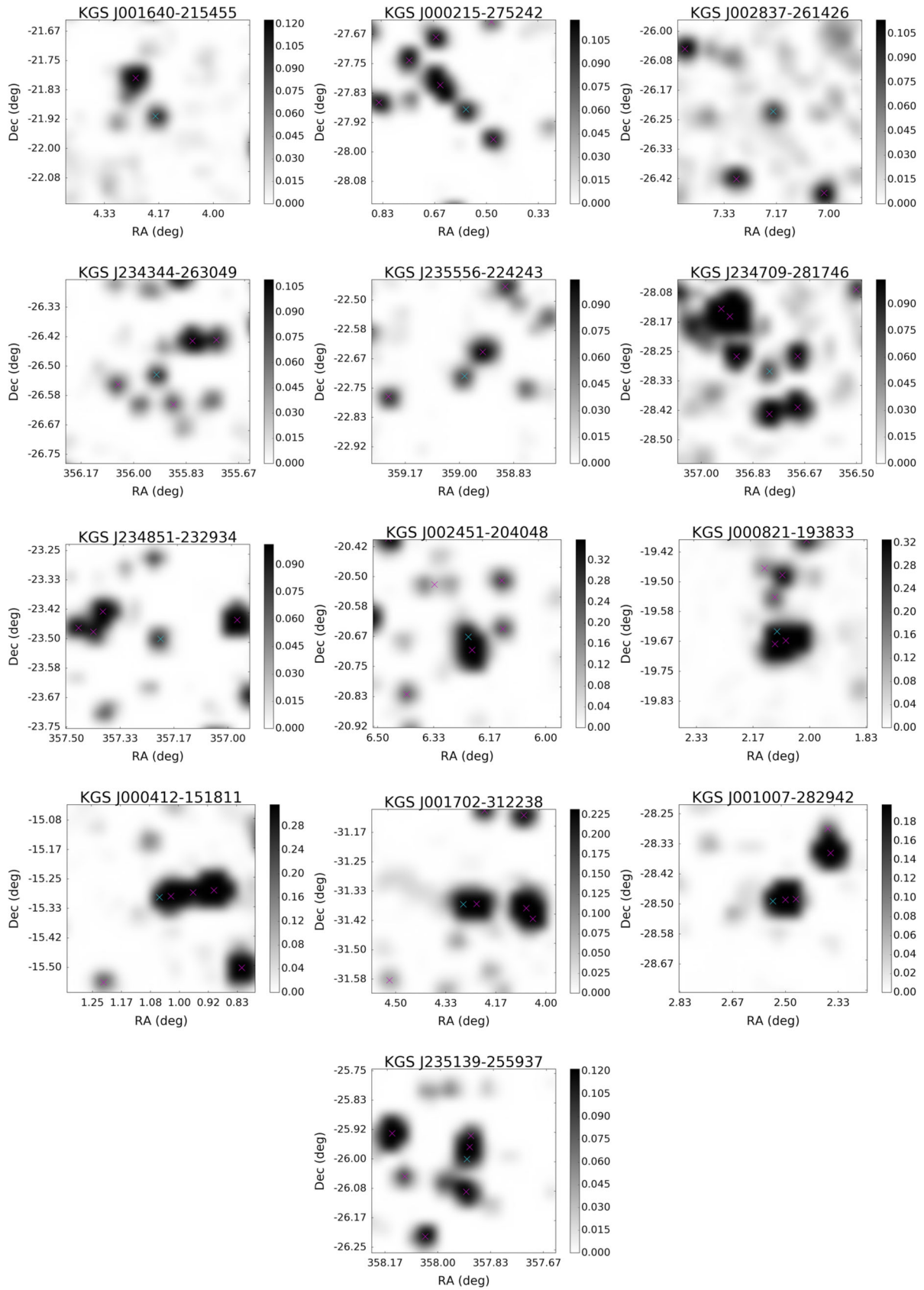
Figure 9. *– continued*

$R_c = 59$ arcmin. Numerous other sources including several cluster members and are also found.

KGS J234851-232934: A 67 mJy source matched to two GALEXASC UV sources within 30 arcsec. The radio source NVSS J234845-232827 at 1.69 arcmin has a low probability of association. Two galaxies and four other UV sources are found within 2.3 arcmin.

KGS J002451-204048: A 240 mJy confused source blended with a 2 Jy source just beyond the 2.3 arcmin search radius. It is coincident with the centre of the galaxy cluster Abell 0027 at only 14 arcsec separation. Several cluster members and numerous other sources are found within the search radius.

KGS J000821-193833: A 217 mJy confused source possibly associated with the galaxy cluster Abell 0002. It is most closely matched to GALEXASC and 2MASX identified galaxies at 7.4 arcsec and 8 arcsec, respectively from the source position. Abell 0002 has a cluster radius $R_c = 14$ arcmin centred 1.2 arcmin from the source position. This source may be associated with the cluster member and radio source PKS 0005-199 but is not matched at 1.67 arcmin separation. Numerous other sources at all wavelengths are found within the search radius.

KGS J000412-151811: A 211 mJy confused source possibly associated with PMN J0004-1518 (cross identified in the NVSS and VLSS) but unmatched at 1.5 arcmin. Many galaxies and UV sources are found within the search radius.

KGS J001702-312239: A 155 mJy confused source. It is most closely matched to 2dFGRS S436Z221, part of the 116-member galaxy group 2PIGG SGP 7135, at 1.2 arcmin from the source position. Several other sources are found within the search radius, including the five other 2PIGG member galaxies within 1.7 arcmin.

KGS J001007-282942: A 132 mJy confused source most closely matched to a UV source at 27 arcsec separation. The galaxy 2dFGRS S279Z095 is found at 0.79 arcmin and is part of the two-member galaxy group 2PIGG SGP 8480. Two NVSS radio sources lie within the search radius but are not matched. The X-ray source 1WGA J0009.9-2829 and numerous galaxies and UV sources are also found.

KGS J235139-255937: An 81 mJy confused source closely matched to 2dFGRS S132Z149 at 22 arcsec separation. This is cross identified as an X-ray and extended IR source, and is a member of the galaxy cluster Abell 2667. The nearest radio source is NVSS J235145-260038 at 1.8 arcmin from the source position. Many other sources at all wavelengths are found within the search radius.

7 CONCLUSIONS

We have presented a catalogue of 7394 extragalactic radio sources in the MWA EoR RA = 0 field at 182 MHz. This survey was motivated by the EoR analysis and the need for an accurate foreground model. The foreground catalogue is used for the purposes of calibration and subtraction, and is the predominant systematic hurdle to making an EoR detection. A catalogue of high precision, reliability, and completeness at low frequencies in the southern sky is required. To this aim, new methods were tested for deconvolution and source finding. An in-depth analysis confirmed source reliability and excluded contamination from noise and side lobe sources.

75 consecutive snapshot observations were processed, covering 2.5 h approximately centred on zenith. These were independently deconvolved using FHD, resulting in an array of centroided positive components. Source finding was done by spatially clustering the deconvolved components into source candidates for each observation. This approach was chosen to reduce errors inherent to the process of producing a restored image and fitting simplified morphological shapes to sources in the image plane. By source finding independently for each snapshot we retain information on the detection frequency, a valuable diagnostic of source reliability. We identified 9490 unique source candidates detected in at least two snapshots.

Radio surveys are afflicted by contamination from both noise and side lobes. This is especially troublesome outside of the typical half-power cutoff of the primary beam response. We wished to push this boundary to 5 per cent beam response while maximizing both completeness and reliability. Given the basic knowledge of how true sources, noise, and side lobes are expected to behave in the data, we used machine learning methods to categorize the source candidates into 10 'reliability' classes based on their observed properties. This gave us a more informative indication of confidence than SNR or detection frequency alone, while remaining entirely self-consistent.

For the purposes of this work, we selected only 7466 robustly detected above 5σ or reliability class $R_{\text{class}} < 7$. These were probabilistically cross-matched to overlapping radio surveys from 74 MHz to 1400 MHz using both positional and broad-band spectral information. Outliers, complex sources, and unmatched sources were flagged and individually investigated. The reliability classification and postage stamp images further aided the decision to include, modify, or exclude a source. Individually inspecting outliers for consistency resulted in the serendipitous identification of variable, peaked or steep spectrum, as well as morphologically complex sources.

Among sources unmatched to another catalogue, 25 are included in the final catalogue. Some of these detections can be attributed to the low surface-brightness sensitivity of the MWA. Others may be ultra steep spectrum HzRG candidates. We identify several possible associations with galaxy clusters or groups.

The final catalogue contains 7394 sources. Compared to NVSS or SUMSS the median absolute offset is only 7 arcsec compared to a 2.3 arcmin beam FWHM. No significant evidence is found for flux scale bias in the catalogue. The median broad-band spectral index is found to be -0.85 but is dependent on the catalogues matched for each source. A trend in median spectral index is observed as a function of frequency, possibly indicating spectral flattening among a sub-population of sources. The median spectral index at 182 MHz is estimated at -0.71 using a second order polynomial fit to sources detected in three or more surveys.

This catalogue provides the most reliable discrete source sky model available to date in the MWA EoR0 field for foreground subtraction. Based on the lessons learned through the creation of this catalogue, particularly the value of repeated snapshots in assessing source reliability, we are currently taking new observations to extend the EoR foreground survey across the southern sky.

REFERENCES

Baars J. W. M., Genzel R., Pauliny-Toth I. I. K., Witzel A., 1977, A&A, 61, 99
Barry N., Hazelton B., Sullivan I., Morales M. F., Pober J. C., 2016, MNRAS, 461, 3135
Bhatnagar S., Cornwell T. J., Golap K., Uson J. M., 2008, A&A, 487, 419
Bowman J. D. et al., 2013, PASA, 30, e031
Breiman L., 2001, Mach. Learn., 45, 5
Budavári T., Szalay A. S., 2008, ApJ, 679, 301
Condon J. J., Cotton W. D., Greisen E. W., Yin Q. F., Perley R. A., Taylor G. B., Broderick J. J., 1998, AJ, 115, 1693
Eddington A. S., 1913, MNRAS, 73, 359
Ester M., peter Kriegel H., S J., Xu X., 1996. in Simoudis E., Han J., Fayyad U., eds, Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining. AAAI Press, p. 226
Freund Y., Schapire R. E., 1995, in Vitányi P., ed., 2nd Euro. Conf., Computational Learning Theory. Springer, Berlin, Heidelberg, p. 23
Hancock P. J., Murphy T., Gaensler B. M., Hopkins A., Curran J. R., 2012, MNRAS, 422, 1812
Heald G. H. et al., 2015, A&A, 582, A123
Hopkins A. M. et al., 2015, PASA, 32, e037
Hurley-Walker N. et al., 2014, PASA, 31, 45
Hurley-Walker N., Johnston-Hollitt M., Ekers R., Hunstead R., Sadler E. M., Hindson L., Hancock P., 2015, MNRAS, 447, 2468
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. I., Cram L. E., Burgess A. M., 1991, The Observatory, 111, 72
Line J. L. B., Webster R. L., Pindor B., Mitchell A., Trott C. M., 2016, PASA, in press
Lonsdale C. et al., 2009, Proc. IEEE, 97, 1497
Mauch T., Murphy T., Buttery H. J., Curran J., Hunstead R. W., Piestrzynski B., Robertson J. G., Sadler E. M., 2003, MNRAS, 342, 1117
Morales M. F., Matejek M., 2009, MNRAS, 400, 1814
Myers S. et al., 2003, ApJ, 591, 575
Offringa A. R., de Bruyn A. G., Biehl M., Zaroubi S., Bernardi G., Pandey V. N., 2010, MNRAS, 405, 155
Offringa A. R. et al., 2016, MNRAS, 458, 1057
Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., 2011, J. Mach. Learn. Res., 12, 2825
Pober J. et al., 2016, ApJ, 819, 8
Roger R. S., Costain C. H., Bridle A. H., 1973, AJ, 78, 1030
Sullivan I. S. et al., 2012, ApJ, 759, 17
Taylor M. B., 2006, in Gabriel C., Arviset C., Ponz D., Enrique S., eds, ASP Conf. Ser. Vol. 351, Astronomical Data Analysis Software and Systems XV. Astron. Soc. Pac., San Francisco, p. 666
Thyagarajan N., Jacobs D. C., Bowman J. D., Barry N., Beardsley A. P., Bernardi G., Briggs F., 2015a, ApJ, 804, 14
Thyagarajan N. et al., 2015b, ApJ, 807, L28
Tingay S. J., Goeke R., Bowman J. D., Emrich D., Ord S. M., Mitchell D. A., Morales M. F., 2013, PASA, 30, 7
Wayth R. B. et al., 2015, PASA, 32, e025
Zhu J., Zou H., Rosset S., Hastie T., 2009, Stat. Interface, 2, 349

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

(http://www.mnras.oxfordjournals.org/lookup/suppl/doi:10.1093/mnras/stw1599/-/DC1).

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.

APPENDIX A: MACHINE LEARNING

Here we expand on the description of the reliability classification summarized in Section 4 and present intermediate results. The classification steps can be broken down as follows; feature selection and standardization, dimensionality reduction, initial cluster finding (unsupervised classification), and building an ensemble classifier. This process makes ample use of the PYTHON software package Scikit-learn (Pedregosa et al. 2011).

A1 Feature standardization and dimensionality reduction

Feature selection is described in Section 4. The features were standardized by subtracting the mean and dividing by the standard deviation to put them on to approximately the same scale. Principle component analysis (PCA) was used to reduce the parameter space prior to fitting a model to the distribution. Fig. A1 shows the fraction of the total variance explained by each PCA component. The first component accounts for nearly 50 per cent of the total variance, while the first thee account for 83 per cent. We select the first three components, beyond which the variance ratio begins to level out. Reducing the parameter space to three dimensions allows for much simpler model fitting and visualization compared to higher dimensions and the 17 per cent information loss is recovered at a later stage.

Each principle component (denoted C_0, C_1, and C_2) is a linear combination of the input features weighted by the set coefficients given in Table A1. The density distribution of the principal components is shown in the top row of Fig. A2.

A2 Initial unsupervised classification

We fit a Gaussian mixture model (GMM) to the data in the reduced parameter space. Various other methods may be used to make this initial classification (e.g. k-means, nearest neighbours, quadratic discriminant analysis) however we found that the distribution was sufficiently represented by a 10 component GMM. At 10 components a sharp minimum is observed in the Bayes information Criterion (BIC; Fig. A3), an indicator of goodness of fit that is sensitive to over-fitting. Every source candidate was labelled according to the GMM class it most probably belonged to. The distribution of classifications in the reduced space is shown in the middle row of Fig. A2.
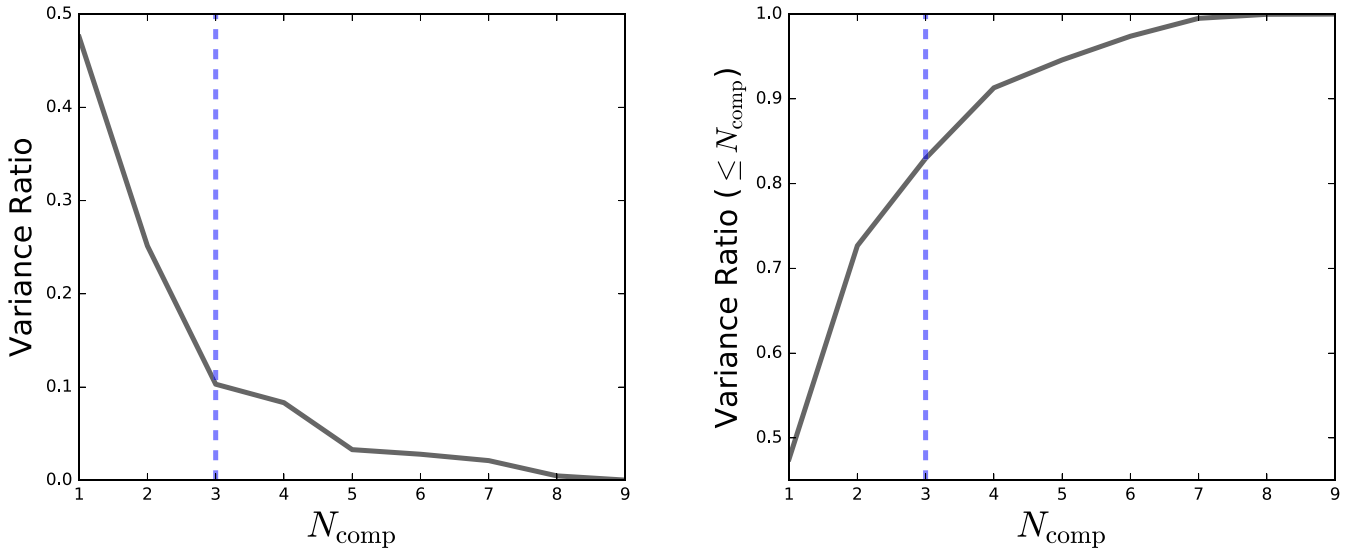
Figure A1. The fraction (left) and cumulative fraction (right) of total feature variance explained by each PCA component. The first component explains nearly half of the variance, while the first three explain 83 per cent.

Table A1. The principle component coefficients of all input features described in Section 4.

Scale	S (Jy) $\log_{10}$	S/σ_S $\log_{10}$	$S/\sigma_{\overline{S}}$ $\log_{10}$	r_{det} linear	N_{det} linear	N_{exp} linear	ρ_N linear	d_{bright} linear	S/S_{bright} $\log_{10}$
C_0	-0.391	0.034	-0.462	-0.458	-0.081	-0.245	-0.479	0.01	-0.355
C_1	0.341	-0.544	-0.181	-0.202	-0.482	0.051	-0.15	0.36	0.356
C_2	0.058	0.111	-0.235	-0.232	-0.049	0.835	-0.064	-0.408	0.062

While the BIC suggests the GMM is a sufficient approximation of the data distribution, it is naive to assume Gaussianity with full knowledge that input feature distributions are non-Gaussian. The boundaries between classes appear to be forced by the Gaussian assumption rather than true to the underlying distribution. We therefore use the GMM classes only as input to train a Decision Tree ensemble classifier.

A3 Decision tree ensemble classifier

Both Random Forest (RF) and AdaBoost (adaptive boosting) classifiers were tested. RF averages over many decision trees created on sub-samples of the data to build a more accurate classifier (Breiman 2001). RF is robust against label noise (mis-classifications on the training set). AdaBoost builds a strong classifier from a decision tree by iteratively adjusting feature weights to focus on the outliers (Freund & Schapire 1995; Zhu et al. 2009). AdaBoost is therefore sensitive to mis-classifications. Because the decision tree favours the most distinguishing features, we used all nine original inputs to minimize information loss.

The source candidates were split randomly 9:1 into training and testing sets. The classifier learns on the training set and is used to predict the test set. This was repeated for 5000 iterations so that each source was independently classified an average of 500 times and assigned to the cluster with the highest mean probability.

The variable input parameters to the RF and AdaBoost classifiers were not fully optimized, but variations on the number of trees or iterations were explored. The RF classifier tended to change the GMM classifications minimally or result in odd boundaries that

were not easily interpretable. The AdaBoost classifier was subject to over-fitting if too many iterations were allowed. An AdaBoost classifier with 50 iterations reduced the artificial footprint of the Gaussian model assumption while maintaining reasonable agreement and an interpretable structure.

The final reliability classifications are shown in the bottom row of Fig. A2. To interpret these we look at their average properties, feature distribution, and spatial distribution. A selection of the input features are shown in Fig. A4 and the median values for each class are listed in Table 5. Lower numbered classes tend to be the most reliable overall in terms of detection rate and SNR. Classes $R0 - R2$ are highly reliable sources observed and detected in nearly all snapshots. Classes $R3 - R6$ capture fainter sources with reliability decreasing further afield. Classes $R5$ and $R6$ appear to identify real sources in high density regions, whereas false side lobe sources are mostly restricted to classes $R8$ and $R9$. Class $R9$ sources stand out, having artificially high SNR due to a very small number of detections. Sources with $R_{class} > 6$ may be real but observed few times near the detection threshold.

The reliability classification is informative but not infallible, as discussed in Section 5.3. Only in the case of a source unmatched to another radio catalogue is the reliability class used to include or exclude it from the final catalogue. Future work will explore the usefulness of this classification in determining an optimal EoR foreground model.

APPENDIX B: BAYESIAN MATCHING

In Section 5 we cross match to multiple other radio catalogues using the Positional Update and Matching Algorithm (Line et al. 2016).
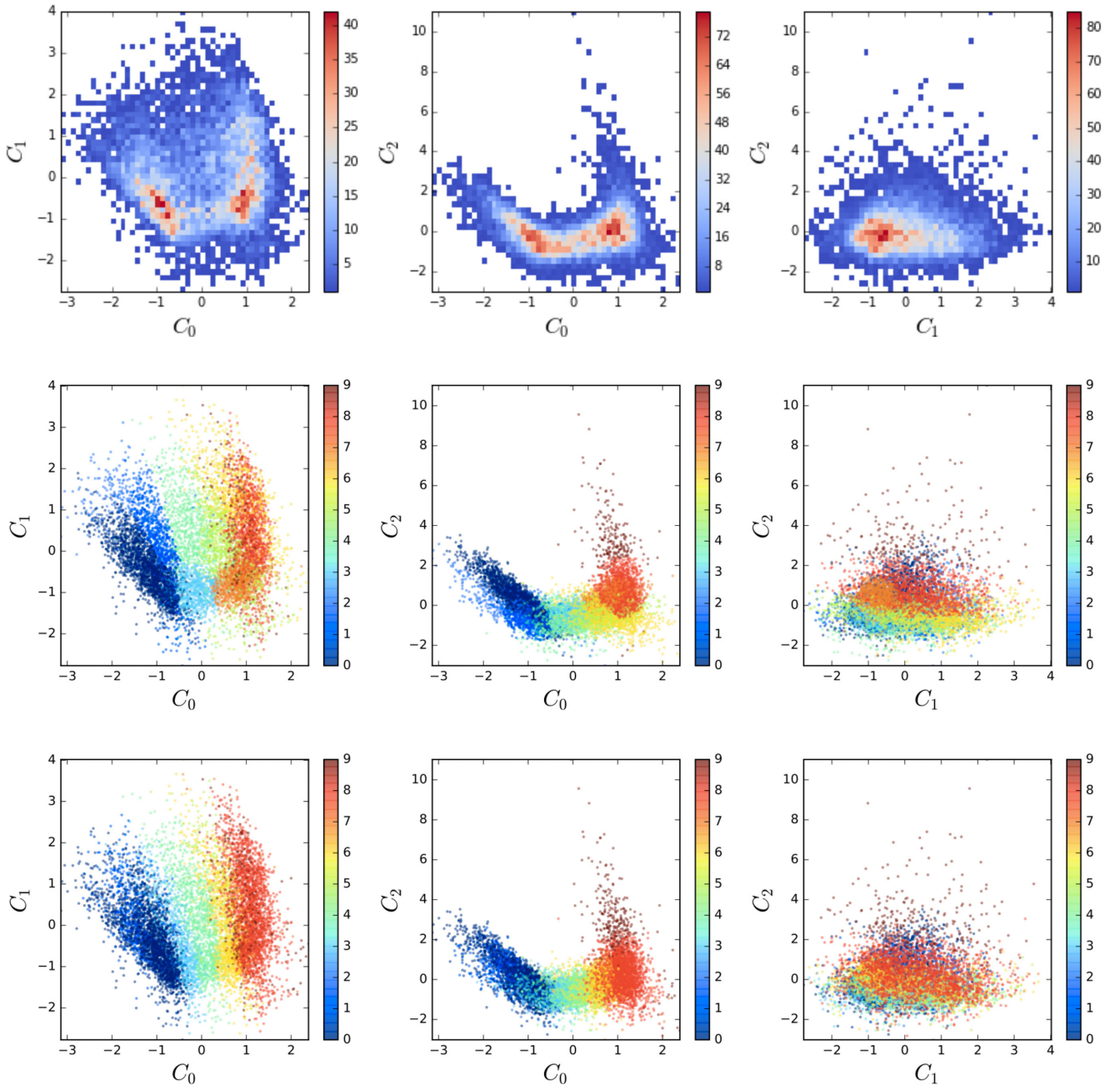
Figure A2. The distributions of the principle components for all source candidates. Top row: the density distribution. Middle row: scatter plot colored by GMM classification. Bottom row: scatter plot colored by the AdaBoost classification.

PUMA uses positional and spectral information to statistically test whether sources from multiple surveys in close proximity to one-another are true matches.

When matching N_{cat} catalogues, it can be shown (see Budavári & Szalay 2008, for further details) that the Bayes Factor is given by

$$B = 2^{N_{cat}-1} \frac{\prod w_i}{\sum w_i} \exp\left(-\frac{\sum_{i<j} w_i w_j \psi_{ij}^2}{2 \sum w_i}\right), \tag{B1}$$

where ψ_{ij} is the angular separation between sources in the ith and jth catalogues, and w_i is the weighting for the ith catalogue. This is given by $w = 1/\sigma^2$, where σ^2 is the astrometric error

(taken to be $\sigma^2 = \sigma_{RA}^2 + \sigma_{Dec}^2$). B can be used to estimated to posterior probability that the catalogues are describing the same source through

$$P(H|D) = \frac{BP(H)}{1 + BP(H)}.$$

The prior, $P(H)$, is given by

$$P(H) = \frac{n_0}{\prod_{i=1}^{N_{cat}-1} n_i}, \tag{B2}$$
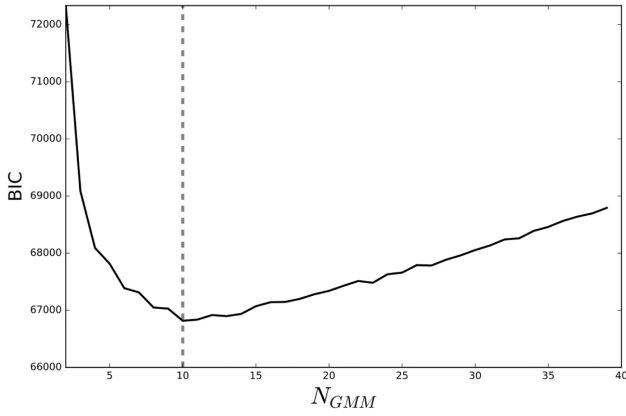
Figure A3. The optimum number of Gaussian components was chosen to minimize the Bayes Information Criterion. A strong minimum is found at 10 components.

where the scaled full sky number of sources in each catalogue is $n_i = 4\pi N_i/\Omega_i$, with N_i equal to the number of sources in the catalogue, and Ω_i the catalogue survey area. n_0 is the scaled full sky number of sources in the base catalogue.

If $P < 0.95$, the SED is fit with a power-law model of the form $\log S \propto \alpha \log \nu$ using linear least squares. The fit is considered good if the reduced Chi-square statistic $\chi^2_{\rm red}$ is less than 10. Due to uncertainty on $\chi^2_{\rm red}$ given the small number of data points and uncertainty on the errors, the fit is additionally considered good if the residuals ϵ are less than 0.1 (B3).

$$\epsilon = \frac{1}{N_{\rm cat}} \sum_{i=1}^{N_{\rm cat}} \left(\frac{|f_i - \langle f_i \rangle|}{f_i} \right) \tag{B3}$$
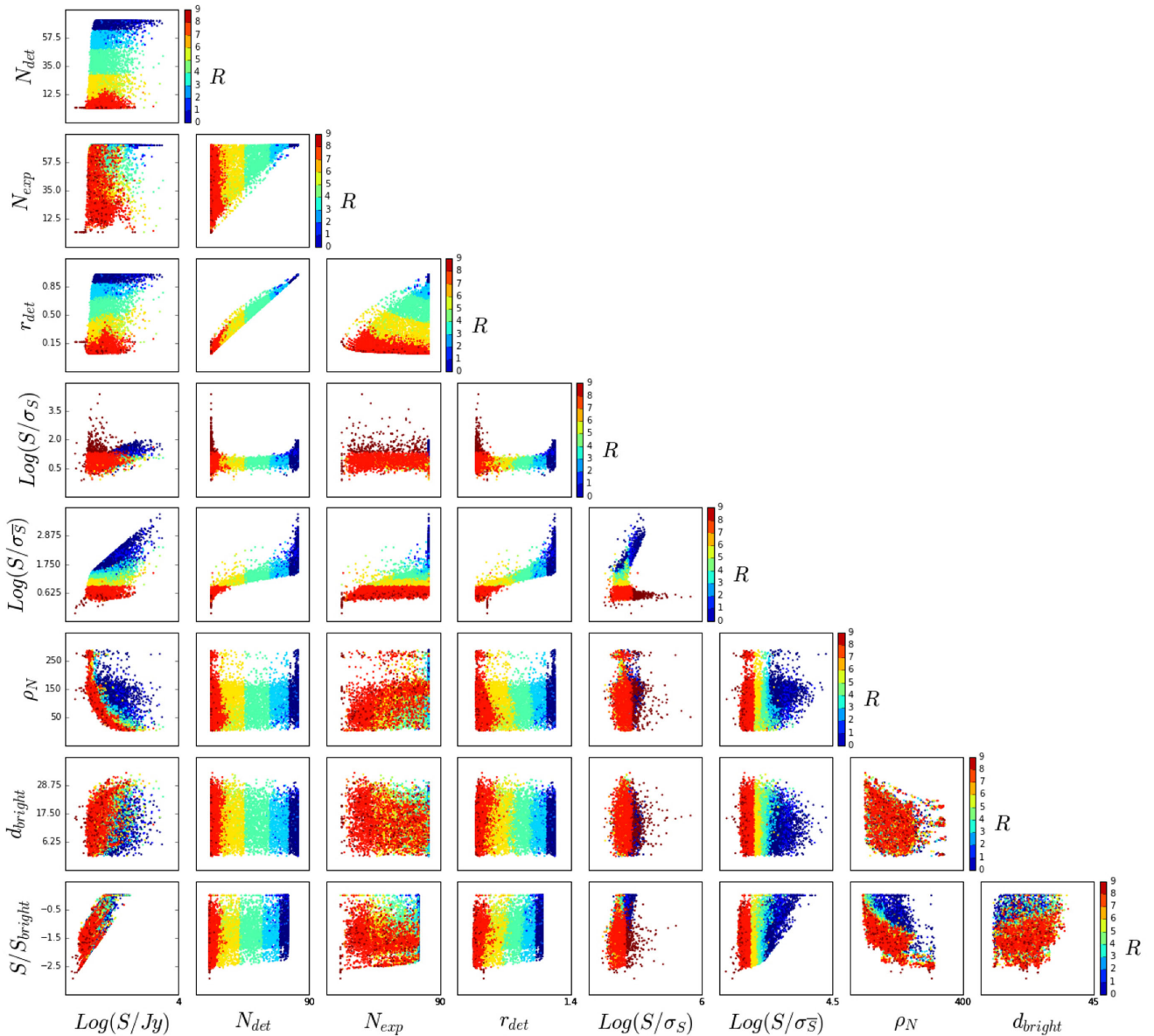
Figure A4. The input features and their final classifications ($R0 - R9$). Blue are the most reliable and red the least.

[1]*University of Washington, Seattle, WA 98195, USA*

[2]*University of Melbourne, Parkville, VIC 3010, Australia*

[3]*ARC Centre of Excellence for All-sky Astrophysics (CAASTRO), Redfern, NSW 2016, Australia*

[4]*University of Washington, eScience Institute, Seattle, WA 98195, USA*

[5]*Arizona State University, Tempe, AZ 85287, USA*

[6]*Brown University, Providence, RI 02912, USA*

[7]*Square Kilometre Array South Africa (SKA SA), Cape Town, 7405, South Africa*

[8]*Department of Physics and Electronics, Rhodes University, Grahamstown, 6139, South Africa*

[9]*The Australian National University, Canberra ACT 0200, Australia*

[10]*MIT Haystack Observatory, Westford, MA 01886, USA*

[11]*Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

[12]*International Centre for Radio Astronomy Research, Curtin University, Bentley, WA 6102, Australia*

[13]*Dunlap Institute for Astronomy and Astrophysics, University of Toronto, Toronto, ON M5S 3H4, Canada*

[14]*University of Sydney, NSW 2006, Australia*

[15]*Harvard–Smithsonian Center for Astrophysics, Cambridge, MA 02138, USA*

[16]*School of Chemical and Physical Sciences, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand*

[17]*University of Wisconsin–Milwaukee, Milwaukee, WI 53211, USA*

[18]*Department of Atmospheric, Oceanic and Space Sciences, University of Michigan, Ann Arbor, MI 48109, USA*

[19]*CSIRO Astronomy and Space Science (CASS), Marsfield, NSW 2122, Australia*

[20]*National Centre for Radio Astrophysics, Tata Institute for Fundamental Research, Pune, Maharashtra 411007, India*

[21]*Netherlands Institute for Radio Astronomy (ASTRON), NL-7991 PD Dwingeloo, Netherlands*

[22]*Raman Research Institute, Bangalore 560080, India*

[23]*National Radio Astronomy Observatory, Charlottesville, VA 22903, USA*

[24]*International Centre for Radio Astronomy Research, University of Western Australia, Crawley, Perth 6009, Australia*

[25]*University of California, Berkeley, Berkeley, CA 94720, USA*

[26]*SKA Organization Headquarters, Jodrell Bank, Macclesfield SK11 9DL, UK*

This paper has been typeset from a TEX/LATEX file prepared by the author.