Square Kilometre Array Science Data Challenge 1: analysis and results

A. Bonaldi [1,2]★ T. An [3] M. Brüggen [4] S. Burkutean [5] B. Coelho,[6] H. Goodarzi,[7] P. Hartley,[1]
P. K. Sandhu,[8] C. Wu,[9] L. Yu,[10] M. H. Zhoolideh Haghighi [7] S. Antón,[6,11] Z. Bagheri,[7,12] D. Barbosa,[6]
J. P. Barraca,[6,13] D. Bartashevich,[6] M. Bergano,[6] M. Bonato [5] J. Brand [5] F. de Gasperin,[4]
A. Giannetti,[5] R. Dodson,[9] P. Jain,[8] S. Jaiswal [3] B. Lao,[3] B. Liu [10] E. Liuzzo,[5] Y. Lu,[3] V. Lukic,[4]
D. Maia,[14] N. Marchili,[5] M. Massardi,[5] P. Mohan [3] J. B. Morgado,[14] M. Panwar,[8] P. Prabhakar,[8]
V. A. R. M. Ribeiro [6,15] K. L. J. Rygl,[5] V. Sabz Ali,[7] E. Saremi,[7] E. Schisano [16] S. Sheikhnezami [17,7]
A. Vafaei Sadr,[20,18] A. Wong[19] and O. I. Wong[9,21,22]

[1] *SKA Organization, Jodrell Bank, Lower Whitington, Macclesfield SK11 9FT, UK*
[2] *Jodrell Bank Centre for Astrophysics, Department of Physics and Astronomy, The University of Manchester, Manchester M13 9PL, UK*
[3] *Shanghai Astronomical Observatory, Key Laboratory of Radio Astronomy, Chinese Academy of Sciences, 80 Nandan Road, Shanghai 200030, China*
[4] *Hamburger Sternwarte, University of Hamburg, Gojenbergsweg 112, D-21029 Hamburg, Germany*
[5] *INAF, Istituto di Radioastronomia, Italian ALMA Regional Centre, Via P. Gobetti 101, Bologna 40129, Italy*
[6] *Instituto de Telecomunicações, Campus Universitário de Santiago, P-3810-193 Aveiro, Portugal*
[7] *School of Astronomy, Institute for Research in Fundamental Sciences (IPM), PO Box 1956836613, Tehran, Iran*
[8] *Department of Physics, Indian Institute of Technology Kanpur, Kanpur 208016, Uttar Pradesh, India*
[9] *ICRAR-M468, UWA, 35 Stirling Hwy, Crawley, WA 6009, Australia*
[10] *CAS Key Laboratory of FAST, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China*
[11] *CIDMA, Departamento de Física, Universidade de Aveiro, Campus Universitário de Santiago, P-3810-193 Aveiro, Portugal*
[12] *Research Institute for Astronomy and Astrophysics of Maraghe, 55177-36698 Maraghe, Iran*
[13] *Universidade de Aveiro, Campus Universitário de Santiago, P-3810-193 Aveiro, Portugal*
[14] *CICGE, Faculdade de Ciências da Universidade do Porto, Observatôrio Astronômico, Alameda do Monte da Virgem, P-4430-146 Vila Nova de Gaia, Portugal*
[15] *Departamento de Física, Universidade de Aveiro, Campus Universitário de Santiago, P-3810-193 Aveiro, Portugal*
[16] *INAF–IAPS, Via Fosso del Cavaliere 100, Rome 00133, Italy*
[17] *Department of Physics, Institute for Advanced Studies in Basic Sciences (IASBS), PO Box 11365-9161, Zanjan, Iran*
[18] *School of Physics, Institute for Research in Fundamental Sciences (IPM), PO Box 19395-5531, Tehran, Iran*
[19] *Department of Systems Design Engineering, University of Waterloo, 200 University Avenue West, Waterloo, ON Canada N2L 3G1, Canada*
[20] *Departement de Physique Theorique and Center for Astroparticle Physics, University of Geneva,1205, Switzerland.*
[21] *ARC Centre of Excellence for Astrophysics in Three Dimensions (ASTRO 3D), Australia*
[22] *CSIRO Astronomy and Space Science, PO Box 1130, Bentley, WA 6102, Australia*

ABSTRACT

As the largest radio telescope in the world, the Square Kilometre Array (SKA) will lead the next generation of radio astronomy. The feats of engineering required to construct the telescope array will be matched only by the techniques developed to exploit the rich scientific value of the data. To drive forward the development of efficient and accurate analysis methods, we are designing a series of data challenges that will provide the scientific community with high-quality data sets for testing and evaluating new techniques. In this paper, we present a description and results from the first such Science Data Challenge 1 (SDC1). Based on SKA MID continuum simulated observations and covering three frequencies (560, 1400, and 9200 MHz) at three depths (8, 100, and 1000 h), SDC1 asked participants to apply source detection, characterization, and classification methods to simulated data. The challenge opened in 2018 November, with nine teams submitting results by the deadline of 2019 April. In this work, we analyse the results for eight of those teams, showcasing the variety of approaches that can be successfully used to find, characterize, and classify sources in a deep, crowded field. The results also demonstrate the importance of building domain knowledge and expertise on this kind of analysis to obtain the best performance. As high-resolution observations begin revealing the true complexity of the sky, one of the outstanding challenges emerging from this analysis is the ability to deal with highly resolved and complex sources as effectively as the unresolved source population.

Key words: methods: data analysis – techniques: image processing – astronomical data bases: miscellaneous – galaxies: statistics – radio continuum: galaxies.

★ E-mail: a.bonaldi@skatelescope.org

1 INTRODUCTION

The Square Kilometre Array (SKA)[1] will be the world's largest radio telescope. The sensitivity and image quality of the SKA define new opportunities for science exploitation but also new challenges for data analysis. The overwhelming volume of raw SKA data means that they typically cannot be delivered to the Principal Investigators or the Key Science Project teams, who instead will have access to calibrated and gridded imaging/non-imaging products, called SKA Observatory (Observatory, in short hereafter) data products. Thus, part of the analysis that, on currently operating radio facilities, research groups typically carry out themselves will be the responsibility of the Observatory. For this change in handover point to be successful, it is necessary to build a good understanding of the needs of the science community and of the nature of the Observatory data products.

One of the actions that has been implemented in order to support the necessary development is the SKA Science Data Challenges (SDCs). These challenges, which are being regularly issued to the community, may consist of real data from currently operating radio facilities or of simulated SKA data. To make each challenge manageable and provide a training opportunity, the full SKA complexity is reached in steps. The goal of each SDC is to exercise some aspects of the analysis that will ultimately be performed on an Observatory data product.

In this paper, we discuss the aims and the outcome of the SKA SDC #1 (SDC1; Bonaldi & Braun 2018). SDC1 addresses source finding, characterization, and classification for radio continuum sources, on SKA MID simulated images. Many source-finding methods have been developed and tested in the literature (e.g. Bertin & Arnouts 1996; Hopkins et al. 2002; Molinari et al. 2011; Hales et al. 2012; Whiting 2012; Whiting & Humphreys 2012; Frean et al. 2014; Mohan & Rafferty 2015; Hancock, Trott & Hurley-Walker 2018; Hale et al. 2019; Lukic, de Gasperin & Brüggen 2019; Wu et al. 2019). Comparisons of performance of some of these methods in the context of ASKAP and VLA observations are presented in Hancock et al. (2012), Vernstrom et al. (2016), and Hopkins et al. (2015). Rather than performing a similar analysis in the context of the SKA, the aim of SDC1 is to facilitate further work on the subject and encourage a larger participation to this field, which in time can lead to new ideas and new methods for the future SKA surveys. Another goal is to familiarize the community with the complexity of the SKA data and the challenges posed by its analysis, and to provide a training opportunity to overcome them using the SDC1 data set as a concrete example.

In line with our goal of community engagement, participation in SDC1 was not restricted to developers of source-finding methods or their most expert users, but instead open to any team that wanted to engage in it. Teams could use their own combination of publicly available and purpose-developed software, with no requirement for the methods to be independent from one another or their list to be exhaustive. In line with our goal of exposing the future SKA challenges, SDC1 constitutes a significant step forward in the complexity of the data set with respect to previous work, in terms of sheer number of sources (of the order of 10^5 per square degree, down to well below the 1σ noise levels), and source properties (we include resolved sources and multicomponent sources with a complex morphology).

While these aspects make SDC1 not straightforward in terms of direct comparison of method performance, they give more of a global view of the outstanding challenges at the scale and complexity of the

SKA, and of the preparedness of the community. By delivering the data set publicly[2] as well as the software to evaluate one own's performance,[3] we leave open an opportunity for the training to continue and the results to improve beyond what is presented in this paper.

The outline of the paper is as follows: Section 2 describes the SDC1 data set and the challenge; Section 3 describes the teams that participated and the methods they used; Section 4 explains how the submissions were evaluated and scored; Section 5 shows the results and finally Section 6 presents our conclusions.

2 SDC1 DEFINITION

2.1 The data set

The SDC1 data set was released on 2019 November 25 and it is available on the SKA astronomers website https://astronomers.skatelescope.org/ska-science-data-challenge-1/. It consists of nine image files, in FITS format. Each file is a simulated SKA continuum image in total intensity at the following three frequencies:

(i) 560 MHz, representative of SKA Mid-Band 1;
(ii) 1.4 GHz, representative of SKA Mid-Band 2;
(iii) 9.2 GHz, representative of SKA Mid-Band 5.

Furthermore, three telescope integration depths per frequency are provided:

(i) 8 h, representative of a single-track observation;
(ii) 100 h, representative of a medium-depth integration;
(iii) 1000 h, representative of a deep integration.

The simulated field is nominally centred at RA = 0 and Dec = −30 for each map. The sky model is a plausible realization of the radio sky at those frequencies, but there is no attempt to make it similar to the actual sky at those coordinates. The nine maps share the same sky model realizations, to allow cross-matching between frequencies and direct comparisons between results for different noise levels.

The simulated observation strategy is that of a single telescope pointing. This means that the sensitivity of the array is maximum at the centre coordinates and decreases towards its outskirts, as described by the primary beam. The field of view (FoV) was chosen for each frequency to contain the primary beam out to the first null. This gives a map size of 5.5, 2.2, and 0.33 deg on a side for 560 MHz, 1.4 GHz, and 9.2 GHz, respectively.

The number of pixels on a side is always 32 768, which gives a pixel size of 0.60, 0.24, and 0.037 arcsec for 560 MHz, 1.4 GHz, and 9.2 GHz, respectively. The imaging resolution in the Gaussian approximation is 1.5, 0.6, and 0.09 arcsec full width at half-maximum (FWHM) for 560 MHz, 1.4 GHz, and 9.2 GHz, respectively. A more accurate description of the resolution (including sidelobes) is given by the synthesized beam. Both the primary beam and the synthesized beam are provided as ancillary data. Along with the images, a catalogue revealing only a fraction of the simulated galaxies was released for each frequency band. This aims to test and debug the procedures, as well as providing a training set for methods requiring training. This catalogue lists all of the embedded sources within 5 per cent of the FoV area and unveils their properties. Full details of how the data set was generated can be found in Bonaldi & Braun (2018); below, we provide a summary of the steps taken.
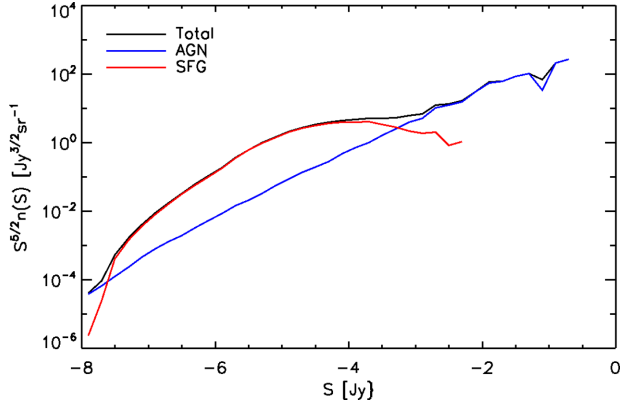
Figure 1. Differential source counts of the SDC1 simulation at 1.4 GHz.

The simulated data set was produced by first using the Tiered Radio Extragalactic Continuum Simulation software[4] (Bonaldi et al. [2019]) to generate a sky model catalogue containing star-forming galaxies (SFGs) and active galactic nuclei (AGNs), with integrated flux densities, sky coordinates, and size and shape information attributed to each source. For SFGs, a redshift-dependent luminosity function is generated by exploiting the tight correlation with star formation rate, the evolution of which is well studied. For AGNs, an evolutionary luminosity function model representing steep-spectrum sources (SS-AGN), flat-spectrum radio quasars, and BL Lac was adopted.

Fig. [1] shows the differential source counts of the simulation at 1.4 GHz. Agreement of the counts with observations, as well as luminosity functions and redshift distributions, is shown in Bonaldi et al. ([2019]). Radio-loud AGNs dominate the counts above the mJy level; the simulation is complete above an integrated flux of $10^{-7.5}$ Jy.

A morphological model representing the properties of each source was then injected on to the simulated field. The size of each source was used to determine whether it would appear to be 'resolved' or 'unresolved' on the sky at each observing frequency, with a cut made at a size $\geq$ or < of 3 pixels, respectively. Images representing the morphological structure of the extended SFG population were generated using a Galsim (Rowe et al. [2015])-based pipeline developed for the SuperCLASS project[5] (Harrison et al. [2020]). Each SFG was modelled using an exponential Sersic profile, projected into an ellipsoid with a given axial ratio and position angle. Extended SS-AGN sources were created from a library of real, high-resolution images to which a set of scaling, rotation, and reflection transformations were applied. We note that some subclasses of morphologies, such as giant radio galaxies or wide-angle tailed galaxies, could be underrepresented in the image library and therefore in the simulation. Resolved flat-spectrum sources were modelled using Gaussian and point component pairs. Once generated, images of extended sources were added as 'postage stamps' to the full sky field. Compact sources from all populations were added to the image as elliptical Gaussian components.

Once added to the sky, all sources underwent a convolution using an FWHM of two pixels in order to produce a gridded sky model, before a primary beam attenuation was applied. Visibility files were generated using the locations of all 197 dishes of the SKA1-Mid and MeerKAT configurations for the 1400 MHz band, while the

Table 1. Noise RMS (nJy per beam) of the simulated maps per frequency and exposure, compared to the classical confusion limit.

	560 MHz	1400 MHz	9200 MHz
8 h	2880	710	430
100 h	810	200	120
1000 h	255	73	38
Confusion	15	0.36	0.0002

560 and 9200 MHz bands used the 133 SKA1-Mid locations.[6] Time and frequency sampling was significantly reduced with respect to full operational sampling rates, but was based on a 30 per cent fractional bandwidth in frequency and hour angle coverage spanning (-4)–$(+4)$ h of Local Sidereal Time. An FFT-based convolution of the natural visibility density grid was used to calculate so-called 'uniform' gridding weights, before application of a Gaussian taper chosen to match the target FWHM of 1.5, 0.60, and 0.09 arcsec in the 560, 1400, and 9200 MHz bands, respectively.

Due to the overwhelming volume of raw data, the SKA data reduction workflow is such that calibration and imaging happen on single-track observations, generating Observatory data products that are accumulated and co-added to produce deeper images. To reflect this, deconvolution has been simulated for all our images down to a depth appropriate for an 8 h observation. Image deconvolution effects were simulated by applying a clipping threshold (three times the expected RMS noise level of an 8 h observation) to the primary beam tapered sky model, the brightness distribution above the threshold was convolved with a Gaussian restoring beam – taking explicit account of the sky model gridding convolution function – and below this threshold was convolved with the relevant dirty beam (again taking explicit account of the gridding convolution). The sum of these two images represents the simulated observed sky model.

In an interferometric observation, both the noise and the sky signal are processed through the same spatial frequency filter. To get noise fields that are an accurate representation of the expected correlated image noise, dirty noise images were generated from the same imaging simulations described above to produce the synthesized beams. The RMS noise amplitude was scaled to represent values appropriate for the three different simulation depths and frequency bands under consideration. Different noise realizations were used in all cases, to ensure that the noise would not correlate among either image depth or between frequency bands. The RMS noise levels are reported in Table [1].

For comparison, the last row of Table [1] also reports the classical confusion limit, computed as a function of the beam size θ and frequency ν following Condon et al. ([2012])

$$\sigma_c = 1.2\,\mu\text{Jy beam}^{-1} \left(\frac{\nu}{3.02\,\text{GHz}}\right)^{-0.7} \left(\frac{\theta}{8''}\right)^{10/3}, \quad (1)$$

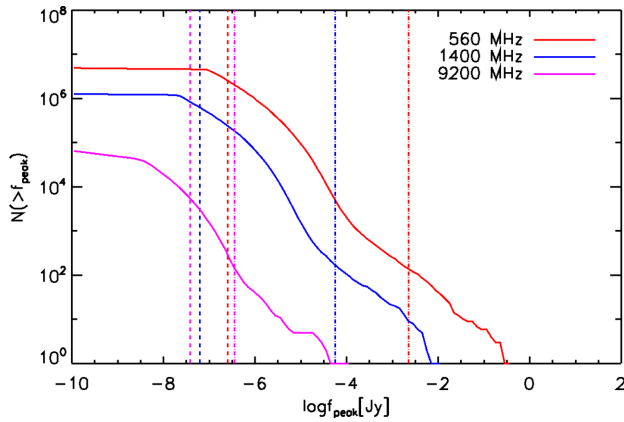which shows that the maps are still noise limited even at the deepest exposure.

The final data products were the sum of the simulated sky model and the relevant noise image described above. It is worth mentioning that the image quality is unrealistically good, because no systematic effects such as calibration errors, pointing errors, or deconvolution errors were injected (see Bonaldi & Braun [2018], for more details).

Fig. [2] shows the cumulative counts for the sources that were injected in the maps at the three considered frequencies. The counts

---

[4]https://github.com/abonaldi/TRECS
[5]http://www.e-merlin.ac.uk/legacy/projects/superclass.html

[6]The overlap between the MeerKAT UHF band, 580–1050 MHz, and SDC1 Band1, 480–650 MHz, is only partial.

Figure 2. Cumulative source counts for the three simulated maps for the full FoV as a function of apparent peak flux. The vertical dashed lines are the 1σ noise levels for the 1000 h exposure.

Table 2. Number of sources in the simulated maps above different noise levels.

Exposure	σ levels	560 MHz	1400 MHz	9200 MHz
1000 h	5	757 985	227 168	657
	7	624 389	155 426	295
	10	410 832	126 442	208
100 h	5	257 472	80 436	123
	7	199 833	47 636	87
	10	114 085	25 849	74
8 h	5	60 472	12 763	45
	7	29 422	8698	40
	10	20 156	3934	29

are shown as a function of the peak apparent flux; they have not been normalized for the different sky areas; therefore, they directly give the total number of sources in the maps for various flux thresholds.

The flattening of the curves for the lowest fluxes is artificial, and is due to a flux threshold applied to the sources for computational reasons. However, the injection of sources goes well below the 1σ noise levels for the deepest exposure, which is indicated by the dashed vertical lines. This means that our data set contains a background noise due to fainter and fainter sources that are present in the sky. Values of the cumulative counts for different S/N levels are tabulated in Table 2. This gives an appreciation for the number of sources that could be detected at various significances; these are of the order of hundreds of thousandths for 560 and 1400 MHz and of hundreds at 9200 MHz for the 1000 h images.

2.2 The challenge

The SKA community was invited to retrieve the SDC1 images and undertake source finding, source classification, and characterization. The results submitted should be catalogues of detected sources, containing the following:

(i) source coordinates (RA, Dec) to locate the centroids and where appropriate the core positions;
(ii) integrated primary-beam-corrected flux density;
(iii) core fraction (it is different from zero only for AGNs);
(iv) major and minor axis size;
(v) major axis position angle;

(vi) source population identification (one of AGN-steep, AGN-flat, SFG).

Potentially challenging aspects of this data set are as follows:

(i) The sheer number of sources (see Table 2), which requires the source extraction and source characterization methods to be automated, efficient, and, ideally, parallel.
(ii) The source density, which allows probing source extraction in a regime of high source crowding.
(iii) Within each SDC1 map, and even more so across the 950–9200 MHz frequency range, sources range from unresolved to fully resolved, revealing in some case a complex morphology. Such diversity could challenge source extraction methods.

The SKA imaging products will be challenging in terms of data sizes as well. For this exercise, we considered only one frequency per band, thus reducing the size to ~4 GB for each map. Although not representative of the full SKA data size, this file size already requires participants to access and analyse portions of the data separately on most computers, and therefore to organize the workflows in a way that can be scaled to even bigger sizes.

The description of how submissions were evaluated and scored is given in Section 4.

3 METHODS

The SDC1 data set was released and advertised publicly on 2018 November 25. Several teams, from various countries in the world, registered their interest on analysing the data set. Of those, nine teams managed to submit results by the deadline of 2019 April 30, and eight participated to this paper. Participation in the challenge was completely open and voluntary. Different teams approached it from a different level of specific expertise and preparedness. Teams were allowed to use their own developed methods as well as publicly available ones. As a result, the list of approaches used is most likely not complete. This section lists the teams in alphabetical order and describes their analyses.

3.1 ARCIt-CACAO (Burkutean, Brand, Massardi, Schisano, Bonato, Liuzzo, Marchili, Giannetti, Rygl)

ARCIt-CACAO is a source detection and classification pipeline developed in the context of the first SKA data challenge.[7] It is entirely PYTHON based, parallelized, and makes use of SCIPY (Virtanen et al. 2020) and ASTROPY (Astropy Collaboration 2013, 2018) routines. The ARCIt-CACAO pipeline is split into three distinctive steps, namely source detection, description, and classification.

For the source detection and description steps, ARCIt-CACAO uses identical pipeline set-ups for all bands and exposure times, the only differences being the band, exposure time, and associated expected theoretical noise, and of course the cleaned images and their primary beams. All other parameter settings were held constant for each ARCIt-CACAO run with a particular combination of frequency and exposure time. This decision was made in order to mimic a general solution approach to the problem rather than a solution tailored to a band-exposure-time combination.

The cleaned input images were split into sub-fields of 1024×1024 pixel^2 so that the source detection and description could be run in parallel on each of them. Although producing a different set of sub-images that overlap with the originally generated

---
[7]https://www.ira.inaf.it/cacao/CACAO_V1/cacao.html

sub-images can be implemented in the ARCIt-CACAO pipeline via an additional tessellation, this was not done for the SKA challenge in order to speed up the overall potential source table generation per run. As the ARCIt-CACAO source detection operates at the sub-image level, this could have resulted in some extended high-flux sources, which have a higher probability of spanning different sub-images, not having been recognized. Solutions from the outer 512 pixels on the sides of the large 32768×32768 pixel^2 input images were rejected to avoid edge effects.

The ARCIt-CACAO pipeline first applies a primary beam correction to the cleaned input sub-fields, using a re-sampled primary beam that matches the spatial extent and resolution of the cleaned sub-images. A Gaussian filter with a kernel size set to the FWHM of the clean beam is applied to the sub-field input images leading to the creation of a binary mask after filtering out values smaller than 0.5σ, where σ corresponds to the expected theoretical noise level input parameter. Islet identification within the binary mask via SCIPY tasks (Virtanen et al. 2020) generates a list of potential source locations and maximum 2D pixel source extent for each potential source within each sub-field.

For each of these islets within the sub-field, the ARCIt-CACAO source description routine generates source description measures from image moment analyses of the sub-field region spatially extracted at the islet region. The position angle and source extent were computed from the eigenvectors and associated eigenvalues of the covariance matrix of the islet region extracted from the primary-beam-corrected sub-field and the flux was determined in the selected region. Potential sources with a maximum extent smaller than the beam size were rejected. In addition, a conservative criterion of source acceptance was placed via a $2\sigma_{pb}$ threshold on the maximum pixel value within each islet, where σ_{pb} corresponds to σ divided by the mean primary beam correction factor within the islet.

For each sub-image, individual detected source lists were created that were then concatenated into the final table containing source position, size, and flux using pandas (Wes McKinney 2010). Source classification was made based solely on the source flux and size distribution property analyses of the originally provided training data sets.

3.2 Engage SKA – Portugal

Using the function `read.fits` from the R package 'astro', we divided the simulated SKA continuum images into smaller ones (of $16\,384 \times 16\,384$ pixels). The primary beam images were re-projected and re-sampled in order to match the dimensions of the latter continuum sub-images, using the function `reproject_interp` of the package 'reproject' from ASTROPY (Astropy Collaboration 2013, 2018). Then, using the CASA (McMullin et al. 2007) task `impbcor`, we obtained the primary-beam-corrected images.

We ran PROFOUND (Robotham et al. 2018; Hale et al. 2019) on the primary-beam-corrected images to pinpoint the sources and determine their properties. The argument `skycut`, defining the S/N threshold, was set to 3, which is the minimum value typically used in extragalactic astronomy. The minimum of pixels required to identify a source was chosen to be 2 (`pixcut=2`), which is the minimum value that allows an orientation of the source to be defined. Finally, the argument `tolerance` was set to 1 in order to maximize the deblending of nearby sources.

With the Tool for OPerations on Catalogues and Tables (TOPCAT) software (Taylor 2005), we searched for the counterparts of the identified sources, via a cross-match of the simulated multifrequency data. Based on the latter information, we classified the sources as either steep-spectrum AGN [those objects that have a spectral index $\log(F_{1400}/F_{9200})$ larger than 0.7]; flat-spectrum AGN (having spectral index <0.7); and as SFGs (if not detected at 9200 MHz).

3.3 hs (Lukic, Brüggen, De Gasperin)

We explored source finding using ConvoSource, a Convolutional Neural Network (CNN) that was trained on a solution map derived from knowledge of the source locations. For the purpose of source finding, the output images that had to be matched are those of the source locations, rather than the original input source maps. For that purpose, we transformed the source locations into an image, the source location map. This map, as well as the original source map, was segmented into smaller square images of 50×50 pixels, which then provide the inputs to the CNN.

Our work was based on the `Keras` package with the `TensorFlow` backend. We used a convolutional network architecture of three consecutive convolutional layers and one dense layer. Altogether, this network has a total of 32 193 parameters. Early stopping was used with a patience of five training epochs. 80 per cent of the data were used for training, and the remaining 20 per cent were used for testing. In order to make the network more robust by reducing overfitting, we placed dropout layer with a dropout fraction of 0.25 between the first and second convolutional layers. The batch size was set to 128 and we also used the binary cross-entropy cost function. We also experimented with augmentation by rotating and flipping images. More details of the algorithm can be found in Lukic et al. (2019).

3.4 ICRAR (Wu, Wong, and Dodson)

The ICRAR team's method is primarily based on the Classifying Radio galaxies Automatically using Neural networks (CLARAN; Wu et al. 2019) version 0.2[8] prototype. In this section, we provide a brief description of CLARAN and our specific approach to SDC1.

CLARAN's primary purpose is to classify extended radio sources within any given image field without knowing a priori the number of independent radio sources within that image – thereby combining the two tasks of source identification with source morphology classification. In recent years, citizen science projects such as Radio Galaxy Zoo (RGZ; Banfield et al. 2015) have helped increase the sample sizes of extreme classes of extended radio galaxies (e.g. Banfield et al. 2016; Kapińska et al. 2017). The many RGZ-enabled discoveries and publications using archival radio observations from the FIRST survey (Becker, White & Helfand 1995) in combination with *WISE* mid-infrared maps (Wright et al. 2010) demonstrate that current methods for radio source classification and our understanding of extended radio galaxies can still be furthered. To this end, we developed CLARAN, a prototype end-to-end deep learning classifier, trained on extended radio sources identified in a 2018 version of the RGZ Data Release 1 catalogue (Wong et al., in preparation). CLARAN employs a deep learning method based on a Faster R-CNN algorithm (Ren et al. 2015) that has been reimplemented in `TensorFlow`. A more complete description of CLARAN can be found in Wu et al. (2019).

The main update between the method as described in Wu et al. (2019) (version 0.1) and the one used in the challenge (version 0.2) is that the default network architecture is now based on the ResNet50 (He et al. 2015) model, instead on the VGG-16 (configuration D;

---

[8] https://github.com/chenwuperth/ClaRAN

Simonyan & Zisserman 2015). Version 0.2 has integrated Feature Pyramid Networks for object detection (Lin et al. 2016), which enables multiscale feature extraction and also supports on-the-fly rotation within the extended image augmentation pipeline. These new updates have attained an improved mean average precision from 82.6 per cent with version 0.1 to 86.1 per cent with version 0.2 for CLARAN's D3 method (which involves the input of an infrared map overlaid with radio contours).

Specific to SDC1, our approach can be generalized into three main steps: (1) the pre-processing and preparation of the training data sets; (2) training CLARAN; and (3) source and flux extraction.

Before we begin the training, we need to prepare and pre-process the ground truth catalogues provided by SDC1 in order to obtain a suitable training set for CLARAN. The pre-processing steps are as follows:

(i) Convert the catalogues to CSV format files;

(ii) Divide images into a set of smaller cut-outs that are 205 pixels by 205 pixels in size;

(iii) Filter for sources from the ground truth catalogues that have fluxes (S) greater than $k \times \sigma$, where σ is the respective RMS of the image field and k ranged from 0.5 to 3.0;

(iv) Determine the bounding box and class label for each filtered source and put into a JSON file with the source and cut-out identifiers.

The pre-processed data set is divided into two: a JSON file for the training set and another one for the testing data set. Using CLARAN version 0.2, we train CLARAN to recognize the types of sources that are presented in SDC1.

After running CLARAN on the full SDC1 data set, we measure the fluxes of the identified sources automatically using the `Imfit` tool from the MIRIAD software suite (Sault, Teuben & Wright 1995). As CLARAN was built specifically for source classification, we needed to estimate the integrated fluxes of the sources with a different tool. While we initially fitted a Gaussian to each radio source identified by CLARAN, we found a systematic underestimation of the true flux and thus we integrated the flux of the radio sources within a flat-top elliptical disc, which had significantly better performance. All materials used by our team's solution for SDC1 are in a publicly accessible Github repository at https://github.com/ICRAR/skasdc1.

3.5 IPM (Goodarzi, Bagheri, Sabz Ali, Saremi, Sheikhnezami, Vafaei Sadr, Zhoolideh Haghighi, Wong)

At IPM, we follow two different paths to find radio sources in the image. The first approach, which we labelled IPM1, used a pipeline called SEXTRACTOR (Bertin & Arnouts 1996). The second approach, labelled IPM2, is a method we have developed from scratch by using image processing tools such as scikit-image (Van der Walt et al. 2014).

SEXTRACTOR is a free pipeline to extract catalogues from astronomical images. Although this routine was originally developed to identify objects from large-scale galaxy-survey data, it can also extract sources from crowded star fields (Bertin & Arnouts 1996). In order to use this package, some parameters such as detection threshold and detection minimum area should be set. We selected these parameters by trial and error and visual inspection. First, we start with the guess values and run the software. Then, the calculated contours were plotted and overlapped with the image. We iterated on the input parameters until the number of detected objects and recognized shapes reached an optimal level, as judged by visual inspection. By using this method, we set the detection threshold at 2.5σ above the local background and choose 3 pixels as the minimum
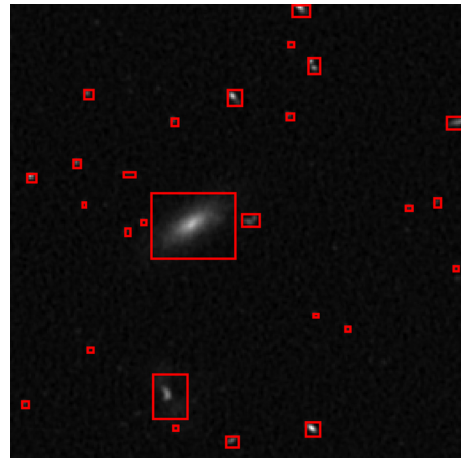
Figure 3. Detected sources with their bounding boxes in a 256 by 256 pixel part of the original image.

number of pixels above threshold value in which a group should have to be detected as a single object.

The SEXTRACTOR software utilizes multithresholding as a de-blending method, so we set the number of de-blending sub-thresholds as 32 (the value recommended in the software manual). Also we selected the minimum contrast parameter for de-blending as 0.005. Since the simulated images were crowded, we selected the option of local background in order to compute background in a rectangular region around the isophotal limits of the object. This helped us to improve the evaluation of the background and the consequent photometry. We should note that our method is the initial approximation for usage of the SEXTRACTOR as source finder; by comparing its results with the true catalogue, it is possible to optimize the input parameters and improve the results. The results of this technique are presented in Section 5 as IPM1.

In the second approach, we started with a denoising step to reduce the noise levels in the images. Among the many denoising methods available, we used the median filter module available in SCIPY (Jones et al. 2001), and by applying a multidimensional median filter on the original images, we removed the background noise. Then, in order to distinguish the sources from the background we have to set a threshold. The best thresholding method for our images, as assessed by trial and error, resulted to be Li thresholding (Li & Tam 1998). After denoising, we make a binary (black and white) image in which radio sources are white objects and background would be black. Finally, we use some modules of the scikit-image (Van der Walt et al. 2014) including `morphology` and `measure.label`, `measure.regionprops` to find the centroids of radio sources alongside their bounding boxes. Once we get bounding boxes, not only we can calculate the properties required by the challenge, but also other interesting features such as Euler number and eccentricity. The results of this technique are presented in Section 5 as IPM2, and for a better visualization of the bounding boxes and recovered sources you can refer to Fig. 3. For future prospects, we investigate training CNNs as proposed in Vafaei Sadr et al. (2019) as denoising method in order to customize the process.

3.6 JLRAT (Yu, Liu)

The size and intensity of the radio sources provided by SDC1 both have a large dynamic range. To deal with these features and detect weak sources, we have designed the JLRAT source finder model
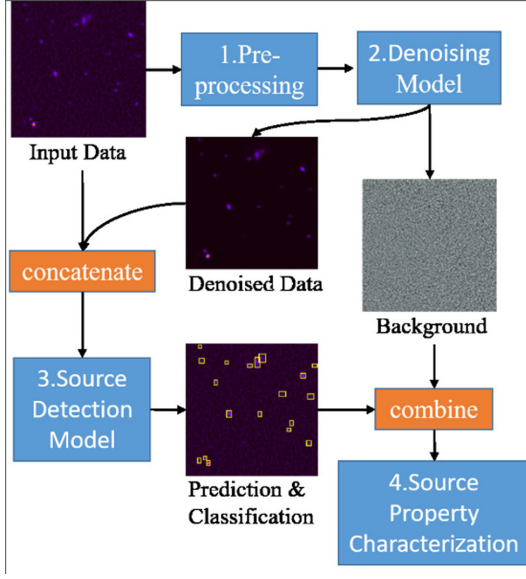
Figure 4. The pipeline of JSFM.

(JSFM) with feature pyramid network. Our method can be described in four main steps: (1) pre-processing, (2) denoising model, (3) source detection and classification, and (4) source property characterization.

The pipeline of our method is demonstrated by Fig. 4. In the first place, the raw image is pre-processed with pixel value scaling, the $\log_{10}$ operation, and zero centring. Then, we build a denoising model, which applies the core idea from Zhang et al. (2017), but without batch normalization (Ioffe & Szegedy 2015) in convolutional layer. By using the residual learning, it generates both smoothed and background images. The RMS of the background data times a scalar is the threshold value for building a binary image that is used to source property characterization. In the third step, the source detection model locates sources in the image and identifies their corresponding classes. Finally, the binary image of each detected source region is fitted to extract the source properties.

The model for the noisy observation S_{obs} is $S_{\mathrm{obs}} = S_{\mathrm{real}} + n$, where S_{real} is a clean image and the noise n is the additive white Gaussian noise with standard deviation σ. The purpose of image denoising is to recover a clean image from the noisy observation S_{obs}. In this section, we use a multiconvolutional layer as our denoising model that adopts a residual learning to train a residual mapping $\mathcal{R}(S_{\mathrm{obs}}) \approx n$. That acquires a background image and generates a denoised image $S_{\mathrm{real}} = S_{\mathrm{obs}} - \mathcal{R}(S_{\mathrm{obs}})$. Formally, the loss function is

$$\ell(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} \left\| \mathcal{R}\left(S_{\mathrm{obs}}^{i}; \Theta\right) - \left(S_{\mathrm{obs}}^{i} - S_{\mathrm{real}}^{i}\right) \right\|_{F}^{2}.$$

A key idea of our method of finding source location is that we evaluate it on different scales. This kind of method has been applied in various scenarios, such as Cai et al. (2016), Lin et al. (2016), Redmon & Farhadi (2018), Zhao et al. (2019), and Zhu et al. (2019). In order to retain the features of relatively small sources, our model does not use many down-sampling operators in the base network.

The source detection model begins with a simple Base Network to extract the base feature. Then, a Feature Pyramid Network is applied to build a multiscale feature pyramid, with two alternative models. The maximum size layer of each feature pyramid is the output that will concatenate with the base feature as the input for the next stage. After that, feature channels concatenate and weight model

first concatenates same scale layers among all feature pyramids, and then weight each scale channels separately through the squeeze-and-excitation method from Hu et al. (2017). The overall architecture of source detection model and feature pyramid model I and II is illustrated in Fig. B1.

In the prediction step, in order to achieve source location regression and classification, six different scale features are connected with five convolution layers, respectively. The detection boxes scale ranges for those features; we follow the setting of the original Single Shot MultiBox Detector from Liu et al. (2016).

After the location of the sources is acquired, the RMS of the background that is generated by the denoising model is used as a benchmark of the threshold value. Then it is multiplied by a scalar to determine the binary image that is used for the source property characterization.

3.7 RADGK (Pritpal, Pankaj, Mohit, Prabhakar)

The RADGK team approached the challenge without having an automated data analysis pipeline. At this stage, we concentrated on understanding the single steps involved in the analysis, by working on the few sources that could be identified by visual inspection and interactive processing. Further steps including automation of our procedures are outlined at the end of this section.

The simulated data set was downloaded at 560 and 1400 MHz and for 1000 h exposure. Sources were identified by visual inspection and the license-based ALADIN (Bonnarel et al. 2000) software has been used to extract their properties, in particular position (core and centroid), peak and integrated flux, major and minor axes, and position angle. We identified 34 sources at 560 MHz and 25 sources at 1400 MHz by visual inspection.

The centroid of each source was calculated using the following expression:

$$\left(x_o, y_o\right) = \left(\frac{\Sigma I_{ij} x_{ij}}{\Sigma I_{ij}}, \frac{\Sigma I_{ij} y_{ij}}{\Sigma I_{ij}} \right), \tag{2}$$

where x_{ij} and y_{ij} indicate the RA and Dec coordinates of the pixel (i, j), respectively, and I_{ij} is the flux density at that location. The list of pixels i, j to consider for each source was selected by cropping the individual sources from the image file in FITS format using the ALADIN® software. Data extraction for each source is obtained in a tabulated form again using ALADIN®, where the first column gives the intensity/pixel and the second and third columns give the position of the pixel in RA and Dec.

We next fitted a 2D elliptical Gaussian function centred at (x_o, y_o) for each detected source, given by

$$I(x, y) = A_o e^{\left\{ -\left[\frac{\left(x-x_o\right)^2}{2\sigma_x^2} + \frac{\left(y-y_o\right)^2}{2\sigma_y^2} + \frac{\beta\left(x-x_o\right)\left(y-y_o\right)}{\sigma_x \sigma_y} \right] \right\}}. \tag{3}$$

Here $I(x, y)$ is the intensity at coordinates (x, y), A_o is the amplitude (maximum intensity), σ_x and σ_y are the standard deviations along the major and minor axes, respectively, and β is the source position angle. All the four parameters (A_o, σ_x, σ_y, and β) have been obtained by fitting the Gaussian function to the source intensity profile. The beam major axis and beam minor axis have been obtained as the FWHM along the major and minor axes, derived from the Gaussian model. In order to obtain the integrated flux, we first correct the primary beam data using the CASA task impbcor and we then integrate the primary-beam-corrected flux over all the pixels associated with the source.

As explained above, so far we have extracted sources only through visual inspection. An automated extraction, which may be applied in future, would consist of the following steps:

(1) The original image will be split into several smaller regions.

(2) The software PYBDSF (Mohan & Rafferty 2015) will be used to generate the catalogue of each sub-region.

(3) After source extraction from PYBDSF, we will identify different types of sources, such as SS-AGN, FS-AGN, and SFG. The visual extraction will be used to validate the automated extraction procedure.

(4) Our current pipeline will be used to extract different source properties.

For extended sources, we may also use the Gaussian mixture model to extract the source information with better accuracy.

3.8 Shanghai (An, Jaiswal, Lu, Mohan, Lao)

The simulated images were analysed through three source-finding algorithms, namely AEGEAN (Hancock et al. 2012, 2018), DUCHAMP (Whiting 2012), and SEXTRACTOR (Bertin & Arnouts 1996). These methods extract contiguous islands of pixels above a specified detection threshold in the image. After applying a suitable fitting algorithm (e.g. two-dimensional Gaussian models), these islands are referred to as astronomical objects. Normally, the detection threshold is expressed in multiples of RMS noise in the image. Each software provides the centroid positions of the extracted sources, their integrated flux densities, and their geometrical properties. The integrated flux densities were finally corrected for primary beam by dividing with the primary beam response at each source coordinate.

The number of sources extracted through DUCHAMP was found to be significantly less in comparison to the other two for almost the same detection threshold level used (5σ); therefore, the source finding and comparison were performed only through AEGEAN and SEXTRACTOR. Extracted sources from AEGEAN and SEXTRACTOR were first cross-matched with the training data set using TOPCAT (Taylor 2005). The extracted source positions from SEXTRACTOR were found to match with the actual source positions more accurately than that from AEGEAN, so we finally used the results obtained with SEXTRACTOR for the final submission. The integrated flux densities estimated above were corrected for the primary beam by dividing with primary beam response (the given ancillary data) at the centre of the source. The source geometrical parameters, i.e. major axis length, minor axis length, and major axis position angle, were estimated by fitting the source brightness distribution with a two-dimensional Gaussian model having elliptical footprints. We have not attempted to fit different models for different sources according to the various source classes (SS-AGN, FS-AGN, or SFG).

4 EVALUATION OF THE SUBMISSIONS

4.1 True and submitted catalogue cross-match

The first step of the evaluation analysis was to cross-match the true and submitted catalogues. We started by performing a positional cross-match to identify true sources within a given radius from each detection (we used 1.5 the estimated source size convolved by the beam).

Given the very high density of sources in the true catalogue (50 sources per square arcmin at 560 MHz on average), this step typically yields multiple matches. The best match among this list has been defined as the source minimizing the sum in quadrature of positional, size, and flux density mismatch, respectively, defined as

$$D = \sqrt{D_{\text{pos}}^2 + D_{\text{size}}^2 + D_{\text{flux}}^2} \tag{4}$$

$$D_{\text{pos}} = \sqrt{(x - x')^2 + (y - y')^2}/S' \tag{5}$$

$$D_{\text{size}} = |S - S'|/S' \tag{6}$$

$$D_{\text{flux}} = |f - f'|/f', \tag{7}$$

where the prime denotes the attributes of the True catalogue; (x, y) are the pixel coordinates corresponding to (RA, Dec); S is the average source size, computed as the mean between major and minor axes, and f is the integrated flux density. We found this method to be much more reliable in identifying the best match than using position only, given the field crowding.

In setting the threshold on D to finally accept the best match as a true positive, we had to consider the probability of a false positive to be associated with a true source by chance. We assessed this likelihood by performing a null test for each submission. This consisted of creating a copy of the submitted catalogue with random coordinates (which we called null-test catalogue) and repeating the cross-match. The matches yielded by this procedure are all chance matches, and provide an estimate of the contamination due to chance matches for that submission. We found that, due to the high number of true sources in the catalogue, chance match could be significant. We set the threshold on D to 5σ of the distribution across all teams, after which this contamination is below 10 per cent for most submissions (see Table 3).

4.2 Scoring metrics

SDC1 addresses two separate aspects: source finding and source classification/characterization. In the following, we define metrics to judge these aspects individually, as well as a global score that considers them both. Each metric distils the performance to a single number per submission set (three frequencies, one depth).

To evaluate the source-finding component of SDC1, we defined scores based on the number of detections N_{d} and the number of matches N_{m}

$$C_{\text{tot}} = \frac{N_{\text{m},\nu 1}}{\text{FoV}_{\nu 1}} + \frac{N_{\text{m},\nu 2}}{\text{FoV}_{\nu 2}} + \frac{N_{\text{m},\nu 3}}{\text{FoV}_{\nu 3}} \tag{8}$$

$$R_{\text{tot}} = \frac{1}{3} \left[\frac{N_{\text{m},\nu 1}}{N_{\text{d},\nu 1}} + \frac{N_{\text{m},\nu 2}}{N_{\text{d},\nu 2}} + \frac{N_{\text{m},\nu 3}}{N_{\text{d},\nu 3}} \right], \tag{9}$$

where $\nu 1$, $\nu 2$, and $\nu 3$ are the three frequency channels for the same depth and FoV_ν is the sky area at frequency ν. The FoV normalizations in equation (8) are necessary to give comparable weight to the three frequency channels despite the very different number of sources in them. R_{tot} is the total reliability averaged between the three frequencies, and as such it is already correctly normalized.

To assess the accuracy of source classification and characterization, we consider errors on all the measured source attributes. These include

$$D_{\text{cf}} = |\text{cf} - \text{cf}'|/0.75 \tag{10}$$

$$D_{\text{PA}} = |\text{PA} - \text{PA}'|/10° \tag{11}$$

$$D_{\text{class}} = 0 \text{ if correctly classified}$$
$$= 1 \text{ otherwise}, \tag{12}$$

Table 3. Number of detections, matched sources, and associated uncertainty (matched sources in the null test) for all the submitted catalogues.

Team	Frequency	Depth	N_{det}	N_m	N_n
ARCIt-CACAO	560 MHz	1000 h	520 166	384 778	22 298
	1400 MHz	1000 h	150 370	143 713	3157
	9200 MHz	1000 h	765	600	2
	560 MHz	100 h	195 935	186 982	7450
	1400 MHz	100 h	58 328	57 012	783
	9200 MHz	100 h	263	115	2
	560 MHz	8 h	50 079	48 317	838
	1400 MHz	8 h	12 468	12 222	68
	9200 MHz	8 h	235	25	3
EngageSKA Portugal	560 MHz	1000 h	422 038	417 909	45 601
	1400 MHz	1000 h	144 147	142 659	43 802
	9200 MHz	1000 h	633	274	32
hs	560 MHz	1000 h	39 602	1030	569
	1400 MHz	1000 h	52 932	10 654	4130
	560 MHz	100 h	37 438	862	207
	1400 MHz	100 h	36 858	3618	1374
	560 MHz	8 h	19 991	241	44
	1400 MHz	8 h	12 708	677	189
ICRAR	560 MHz	1000 h	279 914	259 531	14 677
	1400 MHz	1000 h	41 877	32 694	718
	9200 MHz	1000 h	733	301	3
IPM	560 MHz	1000 h	610 505	18 037	3748
	1400 MHz	1000 h	214 460	15 158	437
	9200 MHz	1000 h	16 368	696	9
	560 MHz	8 h	83 769	5561	132
IPM2	560 MHz	1000 h	517 794	59 410	53 432
	1400 MHz	1000 h	672 544	46 717	56 037
	9200 MHz	1000 h	599 731	14 591	16 984
JLRAT	560 MHz	1000 h	1381 466	906 914	40 779
	1400 MHz	1000 h	391 562	249 051	18 024
	9200 MHz	1000 h	6334	2944	18
RADGK	560 MHz	1000 h	33	3	1
	1400 MHz	1000 h	25	4	1
Shanghai	560 MHz	1000 h	292 670	291 547	12 160
	1400 MHz	1000 h	102 129	101 687	1727
	9200 MHz	1000 h	316	296	1
	560 MHz	100 h	113 659	113 168	2576
	1400 MHz	100 h	32 143	31 898	308
	9200 MHz	100 h	68	60	1
	560 MHz	8 h	21 784	21 582	192
	1400 MHz	8 h	5582	5489	26
	9200 MHz	8 h	17	10	1

where PA is the position angle, cf is the core fraction, and the prime is the true catalogue value, as well as the already defined D_{pos}, D_{size}, and D_{flux} [equations (5)–(7)].

Those errors were used to associate a weight per matched source, w_i, ranging from a minimum of 0 to a maximum of 1 (see Appendix A for more details). Once summed over all matched sources, this yields an 'effective' number of matched sources,

$$\tilde{N}_m = \sum_{i=1}^{N_m} w_i \leq N_m, \qquad (13)$$

which is weighted down by errors in their characterization/classification. We finally defined an accuracy metric as

$$A_{tot} = \frac{\tilde{N}_{m,\nu 1}}{FoV_{\nu 1}} + \frac{\tilde{N}_{m,\nu 2}}{FoV_{\nu 2}} + \frac{\tilde{N}_{m,\nu 3}}{FoV_{\nu 3}}. \qquad (14)$$

The global SDC1 score, to assess both accuracy and source finding, is finally defined as

$$G_{tot} = \frac{B_{\nu 1}}{FoV_{\nu 1}} + \frac{B_{\nu 2}}{FoV_{\nu 2}} + \frac{B_{\nu 3}}{FoV_{\nu 3}}, \qquad (15)$$

where B is the difference between the effective number of matches and the number of false positives, N_f:

$$B = \tilde{N}_m - N_f. \qquad (16)$$

We note that B (and therefore G_{tot}) can become negative, if the number of false detections is larger than the number of matched sources or if the accuracy of characterization/classification is low.

It is worth pointing out that, due to the way that our cross-matching procedure works, the source finding and the source characterization performances are linked. Sources that have been correctly identified by a source-finding pipeline could fail the cross-matching step due
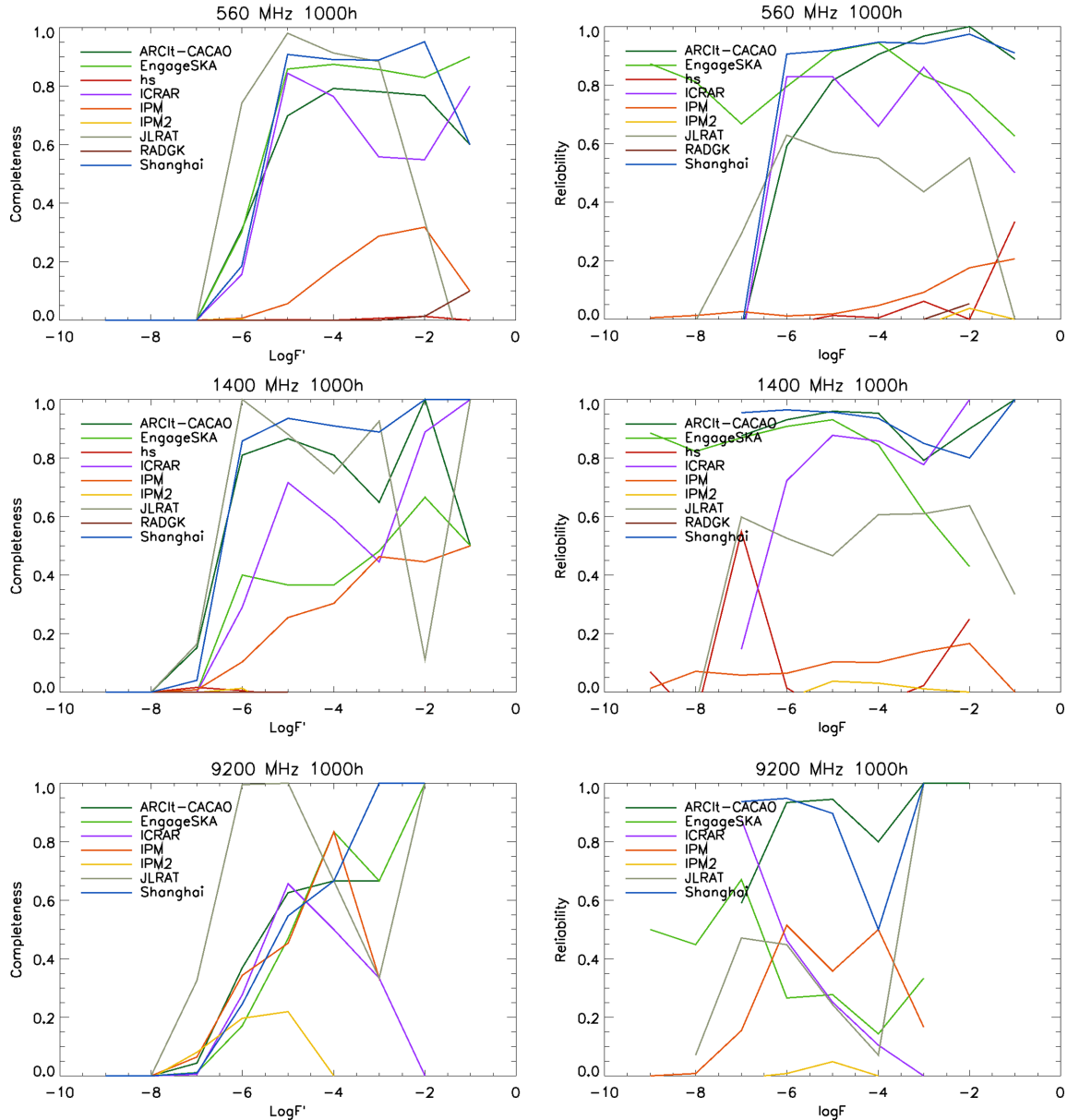
Figure 5. Completeness (left) and reliability (right) as a function of frequency (560 MHz, 1.4 GHz, and 9.2 GHz from top to bottom) for all the 1000 h submission sets.

to a large error on position, flux, or size. In this case, they would be classified as false positives, with an impact on the source-finding metrics as well as the accuracy ones. Although this feature of the evaluation process is not ideal, it could not be avoided. Setting a generous allowance for errors in the cross-match results in a very high contamination from chance matches, which makes every metric meaningless.

In the case of partial submissions (some of the frequency catalogues are missing for a given depth), a score of 0 was awarded for the missing frequency component in all scores.

The Leaderboard[9] for SDC1 was based on G_{tot} achieved by each team on the 1000 h submission by the challenge deadline, 2019 April

30. Since the deadline, teams have been allowed to update their submission for the purpose of this paper.

One of the teams that participated in the challenge (Ox-ICRAR) made the choice not to have their results appear in this paper; however, they are listed in the original Leaderboard. The results presented in the next section consider all the metrics defined in Section 4.2 and additional diagnostic plots.

The scoring procedure described in this section was originally implemented as detailed in Bonaldi & Braun (2018) to produce the challenge Leaderboards and the results of this paper. The released version[2] is an independent implementation of the same procedure that improves computational efficiency and portability, while providing consistent results. The scoring procedure evaluates individual frequency submissions and returns the B_ν terms in equation (15) that contribute to the overall score G_{tot}. The FoV$_\nu$ terms needed to compute G_{tot} are 30.25, 4.84, and 0.112 square deg for 560, 1400, and 9200 MHz, respectively.

---

[9]The Leaderboard is available at https://astronomers.skatelescope.org/ska-s cience-data-challenge-1-results/.

Table 6. C_{tot} and R_{tot} metrics for all teams, in order of decreasing C_{tot}.

Teams	C_{tot}	R_{tot}
IPM2	141 801	0.07
JLRAT	107 705	0.58
ARCIt-CACAO	47 766.1	0.83
EngageSKA	45 734.9	0.80
Shanghai	33 288.6	0.98
ICRAR	18 020.1	0.71
IPM	9937.96	0.05
hs	2235.29	0.07
RADGK	0.9256	0.08

5 RESULTS

Table 3 lists all the results submitted for SDC1. For each Submitted catalogue, we show the number of detected sources, N_d, the number of matches, N_m, and the number of matches obtained for the null catalogue, N_n, providing the estimated number of chance matches in N_m.

Only two teams (ARCIt-CACAO and Shanghai) submitted entries for all three depths, while the rest of them focused on the 1000 h exposure. In the following, the 1000 h results will be considered, unless otherwise specified.

5.1 Source finding

From Table 3, we can observe that the number of detections, N_d, spans orders of magnitudes between different teams, from some tens of sources detected via visual inspection to some hundred thousands detected with automated pipelines; a similarly large range applies to the number of matches, N_m.

Fig. 5 shows completeness and reliability, defined as

$$C(\log F') = (N_m(\log F') - N_n(\log F'))/N_t(\log F') \tag{17}$$

$$R(\log F) = (N_m(\log F) - N_n(\log F))/N_d(\log F), \tag{18}$$

where F is the integrated apparent flux density (before primary beam correction) and the prime denotes the true catalogue value. N_t, N_d, N_m, and N_n are the histograms of the true and submitted catalogues, of their cross-match, and of the cross-match obtained for the null catalogue, respectively. We note that C is measured as a function of the true catalogue entries ($\log F'$) and R as a function of the submitted catalogue entries ($\log F$). This is because the true catalogue and the submitted catalogue contain only $\log F'$ and $\log F$, respectively. Since the cross-matched catalogues in the numerators contain both $\log F$ and $\log F'$, we computed both histograms in order to achieve consistency with the denominators. This always guarantees the correct normalization of C and R even in the presence of errors in the estimation of the flux F. Whenever such errors are significant, this may, however, cause some discrepancy between C and R. Table 6 provides a summary of the source-finding performance, in terms of the C_{tot} and R_{tot} metrics, which is consistent with the view provided by Fig. 5.

The irregular shapes of the completeness and reliability curves for the 9.2 GHz channel are due to the much smaller number of sources, which makes these statistics much noisier. At the lowest frequencies, the curves are more regular, although some present interesting features. Some of the curves show a drop towards the highest fluxes, which is counter-intuitive. We investigate this aspect below.

At difference with most previous analyses (e.g. Hancock et al. 2012; Hopkins et al. 2015; Vernstrom et al. 2016), the SDC1 data
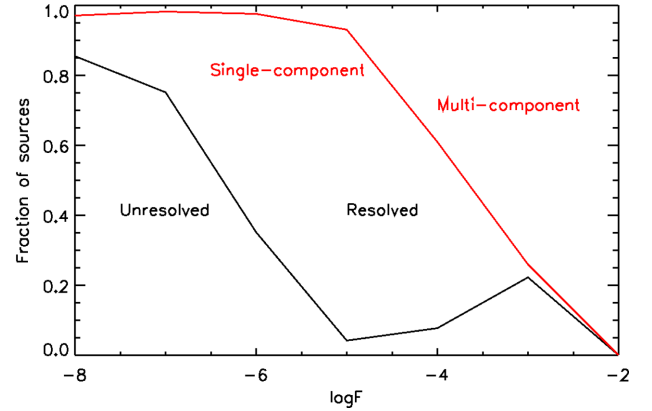
Figure 6. Fraction of resolved/unresolved and single/multi-component sources in the SDC1 map at 1.4 GHz as a function of the apparent flux.

set includes resolved sources and multicomponent sources with a complex morphology as well as unresolved (point-like) sources. In Fig. 6, we show the fraction of SDC1 sources at 1.4 GHz in the resolved/unresolved and multi/single-component categories as a function of the apparent flux. Different categories dominate at different fluxes:

(i) at $F \geq 10^{-3}$ Jy, the majority of sources are AGNs; they typically have a complex morphology and multiple components;

(ii) for $10^{-6} \leq F \leq 10^{-3}$ Jy, the main population is SFGs; they are modelled as single components but they still appear resolved at the 0.6 arcsec resolution of the 1.4 GHz map;

(iii) for $F \leq 10^{-6}$ Jy, the sources become mostly unresolved.

In Fig. 7, we show the completeness at 1400 MHz for the multiple components and the single components separately. Several teams performed quite differently on the two categories, both in terms of shape and of normalization of the curves. A different degree of success of the source-finding methods on the different source categories, dominating at different fluxes, can explain the non-trivial behaviour of completeness and reliability as a function of S/N.

One aspect that played a role in some of these results is the tessellation into smaller sub-images that several teams performed in order to reduce the computational complexity of the analysis (see the tile sizes in Table 4). A given tile size would affect source sizes several times smaller, the exact cut-off depending on details of the tessellation strategy (e.g. whether sub-images are overlapping or not) and of the analysis. Those sources would either not be identified or not characterized well enough; sources consisting of multiple components spanning two tiles would be identified as two independent sources. Table 5 shows the total number of sources above a noise level of 5σ and having largest angular size (LAS) above given sizes. These numbers can be compared to those in Tables 2 and 3 to have an idea of the fraction of sources potentially missed due to tessellation.

We can see that tile sizes of some hundreds of pixels would reduce the number of detections, with some impact on performance possibly happening up to a tile size of around 1000 pixels. ARCIt-CACAO, adopting a 1024 pixel tile, ascribes the features in Fig. 7 to tessellation effects (see Section 3.1 for more details).

Overall, our analysis supports the picture that the higher level of realism in the morphology of the SDC1 sources is responsible for some reduction in performance of the source-finding methods with respect to the ideal case. This is something to bear in mind as future high-resolution observations will expose the full complexity of the real sky. When needed for computational reasons, any tessellation of
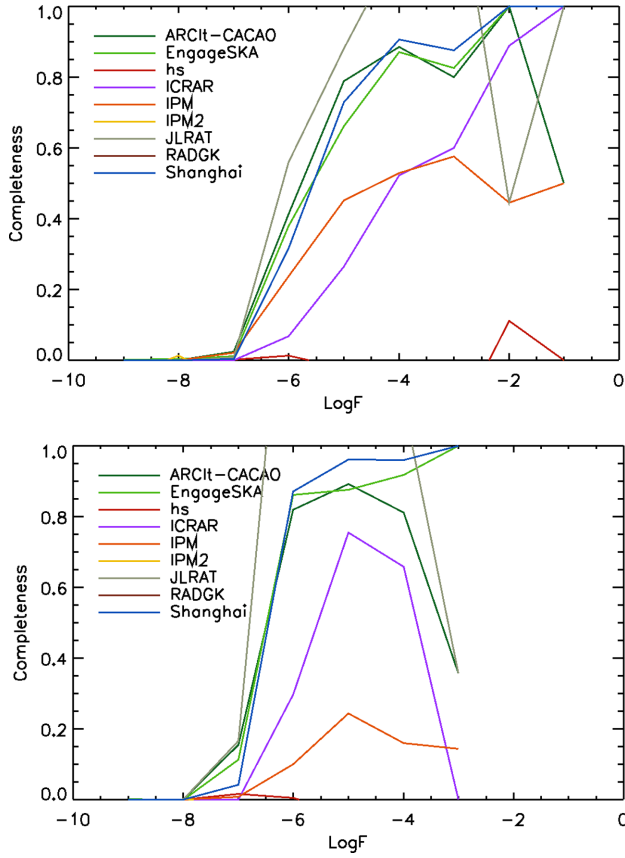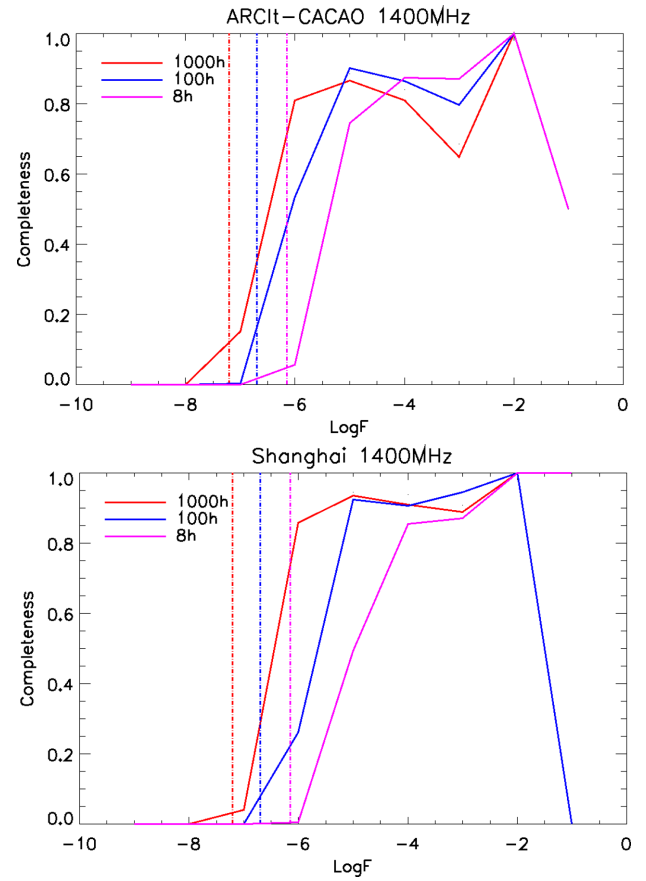
Figure 7. Completeness at 1.4 GHz separately for multicomponent (top) and single-component (bottom) sources.

Figure 8. Completeness as a function of depth for the ARCIt-CACAO (top) and Shanghai (bottom) teams at 1.4 GHz. The vertical dotted lines are the 1σ noise RMS.

Table 4. Size of the tiles used by the teams to divide the original SDC1 images for source detection, ordered by increasing size.

Teams	Side of square tile (pixels)
hs	50
ICRAR	205
JLRAT	320
IPM2	1000
ARCIt-CACAO	1024
EngageSKA	16 384
IPM	Full size
RADGK	Full size
Shanghai	Full size

Table 5. Number of $> 5\sigma$ sources with LAS above given sizes in pixels.

LAS (pixels)	560 MHz	1400 MHz	9200 MHz
25	6561	6555	114
50	1147	1369	45
100	98	107	7
200	20	30	6
500	3	6	2
1000	1	2	1
15 000	1	1	1

the FoV should be designed to cope with the presence of extended and multicomponent sources, by choosing carefully the tile sizes or by performing the analysis at multiple scales and spatial resolutions.

Fig. 8 shows how completeness at 1400 MHz varies with varying depth, from 1000 to 8 h, for the three teams providing those catalogues. All those teams managed to improve their catalogues for a deeper image, which is indicative of a noise-limited rather than a systematics-limited performance. The 50 per cent completeness for both the 100 and 1000 h exposures is at the 4.5σ–14σ depending on the teams.

5.2 Source characterization

Fig. 9 shows the distribution of the errors on all source attributes for all teams. Errors refer to the 560 MHz submission, which provides the largest source statistics; however, the results at other frequencies are similar. By definition, all these errors are positive and therefore the distributions are asymmetric. In all cases, a good performance produces a narrow distribution (small random error) peaked on 0 (no bias).

The left-hand panels show the positional, flux, and size errors D_{pos}, D_{size}, and D_{flux} as defined in equations (5)–(7). These error components are particularly important because of their role in the cross-match procedure. If their sum in quadrature D (equation 4) exceeds the threshold value, the source is discarded from further analysis and classified as a false positive. Some teams have a large
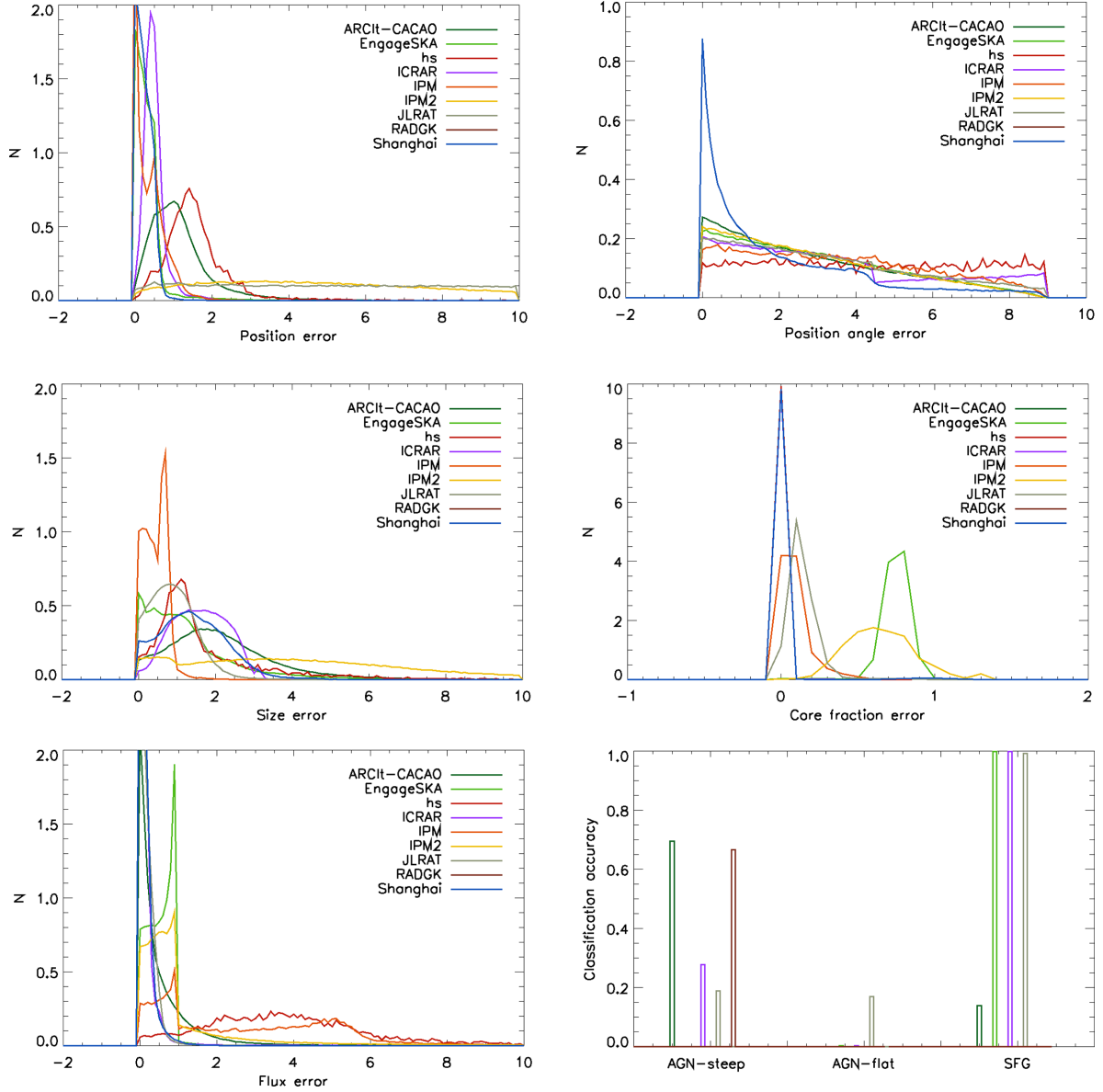
Figure 9. Distribution of errors on the source attributes for all teams on the 560 MHz 1000 h submissions.

random or systematic error on one or more of those attributes, and this is at the root of low scores in Table 6.

Position is generally well recovered, with only a few exceptions. Error on size is generally wider, which reflects the different size definition adopted by different approaches and difficulty in deconvolving the beam size. The flux error varies between different teams from very small (ARCIt-CACAO, ICRAR, Shanghai) to quite large (hs, IPM).

The right-hand panels show the position angle and core fraction errors D_{PA} and D_{cf} and D_{class} as defined in equations (10)–(12) and the accuracy of classification.

Most D_{PA} distributions are relatively flat, indicating that this attribute is not very successfully recovered. Some of the distributions for the core fraction error are bimodal because the true core fraction distribution is bimodal (cf = 0 for SFGs and cf $\neq$ 0 for AGNs). The accuracy of source classification is discussed in the next section.

In Table 7, we show the G_{tot} and A_{tot} scores for all teams, in order of G_{tot}. The accuracy score A_{tot} provides a good summary of the performance shown in Fig. 9. The global score G_{tot} summarizes all metrics

previously introduced, and G_{tot}^* records its value as it was achieved by the challenge deadline (the SDC1 leaderboard is based on G_{tot}^*).

5.3 Source classification

Overall source classification was not probed well enough by this challenge. The simulated observing strategy (a single telescope pointing) did not provide enough multifrequency information for the sources, which could have helped classification. Similarly, the challenge did not include data at other wavelengths, most notably optical and/or IR, which is often used to separate SFGs from AGNs by means of the methods involving optical emission line information such as BPT diagram and radio–IR correlation. These issues, inherent in the challenge design, are now understood and will inform the development of future data challenges featuring a source classification aspect. Since classification contributes only marginally to the final score, the impact of this design issue on the overall challenge is limited.

Table 7. G_{tot} and A_{tot} metrics for all teams, in order of decreasing G_{tot}. G_{tot}^* is the total score achieved by the deadline of 2019 April 30, which determined the SDC1 leaderboard.[7]

Teams	G_{tot}	G_{tot}^*	A_{tot}
Shanghai	19 112.8	−33 226.1	19 419.7
ARCIt-CACAO	17 361.3	2733.58	24 684.6
EngageSKA	16 914.9	16 914.9	20 551.5
ICRAR	5265.56	5265.56	11 691.1
RADGK	−4.584 27	−4.584 27	0.746 315
hs	−9325.29	−9325.29	684.933
JLRAT	−10 625.9	−53 069.4	64 752.6
IPM	−196 237	−196 237	4356.57
IPM2	−533 625	–	28 973.2

Table 8. Source classification statistics. The columns are the true classes and the rows the estimated classes for all teams that performed classification.

		AGN-steep	AGN-flat	SFG
ARCIt-CACAO	AGN-steep	5858	1972	181 325
	AGN-flat	0	0	0
	SFG	1745	1629	51 640
	None	826	1656	138 158
EngageSKA	AGN-steep	6	23	64
	AGN-flat	2	18	369
	SFG	8340	5780	403 547
	None	0	0	0
ICRAR	AGN-steep	1266	6	309
	AGN-flat	4	13	1
	SFG	3293	4233	261 066
	None	0	0	0
JLRAT	AGN-steep	3066	395	3552
	AGN-flat	394	3236	6543
	SFG	12 737	15 471	1268 354
	None	0	0	0
RADGK	AGN-steep	2	0	1
	AGN-flat	1	0	1
	SFG	0	0	0
	None	0	0	0

ARCIt-CACAO, EngageSKA-Portugal, ICRAR, JLRAT, and RADGK attempted source classification. In Table 8, we report how many sources have been classified in total by each team at 560 MHz, and how the true classes (on different columns) have been distributed into the estimated classes (in different rows). The perfect classification would correspond to all objects being in the diagonal of the first three rows. To visualize some of this information, the bottom right panel of Fig. 9 shows the fraction of correctly identified sources (on the diagonal) of each true class (sum of all elements of a column) for AGN-steep, AGN-flat, and SFGs, respectively.

EngageSKA, ICRAR, and JLRAT have almost all SFGs correctly identified. Their strategy has been to classify sources initially as SFGs, since they are the outstanding majority of the total, and to update their class to AGN if that was constrained by the data. This strategy maximizes the total score G_{tot} for their submission. ARCIt-CACAO did not attribute a class to all sources in their submitted catalogue; therefore, they have a lower percentage of correctly identified SFGs relative to the total (around 20 per cent).

Around 70 per cent of steep-spectrum AGNs were reliably classified by ARCIt-CACAO and, albeit on a very small sample, RADGK. All teams struggled with the classification of the flat-spectrum AGNs. The best results here is a 20 per cent correct classification by JLRAT. Flat-spectrum AGNs are compact objects, so they cannot easily be

morphologically distinguished from SFGs. Due to the simulated observing strategy of a single telescope pointing, the full spectral information, which could have helped classification, was available only for the small number of sources detectable at 9.2 GHz.

6 CONCLUSIONS AND DISCUSSION

With the SKA SDC1, we started investigating the properties of SKA continuum imaging products and the issues associated with their analysis. The challenge was meant as a training opportunity, to develop the skills of the astronomical community in the analysis of SKA-like data. Different teams approached it from a different level of expertise in the specific kind of analysis, and a different level of readiness in terms of existing pipelines. In all cases, the challenge constituted a step forward in the understanding of the problem and in specific expertise, which in itself is a huge success. The conclusions from this analysis are as follows:

(i) Eight independent teams, using various approaches (described in Section 3), succeeded in processing the images for source detection, classification, and characterization.

(ii) The high spatial density of sources challenged the standard procedure of confirming source detections through a positional cross-match. Additional source properties (flux and size) were used to distinguish between multiple matches consistent with each source position. The same issue will likely be encountered with real data as the depth of surveys increases.

(iii) Several teams achieved good performance in the source finding and/or source characterization aspect of the challenge and showed complementary strengths and weaknesses, which highlights the importance of applying multiple pipelines to the data.

(iv) The size of the data (4 GB per map) was just a first step towards the full SKA complexity, with a prospect to go further in future exercises. It has been overcome by most teams by splitting the data into sub-regions and analysing them separately. In some cases, the size of those regions was too small to successfully detect and characterize the more extended sources, with some impact on the performance.

(v) Teams that delivered results for different depths (8, 100, and 1000 h) succeeded in getting significantly higher detection rates for longer exposures. At 1.4 GHz, 50 per cent completeness varies across those teams between 4.5σ and 14σ. Most notably, it is stable going from the 100 to 1000 h exposure, which indicates that the challenge due to source crowding has been overcome very well. At 560 MHz, consistently with the lower spatial resolution, crowding might be playing a somewhat bigger role, as the 50 per cent completeness thresholds double from 5σ to 10σ going from medium to long integration. The very good image quality of the SDC1 simulation might have given an optimistic view on the subject of source crowding, which should be reassessed after introducing calibration, pointing, and deconvolution errors.

(vi) Some teams had difficulties in setting an optimal signal-to-noise threshold for source detection, which resulted in either a low completeness or a low reliability. In this respect, the specific expertise within teams in this kind of analysis played a major role. This stresses the importance of building domain knowledge and expertise alongside developing good, publicly available software for radio data analysis.

(vii) The increased realism of the SDC1 simulation (particularly the presence of resolved sources and complex morphology with multiple components) meant a decrease in performance of source-finding methods with respect to the ideal case, where completeness

and reliability approach 100 per cent at the highest S/N. Various approaches were particularly successful on one category of sources, but none was consistently good across all source types. This is something to bear in mind as future high-resolution observations will expose the full complexity of the real sky.

(viii) Accurate integrated flux and size estimation presented some problems. Both aspects are again linked to the complexity and the diversity of the source morphology, and to attempts by the adopted methods to fit them all with a single model. For example, Gaussian fitting is adequate on point sources but less so on resolved ones. Within SDC1, we allowed three alternative size definitions to fully represent the range of morphologies injected. Every conversion from those definitions to the ones used by the participating teams is just an approximation and can lead to a biased result.

(ix) The source classification (into SFG, steep-spectrum and flat-spectrum AGNs) was not probed well enough by this challenge. This specific aspect deserves further investigation, possibly by means of dedicated exercises.

ACKNOWLEDGEMENTS

DATA AVAILABILITY

(i) The SDC1 dataset and the challenge results are available at https://astronomers.skatelescope.org/ska-science-data-challenge-1/.

(ii) The SDC1 scoring package is available at https://pypi.org/project/ska-sdc/.

(iii) Results from the ICRAR team are available at https://github.com/ICRAR/skasdc1.

(iv) Most of the codes used by the participating teams are available publicly (see Section 3 for more details).

REFERENCES

An T., Wu X.-P., Hong X., 2019, Nat. Astron., 3, 1030
Astropy Collaboration, 2013, A&A, 558, A33
Astropy Collaboration, 2018, AJ, 156, 123
Banfield J. K. et al., 2015, MNRAS, 453, 2326
Banfield J. K. et al., 2016, MNRAS, 460, 2376
Becker R. H., White R. L., Helfand D. J., 1995, ApJ, 450, 559
Bertin E., Arnouts S., 1996, A&AS, 117, 393
Bonaldi A., Braun R., 2018, preprint (arXiv:1811.10454)
Bonaldi A. et al., 2019, MNRAS, 482, 2
Bonnarel F. et al., 2000, A&AS, 143, 33
Cai Z., Fan Q., Feris R. S., Vasconcelos N., 2016, preprint (arXiv:1607.07155)
Condon J. J. et al., 2012, ApJ, 758, 23
Frean M., Friedlander A., Johnston-Hollitt M., Hollitt C., 2014, in Niven R. K., Brewer B., Paull D., Shafi K., Stokes B., eds, AIP Conf. Proc. Vol. 1636, Bayesian Inference and Maximum Entropy Methods in Science and Engineering. Am. Inst. Phys., New York, p. 55
Hale C. L., Robotham A. S. G., Davies L. J. M., Jarvis M. J., Driver S. P., Heywood I., 2019, MNRAS, 487, 3971
Hales C. A., Murphy T., Curran J. R., Middelberg E., Gaensler B. M., Norris R. P., 2012, MNRAS, 425, 979
Hancock P. J., Murphy T., Gaensler B. M., Hopkins A., Curran J. R., 2012, MNRAS, 422, 1812
Hancock P. J., Trott C. M., Hurley-Walker N., 2018, PASA, 35, e011
Harrison I. et al., 2020, MNRAS, 495, 1737
He K., Zhang X., Ren S., Sun J., 2015, preprint (arXiv:1512.03385)
Hopkins A. M., Miller C. J., Connolly A. J., Genovese C., Nichol R. C., Wasserman L., 2002, AJ, 123, 1086
Hopkins A. M. et al., 2015, PASA, 32, e037
Hu J., Shen L., Albanie S., Sun G., Wu E., 2017, IEEE Trans. Pattern Anal. Mach. Intell., 42, 2011
Ioffe S., Szegedy C., 2015, PMLR, 37, 448
Jones E., Oliphant T., Peterson P., 2001, SciPy: Open Source Scientific Tools for Python. Available at: http://www.scipy.org/
Kapińska A. D. et al., 2017, AJ, 154, 253
Li C., Tam P., 1998, Pattern Recognit. Lett., 19, 771
Lin T.-Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S., 2016, preprint (arXiv:1612.03144)
Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C.-Y., Berg A. C., 2016, in Leibe B., Matas J., Sebe N., Welling M., eds, Lecture Notes in Computer Science, Vol. 9905. Springer, Cham, p. 21
Lukic V., de Gasperin F., Brüggen M., 2019, Galaxies, 1, 3

[10] http://www.astropy.org

McMullin J. P., Waters B., Schiebel D., Young W., Golap K., 2007, in Shaw R. A., Hill F., Bell D. J., eds, ASP Conf. Ser. Vol. 376, CASA Architecture and Applications. Astron. Soc. Pac., San Francisco, p. 127

Mohan N., Rafferty D., 2015, PyBDSF: Python Blob Detection and Source Finder, Astrophysics Source Code Library, record ascl:1502.007

Molinari S., Schisano E., Faustini F., Pestalozzi M., di Giorgio A. M., Liu S., 2011, A&A, 530, A133

Redmon J., Farhadi A., 2018, preprint (arXiv:1804.02767)

Ren S., He K., Girshick R., Sun J., 2015, preprint (arXiv:1506.01497)

Robotham A. S. G. et al., 2018, MNRAS, 476, 3137

Rowe B. T. P. et al., 2015, Astron. Comput., 10, 121

Sault R. J., Teuben P. J., Wright M. C. H., 1995, in Shaw R. A., Payne H. E., Hayes J. J. E., eds, ASP Conf. Ser. Vol. 77, A Retrospective View of MIRIAD. Astron. Soc. Pac., San Francisco, p. 433

Simonyan K., Zisserman A., 2015, Bengio Y., LeCun Y., in 3rd International Conference on Learning Representations. ICLR, San Diego, CA

Taylor M. B., 2005, in Shopbell P., Britton M., Ebert R., eds, ASP Conf. Ser. Vol. 347, TOPCAT & STIL: Starlink Table/VOTable Processing Software. Astron. Soc. Pac., San Francisco, p. 29

Vafaei Sadr A., Vos E. E., Bassett B. A., Hosenie Z., Oozeer N., Lochner M., 2019, MNRAS, 484, 2793

Van der Walt S. et al., 2014, PeerJ, 2, e453

Vernstrom T., Scott D., Wall J. V., Condon J. J., Cotton W. D., Perley R. A., 2016, MNRAS, 461, 2879

Virtanen P. et al., 2020, Nat. Methods, 17, 261

Wes M., 2010, in van der Walt S., Millman J., eds, Proc. 9th Python in Sci. Conf. SciPy, Texas, p. 56

Whiting M. T., 2012, MNRAS, 421, 3242

Whiting M., Humphreys B., 2012, PASA, 29, 371

Wright E. L. et al., 2010, AJ, 140, 1868

Wu C. et al., 2019, MNRAS, 482, 1211

Zhang K., Zuo W., Chen Y., Meng D., Zhang L., 2017, IEEE Trans. Image Process., 26, 3142

Zhao Q., Sheng T., Wang Y., Tang Z., Chen Y., Cai L., Ling H., 2019, in Proc. AAAI Conf. Artif. Intell., Vol. 33. AAAI Press, Palo Alto, California, p. 9259

Zhu Z., He M., Dai Y., Rao Z., Li B., 2019, 14th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE, Xi'an, China, p. 1789

## APPENDIX A: DEFINITION OF SCORES PER ATTRIBUTE AND PER SOURCE

The definition of the A_{tot} and G_{tot} metrics in Section 4.2 contained weights w_i. Each of these weights ranges from 0 to 1 and it quantifies the accuracy of the characterization and classification of the source i. In this appendix, we give details on how the weights w_i are computed.

The source properties considered for computing w_i are seven: position (the best between core and centroid position if both are present), flux density, core fraction, major axis, minor axis, position angle, and class. For all of them except the source class, the score per source i and per attribute j, w_i^j, is defined as

$$w_i^j = \frac{1}{7} \min \left\{ 1, \frac{D_i^j}{\text{thr}^j} \right\},\tag{A1}$$

Table A1. Definition of errors and thresholds for source attributes.

Attribute	Error	Threshold
Position	equation (5)	0.3
Flux density	equation (7)	0.1
Major axis	equation (6)	0.3
Minor axis	equation (6)	0.3
Position angle	equation (11)	1
Core fraction	equation (10)	0.05
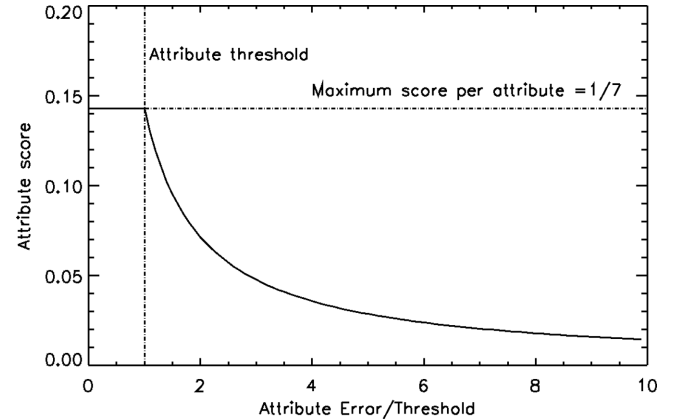
Figure A1. Score awarded per attribute as a function of attribute error/threshold.

where D_i^j is the error on the attribute j for the source i and thr^j is a threshold set on that attribute for all sources. The definitions of errors and thresholds for these source properties are in Table A1.

The behaviour of w_i^j is shown in Fig. A1. The maximum score of 1/7 is awarded if the error on the considered attribute is below the threshold; above the threshold, the score decreases and tends to 0 for $D_i^j \gg \text{thr}^j$.

The thresholds have been calibrated on the global error statistics for the submissions, so that the full range of scores from 0 to 1/7 are awarded on the full sample of results. In the case of the source class, w_i^j is either 1/7 or 0 depending on whether the source has been classified correctly or not. The final score per source is finally

$$w_i = \sum_{j=1}^{7} w_i^j.\tag{A2}$$

The 1/7 normalization of w_i^j guarantees that the maximum value of w_i is 1; this is awarded whenever all attributes of that source have been estimated with an error lower than the set threshold and the class has been correctly identified.

APPENDIX B: JLRAT ADDITIONAL FIGURES
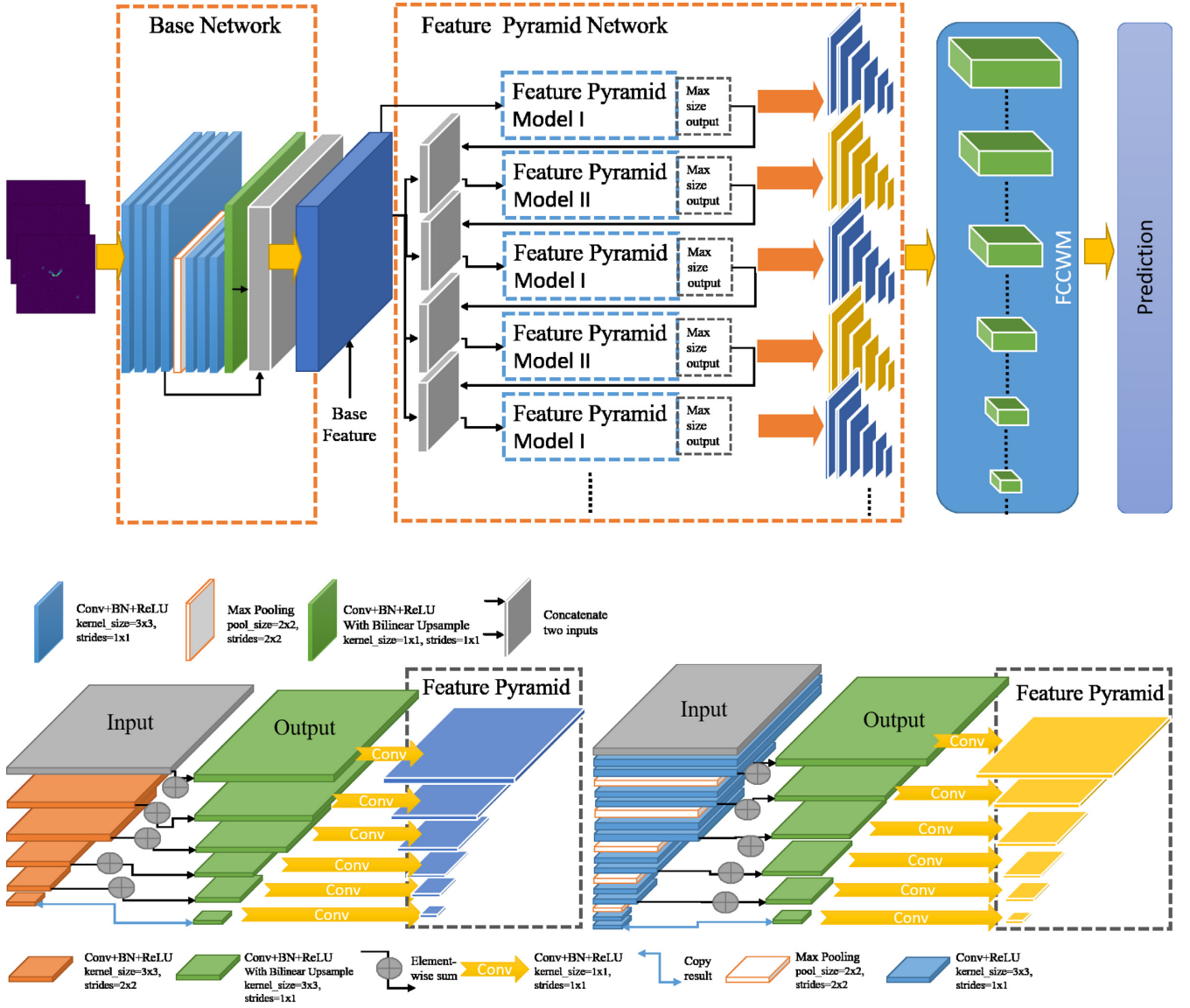
Figure B1. *Top:* the Source Detection Model of JSFM. *Bottom:* JSFM feature pyramid models I (left) and II (right).

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