



## Research Article

# WALLABY Pilot Survey: Public data release of $\sim 1800$ H I sources and high-resolution cut-outs from Pilot Survey Phase 2

C. Murugesan<sup>1,2</sup>, N. Deg<sup>3</sup>, T. Westmeier<sup>2,4</sup>, A. X. Shen<sup>1</sup>, B.-Q. For<sup>2,4</sup>, K. Spekkens<sup>3,5</sup>, O. I. Wong<sup>1,2,4</sup>, L. Staveley-Smith<sup>2,4</sup>, B. Catinella<sup>2,4</sup>, K. Lee-Waddell<sup>1,4,6</sup>, H. Dénes<sup>7</sup>, J. Rhee<sup>4</sup>, L. Cortese<sup>2,4</sup>, S. Goliath<sup>8</sup>, R. Halloran<sup>9</sup>, J. M. van der Hulst<sup>10</sup>, P. Kamphuis<sup>11</sup>, B. S. Koribalski<sup>12,13</sup>, R. C. Kraan-Korteweg<sup>14</sup>, F. Lelli<sup>15</sup>, P. Venkataraman<sup>16</sup>, L. Verdes-Montenegro<sup>17</sup> and N. Yu<sup>18,19</sup>

<sup>1</sup>ATNF, CSIRO, Space and Astronomy, PO Box 1130, Bentley, WA 6102, Australia, <sup>2</sup>ARC Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), Australia, <sup>3</sup>Department of Physics, Engineering Physics, and Astronomy, Queen's University, Kingston, ON, K7L 3N6, Canada, <sup>4</sup>International Centre for Radio Astronomy Research (ICRAR), The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia, <sup>5</sup>Department of Physics and Space Science, Royal Military College of Canada, PO Box 17000, Station Forces, Kingston, Ontario, Canada, K7K 7B4, <sup>6</sup>International Centre for Radio Astronomy Research (ICRAR), Curtin University, Bentley, WA 6102, Australia, <sup>7</sup>School of Physical Sciences and Nanotechnology, Yachay Tech University, Hacienda San José S/N, 100119, Urcuquí, Ecuador, <sup>8</sup>Canadian Astronomy Data Centre, NRC Herzberg, 5071 West Saanich Road, Victoria, British Columbia, Canada, V9E 2E7., <sup>9</sup>Queens University, 99 University Ave, Kingston, ON, K7L3N6, Canada, <sup>10</sup>Kapteyn Astronomical Institute, P.O. Box 800, 9700AV Groningen, The Netherlands, <sup>11</sup>Ruhr University Bochum, Faculty of Physics and Astronomy, Astronomical Institute (AIRUB), 44780 Bochum, Germany, <sup>12</sup>Australia Telescope National Facility, CSIRO, Space and Astronomy, P.O. Box 76, Epping, NSW 1710, Australia, <sup>13</sup>School of Science, Western Sydney University, Locked Bag 1797, Penrith, NSW 2751, Australia, <sup>14</sup>Department of Astronomy, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa, <sup>15</sup>INAF, Arcetri Astrophysical Observatory, Largo E. Fermi 5, Florence 50125, Italy, <sup>16</sup>CIRADA, Dunlap Institute for Astronomy and Astrophysics, University of Toronto, Toronto, ON M5S 3H4, Canada, <sup>17</sup>Instituto de Astrofísica de Andalucía (CSIC), Spain, <sup>18</sup>National Astronomical Observatories, Chinese Academy of Sciences, 20A Datun Rd, Chaoyang District, Beijing 100101, China and <sup>19</sup>Key Laboratory of Radio Astronomy and Technology, Chinese Academy of Sciences, 20A Datun Rd, Chaoyang District, Beijing 100101, China

## Abstract

We present the Pilot Survey Phase 2 data release for the Wide-field ASKAP L-band Legacy All-sky Blind survey (WALLABY), carried-out using the Australian SKA Pathfinder (ASKAP). We present 1760 H I detections (with a default spatial resolution of  $30''$ ) from three pilot fields including the NGC 5044 and NGC 4808 groups as well as the Vela field, covering a total of  $\sim 180$  deg<sup>2</sup> of the sky and spanning a redshift up to  $z \simeq 0.09$ . This release also includes kinematic models for over 126 spatially resolved galaxies. The observed median rms noise in the image cubes is 1.7 mJy per  $30''$  beam and 18.5 kHz channel. This corresponds to a  $5\sigma$  H I column density sensitivity of  $\sim 9.1 \times 10^{19} (1+z)^4$  cm<sup>-2</sup> per  $30''$  beam and  $\sim 20$  km s<sup>-1</sup> channel and a  $5\sigma$  H I mass sensitivity of  $\sim 5.5 \times 10^8 (D/100 \text{ Mpc})^2 M_{\odot}$  for point sources. Furthermore, we also present for the first time  $12''$  high-resolution images (“cut-outs”) and catalogues for a sub-sample of 80 sources from the Pilot Survey Phase 2 fields. While we are able to recover sources with lower signal-to-noise ratio compared to sources in the Public Data Release 1, we do note that some data quality issues still persist, notably, flux discrepancies that are linked to the impact of side lobes associated with the dirty beams due to inadequate deconvolution. However, in spite of these limitations, the WALLABY Pilot Survey Phase 2 has already produced roughly a third of the number of HIPASS sources, making this the largest spatially resolved H I sample from a single survey to date.

**Keywords:** zgalaxies: evolution; galaxies: fundamental parameters; galaxies: ISM; galaxies: kinematics and dynamics

(Received 8 May 2024; revised 16 August 2024; accepted 22 August 2024)

## 1. Introduction

The role of neutral hydrogen (H I) gas as the primary fuel for star formation in galaxies is now well established. Several surveys utilizing single-dish (e.g. Meyer et al. 2004; Koribalski et al. 2004; Wong et al. 2006; Giovanelli et al. 2005; Catinella et al. 2010) and interferometric observations (e.g. van der Hulst et al. 2001;

Verdes-Montenegro et al. 2005; Walter et al. 2008; Begum et al. 2008; Heald et al. 2011; Cappellari et al. 2011; Hunter et al. 2012; Ott et al. 2012; Koribalski et al. 2018) have shown the significance of the H I gas in understanding galaxy evolution. While significant progress has been made in studying galaxy evolution through resolved H I observations, a thorough perspective of the H I gas distribution in galaxies, its statistical properties and its relation to star formation necessitates more resolved observations of tens of thousands of galaxies from unbiased surveys.

The Wide-field ASKAP L-band Legacy All-sky Blind survey (WALLABY; Koribalski et al. 2020) is already contributing on this front and is expected to detect over  $\sim 200,000$  sources out to a redshift of  $z \sim 0.1$  covering the majority of the southern sky using the Australian SKA Pathfinder (ASKAP; Hotan et al. 2021) telescope. This is almost a factor of 10 better than the number

**Corresponding author:** T. Westmeier; Email [tobias.westmeier@uwa.edu.au](mailto:tobias.westmeier@uwa.edu.au)

**Cite this article:** Murugesan C, Deg N, Westmeier T, Shen AX, For B-Q, Spekkens K, Wong OI, Staveley-Smith L, Catinella B, Lee-Waddell K, Dénes H, Rhee J, Cortese L, Goliath S, Halloran R, van der Hulst JM, Kamphuis P, Koribalski BS, Kraan-Korteweg RC, Lelli F, Venkataraman P, Verdes-Montenegro L and Yu N. (2024) WALLABY Pilot Survey: Public data release of  $\sim 1800$  H I sources and high-resolution cut-outs from Pilot Survey Phase 2. *Publications of the Astronomical Society of Australia* 41, e088, 1–23. <https://doi.org/10.1017/pasa.2024.91>

of sources detected in ALFALFA (Giovanelli et al. 2005; Haynes et al. 2018). In addition, WALLABY will be able to resolve tens of thousands of galaxies with a default resolution of 30", while also producing higher-resolution 12" 'cut-outs' for a select sub-sample of galaxies. The 12" data products will become part of regular full WALLABY survey data releases. The aim is to image all HIPASS sources ( $N \sim 5000$ ) in high resolution, in addition, the WALLABY team is compiling a catalogue of galaxies selected based on their optical properties which we also intend to image at 12" resolution. As such, WALLABY will deliver 12" data products for thousands of galaxies in its first 5-year survey period. Some of the main goals that can be achieved with the higher resolution data include but are not limited to:

- Studying the H I morphology of galaxies at higher resolution and detailed kinematic studies of local galaxies by accurately modelling the H I distribution. In addition, the higher resolution also allows for complementary studies with IFU observations. This will also enable us to look for kinematical misalignment between the H I gas and/or the ionised gas and stars in galaxies (e.g. Wong et al. 2015; Bryant et al. 2019).
- Dynamical scaling laws of rotation supported galaxies and resolved angular momentum studies (e.g., McGaugh et al. 2000; Verheijen & Sancisi 2001; Lelli et al. 2017; Murugesan et al. 2020; Kurapati et al. 2018; Mancera Piña et al. 2021; Sorgho et al. 2024) and tracing the effects of non-axisymmetric potentials such as bars and bulges on the H I gas in galaxies (e.g. Masters et al. 2012; Murugesan et al. 2023). With the higher resolution, we may also be able to trace warps in the discs of galaxies more accurately (Garcia-Ruiz et al., 2002).
- Probing the dynamics of galaxies using reliable and robust rotation curves derived from the higher resolution data (de Blok et al. 2008; Lelli et al. 2012). This will also enable us to probe the dark matter distribution in local galaxies and additionally address the core-cusp problem relating to dwarf galaxies (Katz et al., 2017).
- Studying the star formation properties and star formation laws pertaining to the high column density H I gas ( $N_{\text{HI}} \geq 10^{20} \text{ cm}^{-2}$ ). In addition, we may also be able to probe the H I gas and star formation properties of well resolved local dwarf galaxies (e.g. Roychowdhury et al. 2014; Bacchini et al. 2020).

These science cases highlight the need for high-resolution H I imaging of targeted (and potentially interesting) galaxies. As such, WALLABY will truly pave the way for high-resolution H I studies of local galaxies to an unprecedented scale by imaging thousands of galaxies at 12" resolution.

For more specific details on the WALLABY survey, we refer the reader to the original WALLABY paper (Koribalski et al., 2020). We summarise some important updated WALLABY survey parameters for the next 5-year period in Table 1. Pre-pilot and pilot surveys were conducted to assess ASKAP data quality and to plan full survey strategies. The targeted fields of the pre-pilot surveys are listed in the following WALLABY pre-pilot survey papers by For et al. (2021), Wong et al. (2021) and Murugesan et al. (2021), while the details of the public data release of the

**Table 1.** Important updated WALLABY survey parameters

Parameter	Value
Sky coverage	14000 deg <sup>2</sup>
Declination range (core survey)	$-60^\circ \leq \delta \leq -15^\circ$
No. of ASKAP tiles	552
Integration time per tile (h)	16
Maximum baseline – default [high-resolution] (km)	2 [6]
Spatial resolution (FWHM) – default [high-resolution]	30" [12"]
Redshift range	0 – 0.1
Observed bandwidth (MHz)	1151.5 – 1439.5
Frequency resolution (kHz)	18.5
Velocity resolution (at $z = 0$ ) ( $\text{km s}^{-1}$ )	4
RMS flux density sensitivity (per 0.1 MHz/20 $\text{km s}^{-1}$ )	0.7 mJy
RMS column density sensitivity (per 0.1 MHz/20 $\text{km s}^{-1}$ )	$1.7 \times 10^{19} \text{ cm}^{-2}$

Pilot Survey Phase 1 (hereafter Phase 1 or PDR1) observations are described in Westmeier et al. (2022) and Deg et al. (2022).

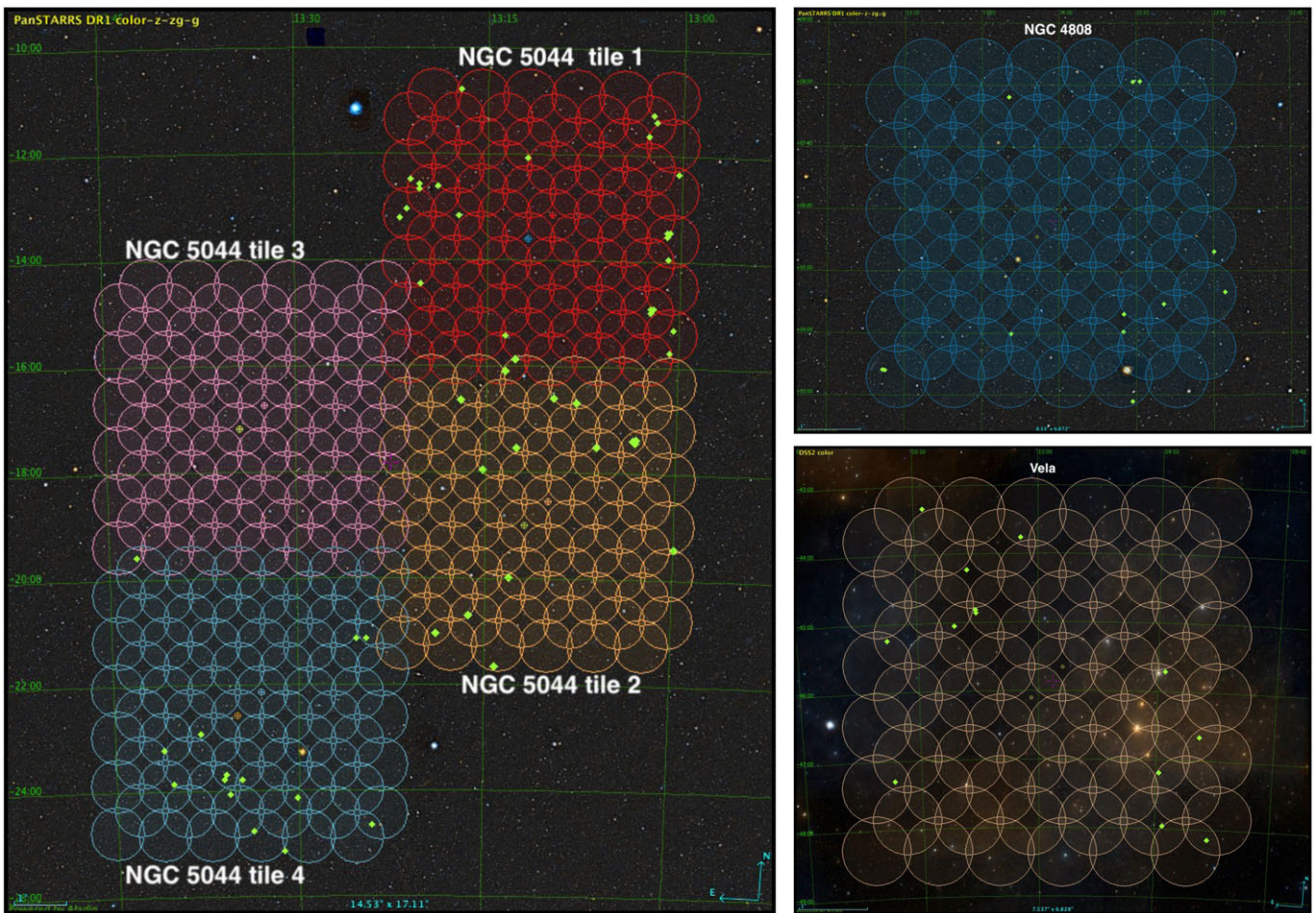
In this paper, we present the public data release of the H I catalogues and associated data products from the WALLABY Pilot Survey Phase 2 (hereafter also Phase 2 or PDR2) observations. Section 2 gives details of the targeted fields, observations, data reduction, and briefly introduces the methods employed for the validation of the observations. In Section 3, we highlight the source finding strategy and provide specific notes for each target field. In Section 4, we present the general properties of the detected 30" sample. Section 5 introduces the high-resolution 12" data, the data reduction pipeline, and characteristics of the sources. In Section 6, we describe an observed flux discrepancy in the WALLABY data and give details on the simulation studies undertaken to uncover the origins of this flux discrepancy. Section 7 describes the kinematic modelling pipeline and presents the kinematic models along with some comparisons between the 30" and 12" models. Finally, Section 8 provides details on how to access the data, while in Section 9 we provide a summary and the future goals of the WALLABY survey.

## 2. Observations and data reduction

The data used in this work have been acquired via ASKAP observations of the WALLABY Pilot Survey Phase 2 fields – NGC 4808, NGC 5044, and the Vela group. Located at the Inyarrimanha Ilgari Bundara, the Murchison Radio-astronomy Observatory (MRO), ASKAP (Hotan et al., 2021) is a state-of-the-art radio interferometer comprising of 36 12-meter antennas, equipped with Mk II phased array feeds (PAFs; De-Boer et al. 2009; Chippendale et al. 2010; Hotan et al. 2014). ASKAP is able to form 36 beams simultaneously on the sky using the advantage of the PAF, thus covering a very large area on the sky in a single pointing. For WALLABY, the 36 beams are typically arranged in the form of  $6 \times 6$  square footprints (see Figure 1). The simultaneous field of view (FOV) of ASKAP is  $\sim 30 \text{ deg}^2$  at 1.4 GHz. For the WALLABY survey observations, two  $6 \times 6$  square footprints (footprint A and B) are interleaved to attain the required uniform sensitivity across the field. A combination of both footprints A and B is referred to as a tile.

**Table 2.** Details of the observations. Col (1): Name of the field; Col (2): tile/footprint; Col (3): ASKAP Scheduling block identifier (SBID) used to tag the data in CASDA; Col (4): Date of observation; Col (5) - (6): RA and Dec of the centre of the footprint, respectively, in J2000; Col (7): Phase rotation of the footprint on the sky in deg; Col (8): Number of antennas used; Col (9): Flagged fraction. <sup>a</sup>EMU-POSSUM-WALLABY commensal field; <sup>b</sup>GASKAP-WALLABY commensal field.

Field	Footprint	SBID	Date	RA (J2000)	Dec (J2000)	$\phi$ (deg)	No. of Ant.	Flagged fraction
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
NGC 5044	1A	33879	2021 Nov 23	13 <sup>h</sup> 10 <sup>m</sup> 00 <sup>s</sup>	-13°09'04"	0.00	34	~ 10%
	1B	34302	2021 Dec 9	13 <sup>h</sup> 11 <sup>m</sup> 51 <sup>s</sup>	-13°36'02"	-0.11	33	~ 20%
	2A	34166	2021 Dec 4	13 <sup>h</sup> 10 <sup>m</sup> 00 <sup>s</sup>	-18°32'53"	0.00	32	~ 20%
	2B	34275	2021 Dec 8	13 <sup>h</sup> 11 <sup>m</sup> 54 <sup>s</sup>	-18°59'50"	-0.15	31	~ 20 – 25%
	3A <sup>a</sup>	31536	2021 Sep 1	13 <sup>h</sup> 32 <sup>m</sup> 30 <sup>s</sup>	-16°45'00"	-1.20	33	~ 10%
	3B <sup>a</sup>	40905	2022 May 23	13 <sup>h</sup> 34 <sup>m</sup> 25 <sup>s</sup>	-17°11'24"	-1.34	33	~ 10%
	4A	25701	2021 Apr 4	13 <sup>h</sup> 33 <sup>m</sup> 00 <sup>s</sup>	-22°08'48"	-1.20	30	~ 25 – 30%
	4B	25750	2021 Apr 5	13 <sup>h</sup> 34 <sup>m</sup> 59 <sup>s</sup>	-22°35'12"	-1.39	30	~ 25 – 30%
NGC 4808	A	33681	2021 Nov 18	12 <sup>h</sup> 59 <sup>m</sup> 37 <sup>s</sup>	+06°00'00"	0.00	34	~ 5 – 10%
	B	37604	2022 Feb 23	13 <sup>h</sup> 01 <sup>m</sup> 26 <sup>s</sup>	+05°32'59"	0.05	34	~ 5 – 10%
Vela <sup>b</sup>	A	33521	2021 Nov 14	09 <sup>h</sup> 58 <sup>m</sup> 00 <sup>s</sup>	-45°39'17"	0.00	34	~ 10%
	B	33596	2021 Nov 16	10 <sup>h</sup> 00 <sup>m</sup> 36 <sup>s</sup>	-46°06'10"	-0.47	34	~ 10%



**Figure 1.** The ASKAP footprints covering the Pilot Phase 2 fields overlaid on top of their PanSTARRS composite optical images. The green points show the location of the HIPASS sources imaged with a 12" resolution for the high-resolution cut-outs.

The observations of the various Phase 2 fields were carried out between April 2021 and May 2022 (for exact observing dates refer to Table 2) with an integration time of  $\sim 8$  h for each footprint and thus a total on-source time of  $\sim 16$  h per tile. During the observations, most of the 36 antennas were used to correlate the data, although a few antennas were flagged as bad during the data reduction process (for details, refer to Table. 2).

We note that the observations were carried out in the frequency range of 1152 – 1440 MHz, with a total bandwidth of 288 MHz, consisting of 15,552 channels corresponding to a spectral resolution of 18.5 kHz. As with Phase 1 observations, we note that only the upper half of the band above  $\sim 1300$  MHz has been processed as the observations below this frequency are severely affected by Radio Frequency Interference (RFI) due to Global Positioning System/Global Navigation Satellite System (GPS/GNSS).

### 2.1. Field selection

For the Phase 2 observations, each ASKAP Science Survey Project (SSP) was allocated a total of 100 h of observing time. In Figure 1, we show the targeted pilot Phase 2 fields. The field selection was decided based on the following criteria:

**Scientific merit** – The Phase 2 fields were chosen on their merit, ensuring that multi-wavelength data is readily available, and in addition, have the potential to maximise the science goals, which include probing large-scale structures in the zone of avoidance (ZOA) and investigating environmental effects on galaxy groups.

**Commensality with other ASKAP Science Survey Teams** – WALLABY is commensal with other ASKAP surveys such as the Evolutionary Map of the Universe (EMU; Norris *et al.* 2011) survey, Polarisation Sky Survey of the Universe's Magnetism (POSSUM; Gaensler *et al.* 2010), the Galactic ASKAP Survey (GASKAP; Dickey *et al.* 2013), and the Commensal Real-time ASKAP Fast Transients Survey (CRAFT; Macquart *et al.* 2010). The NGC 5044 tile 3 field was chosen to be the EMU-POSSUM-WALLABY three-way commensal field. While the Vela field was chosen to be commensal with GASKAP, wherein observations in the Galactic range ( $V_{\text{sys}} 500 \text{ km s}^{-1}$ ) were reduced in 'zoom mode' with a full spectral resolution of  $2 \text{ km s}^{-1}$ .

**Source finding strategy** – The NGC 5044 fields were targeted as they cover a contiguous region on the sky (see Figure 1). Observing overlapping fields/tiles was necessary so as to test our source finding strategy in preparation for the full survey, which will involve running the source finding pipeline on contiguous adjacent fields. Refer to Section 3.1 for more details on the source finding strategy.

### 2.2. Default 30" WALLABY data reduction pipeline

For a detailed description of the data reduction process of the default 30" WALLABY observations, we refer the reader Westmeier *et al.* (2022). We describe very briefly the different stages of the default pipeline. We note that each of the steps below are performed independently for each ASKAP beam before they are mosaicked to form the final image cubes. First, the pipeline runs an automated flagging procedure which identifies bad antennas and flags the bad data for each beam. After the flagging procedure, the pipeline proceeds to perform the bandpass calibration, followed by imaging the continuum. Then using the component and sky models derived from the continuum imaging, continuum subtraction is performed in the UV-domain. The

next steps involve imaging each ASKAP beam separately, which also includes the deconvolution step, where the data is cleaned to a peak residual flux density of 3.5 mJy, followed by a deeper cleaning (within the pixels corresponding to the identified clean components) to a residual peak flux density threshold of 0.5 mJy. This is then followed by restoring the clean components convolved with a 30" Gaussian beam and adding back the residuals to form the image cubes. After the restoring phase, an image-based continuum subtraction routine is performed. A primary beam correction is then performed after which all the beams are mosaicked together to form two footprint (A and B) image cubes, which are then mosaicked to form the final full sensitivity image cube. We note that the main change in the data reduction pipeline for Phase 2 is the use of holographic measurements of the actual ASKAP primary beams used for the observations (Hotan, 2016) for the primary beam correction, as opposed to the static Gaussian primary beam correction that was used for the Phase 1 data reduction. The introduction of the holography models for the correction provides more accurate primary beam model weights leading to more accurate flux recovery from detections across the entire FOV compared to the flux based on the static Gaussian primary beam model.

### 2.3. Data quality assessment and validation

RFI and antenna flagging are performed on a beam-by-beam basis. The overall flagged visibility fraction ranges from 5 to 30% across all beams, and typically all 36 antennas were utilised for all beams.

We evaluate the data quality of each footprint image cube based on a set of metrics. These metrics were established based on the data in the WALLABY early science field of M83 and pre-pilot field of Eridanus (see For *et al.* 2019; For *et al.* 2021), which include RMS, minimum and maximum flux densities, 1 percentile noise level and median absolute deviation of median flux (MADMF). Each set consists of values for three types of image cubes, i.e. before and after continuum subtraction image cubes as well as a residual cube.

The broadband RFI/artefacts are evaluated with the MADMF statistic. This metric is sensitive to strong artefacts. The distribution of flux density values for all voxels in each beam at the 1 percentile level indicates any bandpass calibration and/or sidelobe issues. All these metrics and observation information are presented in a HTML style summary report for each footprint. The report of each footprint and description of each metric is available at the CSIRO ASKAP Science Data Archive (CASDA Huynh *et al.* 2020).

Following this, a quality checking pipeline verifies that the footprints in CASDA meet the data quality requirements as mentioned above. The pipeline is executed when a new observation is available on CASDA. We run the Source Finding Application (SoFiA; Serra *et al.* 2015; Westmeier *et al.* 2021) described in detail in Section 3 on the mosaicked image cubes to generate moment 0 images of the field. Then, we verify by eye that there are no significant artefacts in the source finding output. Footprints that show significant artefacts from the source finding run are rejected by the team and marked to be re-observed. Accepted footprints are recorded in a database (for more details see Appendix C), which is used by the main source finding pipeline.

## 3. Source finding and parametrisation

Source finding on the final image cubes was performed using the Source Finding Application (SoFiA; Serra *et al.* 2015; Westmeier

et al. 2021) version 2.3.1. For this purpose, each tile was split into sub-regions of approximately  $1500 \times 1500$  spatial pixels and 1400 spectral channels for parallel processing on multiple nodes of the Nimbus computing cluster at the Pawsey Supercomputing Centre in Perth. In total, the frequency range of 1305 – 1418 MHz, corresponding to a recession velocity range of  $500 \lesssim cz \lesssim 26,500 \text{ km s}^{-1}$ , was searched for H I emission.

Each sub-region was first multiplied by the square root of the associated weights cube to normalise the noise across the data cube. This was followed by automatic flagging of artefacts and the positions of radio continuum sources with flux densities  $150 \text{ mJy}$  in the Rapid ASKAP Continuum Survey (RACS McConnell et al. 2020) catalogue. A circle with a radius of 5 pixels (or  $30''$ ) was flagged around the position of each such continuum source, flagging the entire frequency range (including any H I emission) within those pixels, creating a circular hole in the affected area. If an H I detection is affected by flagging, the flag parameter in the catalogue is set accordingly to alert users of the fact that the detection is adjacent to flagged pixels. Additional noise normalisation in a running window of size  $51 \times 51$  spatial pixels and 51 spectral channels was carried out to normalise any remaining noise variation that was not accurately reflected by the weights cube. In addition, a robust first-order polynomial was fitted to each spectrum to remove any remaining low-level continuum residuals, and SoFiA's ripple filter was employed to remove any low-level bandpass ripples due to RFI.

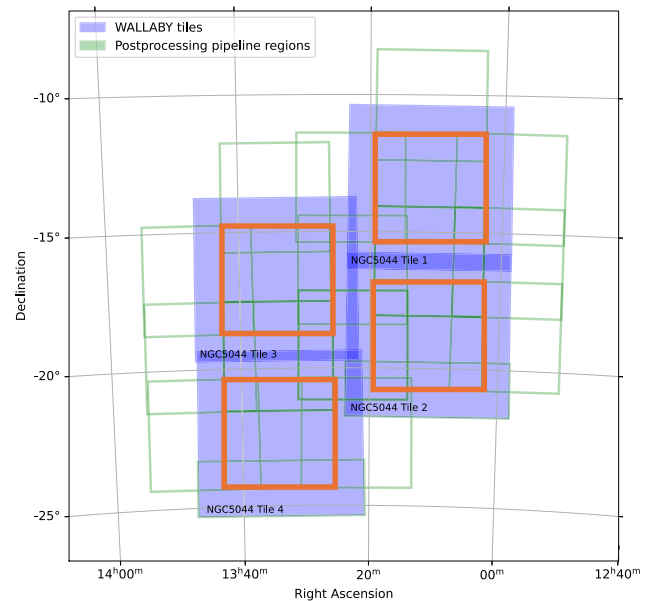
After these preconditioning steps, SoFiA's 'smooth and clip' (S+C) algorithm was used to detect emission above a threshold of 3.8 times the noise level in each smoothing iteration. Smoothing kernels of 0, 5, and 10 spatial pixels and 0, 3, 7, 15, and 31 spectral channels were employed to boost the signal-to-noise ratio of faint, extended H I emission on spatial scales of up to 1 arcmin and velocity widths of up to about  $120 \text{ km s}^{-1}$  (at  $z \approx 0$ ). All detected pixels were then merged into coherent detections across a merging length of 2 spatial pixels and 3 spectral channels, with detections smaller than 5 pixels or channels discarded. Next, SoFiA's reliability module was used to discard all detections with a reliability of less than 0.7 or an integrated signal-to-noise ratio of less than 3. In addition, detections with a total of less than 300 spatial and spectral pixels were also discarded as unreliable.

The remaining detections were then parameterised before SoFiA generated the final source catalogue and output products, including cubelets, moment maps, and integrated spectra for all detections. Table 5 in Appendix B lists some important SoFiA parameter values used for the  $30''$  source finding runs.

### 3.1. Source finding strategy

A pipeline has been developed through the Australian SKA Regional Centre (AusSRC) to run the source finding for the WALLABY survey. The pipeline communicates with external databases such as CASDA and the WALLABY database to automatically check for new footprints (and tiles) that have been uploaded on to CASDA. When a new observing tile has been deposited in CASDA, the pipeline mosaics overlapping regions of adjacent tiles outside the central  $4^\circ \times 4^\circ$  (orange boxes in Figure 2) and executes the source finding application.

The Vela and NGC 4808 fields were covered by only a single ASKAP tile each, and hence source finding was performed on the full tile at once. For the four-tile NGC 5044 mosaic we instead employed a staged source finding approach to account



**Figure 2.** Strategy for source finding in the NGC 5044 field which has overlapping regions. Tiles are shown as blue-shaded regions while each orange box corresponds to a central  $\sim 4^\circ \times 4^\circ$  area, where the source finding is performed. For the NGC 5044 field central regions are processed when both footprints have been observed, and overlapping regions are processed when adjacent tiles are completed. The light green boxes represent  $\sim 4^\circ \times 4^\circ$  areas where source finding is run when appropriate adjacent tiles are available (or become available in the future).

for the gradual completion and release of observations for this field. Figure 2 gives a visual representation of the source finding strategy adopted for the NGC 5044 field. First, SoFiA was run on the central  $4^\circ \times 4^\circ$  region (orange boxes) of each individual NGC 5044 tile. This central region corresponds to the area across which the noise level is roughly constant in an individual tile. Beyond the central  $4^\circ \times 4^\circ$  region, the noise in the outskirts of the tile typically tend to increase by a factor of two or more (see Appendix A in Westmeier et al. 2022). Once adjacent tiles became available, SoFiA was then additionally run on the overlapping regions between those tiles in steps of  $4^\circ \times 4^\circ$  regions (green boxes), to gradually build up a source catalogue of the entire NGC 5044 field. This staged source finding approach will also be applied to the full WALLABY survey in the future. The NGC 5044 mosaic provided us with the opportunity to develop and test this approach in anticipation of the full survey observations.

Detections from the source finding pipeline are uploaded into a database, and WALLABY sources are then manually accepted following visual inspection. For more details on the manual workflow, refer to Appendix C.

### 3.2. Notes on individual fields

In this section, we present some pertinent notes on the individual fields. We note that while due care has been taken to avoid artefacts and false positives in the final source catalogues through visual inspection of all raw detections from SoFiA, we caution that some false positives may still remain in the final source catalogue as well as the issue of blended sources and/or sources broken into separate detections. This is true for all three Phase 2 fields. Where possible, comments are made in the source catalogue highlighting

such issues. In addition, we also added ‘multiplet’ and ‘component’ tags to mark such cases.

### 3.2.1. NGC 5044

The data quality for the NGC 5044 tiles is good overall with only a few artefacts still present in the final mosaicked image cube. This release consists of the source finding detections from the four NGC 5044 tiles covering  $4 \times 30 \sim 120 \text{ deg}^2$  and spanning a velocity range of  $cz \sim 500 - 26,500 \text{ km s}^{-1}$  ( $z0.089$ ) using the full RFI-free higher frequency band available for WALLABY. Some artefacts still remain in the data cube particularly related to faint continuum residuals and sidelobes that have affected the northern edge of tile 1, the southern edge of tile 2 and a small region of tile 4 of the NGC 5044 mosaic. We note that this may have reduced the completeness of the source finding runs in the affected regions.

After the source finding run, all detections were visually inspected and obvious artefacts were removed following which 1326 detections remain. We note that the NGC 5044 tile 4 was the only Phase 2 tile for which a Gaussian primary beam model was used for primary beam correction instead of using a holography model, due to which we anticipate minor flux-related issues such a potential increase in flux by about  $\sim 15 - 20\%$  for sources that lie further away from the beam centre and/or close to the edge of the tile/footprint. For tiles 1, 2 and 3 of the NGC 5044 field, the holography-based primary beam correction was performed.

### 3.2.2. NGC 4808

The data release for the NGC 4808 field covers  $30 \text{ deg}^2$  of the sky with a velocity range of  $cz \sim 500 - 26,500 \text{ km s}^{-1}$  ( $z0.089$ ). There were no major issues identified with the NGC 4808 field and the data quality is overall good, with very few artefacts in the image cube. The source finding run resulted in the retention of 231 detections following removal of few faint artefacts.

### 3.2.3. Vela

The Vela field covers  $30 \text{ deg}^2$  with a redshift range of  $cz \sim 500 - 25,400 \text{ km s}^{-1}$  ( $z0.085$ ). As mentioned in Section 2.1, this field was observed commensally with the GASKAP-H I project in spectral zoom mode and processed at the full spectral channel width of 9.26 kHz. After this, the extragalactic frequency range of the data was re-binned to the default WALLABY spectral resolution of 18.5 kHz prior to spectral imaging. However, due to flagging preceding binning, some faint RFI from global navigational satellite systems was not fully flagged in the higher spectral resolution data, which has resulted in a significant number of false detections at frequencies of  $\nu \simeq 1380 \text{ MHz}$  and  $\nu \simeq 1310 \text{ MHz}$ . This therefore has resulted in the reliability of detections at those frequencies to be reduced which may have resulted in some genuine H I sources being omitted by SoFiA. Overall, 203 detections are retained after visual inspection and removing artefacts and false positives.

## 4. Source characterization

In this section, we highlight some characteristics of the source properties from the Phase 2 source finding runs, such as the distribution of the signal-to-noise-ratio (SNR) of the detected sources, size distribution, H I mass distribution as well as their H I mass – distance plot. We also compare the Phase 2 source properties with the Phase 1 detections in order to highlight the significant improvement in the data quality.

Panel a) in Figure 3 shows the distribution of the barycentric redshift for the Phase 2 detections (in blue) compared to the redshift distribution of sources in Phase 1. We find that the median redshift of the sources in Phase 2 is  $\sim 0.027$  ( $cz \sim 8094 \text{ km s}^{-1}$ ). The median redshift of sources in the NGC 5044 field is  $\sim 0.025$  ( $cz \sim 7495 \text{ km s}^{-1}$ ), the NGC 4808 field is  $\sim 0.039$  ( $cz \sim 11692 \text{ km s}^{-1}$ ) and the Vela field is  $\sim 0.04$  ( $cz \sim 11992 \text{ km s}^{-1}$ ). We see the clumping in redshifts in two distinct peaks in Figure 3. In comparison, the Phase 1 sources were mainly from nearby groups and clusters and as such, show a median barycentric redshift of  $\sim 0.014$  ( $cz \sim 4197 \text{ km s}^{-1}$ ).

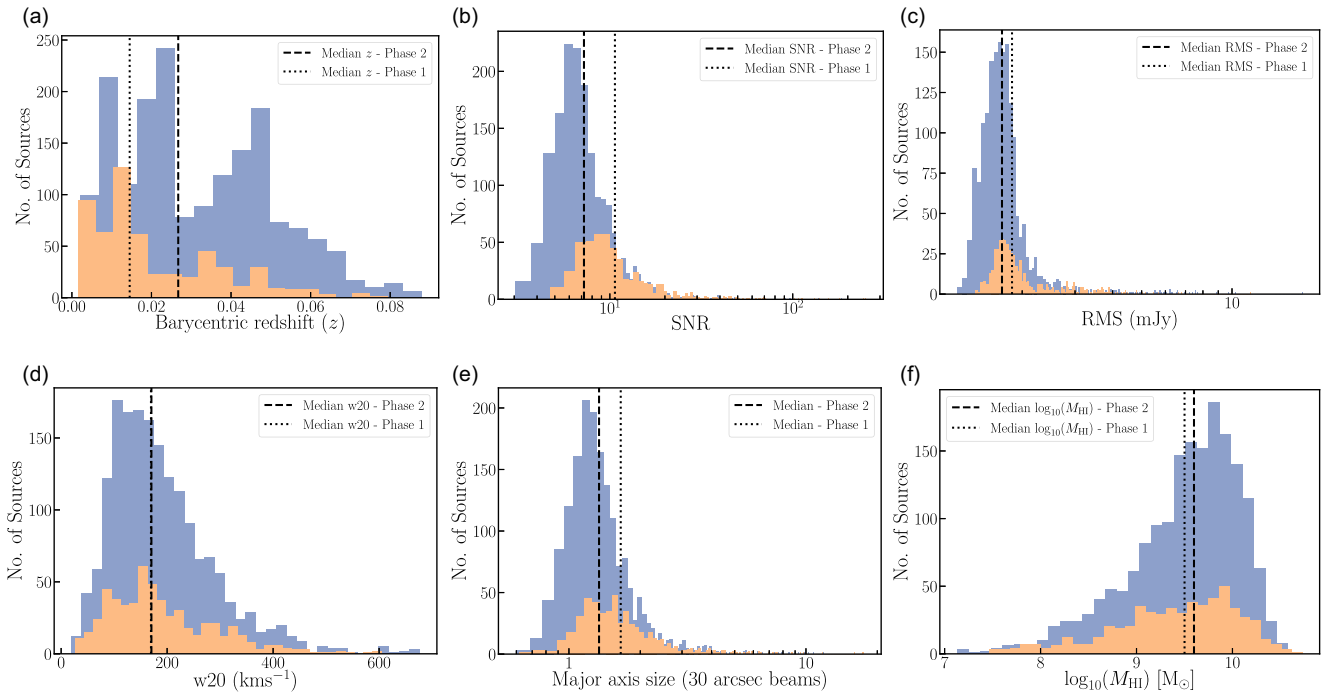
In panel b) of Figure 3 we show the SNR (defined as the ratio of the integrated flux to the uncertainty in the integrated flux measured by SoFiA) of the detected sources for both the Phase 1 and Phase 2 fields. As reported in Westmeier *et al.* (2022), the peak of the SNR for the Phase 1 data is  $\sim 9$  (with median  $\sim 11$ ), while the peak of the SNR distribution for the Phase 2 detections is  $\sim 6$  (with median  $\sim 7$ ). This significant improvement in detecting low SNR sources in Phase 2 from the source finding runs can mainly be attributed to the following reasons – a) the overall data quality of the Phase 2 observations has improved significantly compared to Phase 1 data mainly because the fields targeted in Phase 2 were chosen specifically to avoid continuum sources brighter than 2 Jy. This leads to better data quality with fewer continuum-related artefacts, leading to the source finding runs being more complete out to low SNR; b) the on-dish calibrators were switched off for Phase 2, as they had caused a lot of RFI in the Phase 1 data, particularly in the corner beams; c) the SoFiA settings were fine-tuned based on the experience participating in the SKA Science Data Challenge 2 (Hartley *et al.*, 2023), which has also contributed to a higher completeness of the catalogue in Phase 2.

Panel c) in Figure 3 shows the distribution of the local rms noise in the image cubes for both Phase 1 and 2 sources. The median rms in the image cubes for Phase 2 is  $\sim 1.7 \text{ mJy per } 30'' \text{ beam}$  and  $18.5 \text{ kHz}$  ( $\sim 4 \text{ km s}^{-1}$ ) channel width, which is close to the expected theoretical rms noise in the image cube for WALLABY (Koribalski *et al.*, 2020). This translates to a  $5\sigma$  H I column density ( $N_{\text{HI}}$ ) sensitivity of  $\sim 9.1 \times 10^{19} (1+z)^4 \text{ cm}^{-2}$  per  $30''$  beam and  $\sim 20 \text{ km s}^{-1}$  channel, and a  $5\sigma$  H I mass sensitivity of  $\sim 5.5 \times 10^8 (D/100 \text{ Mpc})^2 M_{\odot}$  for point sources, where  $D$  is the Hubble distance to the source.

In terms of the line width of the detections in Phase 2, we show the distribution of the  $w_{20}$  line-widths (defined as the spectral width corresponding to 20% of the peak flux in the integrated spectrum) for both the Phase 1 and Phase 2 samples in panel d) in Figure 3. The median  $w_{20}$  value for both the samples is  $\sim 170 \text{ km s}^{-1}$ .

Panel e) in Figure 3 shows the distribution of the major axis size of the ellipse fit to the moment 0 map of the detections. It can be seen that the median size of sources detected in Phase 2 is  $\sim 1.3$  beams, at the nominal  $30''$  resolution, compared to a median value of  $\sim 1.6$  for the Phase 1 detections. This means that WALLABY has managed to detect a larger number of marginally resolved galaxies in Phase 2, primarily because the median redshift of Phase 2 detections is a factor of two higher than the median value for Phase 1 observations.

Panel f) in Figure 3 shows the distribution of the H I mass for all pilot Phase 1 and 2 detections. The H I mass is computed using equation 7 in the PDR1 paper (Westmeier *et al.*, 2022). We observe that the Phase 2 detections have a median H I mass of  $\log_{10} (M_{\text{HI}}/M_{\odot}) \sim 9.6$  which is consistent with the median H I



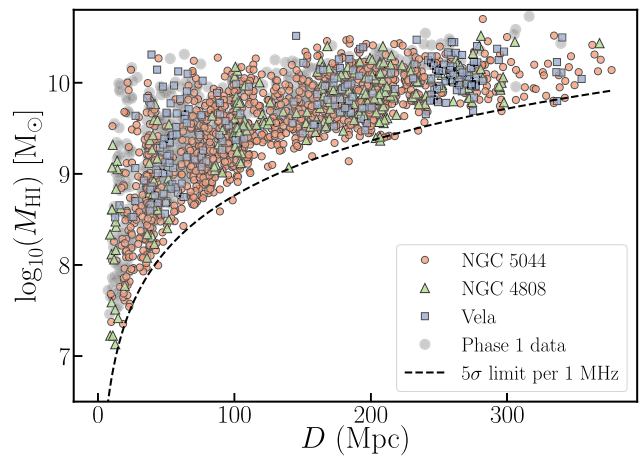
**Figure 3.** **a**) Distribution of the barycentric redshifts of the Phase 2 sources (blue) compared to the Phase 1 detections (orange). **b**) Histogram of the Signal-to-noise (SNR) for both the Phase 2 and Phase 1 detections. **c**) Local noise distribution in the images cubes for the Phase 2 and Phase 1 detections. **d**) Distribution of the  $w_{20}$  H I line-width distribution. **e**) Histogram of the major axis size (in units of  $30''$  beams) for the two samples. **f**) The H I mass distribution for the Phase 2 and Phase 1 samples. In all plots, the dashed and dotted black lines represents the median value of the distribution for the Phase 2 and Phase 1 detections, respectively.

mass value of  $\log_{10}(M_{\text{HI}}/M_{\odot}) \sim 9.5$  for the pilot Phase 1 detections. The phase 2 median H I mass is slightly higher than the Phase 1 median H I mass, which is expected from the higher median redshift of the Phase 2 sample. We note that we make use of the Hubble distance,  $D = \nu H_0$ , of the sources to estimate their H I mass. Where  $\nu$  is the measured barycentric velocity and  $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$  is the Hubble constant. We caution that this distance is only an approximation and will be prone to large errors of up to  $\sim 20\%$  due to effects of peculiar velocities in the local Universe, as well as systematic errors from using barycentric redshifts (Strauss & Willick 1995; Willick et al. 1997). We have used the Hubble distances for this release to remain consistent with the distance estimates used in Phase 1. However, going forward, for the full survey, the WALLABY team plans to apply more sophisticated flow models and correct the redshifts appropriately before measuring derived quantities such as distances and H I masses.

Figure 4 shows the distribution of the H I mass of the detections from both pilot Phase 1 (grey points) and Phase 2 (color-coded by the different fields) as a function of their measured Hubble distance ( $D = \nu H_0$ ). Also plotted is the  $5\sigma$  H I mass detection threshold (dashed black line) measured across a 1 MHz frequency bandwidth and assuming the median local RMS noise level of  $\sim 1.71 \text{ mJy}$  in the image cubes derived from the SoFiA runs. The  $5\sigma$  H I mass detection threshold is computed as follows:

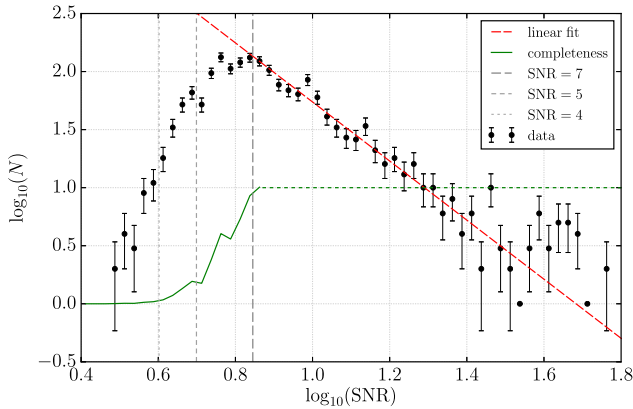
$$\frac{M_{\text{HI}}(5\sigma)}{M_{\odot}} = \frac{5 \times 49.7 \times \left(\frac{\sigma}{\text{Jy Hz}}\right)}{\sqrt{\Delta\nu/d\nu}} \left(\frac{D}{\text{Mpc}}\right)^2 \quad (1)$$

where  $\sigma = 1.71 \times 10^3 \text{ Jy Hz}$  and  $\Delta\nu = 1000 \text{ kHz}$  is the 1 MHz channel width and  $d\nu = 18.5 \text{ kHz}$  is the default spectral resolution.



**Figure 4.** The H I mass plotted against the estimated Hubble distance for the combined Pilot Phase 2 sample. The orange circles represent the NGC 5044 field, green triangles the NGC 4808 field and the purple squares the Vela field. The grey circles in the background represent the Phase 1 detections. The dashed black line represents the  $5\sigma$  H I mass threshold as a function of distance, assuming a 1 MHz frequency band width.

We see that our completeness at  $5\sigma$  is close to zero in accordance with Figure 5 in Section 4.1. As with the Phase 1 detections, we find large-scale clustering at various distances corresponding to the different groups detected in the Phase 2 fields. For example, for the NGC 4808 field, we find galaxies clustered at  $\sim 30 \text{ Mpc}$ ,  $\sim 100 \text{ Mpc}$  and another over-density close to  $\sim 200 \text{ Mpc}$ . Similarly, for the Vela field, we find an over-density of galaxies corresponding to a distance of  $\sim 50 \text{ Mpc}$ , at  $\sim 180 \text{ Mpc}$  and another at  $\sim 260$



**Figure 5.** Histogram of the number of detected sources,  $N$ , as a function of integrated signal-to-noise ratio, SNR, in double-logarithmic space in bins of  $\Delta \log_{10}(\text{SNR}) = 0.025$  (black data points). The error bars correspond to  $\sqrt{N}$ . The red, dashed line shows the result of a linear fit in the range of  $0.9 \log_{10}(\text{SNR}) 1.4$ . The resulting completeness, defined as the observed source count divided by the fit, is shown as the green, solid curve at  $\text{SNR} \lesssim 7$  where incompleteness effects are evident.

Mpc. The over-density at  $\sim 260$  Mpc in Vela field is particularly interesting as it lies in the Zone of Avoidance (ZOA) and as such there are limited redshifts. However, a few previous optical studies (e.g. Hudson et al. 2004; Hoffman et al. 2015) have hinted at the existence of a large over-density corresponding to a systemic velocity of  $\sim 18000 \text{ km s}^{-1}$  (roughly a distance of 260 Mpc). This was later confirmed by Kraan-Korteweg et al. (2017), who measured the spectra from  $\sim 4500$  galaxies to map the composition and structure of the over-density. Studying and understanding this large-scale structure will add immensely to our knowledge of modelling bulk flows in the local Universe, as well as mapping the large-scale structures in the ZOA.

#### 4.1. Completeness

In order to estimate the completeness of the source catalogue, we plot in Figure 5 the number of sources,  $N$ , as a function of integrated signal-to-noise ratio (SNR) in double-logarithmic space. As before the SNR is defined here as the ratio of the integrated flux and the statistical uncertainty of the integrated flux measurement within the source mask produced by SoFiA. As expected from an untargeted survey, the source count follows an almost perfect power-law with a turnover at  $\text{SNR} \lesssim 7$ . Under the assumption that the intrinsic population continues to follow a power law at low SNR and that the turnover therefore is entirely caused by incompleteness, we can estimate completeness as a function of SNR. We do this by fitting a straight line to the data points in the range of  $0.9 \log_{10}(\text{SNR}) 1.4$  (red, dashed line) which yields a slope, and hence power-law exponent, of  $-2.54$ . The completeness of our source catalogue as a function of SNR can then be estimated by dividing the number of detected sources in each bin by the expected number of sources predicted by the power-law fit.

The resulting completeness curve is shown as the green, solid line in Figure 5. We reach 100% completeness at  $\text{SNR} \approx 7$  beyond which we do not plot the actual completeness curve any more, as it would eventually show a large scatter around 1 due to stochastic errors as a result of low source counts at high SNR. 50% completeness is reached at  $\text{SNR} \approx 5.5$  below which our completeness rapidly declines to near zero at  $\text{SNR} \approx 4$ .

#### 5. High-resolution $12''$ cut-outs

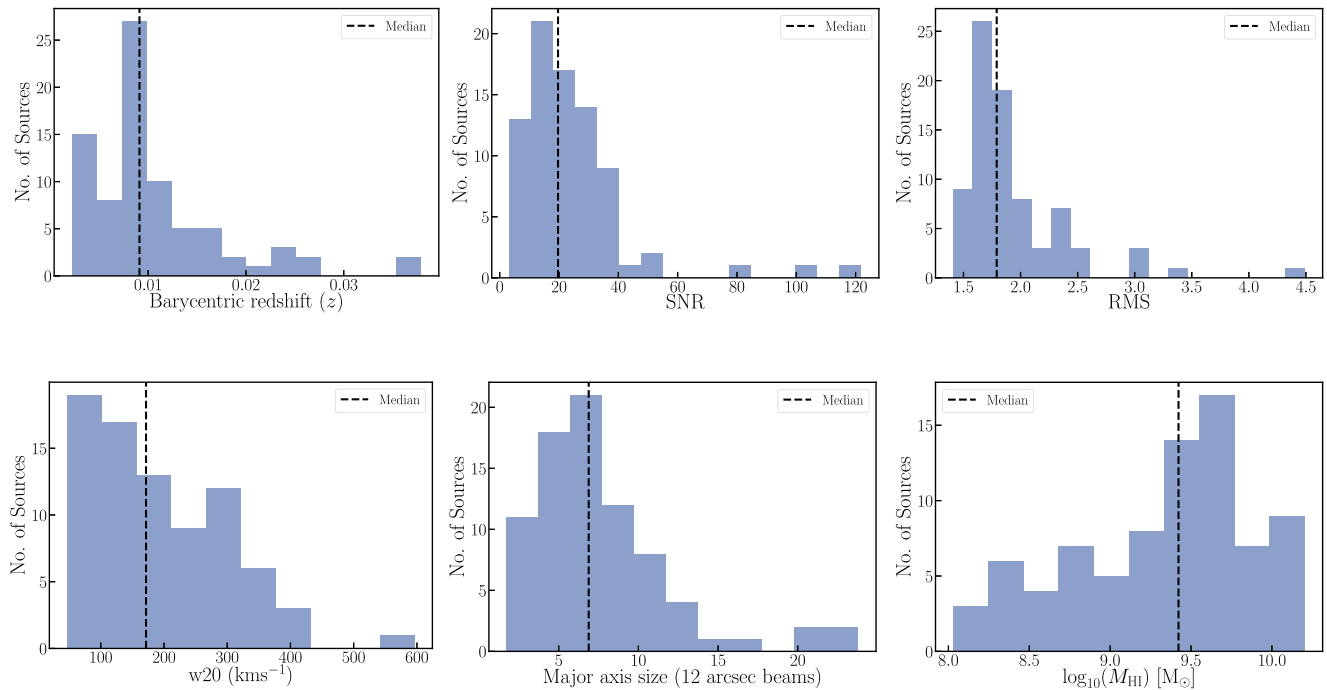
One of the objectives of the WALLABY survey is to generate high-resolution ( $12''$ ) cut-outs for a sub-sample of galaxies. We use the calibrated visibility data derived from the default ASKAP spectral-line processing pipeline (Guzman et al. 2019; Whiting 2020) to image a sub-sample of galaxies at high angular resolution. As mentioned earlier the default spatial resolution of the WALLABY survey is  $30''$ , which was determined to be the optimal resolution that gives a good compromise between resolution, sensitivity, and computational resources required to process large volumes of data. In contrast, the computational resources required to image the data in the full  $12''$  resolution will be significantly higher due to the additional baselines and increasing image sizes. However, it is still possible to image a sub-sample of the WALLABY detections in high-resolution by limiting the bandwidth to be imaged to a few hundred channels and only encompassing the velocity range of the target galaxies. This way, we drastically reduce the computing and storage requirements to process the data. We tested this functionality in preparation for the full WALLABY survey in Phase 2.

For Phase 2, we selected all HIPASS sources from the three fields. We targeted HIPASS sources, as these are likely to be detected in the WALLABY data and also as they are well resolved (tens of  $12''$  beams across the major axis). We note that for the full WALLABY survey, apart from the HIPASS targets, some optically-selected target galaxies are also expected to be included. We emphasise here that since the target galaxies for the high-resolution cut-outs are HIPASS galaxies and therefore H I-selected, this will naturally introduce biases in the sample, which the users need to consider and account for while using the data for their analysis.

To perform the high-resolution imaging making use of the full visibility including the longest baselines, we split out individual ASKAP primary beams containing (and surrounding) our target sources. We split out 250 channels ( $\sim 4.6$  MHz) encompassing the velocity range of the source. For the WALLABY channel width of  $\sim 4 \text{ km s}^{-1}$ , this translates to a total velocity range of  $\sim 1000 \text{ km s}^{-1}$ , which is sufficient to contain the emission from even the most massive and rotationally-dominated galaxies. We split out only 250 channels mainly to bring down the storage and processing costs required for each source. We split out up to 3 PAF beams from each footprint for each source, i.e. up to a total of 6 beams for a single source from both footprints. Each calibrated visibility data set of 250 channels for each beam is  $\sim 15$ GB in size, therefore the total storage cost for each source for 6 beams is  $\sim 90$ GB. The split-out visibilities are then uploaded on to CASDA. The splitting of the visibilities described above is performed automatically whenever a new field has been observed and processed.

The relevant visibilities for each source are then downloaded from CASDA and used to make the high-resolution image cubes using the “high-resolution” imaging pipeline (hereafter high-res pipeline). All data have been reduced on Pawsey Supercomputing Facility’s dedicated High Performance Computing clusters. We make use of ASKAPSoft to process and image the cut-outs. We now describe the various stages of the high-res pipeline. The pipeline is a Python script that reads in a catalogue of sources that need to be imaged, and a user-defined configuration file containing essential information such as the location of the split-out calibrated visibility, holography, and footprint data. The main





**Figure 6.** Plots show the source properties of the 12'' detections in the Phase 2 sample. *Top left:* Distribution of the barycentric redshifts of the 12'' detections. Histogram of the Signal-to-noise (SNR) of the 12'' detections. Local rms noise distribution in the image cubes. Distribution of the  $w_{20}$  H I line-width distribution. Histogram of the major axis size (in units of 12'' beams). The H I mass distribution. In all plots, the dashed black line represents the median value of the distribution.

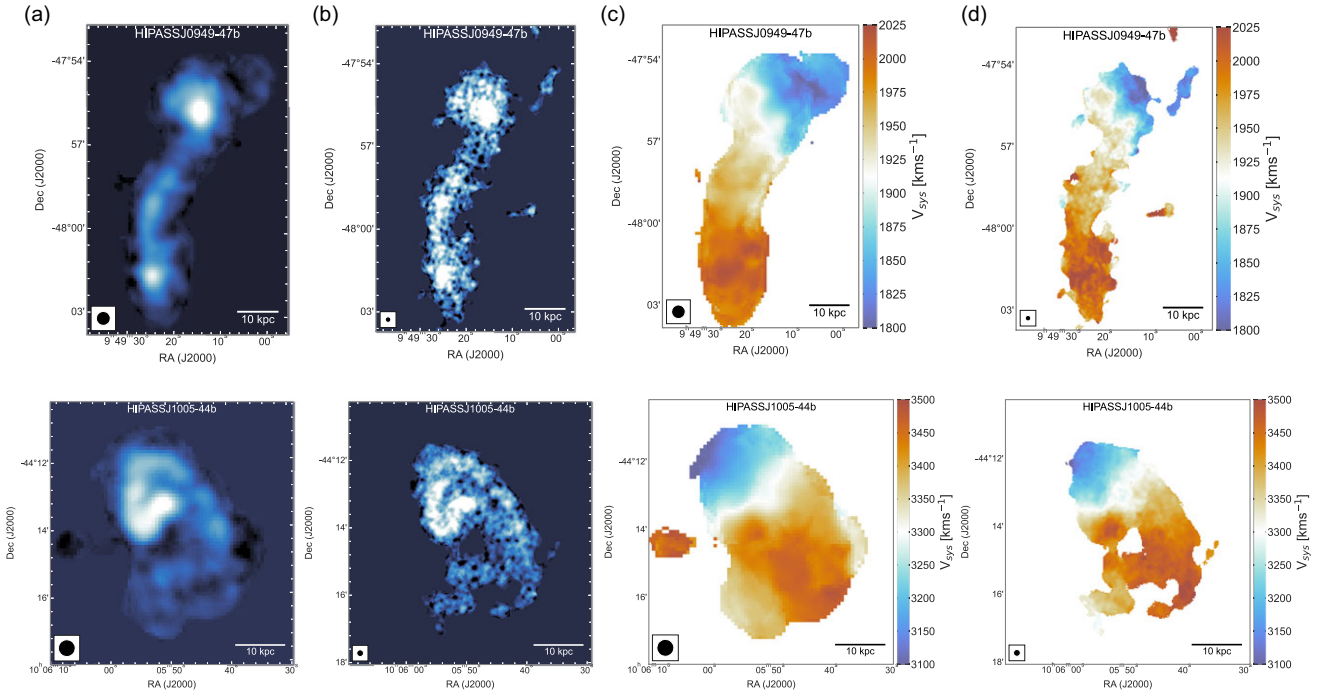
Python pipeline job then creates all the necessary bash scripts such as the parsets and the corresponding slurm job submission scripts for each task (e.g. imager, imcontsub, etc). These jobs for the various tasks are then submitted as dependencies for each beam for each individual source in a parallel framework.

The imaging is first carried out beam-by-beam and then all beams are mosaicked to produce the final image cube for the individual sources. The first step is to image the visibilities for each beam using the `cimager` task in ASKAPSoft. We make an image of size  $384 \times 384$  pixels, with a pixel size of  $2''$ . We use a Wiener filter with a robust parameter value set to 0.5 and apply a Gaussian taper of  $12''$  to achieve a synthesised beam close to  $12''$ . The spectral resolution is kept at 18.5 kHz ( $\sim 4 \text{ km s}^{-1}$ ). In addition, the deconvolution process is also performed within the task `cimager`. For more details on the ASKAPSoft parameters used for the imaging, refer to Table 4 in Appendix A. The imaging step is then followed by the image-based continuum subtraction using the task `imcontsub`. The pipeline then performs the primary beam correction using the holography model with the task `linmos`. These steps are performed for each of the 6 beams that encompass the target HIPASS source. As the final step, all 6 beams are mosaicked to form the final ‘mosaicked’ cube for the source. This is again performed using the mosaicking task `linmos`. The above workflow is adopted for all sources, and a number of jobs are submitted on the cluster to simultaneously image the data for multiple sources at any given time. We now present an overview of the cut-outs sample, and give details of the quality of the data, including the typical SNR of the detections, size distribution, and their H I mass range. In addition, we also compare the properties of the 12'' detections with their corresponding 30'' counterparts.

### 5.1. 12'' imaging results

A total of 73 HIPASS target galaxies were imaged in high-resolution as part of the Pilot Survey Phase 2. We note that in the majority of cases the target HIPASS galaxy is the only genuine detection in the image cube. However, in a few cases, source finding on some target HIPASS image cubes resulted in the detection of genuine smaller sources surrounding the target HIPASS galaxy. Once the source finding is complete, each tentative detection is visually examined to verify if it is a genuine source and is then added to the final source catalogue. A total of 80 sources are detected from the source finding runs from all three Phase 2 fields combined. Most detections in the high-resolution image cubes are also detected in the default 30'' data cubes, however, in some cases it is observed that a 30'' source in the default WALLABY catalogue is split-up into multiple components, with each component being a genuine nearby galaxy in the vicinity of a large galaxy. In such cases, each 12'' component is assigned a unique WALLABY name.

In Figure 6, we plot some of the source characteristics of the 12'' sample. Given that most targeted 12'' sources are HIPASS detections, the redshift distribution of the sample ranges from  $0.002z0.04$ , with a median  $z \sim 0.01$ . We find that the median SNR of the 12'' detections is  $\sim 20$ , while the rms in the local image cubes of the 12'' detections is found have a median value of  $\sim 1.8$  mJy, which is close to the expected theoretical rms of 1.75 mJy (using `robust=0.5` and all baselines including 6 km). This translates to a  $5\sigma$  H I column density ( $N_{\text{HI}}$ ) sensitivity limit of  $\sim 6 \times 10^{20} (1+z)^4 \text{ cm}^{-2}$  assuming a  $12''$  beam and a  $20 \text{ km s}^{-1}$  channel width, which is a factor of 6.6 higher compared to the 30'' data. This is a natural compromise between sensitivity and spatial resolution that is associated with higher-resolution observations and we advice



**Figure 7.** The comparison of moment 0 and moment 1 maps for two galaxies (top: HIPASS J0949-47b, bottom: HIPASS J1005-44b) with a resolution of 30'' and 12''. In each row, panels (a) and (c) show the moment 0 and 1 maps with a resolution of 30'' while panels (b) and (d) show the corresponding 12'' maps. At the bottom of each figure, we show the respective beam size as black circles and a scale bar set to 10 kpc.

the user to be cognisant of this compromise in sensitivity when dealing with the high-resolution data.

We also note that the distribution of the  $w_{20}$  H I line-width for the 12'' detections ranges from  $46w_{20}$  ( $\text{km s}^{-1}$ )597, with a median  $w_{20} \sim 172 \text{ km s}^{-1}$ , which indicates that the majority of the high-resolution sources are likely to be rotationally-supported late-type galaxies. We examined the moment maps and the corresponding optical image for the obvious outlier (with  $w_{20} \approx 597 \text{ km s}^{-1}$ ) and find that the H I emission is much more extended compared to the optical disk, along with kinematic warps and other signatures indicating that this galaxy is likely undergoing an interaction and may have accreted H I gas from a gas-rich low-mass companion. As the SoFiA mask encompasses all the H I emission, it results in considerably broadening the velocity width of this detection. Most of the 12'' detections are well resolved with their major axis size typically spanning  $\sim 7$  (12'') beams across. Finally, we note that the H I mass distribution of the high-resolution sample is  $8.0 \leq \log_{10} \left( \frac{M_{\text{HI}}}{M_{\odot}} \right) \leq 10.2$ , with a sample median of  $\log_{10} \left( \frac{M_{\text{HI}}}{M_{\odot}} \right) \sim 9.42$ .

Figure 7 shows moment 0 (intensity) and 1 (velocity) maps of two interacting system of galaxies in the default 30'' and 12''-resolution. From the images it is very clear that finer details in the H I morphology begin to show-up in the high-resolution images. The high-resolution moment maps highlight the distribution of the high-column density H I gas in the galaxies, which are otherwise washed-out in the 30'' images. In addition, in Figure 8 we show the 30'' and 12'' resolution H I contours overlaid on top of a composite (g,z,i-band) DESI Legacy Survey image for the galaxy NGC 5054. The two contours show H I column densities of  $2.4 \times 10^{20} \text{ cm}^{-2}$  (light orange) and  $7.2 \times 10^{20} \text{ cm}^{-2}$  (dark orange), respectively. The contours correspond to a SNR of 4 and 10 in the 30'' image, respectively, while corresponding to a SNR of 2 and 6

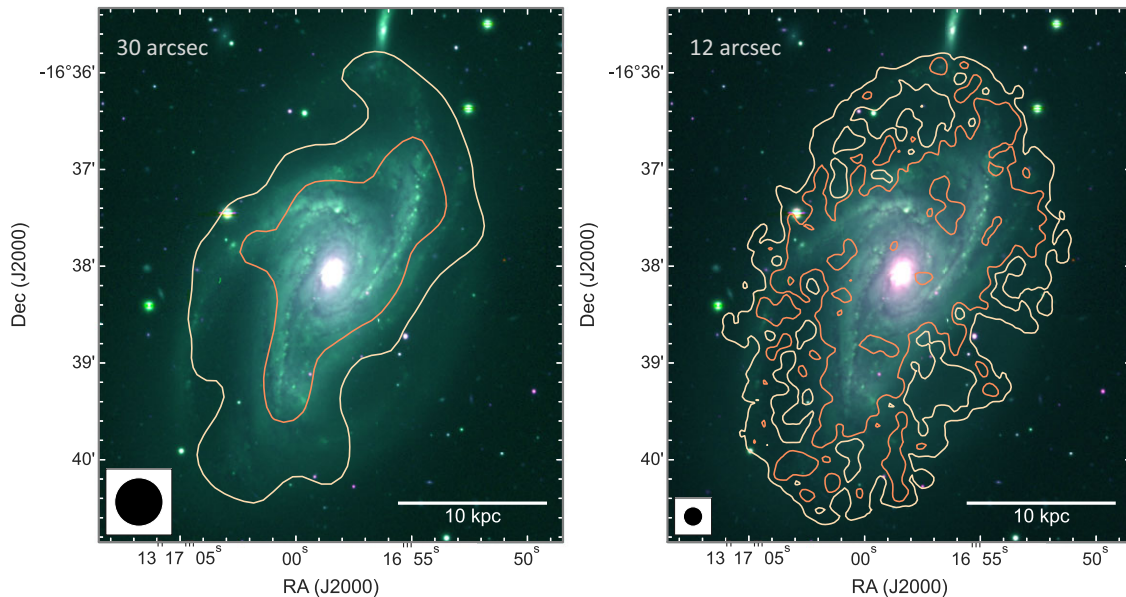
in the 12'' image. Compared to the 30'' resolution H I contours, the 12'' resolution contours clearly trace the high-column density H I gas along the spiral arms in NGC 5044, allowing us to study both the H I gas and star formation properties at a much higher resolution. A factor of  $\sim 3$  improvement in resolution will significantly aid in studies directed towards understanding the distribution of the high-column density gas in galaxies and also enable us to more accurately probe the connection between H I gas, star formation and star formation laws. In addition, the higher resolution enables us to model the kinematics of the H I gas more accurately.

## 5.2. Data quality and known issues with the high-resolution data

We do note that while the overall quality of the 12'' data is good, there were some issues identified with the imaging pipeline as well as the data products. We list below some of the known issues with the cut-outs in this data release.

**Flux discrepancy:** We note that the 12'' sources show a higher integrated flux compared to their 30'' counterparts. The flux of the 12'' sources is on average  $\sim 15\%$  higher compared to their 30'' counterparts. We present a more thorough discussion on this flux discrepancy in Section 6 and also highlight the likely origins of the discrepancy.

**Different synthesized beam size:** Some sources from the NGC 5044 field (tile4) have a different angular resolution. These data sets have a synthesized beam of  $\sim 17''$  instead of 12''. There are 14 such sources. This is because a slightly different tapering was applied during the imaging stage. The visibilities for these sources were not stored as the observations for the NGC 5044 tile 4 were carried-out before the scheme of storing visibilities on to CASDA was introduced. As such, the visibilities for these sources were



**Figure 8.** 30'' resolution H I contours overlaid on top of a composite (g,z,i) DESI Legacy Survey image of the galaxy NGC 5054. Corresponding 12'' resolution H I contours. In both cases the contours levels are set at column densities of  $2.4 \times 10^{20} \text{ cm}^{-2}$  (light orange) and  $7.2 \times 10^{20} \text{ cm}^{-2}$  (dark orange).

unfortunately unavailable to be re-imaged to a 12'' resolution. We have included a comment in the source catalogue for all relevant affected sources to highlight this.

**Unreliable spectra:** 7 sources in the 12'' data show bad spectra. These are typically edge-on galaxies with large spectral widths. Given that only 250 channels are split-out for the high-resolution imaging, we suspect that there were not enough line-free channels for the image-based continuum subtraction routine in ASKAPSoft to properly perform the continuum subtraction, leading to over-subtraction. Sources affected by this issue have a **qflag = 128** in the source catalogue.

**No default 30'' WALLABY cross-match:** We note that 6 sources in the cut-outs source catalogue do not have a corresponding default 30'' WALLABY detection. Upon further examination, it was found that the missing sources in the 30'' WALLABY catalogue are due to one of the following reasons.

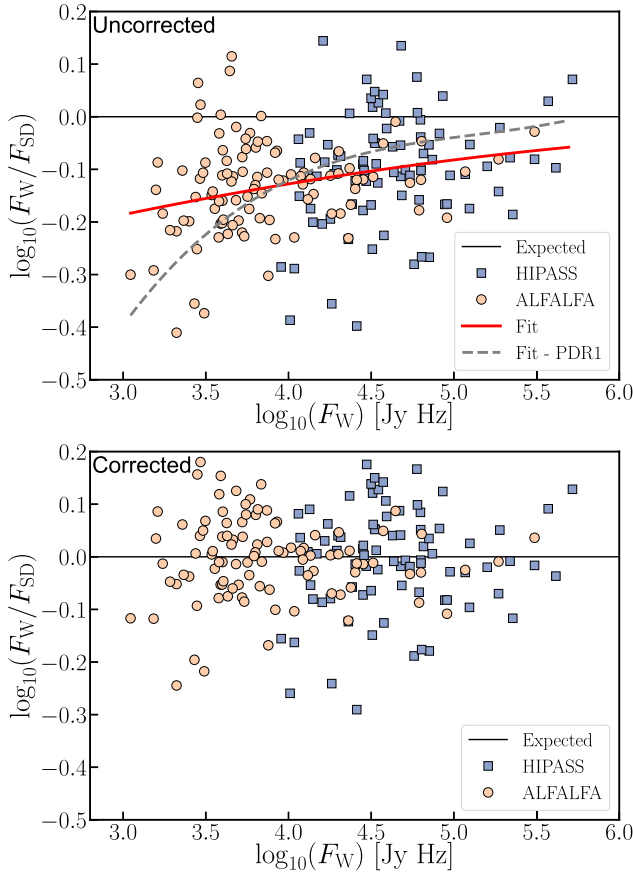
- Source lies in the Galactic velocity range. The default 30'' source finding runs are only performed on the extragalactic velocity range ( $cz \sim 500 - 26500 \text{ km s}^{-1}$ ) and as a consequence all sources below a velocity of  $cz < 500 \text{ km s}^{-1}$  are excluded from the source finding runs. Two sources are missing due to this limitation.
- Source is in the corner of a footprint. The SoFiA source finding runs are only performed on the inner  $4^\circ \times 4^\circ$  area of the mosaicked footprint as the outer edges of the footprint suffer from lower SNR and sensitivity as the noise increases by a factor of two. For this reason, some sources in the outer parts of the specific footprint may have been omitted in the current default 30'' source finding run. These sources will however be added to the catalogue whenever overlapping footprints are subsequently processed and available for source finding. Three sources are missed due to this.
- Very faint sources near the detection threshold may be missed in the global 30'' source finding, as the

completeness curve is known to gradually decrease below an SNR of  $\sim 7 - 8$  (see Figure 5). Since the high-resolution source finding involves checking and verifying each individual detection, in some cases it is possible to detect sources close to the detection threshold of the source finding runs. One source is missed due to this issue.

## 6. Flux discrepancy

As mentioned in the previous sections, the 30'' detections are observed to have a flux deficiency of  $\sim 15\%$  compared to single-dish observations (see upper panel in Figure 9). The flux discrepancy between the 30'' detections and their single-dish observations was also observed in the Phase 1 release. A number of independent effects may be contributing to this issue including flux that is genuinely missed in the 30'' data from extended diffuse emission; inadequate deconvolution threshold, which leads to contributions from uncleaned flux and negative side lobes associated with the 30'' dirty beam (see Westmeier et al. 2022); as well as systematic flux offsets due to the different procedures implemented for the primary beam correction between Phase 1 and Phase 2. While it is difficult to estimate the typical fraction of flux that is missed from diffuse emission, it is easier to resort to simulations of ASKAP observations in order to specifically understand the impact of inadequate deconvolution thresholds on the measured flux. We discuss this in more detail in Section 6.1.

In addition, we observe a flux discrepancy between the 12'' and the 30'' sources, wherein the 12'' detections typically show  $\sim 15\%$  higher flux compared to their corresponding 30'' detections (see lower panel in Figure 10), while being consistent (albeit coincidentally) with their single-dish detections (see upper panel in Figure 10). We examined the ASKAP dirty beams for the 12'' and 30'' observations and find that there are significant positive side-lobes associated with the 12'' beam, whereas there are significant negative sidelobes observed in the 30'' dirty beams (see Figs. 15 and 16 in Appendix E). This observation is true for all declination

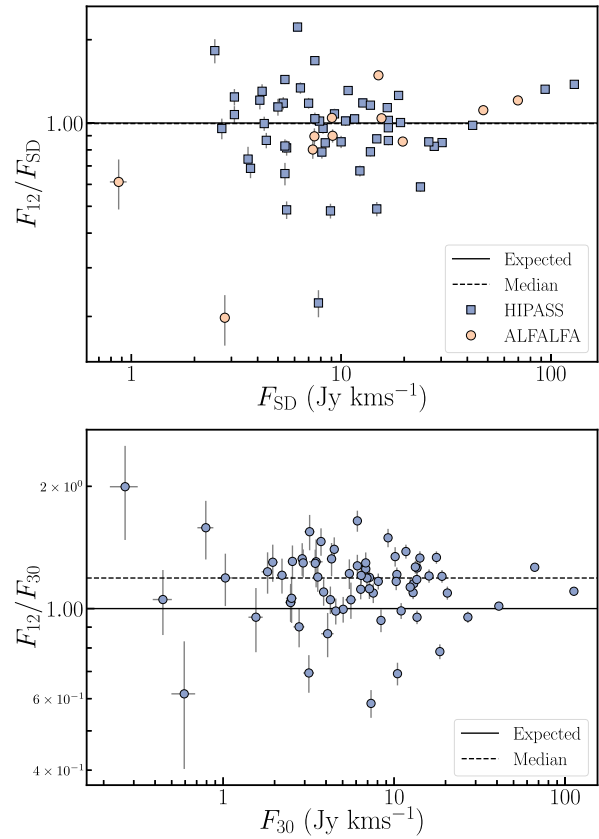


**Figure 9.** *Top:* The ratio of the WALLABY 30'' integrated flux to the single-dish integrated flux plotted against the WALLABY integrated flux for those galaxies which have a corresponding single-dish cross-match, either in ALFALFA and/or HIPASS. For the NGC 5044 and Vela fields, we use the HIPASS data and for the NGC 4808 field, we use the ALFALFA data for the flux comparison. *Bottom:* Similar plot as above, but now the WALLABY fluxes have been corrected using a polynomial fit to the data. The horizontal black line represents a flux ratio of one in both cases.

ranges ( $-47^\circ \delta +8^\circ$ ) covered by Phase 2 and could potentially lead to flux offsets, as any uncleaned flux in the residual maps impacted by the sidelobes of the dirty beams will contribute to the final image cube. In order to study the impact of the side lobes of the dirty beams for the two resolutions we resorted to simulations of ASKAP observations. In the subsection below we describe the details of the simulations and their main outcomes.

### 6.1. Simulating ASKAP observations

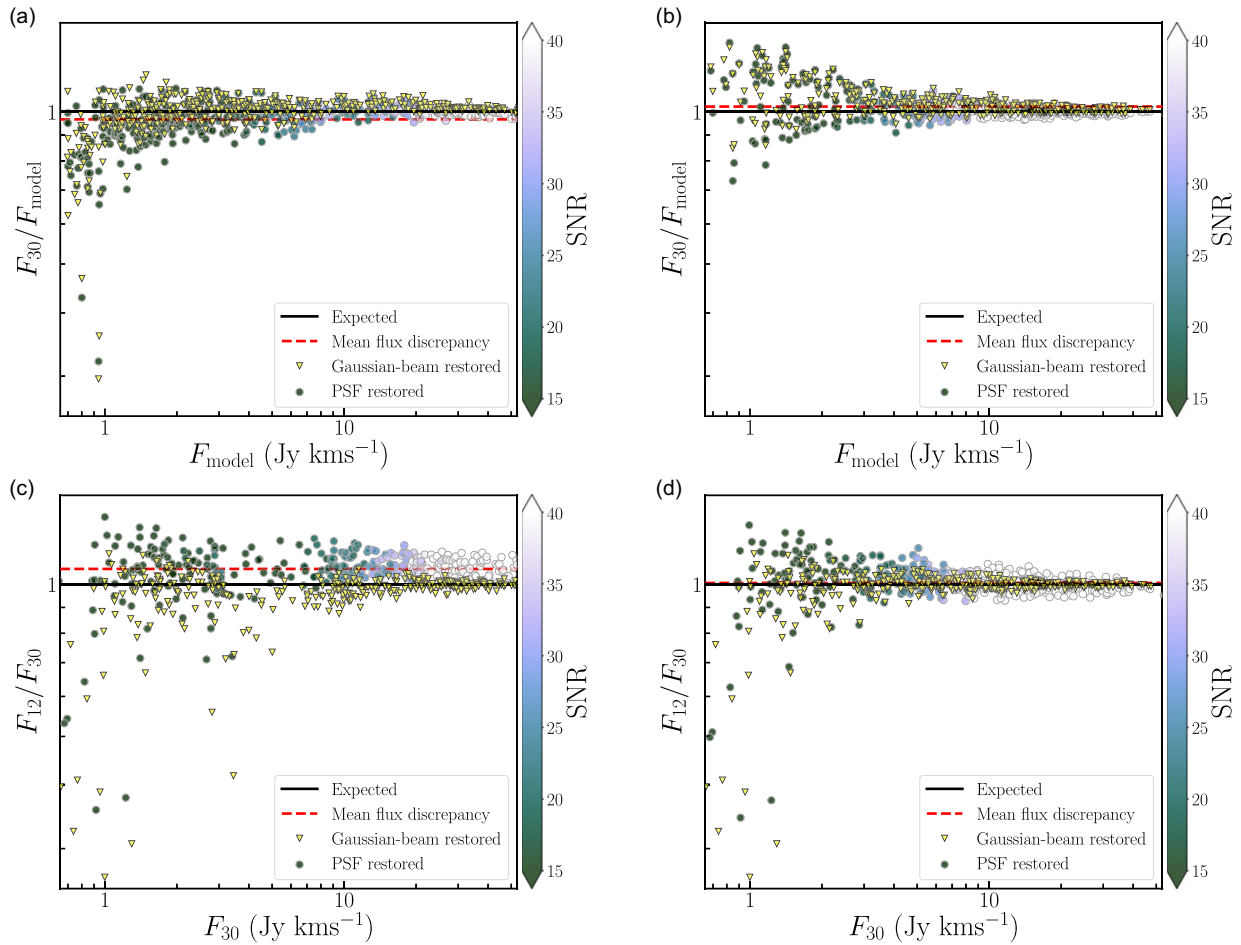
In order to generate mock ASKAP observations, we make use of the MIRIAD software package (Sault *et al.*, 1995). We use the MIRIAD task UVGEN to generate the mock visibilities. UVGEN takes in details of the mock observations, such as the positions of the ASKAP antennas, the correlator setup, the frequency of the observations, the RA and Dec of the pointing, hours of observation, integration time, the latitude of the observatory and the system temperature of ASKAP. This then generates the expected visibilities. We note that we generated mock visibilities at six different declinations (+8, +2, -11, -19, -24 and  $-45^\circ$ ) to represent the declination range observed in the Phase 2 fields. We generated  $4 \times 8$  h mock ASKAP observations to approximately emulate the ASKAP beams which are processed and imaged independently



**Figure 10.** Plot shows the ratio of the integrated flux of the 12'' ( $F_{12}$ ) to the 30'' flux ( $F_{30}$ ) for the overlapping sample. The black solid line represents the expected one-to-one line, and the dashed black line represents the median value of the  $\frac{F_{12}}{F_{30}}$  ratio.

before mosaicking them to form the final image cube with an rms noise close to  $\sim 1.6$  mJy per beam per channel, representative of the noise in typical WALLABY image cubes.

We Fourier-transform the mock visibilities to generate the dirty beam and dirty images at the 12'' and 30'' resolutions, using the MIRIAD task INVERT. We note that we were unable to yield a synthesised beam  $\sim 12''$  using a robust value of 0.5 and thus set the robust weighting to 0 and applied appropriate Gaussian tapering to the mock visibility data in order to achieve nearly 12'' and 30'' dirty beams. Next, we generate  $\sim 60$  mock galaxies per declination range, of varying size, surface densities, integrated flux, position angle and inclination angles using BBarolo's (Di Teodoro & Fraternali, 2015) GALMOD task. The size and the flux range of the model galaxies is set appropriately to reflect the respective range observed in the real 30'' and 12'' WALLABY detections. The model galaxy is then convolved with the 12'' and 30'' dirty beams, respectively, using the MIRIAD task CONVOL. The convolved galaxy models are then injected into the ASKAP observations dirty image cube (noise cube) using the task MATHS. This is then followed by the deconvolution step using MIRIAD's CLEAN task. We set the clean thresholds to match the settings used in the default ASKAPSoft imaging pipeline. This corresponds to a clean threshold in the minor cycle of 3.5 mJy ( $\sim 2\sigma$ ) and an additional second deep clean threshold of 0.5 mJy ( $\sim 0.3\sigma$ ). We set the maximum number of iterations (`n_iters`) for the minor clean cycles to be 800, again replicating the ASKAPSoft pipeline settings.



**Figure 11. a):** Circles show the ratio of the integrated flux of the injected model source convolved with the 30'' PSF ( $F_{30}$ ) to the total flux of the injected model galaxy ( $F_{\text{model}}$ ) for over 350 simulated galaxies in the declination range  $-47^\circ \leq \delta \leq +8^\circ$ . The data was cleaned to a residual flux threshold of 3.5 mJy in the minor CLEAN cycles. The inverted yellow triangles represent the flux ratio of the model sources convolved with a perfect 30'' Gaussian beam to that of the total flux of the injected source into the image cubes. **b):** Same as panel a), but now the sources were cleaned deeper to a residual flux threshold of 0.9 mJy. **c):** Shows the ratio of the integrated flux from the 12'' and 30'' model sources injected into to the image cubes and cleaned to a residual flux threshold of 3.5 mJy. **d):** Same as panel c), but now cleaned to a deeper residual flux threshold of 0.9 mJy. The points are color-coded based on the SNR of the 30'' detections. The black solid line represents the expected one-to-one ratio, while the dashed red line shows the mean flux discrepancy of the distribution.

After the cleaning stage, we restore the images using the task RESTOR. This takes in the dirty beam, the dirty image (with the injected model galaxy) and the clean components from the deconvolution step and generates the residual map as well as the final restored image cube by convolving the clean components with 12'' and 30'' Gaussian beams, respectively, and adding them to the residual map. We then mosaic all four simulated image cubes to generate the final image cube on which source finding is performed using SoFiA.

In addition, we also convolved the model galaxies with a Gaussian of full width at half maximum (fwhm) of 12'' and 30'' which will represent perfectly "cleaned" data. The fluxes derived from these Gaussian beam convolved data sets will not have the influence of the sidelobes that is typically observed in interferometric data sets where some uncleaned flux may remain, impacting the final measured flux as well as the quality of the final image cubes.

We run SoFiA on the simulated 30'' image cubes by using the default parameter settings currently used for the WALLABY source finding. We appropriately change some SoFiA parameters

for the high-resolution data sets as the default 30'' parameters are not optimal for source finding in the 12'' image cubes. For a list of important SoFiA parameters used for source finding on the 30'' and 12'' data-sets we refer the reader to Table 5 in Appendix B.

### 6.2. Origin of the flux discrepancy

We now report the main observations from our simulation experiment and delve into the details of the origin of the flux discrepancy that is observed between the single-dish, 30'' and 12'' detections.

Panel a) in Figure 11 shows a plot of the ratio of the 30'' integrated flux to the total flux of the injected model galaxy ( $\frac{F_{30}}{F_{\text{model}}}$ ) against the injected model galaxy flux ( $F_{\text{model}}$ ). The detections are color-coded based on their SNR values, which are computed as the total flux divided by the uncertainty in the flux measurement, both of which are provided by SoFiA in the source catalogue. As mentioned before, the CLEAN thresholds are set according to the default ASKAPSoft pipeline which is cleaning to a residual peak flux density threshold of 3.5 mJy. We find that indeed the measured flux of the mock galaxies in the 30'' resolution is consistently

lower than the flux of the injected model galaxies. The mean flux discrepancy is  $\sim 4\%$ , while the discrepancy becomes more and more prominent towards the low-flux and low-SNR regime, where the offset can be as much as 15%, corresponding closely to the discrepancy observed in the real data.

Following this, we explored the observed flux discrepancy between the 12'' and 30'' data. Panel c) in Figure 11 shows the ratio of the 12'' to 30'' fluxes ( $\frac{F_{12}}{F_{30}}$ ) plotted against the 30'' fluxes of injected model galaxies. We find that the 12'' fluxes are typically  $\sim 10\%$  higher compared to the 30'' fluxes, which is comparable to what is observed in real ASKAP observations (about  $\sim 15\%$ ). From these simulations, it is evident that the impact of the positive sidelobes on the uncleaned flux in the 12'' resolution is quite significant. A similar observation has also been reported in Radcliffe *et al.* (2024), who undertook a simulation study to investigate the impact of asymmetric PSFs from interferometers on the recovered flux from sources. Their study points to the fact that non Gaussian dirty beams (PSFs) lead to consistent flux offsets compared to the flux of the injected model source. We note that our simulations are in agreement with their observations, wherein the flux is more discrepant for the marginally resolved, low SNR sources, while being somewhat more consistent for extended sources.

As noted previously, this flux discrepancy may be an effect of incomplete cleaning, which results in the uncleaned flux being included into the final image cubes, impacting the final measured integrated flux. In order to investigate the impact of deeper cleaning, we cleaned the data to a peak residual flux threshold of  $\sim 0.5\sigma = 0.9$  mJy in the minor cycles with an additional deep cleaning threshold set to 0.1 mJy. The deeper deconvolution thresholds lead to better flux recovery for both the 30'' and 12'' data sets, with the measured and injected fluxes in better agreement. This is shown in panels b) and d) in Figure 11.

To summarise, we note the following observations from the simulation study.

- We find that the 30'' integrated flux is consistently lower by about 4% compared to the integrated flux of the input model galaxy. However, at the low flux (or SNR) end, the flux discrepancy can be as high as 15%, consistent with observations. This effect is observed in panel a) in Figure 11 (also Figure 9), where we observe that sources with an SNR 20 show a higher flux discrepancy.
- We find that the 12'' fluxes are on average consistently higher than the 30'' fluxes by  $\sim 4 - 10\%$ , depending on a number of factors, including the SNR of the data, as well as how extended and bright the source is.
- The 12'' and 30'' fluxes for the Gaussian beam convolved data sets are consistent with each other for the high SNR (SNR 20), while not surprisingly the 12'' fluxes for the marginally resolved and/or low SNR sources is typically lower than the 30'' fluxes.
- We note that by cleaning deeper in both the 30'' and 12'' data sets, we recover most of the flux, almost completely resolving the flux discrepancy. This suggests that deeper CLEANing thresholds are essential to fully recover the flux from WALLABY observations.

We do acknowledge that while care was taken to carry-out the simulations to reflect as closely as possible the ASKAPSoft pipeline, there are many subtle differences that might still impact

the way the data is processed and hence the recovered fluxes. Such effects are likely to impact sources in the low-flux (-SNR) regime more than well resolved and higher flux sources. However, despite these caveats, the simulations do highlight the importance of deeper cleaning of the ASKAP observations in order to recover fluxes properly.

### 6.3. Correcting the fluxes

We showed in the previous section through simulations that in order to properly recover the total flux from a source, we need to clean much deeper (potentially  $0.5\sigma$  or deeper) than the current deconvolution thresholds set in ASKAPSoft, in addition to using source masks generated from shallower CLEAN runs for further cleaning. Another important point is the implementation of joint deconvolution routines that enable the visibility data from all ASKAP primary beams be jointly imaged and deconvolved, so that the data can be cleaned to the appropriate deeper CLEAN thresholds. However, such a system is still not in place in the current ASKAPSoft pipeline and furthermore will require significantly more computational resources to process the data. This means that one cannot fully clean the data and therefore the impact of the sidelobes on the residual flux will remain an issue for both the 12'' and 30'' data-sets. However, the WALLABY team is currently testing newly implemented parameters in ASKAPSoft's deconvolution algorithm and optimal CLEAN thresholds will be implemented for the full WALLABY survey accordingly.

In the interim, we statistically correct the 30'' fluxes of Phase 2 sources by appropriately scaling their integrated fluxes to their corresponding single dish values from ALFALFA and HIPASS. We find that the data is best fit by a second order polynomial of the form

$$\log_{10} \left( \frac{F_W}{F_{SD}} \right) = -0.006448\phi^2 + 0.103635\phi - 0.439071 \quad (2)$$

where  $\phi = \log_{10} \left( \frac{F_W}{\text{Jy Hz}} \right)$ . The red line in the upper panel in Figure 9 shows the fit to the data, while the bottom panel shows the corrected fluxes. We see that this seems to systematically bring the flux level up, making it more consistent with the single-dish data. We note that we resorted to the new polynomial fit for the Phase 2 data, as the third order polynomial fit from PDRI (dashed grey line in the upper panel in Figure 9) does not fit the data very well and seems to over-correct the fluxes in the low-flux end. The reason for the very different flux offsets observed between Phase 1 and Phase 2 data is likely stemming from the fact that Phase 2 observations utilised the holography-based primary beam correction as opposed to the use of a Gaussian primary beam correction in Phase 1, which will lead to systematic offsets in the flux.

We note that these corrections have not been applied to the data products for each source included as part of this public data release, however, we have included the corrected fluxes in the catalogue and advise the users to be aware of this issue and apply the necessary correction to the fluxes when using the image cubes and moment maps for any analysis. The keywords **f\_sum\_corr** and **err\_f\_sum\_corr** in the source catalogue represent the corrected flux and error on the corrected flux, respectively. Similarly, the keyword **log\_m\_hi\_corr** represents the H I mass derived from the corrected flux and using the Hubble distance to the source.

## 7. Kinematic modelling

One of the goals of WALLABY is to generate kinematic models for as many galaxies as possible. For Phase 1 [Deg et al. (2022) Deg, Spekkens, Westmeier, Reynolds, Venkataraman, Goliath, Shen, Halloran, Bosma, Catinella, de Blok, D aes, DiTeodoro, Elagali, For, Howlett, J\_sa, Kamphuis, Kleiner, Koribalski, Lee-Waddell, Lelli, Lin, Murugesan, Oh, Rhee, Scott, Staveley-Smith, van der Hulst, Verdes-Montenegro, Wang, Wong] developed the WALLABY Kinematic Analysis Proto-Pipeline (WKAPP<sup>a</sup>) that is optimized for the low resolution and signal-to-noise (S/N) of the standard 30'' data. It uses a combination of two different tilted ring (TR) modelling algorithms to generate reliable kinematic models from observed source cubelets. It was used to generate the 109 kinematic models of WALLABY Phase 1 and we use it here on both the 30'' and 12'' source cubelets.

Tilted-ring modelling treats a galaxy as a series of nested rings described by a number of observational parameters (center, systemic velocity, position angle, and inclination angle) and intrinsic ones (surface density, disk thickness, rotation velocity, and velocity dispersion). This technique, introduced by (Rogstad et al., 1974), was first developed for 2D images and has been adapted to work with 3D data cubes. There are a number of advantages to working in 3D including the ability to apply more complex models to well resolved, high S/N data (see for instance J ozsa et al. 2009; Khoperskov et al. 2014; Di Teodoro & Peek 2021; J ozsa et al. 2021). More relevant to the WALLABY context, 3D TR algorithms are also able to model galaxies at lower spatial resolutions across a wider range of disk geometries than equivalent 2D algorithms (e.g. Kamphuis et al., 2015; Di Teodoro & Fraternali, 2015; Lewis, 2019; Jones et al., 2021).

While a full description of WKAPP is found in Deg et al. (2022) we will briefly describe the key points here. WKAPP combines fits from two different 3D TR algorithms to generate its models – Fully Automated TiRiFiC (FAT, Kamphuis et al. 2015), which itself is built on the Tilted Ring Fitting Code (TiRiFiC; J ozsa et al. 2007); and the 3D-Based Analysis of Rotating Objects From Line Observations (BBAROLO; Di Teodoro & Fraternali 2015). Both codes are run in a ‘flat-disk’ mode, where the observed geometry is constant across all rings. Deg et al. (2022) found that the differences between the two codes tended to be larger than the reported uncertainties of either algorithm. As such, WKAPP uses half the difference between the models as the better estimate of the model uncertainty, which is applied to all galaxies with either a SOFIA  $ell\_maj \geq 2$  beams or an integrated  $\log(S/N) \geq 1.25$ . The fits for each code are compared, and if both fits are reasonable, the two are averaged to create the final kinematic model.

In this section we describe the results of applying WKAPP to both the 30'' and 12'' data. Section 7.1 focuses on the 30'' Phase 2 data and how the models compare to Phase 1. Section 7.2 focuses on the 12'' data and how those models compare to the 30'' models.

### 7.1. Normal resolution modelling

While Phase 2 contains many more detections than Phase 1 (~1800 unique detections compared to the ~600 unique detections of Phase 1), these galaxies tend to be further away and smaller in size than the Phase 1 observations. Of these Phase 2 galaxies,

**Table 3.** The number of sources, attempts, and successful models in each release (where TR refers to Team Release). Note that there are no double sources in the 12'' data so a ‘Unique’ 12'' row is the same as the ‘Total’ 12'' row.

Source Release	<i>N</i> sources	<i>N</i> attempted	<i>N</i> models
NGC 4808 TR1	231	38	18
NGC 5044 TR1	353	55	23
NGC 5044 TR2	630	76	36
NGC 5044 TR3	1326	185	85
Vela TR1	203	43	17
Total 30''	2743	397	179
Unique 30''	1827	275	127
NGC 4808 High-Res TR1	12	12	5
NGC 5044 High-Res TR1	55	54	21
Vela High-Res TR1	13	13	4
Total 12''	80	79	30

only 275 have the requisite size and S/N to attempt kinematic modelling. For comparison, Phase 1 contains 209 unique galaxies that satisfy the  $ell\_maj \geq 2$  or  $\log(S/N) \geq 1.25$  criteria. Table 3 lists the sources, the galaxies that satisfied the modelling attempt criteria, and the total number of kinematic models for Phase 2. Considering only those galaxies where kinematic modelling is attempted, the 30'' Phase 2 sources have a success rate of 45%, which is comparable to Phase 1.

Using WKAPP, the final catalogue of Phase 2 kinematic models contains 127 unique galaxies. The left-hand panels of Figure 12 shows the kinematic models for these 127 models. These models span a wide range of rotation velocities and extents, including a number of low-mass dwarfs. It is worth noting that the surface densities are calculated through ellipse-fitting on the moment 0 map using the averaged model geometry. As such the deprojected profiles shown in Figure 12 have not been corrected for beam smearing effects, which may be important for some applications.

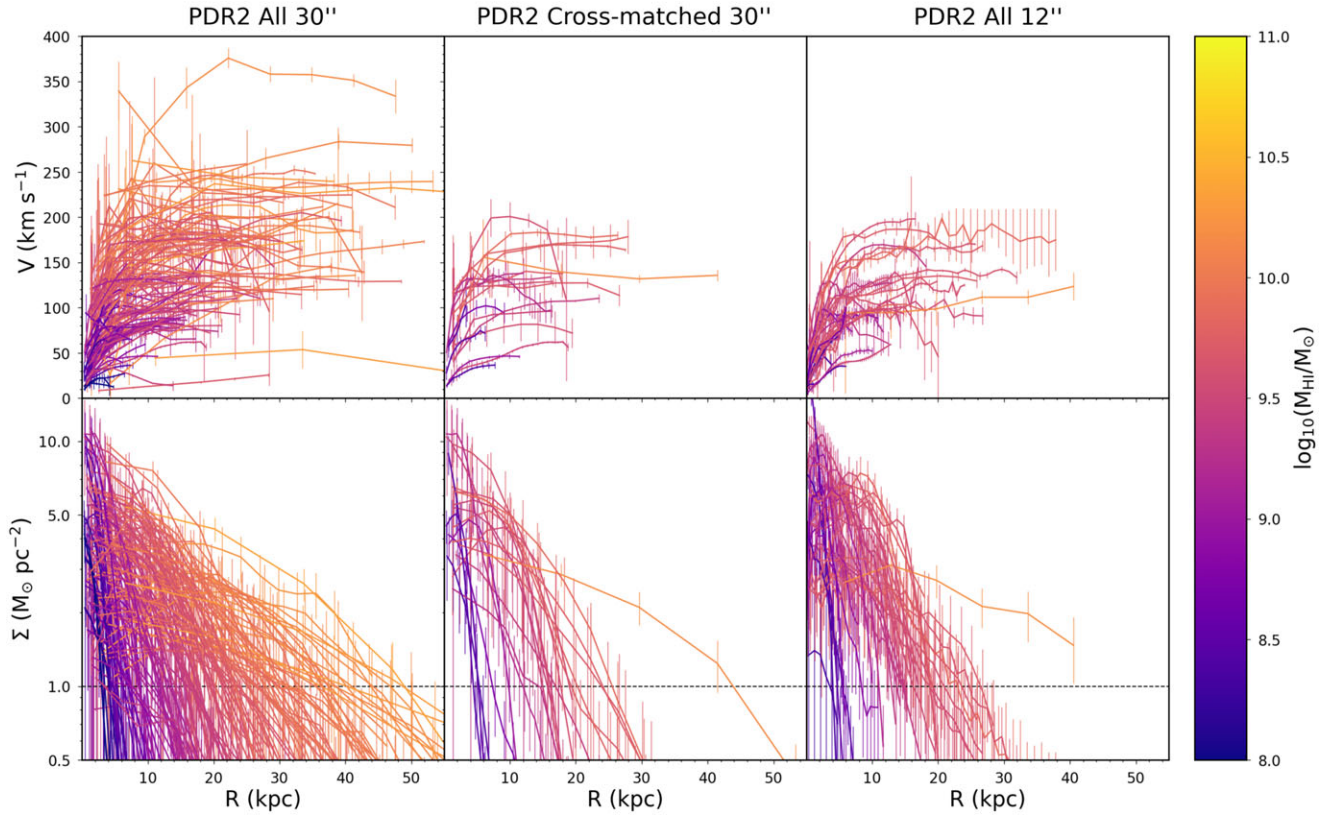
### 7.2. High resolution kinematic modelling

WKAPP was developed with an eye towards the 30'' WALLABY data, and it is not clear that this approach is appropriate for the higher resolution 12'' data. Nonetheless, we have applied the proto-pipeline data to the high resolution data and obtained 27 kinematic models from the 80 detections. The rotation curves and deprojected surface density profiles for these models are shown in the right-hand panels of Figure 12.

Comparing the left and right panels of Figure 12, it is clear that the distribution of 12'' and 30'' models are not the same. The 12'' models are biased towards higher rotation velocities and H I masses compared to the 30'' models (which show a wider range of velocities and H I masses). This is evidenced by the fact that very few modelled galaxies in the 12'' resolution have rotation velocities lower than 80 km s<sup>-1</sup> in the outer/flat parts of their rotation curves. This is likely due to HIPASS preferentially finding relatively nearby gas-rich galaxies, which are typically observed to have higher rotation velocities.

The middle panel of Figure 12 shows the 30'' models for galaxies that also have a 12'' model. Some galaxies in the 12'' sample that are successfully modelled do not have an equivalent 30'' model.

<sup>a</sup>WKAPP is available at <https://github.com/CIRADA-Tools/WKAPP>



**Figure 12.** The rotation curves (top row) and deprojected surface density profiles (bottom row) for Phase 2. The left-hand panels shows the models for all 30'' data while the right-hand panels show the models for the 12'' data. The middle column shows the 30'' models for galaxies that also have a model from their 12'' data. The dashed horizontal line in the surface density panels is at  $1 M_{\odot} \text{pc}^{-2}$ , which is the standard value used to define  $R_{\text{HI}}$ .

Thus, while there are 27 12'' models, there are only 18 cross-matched 30'' models. Comparing the middle and right columns of Figure 12 reveals that the rotation curves are broadly equivalent for the cross-matched models. However the 12'' models tend to be truncated relative to the 30'' data.

To gain a better understanding of the Phase 2 models and the 12'' data, Figure 13 shows the SOFIA *ell\_maj* parameter and integrated S/N for the Phase 2 sources. As noted in Deg *et al.* (2022), there is a clear relationship between the size and the integrated S/N. In the left-hand panel that shows all Phase 2 sources, it is clear that there is a diagonal limit above which no successful models are generated. For a given size, a higher S/N leads to a higher chance of kinematically modelling a galaxy. Conversely, larger galaxies with the same S/N as smaller galaxies are more difficult to model with WKAPP.

Focusing on the 12'' detections and the cross-matched 30'' sources reveals a number of interesting behaviours. Firstly, the majority of the cross-matched sources have approximately the same  $\log(S/N)$ . Secondly, the approximate size of the modelled disc has not increased by a factor of 2.5. This is expected as the smaller beam size results in a worse column density sensitivity, which means that the most extended gas will be below the noise limit. This, combined with the fact that the beam smearing effects are minimised in the 12'' resolution also explains the decreased radial extent of the 12'' kinematic models seen in Figure 12.

A third, and perhaps more important result is apparent in Figure 13. Only 8 galaxies do not have a 30'' kinematic model and a successful 12'' model (indicated by blue lines in Figure 13). By

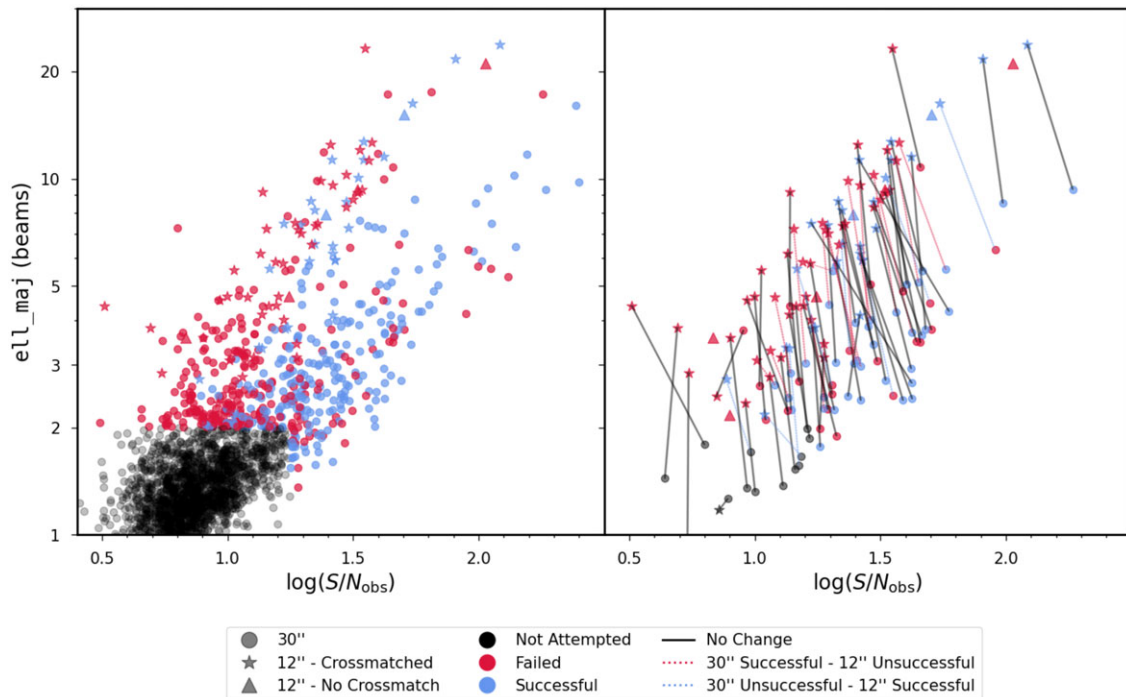
contrast there are 18 sources that were successfully modelled using their 30'' data that were not modelled with the 12'' data. These results show that, for WKAPP, the increased noise of the 12'' data leads to poorer results in terms of kinematic modelling despite the increased resolution. It is important to note here that WKAPP is being run in precisely the same way for the 12'' as for the 30'' data. If a more tailored approach were adopted it is possible that the kinematic modelling would be significantly more successful. Additionally, the increased resolution brings many of the galaxies into the regime where various 2D algorithms are applicable. Collapsing the data to moment maps effectively increases the S/N and may lead to greater success than the 3D approach of WKAPP. These ideas will be explored in a future work.

## 8. Data access

The WALLABY Pilot Survey Phase 2 data and associated catalogues are available to the public through the CSIRO ASKAP Science Data Archive (CASDA) and the Canadian Astronomy Data Centre (CADDC). The data release is similar to Public Data Release 1 and includes all the 30'' source data products, kinematic models and respective catalogues. In addition, in this release we are also including the high-resolution 12'' data products, kinematic models and catalogue. We also provide descriptions and details on the various data products, data quality issues and list details of the various column names in the catalogues.

We note that the source catalogue in this release also includes all detections from the Public Data Release 1 from Westmeier





**Figure 13.** The size and integrated S/N of the Phase 2 sources. The circles show the 30'' detections, while the stars and triangles shows the 12'' detections. The different 12'' symbols indicate whether there is a cross-matched 30'' source for the 12'' source (stars) or not (triangles). The black, red, and blue points indicate galaxies where kinematic modelling was not attempted, attempted and failed, or successfully modelled, respectively. The left-hand panel shows all Phase 2 detections, while the right-hand panel only shows the 12'' sources and their cross-matched 30'' counterpart (if a cross-matched source exists). In the right-hand panel the lines connect the cross-matched sources. Occasionally a 30'' source is broken into two different sources and will have two lines originate from the source. If the kinematic modelling result has not changed (failed for both or successful for both), the line is black. If the 30'' source is kinematically modelled while the 12'' source is not the line is red, and when the situation is reversed the line is blue.

et al. (2022) for easy accessibility to both DR1 and DR2 detections. Furthermore, the new catalogue will include all the relevant updated columns such as the corrected fluxes (`f_sum_corr` and `err_f_sum_corr`) as well as the corrected H I masses (`log_mi_hi_corr`) for both the DR1 and DR2 samples making it convenient for the user to use the corrected values.

The combined footprint A and B mosaics are available on CASDA via <https://doi.org/10.25919/hg66-4v60>. These are very large (typically  $\sim 600$  GB) and we recommend that users interact with these via the CASDA cutout service. These cutouts can be made either through the CASDA Data Access Portal (DAP) or by using the Simple Image Access Protocol (SIAP) coupled with the Serverside Operations for Data Access (SODA) protocol. In the second case, the user interacts using a Python script or Jupyter notebook to select the region and channel range of interest. Additionally, the CASDA module of the Astropy Astroquery package<sup>b</sup> can also generate cutouts.

Users can access the 30'' data via CASDA using the following links for the various types of data-sets. a) Source data products (including moment maps, cubelet, channel map, source mask, and spectra) and complete source catalogue: <https://doi.org/10.25919/qw7w-tn96>; b) 30'' kinematic modelling data products and catalogue: <https://doi.org/10.25919/7w8n-9h19>.

The 12'' source data products, which includes all SoFiA source data products (moment maps, cubelet, channel map, source mask and spectra), kinematics models and catalogue

(including kinematic modelling parameter values) can be accessed via CASDA using the following link: <https://doi.org/10.25919/47tr-k441>.

All the above data products for both the 30'' and 12'' data can be accessed via CADC through a TAP service using ADQL queries. For more details on how to access the data through CADC we refer the reader to the Public Data Release 1 papers (Westmeier et al. 2022; Deg et al. 2022). Furthermore, users can also get detailed instructions and links to the data releases through WALLABY's data access page<sup>c</sup>.

## 9. Summary and future

In this data release paper, we present the catalogue, data products including moment maps and spectra for over 1800 galaxies from the WALLABY Pilot Survey Phase 2. The observations were carried out on three selected fields which include the NGC 5044, NGC 4808, and Vela groups. The total observed sky area is  $\sim 180$  deg<sup>2</sup> and the redshift limit corresponding to  $z \sim 0.09$ . The median rms noise levels in the data cubes is  $\sim 1.7$  mJy, which is close to the expected theoretical noise for the WALLABY observations. This translates to a  $5\sigma$  column density sensitivity of  $\sim 9.1 \times 10^{19} (1+z)^4$  cm<sup>-2</sup> assuming a 30'' beam and a 20 km s<sup>-1</sup> channel width.

In addition to the default 30'' data products, in Phase 2 we have also presented the high-resolution 12'' cut-outs of select HIPASS galaxies demonstrating the true potential of WALLABY to produce high spatial and spectral resolution H I observations

<sup>b</sup><https://astroquery.readthedocs.io/en/latest/casda/casda.html>

<sup>c</sup><https://wallaby-survey.org/data/>

of several thousand galaxies (including all HIPASS galaxies) in the 5-year survey period, thereby forming the largest sample of high spatial resolution H I maps of galaxies until the SKA-mid begins observations. As such, these high-resolution cut-outs carry immense legacy value.

We highlighted the significant improvement in the quality of the data compared to Phase 1 which is mainly attributed to the fact that the Pilot Phase 2 fields were selected in a way as to avoid bright continuum sources, but also due to the introduction of the holography-based primary beam correction for the ASKAP observations, which results in more accurate fluxes for the sources. It is to be noted that there is ongoing work to implement appropriate “peeling” techniques into the ASKAPSoft data reduction pipeline in order to properly subtract residual continuum that is associated with bright continuum sources, likely improving the quality of the data significantly. While the data quality in general is very good, we note the observed flux discrepancy in the ASKAP observations. The issue was first highlighted in the Phase 1 paper (Westmeier *et al.*, 2022), wherein the integrated flux of the 30'' WALLABY detections were observed to be  $\sim 15\%$  lower than the corresponding single-dish flux. This was alluded to improper deconvolution and the impact of residual sidelobes still present in the image cubes. In order to fully understand this issue, we undertook simulations of ASKAP observations and injected model galaxies by varying their properties such as flux and size and find that up on performing the source finding using SoFiA, the simulated galaxies in the 30'' resolution are indeed observed to show consistently lower flux compared to the flux of the injected model galaxies. We attribute this to the contribution of the uncleaned flux in the data, which is impacted by the severe negative sidelobes that systematically brings down the integrated flux. We also note that marginally-resolved and/or low-SNR sources are more severely impacted by this.

Furthermore, we also note that the integrated flux of the 12'' sources is observed to be consistently higher than their 30'' counterparts, which again is attributed to the impact of uncleaned flux in the data and which has the imprint of the highly non-Gaussian 12'' ASKAP dirty beam with strong positive sidelobes. This uncleaned flux therefore artificially boosts the flux of the 12'' detections to about  $\sim 15\%$  depending on a number of factors including the SNR and spatial extent of the source. In order to minimise the impact of the uncleaned flux on the data, going forward for the full WALLABY survey, it is necessary to set appropriate cleaning thresholds and making sure that the thresholds are reached during the clean cycles. In addition, a two stage cleaning approach involving a shallow clean followed by a deeper cleaning using a source mask might result in better flux recovery. This will of course considerably increase the time and resources required to process the data, however, such a scheme may be implemented in the ASKAPSoft spectral-line imaging pipeline given the ASKAP observations are now being processed in the new upgraded Pawsey HPC *Setonix*, which is capable of handling large data volumes.

**Acknowledgement.** We would like to thank the anonymous referee for their useful comments which improved the clarity of this paper. We would also like to sincerely thank Minh Huynh (CSIRO) for all the efforts towards releasing the data onto CASDA.

This scientific work uses data obtained from Inyarrimanha Ilgari Bundara/the Murchison Radio-astronomy Observatory. We acknowledge the Wajarri Yamaji People as the Traditional Owners and native title holders of the Observatory site. CSIRO's ASKAP radio telescope is part of the

Australia Telescope National Facility (<https://ror.org/05qajvd42>). Operation of ASKAP is funded by the Australian Government with support from the National Collaborative Research Infrastructure Strategy. ASKAP uses the resources of the Pawsey Supercomputing Research Centre. Establishment of ASKAP, Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory and the Pawsey Supercomputing Research Centre are initiatives of the Australian Government, with support from the Government of Western Australia and the Science and Industry Endowment Fund.

This research used the facilities of the Canadian Astronomy Data Centre operated by the National Research Council of Canada with the support of the Canadian Space Agency.

The Canadian Initiative for Radio Astronomy Data Analysis (CIRADA) is funded by a grant from the Canada Foundation for Innovation 2017 Innovation Fund (Project 35999) and by the Provinces of Ontario, British Columbia, Alberta, Manitoba and Quebec, in collaboration with the National Research Council of Canada, the US National Radio Astronomy Observatory and Australia's Commonwealth Scientific and Industrial Research Organisation.

This paper includes archived data obtained through the CSIRO ASKAP Science Data Archive, CASDA (<http://data.csiro.au>).

WALLABY acknowledges technical support from the Australian SKA Regional Centre (AusSRC) and Astronomy Data And Computing Services (ADACS).

This research has made use of the NASA/IPAC Extragalactic Database (NED), which is funded by the National Aeronautics and Space Administration and operated by the California Institute of Technology.

Parts of this research were supported by the Australian Research Council Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO 3D), through project number CE170100013.

KS acknowledges funding from the Natural Sciences and Engineering Research Council of Canada.

LC acknowledges support from the Australian Research Council via the Discovery Project funding scheme (DP210100337)

NY acknowledges the fellowship of the China Postdoctoral Science Foundation (grant: 2022M723175, GZB20230766).

PK is partially supported by the BMBF project 05A23PC1 for D-MeerKAT.

LVM acknowledges financial support from the grant CEX2021-001131-S funded by MCIU/AEI/ 10.13039/501100011033, from the grant PID2021-123930OB-C21 funded by MCIU/AEI/ 10.13039/501100011033 and by ERDF/EU

**Data availability.** Links to all data products such as moment maps, cubes and associated catalogs used in this work are listed in the Data Access section (Section 8).

## References

- Bacchini, C., Fraternali, F., Pezzulli, G., & Marasco, A. 2020, *A&A*, **644**, A125
- Begum, A., Chengalur, J. N., Karachentsev, I. D., Sharina, M. E., & Kaisin, S. S. 2008, *MNRAS*, **386**, 1667
- Bryant, J. J., Croom, S. M., van de Sande, J., *et al.* 2019, *MNRAS*, **483**, 458
- Cappellari, M., Emsellem, E., Krajnović, D., *et al.* 2011, *MNRAS*, **413**, 813
- Catinella, B., Schiminovich, D., Kauffmann, G., *et al.* 2010, *MNRAS*, **403**, 683
- Chippendale, A. P., O'Sullivan, J., Reynolds, J., *et al.* 2010, in *Phased Array Systems and Technology* (ARRAY), 648–652
- de Blok, W. J. G., Walter, F., Brinks, E., *et al.* 2008, *AJ*, **136**, 2648
- DeBoer, D. R., Gough, R. G., Bunton, J. D., *et al.* 2009, *Proceedings of the IEEE*, **97**, 1507
- Deg, N., Spekkens, K., Westmeier, T., *et al.* 2022, *PASA*, **39**, e059
- Di Teodoro, E. M., & Fraternali, F. 2015, *MNRAS*, **451**, 3021
- Di Teodoro, E. M., & Peek, J. E. G. 2021, *ApJ*, **923**, 220
- Dickey, J. M., McClure-Griffiths, N., Gibson, S. J., *et al.* 2013, *PASA*, **30**, e003
- For, B. Q., Staveley-Smith, L., Westmeier, T., *et al.* 2019, *MNRAS*, **489**, 5723
- For, B. Q., Wang, J., Westmeier, T., *et al.* 2021, *MNRAS*, **507**, 2300

- Gaensler, B. M., Landecker, T. L., Taylor, A. R., & POSSUM COLLABORATION. 2010, in American Astronomical Society Meeting Abstracts, Vol. 215, [American Astronomical Society Meeting Abstracts #215, 470.13](#)
- García-Ruiz, I., Sancisi, R., & Kuijken, K. 2002, *A*, **394**, 769
- Giovanelli, R., Haynes, M. P., Kent, B. R., et al. 2005, *AJ*, **130**, 2598
- Guzman, J., Whiting, M., Voronkov, M., et al. 2019, ASKAPsoft: ASKAP science data processor software, Astrophysics Source Code Library, record ascl:1912.003, [ascl:1912.003](#)
- Hartley, P., Bonaldi, A., Braun, R., et al. 2023, *MNRAS*, **523**, 1967
- Haynes, M. P., Giovanelli, R., Kent, B. R., et al. 2018, *ApJ*, **861**, 49
- Heald, G., Józsa, G., Serra, P., et al. 2011, *A&A*, **526**, A118
- Hoffman, Y., Courtois, H. M., & Tully, R. B. 2015, *MNRAS*, **449**, 4494
- Hotan, A. 2016, ACES Memo, 11
- Hotan, A. W., Bunton, J. D., Harvey-Smith, L., et al. 2014, *PASA*, **31**, e041
- Hotan, A. W., Bunton, J. D., Chippendale, A. P., et al. 2021, *PASA*, **38**, e009
- Hudson, M. J., Smith, R. J., Lucey, J. R., & Branchini, E. 2004, *MNRAS*, **352**, 61
- Hunter, D. A., Ficut-Vicas, D., Ashley, T., et al. 2012, *AJ*, **144**, 134
- Huynh, M., Dempsey, J., Whiting, M. T., & Ophel, M. 2020, in Astronomical Society of the Pacific Conference Series, Vol. 522, Astronomical Data Analysis Software and Systems XXVII, ed. P. Ballester, J. Ibsen, M. Solar, & K. Shorridge, 263
- Jones, G. C., Vergani, D., Romano, M., et al. 2021, *MNRAS*, **507**, 3540
- Józsa, G. I. G., Kenn, F., Klein, U., & Oosterloo, T. A. 2007, *A&A*, **468**, 731
- Józsa, G. I. G., Oosterloo, T. A., Morganti, R., Klein, U., & Erben, T. 2009, *A&A*, **494**, 489
- Józsa, G. I. G., Thorat, K., Kamphuis, P., et al. 2021, *MNRAS*, **501**, 2704
- Kamphuis, P., Józsa, G. I. G., Oh, S. H., et al. 2015, *MNRAS*, **452**, 3139
- Katz, H., Lelli, F., McGaugh, S. S., et al. 2017, *MNRAS*, **466**, 1648
- Khoperskov, S. A., Moiseev, A. V., Khoperskov, A. V., & Saburova, A. S. 2014, *MNRAS*, **441**, 2650
- Koribalski, B. S., Staveley-Smith, L., Kilborn, V. A., et al. 2004, *AJ*, **128**, 16
- Koribalski, B. S., Wang, J., Kamphuis, P., et al. 2018, *MNRAS*, **478**, 1611
- Koribalski, B. S., Staveley-Smith, L., Westmeier, T., et al. 2020, *Ap&SS*, **365**, 118
- Kraan-Korteweg, R. C., Cluver, M. E., Bilicki, M., et al. 2017, *MNRAS*, **466**, L29
- Kurapati, S., Chengalur, J. N., Pustilnik, S., & Kamphuis, P. 2018, *MNRAS*, **479**, 228
- Lelli, F., McGaugh, S. S., Schombert, J. M., & Pawlowski, M. S. 2017, *ApJ*, **836**, 152
- Lelli, F., Verheijen, M., Fraternali, F., & Sancisi, R. 2012, *A&A*, **537**, A72
- Lewis, C. 2019, PhD thesis, Queen's University at Kingston, [Canada](#)
- Macquart, J.-P., Bailes, M., Bhat, N. D. R., et al. 2010, *PASA*, **27**, 272
- Mancera Piña, P. E., Posti, L., Fraternali, F., Adams, E. A. K., & Oosterloo, T. 2021, *A&A*, **647**, A76
- Masters, K. L., Nichol, R. C., Haynes, M. P., et al. 2012, *MNRAS*, **424**, 2180
- McConnell, D., Hale, C. L., Lenc, E., et al. 2020, *PASA*, **37**, e048
- McGaugh, S. S., Schombert, J. M., Bothun, G. D., & de Blok, W. J. G. 2000, *ApJL*, **533**, L99
- Meyer, M. J., Zwaan, M. A., Webster, R. L., et al. 2004, *MNRAS*, **350**, 1195
- Murugesan, C., et al. 2020, *MNRAS*, **496**, 2516
- Murugesan, C., Kilborn, V. A., For, B. Q., et al. 2021, *MNRAS*, **507**, 2949
- Murugesan, C., Džudžar, R., Bagge, R., et al. 2023, *PASA*, **40**, e018
- Norris, R. P., Hopkins, A. M., Afonso, J., et al. 2011, *PASA*, **28**, 215
- Ott, J., Stilp, A. M., Warren, S. R., et al. 2012, *AJ*, **144**, 123
- Radcliffe, J. F., Beswick, R. J., Thomson, A. P., Njeri, A., & Muxlow, T. W. B. 2024, *MNRAS*, **527**, 942
- Rogstad, D. H., Lockhart, I. A., & Wright, M. C. H. 1974, *ApJ*, **193**, 309
- Roychowdhury, S., Chengalur, J. N., Kaisin, S. S., & Karachentsev, I. D. 2014, *MNRAS*, **445**, 1392
- Sault, R. J., Teuben, P. J., & Wright, M. C. H. 1995, in Astronomical Society of the Pacific Conference Series, Vol. 77, Astronomical Data Analysis Software and Systems IV, ed. R. A. Shaw, H. E. Payne, & J. J. E. Hayes, 433
- Serra, P., Westmeier, T., Giese, N., et al. 2015, *MNRAS*, **448**, 1922
- Sorgho, A., Verdes-Montenegro, L., Hess, K. M., et al. 2024, *MNRAS*, **528**, 1630
- Strauss, M. A., & Willick, J. A. 1995, **261**, 271
- van der Hulst, J. M., van Albada, T. S., & Sancisi, R. 2001, in Astronomical Society of the Pacific Conference Series, Vol. 240, Gas and Galaxy Evolution, ed. J. E. Hibbard, M. Rupen, & J. H. van Gorkom, 451
- Verdes-Montenegro, L., Sulentic, J., Lisenfeld, U., et al. 2005, *A&A*, **436**, 443
- Verheijen, M. A. W., & Sancisi, R. 2001, *A&A*, **370**, 765
- Walter, F., Brinks, E., de Blok, W. J. G., et al. 2008, *AJ*, **136**, 2563
- Westmeier, T., Kitaëff, S., Pallot, D., et al. 2021, *MNRAS*, **506**, 3962
- Westmeier, T., Deg, N., Spekkens, K., et al. 2022, *PASA*, **39**, e058
- Whiting, M. T. 2020, in Astronomical Society of the Pacific Conference Series, Vol. 522, Astronomical Data Analysis Software and Systems XXVII, ed. P. Ballester, J. Ibsen, M. Solar, & K. Shorridge, 469
- Willick, J. A., Courteau, S., Faber, S. M., et al. 1997, *ApJSS*, **109**, 333
- Wong, O. I., Schawinski, K., Józsa, G. I. G., et al. 2015, *MNRAS*, **447**, 3311
- Wong, O. I., Meurer, G. R., Bekki, K., et al. 2006, *MNRAS*, **370**, 1607
- Wong, O. I., Stevens, A. R. H., For, B. Q., et al. 2021, *MNRAS*, **507**, 2905

## Appendix A. ASKAPsoft imaging parameters for the 30" and 12" data reduction

In Table 4 we list some relevant ASKAPsoft imaging parameters used to process both the default 30" and 12" data. For more details on the definition of each of the parameters we refer the reader to the ASKAPsoft User Documentation.<sup>d</sup>

<sup>d</sup><https://www.atnf.csiro.au/computing/software/askapsoft/sdp/docs/current/index.html>

**Table 4.** Important ASKAPSoft imaging, pre-conditioning, deconvolution and tapering parameters for the 30'' and 12'' data processing

Parameter	30''	12''	Description
SPECTRAL_IMAGE_MAXUV	2000	7000	Maximum UV distance (in metres) to apply in the data selection step
SPECTRAL_IMAGE_MINUV	12	12	Minimum UV distance (in metres) to apply in the data selection step
NUM_PIXELS_SPECTRAL	1536	384	No. of spatial pixels along the side for the image cubes
CELLSIZE_SPECTRAL	6	2	The spatial pixel size for the image cubes
CLEAN_SPECTRAL_SCALES	[0, 20, 60, 120]	[0, 10, 30, 60]	Set of scales (in pixels) to use with the multi-scale clean
CLEAN_SPECTRAL_PSFWIDTH	3172	256	The width of the psf patch used in the minor cycle (in pixels)
CLEAN_SPECTRAL_THRESHOLD_MINORCYCLE	[45%, 3.5mJy, 0.5mJy]	[45%, 3.5mJy, 0.5mJy]	Threshold for the minor cycle loop
CLEAN_SPECTRAL_THRESHOLD_MAJORCYCLE	0.5mJy	0.5mJy	The target peak residual. Major cycles stop if this is reached
PRECONDITIONER_LIST_SPECTRAL	[Wiener, GaussianTaper]	[Wiener, GaussianTaper]	List of preconditioners to apply
PRECONDITIONER_SPECTRAL_WIENER_ROBUSTNESS	0.5	0.5	Robustness value for the Wiener filter
PRECONDITIONER_SPECTRAL_GAUSS_TAPER	[30arcsec, 30arcsec, 0deg]	[12arcsec, 12arcsec, 0deg]	Size of the Gaussian taper FWHM (in arcsec) and position angle in degrees

## Appendix B. SoFiA parameters for the 30'' and 12'' source finding runs

In Table 5, we list some important SoFiA parameter values used for the source finding runs for the 30'' and 12'' data sets.

**Table 5.** SoFiA parameter values for the 30'' and 12'' source finding runs.

Parameter	30''	12''
scaleNoise.windowZ	51	51
scaleNoise.windowXY	51	75
scfind.kernelsXY	0, 5, 10	0, 6, 12
scfind.kernelsZ	0, 3, 7, 15, 31	0, 3, 7, 15, 31
scfind.threshold	3.8	4.0
scfind.replacement	2.0	1.4
linker.radiusXY	2	2
linker.radiusZ	3	3
linker.minSizeXY	5	6
linker.minSizeZ	5	5
reliability.threshold	0.7	0.7
reliability.scaleKernel	0.2	0.2
reliability.minSNR	3.0	3.0

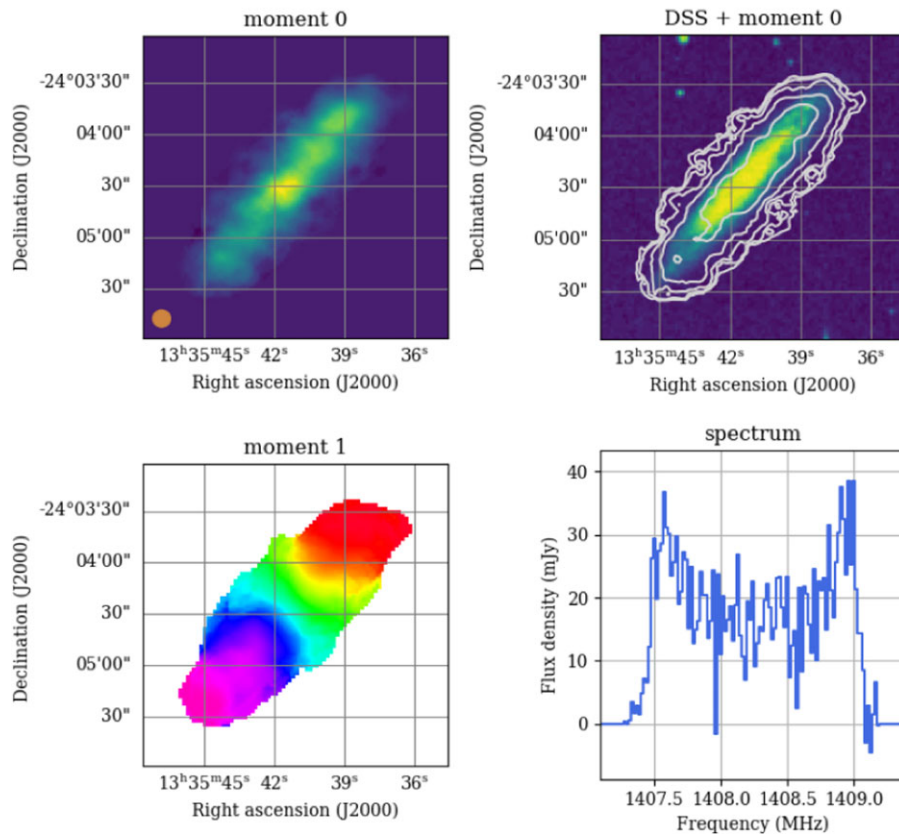
## Appendix C. Manual inspection workflow

A source finding pipeline run generates detections and associated data products, which are then added to a database. The database is populated through a manual process whereby all detections are visually examined by the WALLABY team to ensure that artefacts, false detections, and duplicates are removed. A web portal has been developed for conveniently executing the various stages of this inspection workflow.

To determine whether a detection is a real source, a WALLABY team member is presented with key detection properties (such as flux, RA, Dec) in a table, and a visual summary of data products (moment 0 and 1 maps, spectra, and an overlay of the moment 0 H I contours on to a Digitized Sky Survey (DSS) optical image). An example of the summary figure is shown in Figure 14. Detections that pass this first check are selected as potential genuine sources.

In the case where there is a staggered approach to selecting sources, for example, overlapping regions of the sky subsequently processed by the source finding pipeline may give rise to duplicate detections of already accepted sources. This is the case for the NGC 5044 field, where overlapping regions are shown in Figure 2 in the darker shade of green in the corners of the central  $4^\circ \times 4^\circ$  processing regions. In such cases, an external cross matching routine is performed by the inspection workflow allowing

WALLABY J133541–240427



**Figure 14.** Summary figure presenting the moment 0, moment 1 map, spectra, and optical DSS image of a source. These summary figures, along with properties of the detection from the source finding application are used by the WALLABY team to identify and remove false detections.

the WALLABY team to identify and handle potential duplicate detections. The spatial and spectral locations of the detections from each new source finding run are compared against accepted source entries in the catalogue. If they are within a tight spatial or spectral threshold ( $\Delta_{\text{spat}} \pm 5''$ ,  $\Delta_{\text{spec}} \pm 0.05$  MHz), they are automatically marked as duplicates and are removed from the database. If the candidate is within a lenient spatial and spectral threshold ( $\Delta_{\text{spat}} \pm 90''$ ,  $\Delta_{\text{spec}} \pm 2$  MHz), they are marked for an additional visual inspection step, before being accepted as a genuine detection and assigned a WALLABY source name. Once

this workflow is completed, the accepted sources are ready for release.

**Appendix D. Output source catalogue**

Table 6 provides details of all the parameters that are included in the source catalogue for all PDR2 detections. Two additional parameters are listed in the 12'' source catalogue that represent the 12'' integrated flux corrected to the original 30'' integrated flux and the associated statistical uncertainty.

**Table 6.** List of parameters in the source catalogue.

Name	Type	Unified Content Descriptor	Units	Description
name	char	meta.id; meta.main	–	WALLABY source name (WALLABY Jhhmmss+/-ddmmss)
ra	double	pos.eq.ra; meta.main	deg	Right ascension (J2000) of centroid position
dec	double	pos.eq.dec; meta.main	deg	Declination (J2000) of centroid position
freq	double	em.freq; meta.main	Hz	Barycentric frequency of centroid position
f_sum	double	phot.flux; meta.main	Jy*Hz	Integrated flux within 3D source mask
err_f_sum	double	stat.error; phot.flux	Jy*Hz	Statistical uncertainty of integrated flux
f_sum_corr	double	phot.flux; meta.main	y*Hz	Integrated flux within 3D source mask corrected to the single-dish integrated flux
err_f_sum_corr	double	stat.error; phot.flux	Jy*Hz	Statistical uncertainty of corrected integrated flux

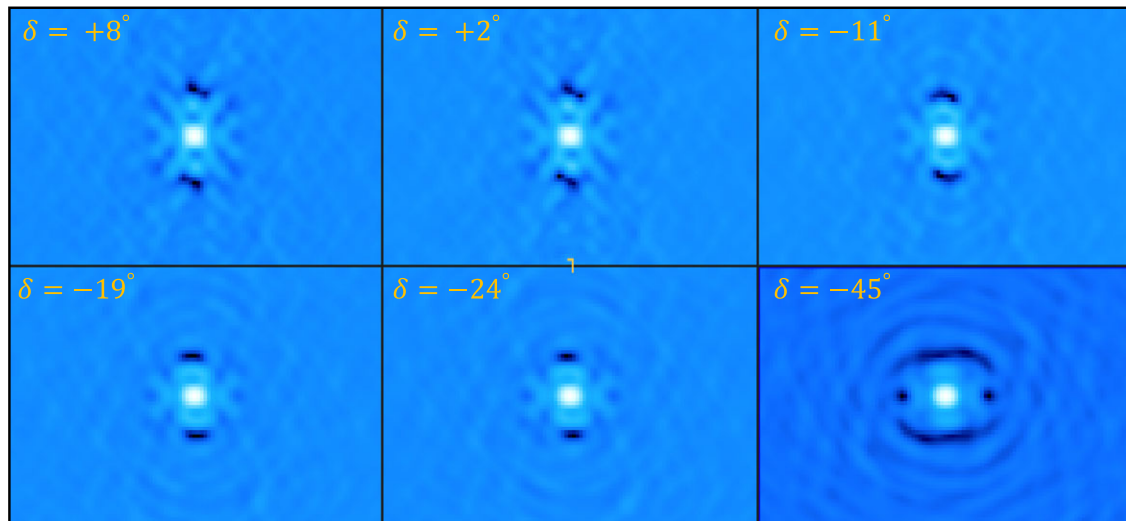
**Table 6.** (Continued)

Name	Type	Unified Content Descriptor	Units	Description
rms	double	instr.det.noise	Jy/beam	Local RMS noise level near source
w20	double	spect.line.width; meta.main	Hz	Spectral line width at 20% of the peak (w20)
w50	double	spect.line.width	Hz	Spectral line width at 50% of the peak (w50)
kin_pa	double	pos.posAng	deg	Position angle of kinematic major axis
rel	double	stat.probability	–	Statistical reliability of detection from 0 to 1
qflag	double	meta.code.qual	–	Source finding quality flag
kflag	double	meta.code	–	Kinematic model flag
n_pix	double	meta.number; instr.pixel	–	Number of pixels in 3D source mask
f_min	double	phot.flux.density; stat.min	Jy/beam	Lowest flux density value within 3D source mask
f_max	double	phot.flux.density; stat.max	Jy/beam	Highest flux density value within 3D source mask
ell_maj	double	phys.angSize	pix	Major axis size of ellipse fitted to moment 0 map
ell_min	double	phys.angSize	pix	Minor axis size of ellipse fitted to moment 0 map
ell_pa	double	pos.posAng	deg	Position angle of ellipse fitted to moment 0 map
ell3s_maj	double	phys.angSize	pix	Same as ell maj but > 3 sigma pixels only and equal weight
ell3s_min	double	phys.angSize	pix	Same as ell min but > 3 sigma pixels only and equal weight
ell3s_pa	double	pos.posAng	deg	Same as ell pa but > 3 sigma pixels only and equal weight
x	double	pos.cartesian.x	pix	Centroid position in x
err_x	double	stat.error;pos.cartesian.x	pix	Statistical uncertainty of centroid position in x
y	double	pos.cartesian.y	pix	Centroid position in y
err_y	double	stat.error;pos.cartesian.y	pix	Statistical uncertainty of centroid position in y
z	double	pos.cartesian.z	pix	Centroid position in z
err_z	double	stat.error;pos.cartesian.z	pix	Statistical uncertainty of centroid position in z
x_min	double	pos.cartesian.x; stat.min	pix	Lower end of bounding box in x
x_max	double	pos.cartesian.x; stat.max	pix	Upper end of bounding box in x
y_min	double	pos.cartesian.y; stat.min	pix	Lower end of bounding box in y
y_max	double	pos.cartesian.y; stat.max	pix	Upper end of bounding box in y
z_min	double	pos.cartesian.z; stat.min	pix	Lower end of bounding box in z
z_max	double	pos.cartesian.z; stat.max	pix	Upper end of bounding box in z
comments	char	meta.note	–	Comments on individual sources
team_release	char	meta.dataset; meta.main	–	Internal WALLABY team release identifier
dist_h	double	pos.distance	Mpc	Local Hubble distance derived from the barycentric source frequency
log_m_hi	double	phys.mass	$M_{\odot}$	H I mass $\log(M_{\text{HI}}/M_{\odot})$ derived from the local Hubble distance
log_m_hi_corr	double	phys.mass	$M_{\odot}$	H I mass $\log(M_{\text{HI}}/M_{\odot})$ derived using the single-dish corrected integrated flux
f_sum_corr_30	double	phot.flux; meta.main	Jy*Hz	Integrated 12'' flux within 3D source mask corrected to the 30'' integrated flux
err_f_sum_corr_30	double	stat.error; phot.flux	Jy*Hz	Statistical uncertainty of 12'' integrated flux corrected to the 30'' integrated flux

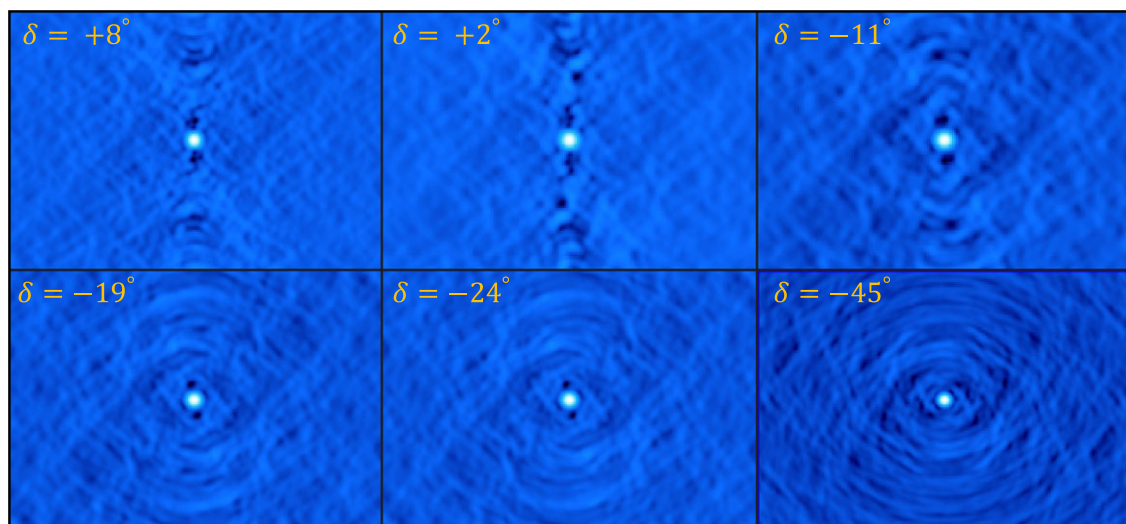
### Appendix E. Dirty beams (PSFs) from simulations

In Figure 15 we show the simulated PSFs of the 12'' beam for various declinations using MIRIAD tasks. We used a robust parameter of 0 and applied appropriate tapering in order to generate a dirty beam that is approximately 12''. The positive sidelobes associated

with the 12'' dirty beam is very significant, making the beam highly non-Gaussian-like. In the case of the 30'' PSFs (see Figure 16), while the central part of the beam is more Gaussian-like, there are significant negative sidelobes associated with the dirty beams.



**Figure 15.** The 12'' dirty beams for various declinations from the simulations.



**Figure 16.** The 30'' dirty beams for various declinations from the simulations.