Bubble size statistics during reionization from 21-cm tomography

Sambit K. Giri,[1]⋆ Garrelt Mellema,[1] Keri L. Dixon[2] and Ilian T. Iliev[2]

[1]*Department of Astronomy and Oskar Klein Centre, Stockholm University, AlbaNova, SE-106 91, Stockholm, Sweden*
[2]*Astronomy Centre, Department of Physics & Astronomy, Pevensey III Building, University of Sussex, Falmer, Brighton, BN1 9QH, United Kingdom*

ABSTRACT

The upcoming SKA1-Low radio interferometer will be sensitive enough to produce tomographic imaging data of the redshifted 21-cm signal from the Epoch of Reionization. Due to the non-Gaussian distribution of the signal, a power spectrum analysis alone will not provide a complete description of its properties. Here, we consider an additional metric which could be derived from tomographic imaging data, namely the bubble size distribution of ionized regions. We study three methods that have previously been used to characterize bubble size distributions in simulation data for the hydrogen ionization fraction – the spherical-average (SPA), mean-free-path (MFP) and friends-of-friends (FOF) methods – and apply them to simulated 21-cm data cubes. Our simulated data cubes have the (sensitivity-dictated) resolution expected for the SKA1-Low reionization experiment and we study the impact of both the light-cone (LC) and redshift space distortion (RSD) effects. To identify ionized regions in the 21-cm data we introduce a new, self-adjusting thresholding approach based on the K-Means algorithm. We find that the fraction of ionized cells identified in this way consistently falls below the mean volume-averaged ionized fraction. From a comparison of the three bubble size methods, we conclude that all three methods are useful, but that the MFP method performs best in terms of tracking the progress of reionization and separating different reionization scenarios. The LC effect is found to affect data spanning more than about 10 MHz in frequency ($\Delta z \sim 0.5$). We find that RSDs only marginally affect the bubble size distributions.

Key words: methods: statistical – techniques: image processing – dark ages, reionization, first stars – early Universe – radio lines: galaxies.

1 INTRODUCTION

Before the completion of hydrogen reionization, which happened around a redshift of about 6, the intergalactic medium (IGM) contained substantial amounts of neutral hydrogen. Its spin-flip transition can produce an observable signal at the rest frame wavelength of 21 cm. This signal constitutes the most direct probe of the reionization process as it depends directly on the distribution of neutral hydrogen (Furlanetto, Oh & Briggs 2006).

Detection of this signal has therefore been one of the key science drivers for a new generation of low-frequency radio interferometers, such as the Giant Metrewave Radio Telescope (GMRT; e.g. Paciga et al. 2011), the Low Frequency Array (LOFAR; e.g. Harker et al. 2010), the Murchison Widefield Array (MWA; e.g. Lonsdale et al. 2009) and the Precision Array for Probing the Epoch of Reionization (PAPER; e.g. Parsons et al. 2010). Detection of the signal is highly non-trivial due to very strong foreground signals as well as calibration challenges caused by ionospheric activity and instru-

mental effects. None of these arrays has yet achieved a detection, but major progress has been made and some useful upper limits have been established (Jacobs et al. 2015; Patil et al. 2017).

The main aim of this first generation of telescopes is to measure the (spherically averaged) power spectrum of the 21-cm signal. Extracting the power spectrum requires less signal to noise than producing images and power spectra also appear to form an excellent statistical measure of the properties of the signal as they are sensitive to both the evolution of reionization and the nature of the ionizing sources (e.g. Furlanetto, Zaldarriaga & Hernquist 2004a; Lidz et al. 2008; Iliev et al. 2012; Greig & Mesinger 2015). Therefore, a considerable amount of effort has been invested into deriving the power spectra from models and simulations (e.g. Mellema et al. 2006b; McQuinn et al. 2007; Mesinger, Furlanetto & Cen 2011) and devising strategies to extract them reliably from the interferometer data (Liu & Tegmark 2011; Dillon, Liu & Tegmark 2013; Trott et al. 2016).

The simulations have shown that the shape of the probability distribution function (PDF) of the 21-cm signal during reionization is far from Gaussian (e.g. Furlanetto, Zaldarriaga & Hernquist 2004b; Mellema et al. 2006b; Ichikawa et al. 2010). For this reason, the

⋆ E-mail: sambit.giri@astro.su.se

spherically averaged power spectrum does not fully describe its statistical properties. This non-Gaussianity can be studied using higher order statistics such as one-point skewness and kurtosis (see e.g. Watkinson & Pritchard 2014) or the full bispectrum and trispectrum (Watkinson et al. 2017, and references therein). However, such statistics are notoriously difficult to interpret physically (see e.g. Shimabukuro et al. 2016). This forms the main motivation for the ambition to map the 21-cm signal tomographically at many different frequencies with the future Square Kilometre Array (SKA; Mellema et al. 2013). Such tomographic data will show both the sizes and shapes of ionized regions as well as the density fluctuations in neutral regions and should thus give a clear view of how reionization progressed.

In order to interpret tomographic data sets in terms of, for example, the properties and distribution of the sources of ionizing photons, statistical tools for comparison to simulation results are needed. How does one characterize the tomographic data of the 21-cm signal? This question is not easily answered as there are no similar data sets in cosmology. The cosmic microwave background (CMB) is not tomographic and has a PDF which is very close to a Gaussian distribution. Galaxy redshift surveys do provide tomographic data sets, but they deal with discrete objects. However, results of those surveys can be transformed into galaxy density fields using density field estimators (e.g. Schaap & van de Weygaert 2000). The algorithms developed for finding voids in those fields may provide some useful insights on how to deal with tomographic image data (see e.g. Nadathur & Hotchkiss 2015, and references therein).

One obvious quantity in the context of reionization is the bubble size distribution (BSD), which describes how many ionized regions of a given size exist in the data. Simulations have shown that this measure describes the progression of reionization as larger and larger ionized regions appear the more reionized the Universe becomes (e.g. Furlanetto et al. 2004a; Mellema et al. 2006b; Mesinger & Furlanetto 2007). It also has appeal as a measure which appears to have a simple physical interpretation. We will therefore in this paper focus on BSDs obtained from 21-cm tomographic data sets.

BSDs have been studied before in the context of comparing simulation results. One obvious conclusion from 3D simulations of reionization is that the morphology of the ionized regions is highly complex. Although cartoon versions of the process occasionally depict the ionized regions as easily identified spherical bubbles, the regions in reality have highly irregular morphologies and a complex connectivity in three dimensions. Therefore, there is no unique way to define the BSDs and different methods will give very different answers.

For example, if one focuses entirely on connectivity, using the friends-of-friends (FOF) algorithm introduced in Iliev et al. (2006a), one will find that well before the end of reionization most of the ionized volume becomes contained in one large connected region. The scenario is very different from the result one obtains if one focuses on the largest spherical volume which fits inside the distribution of ionized regions, a method first introduced by Zahn et al. (2007) and known as the spherical-average (SPA) method. Yet another result is obtained if one finds the distribution of the distances to the edge of an ionized region from a large collection of random points and random directions, a method developed by Mesinger & Furlanetto (2007), also known as the mean-free-path (MFP) method.

Each of these methods has its own definition of bubble size and measure essentially different things. In order to compare inferences from different methods, one should be aware of this difference. Friedrich et al. (2011) and recently Lin et al. (2016) have analysed

these methods in the context of characterizing the differences between the ionization fraction results of different simulations. Those authors have pointed out the various characteristics, advantages and disadvantages of the different methods.

Here, we will take these methods and instead apply them to (simulated) 21-cm data. We will address the impact of resolution and the fact that 3D tomographic data will be in the form of light-cones (LCs) where the signal evolves along the frequency axis. In addition, we will consider the effect of redshift space distortions (RSDs; Jensen et al. 2013, 2016) on the BSDs. For now, we will not take into account other observational effects such as noise and calibration errors. We will restrict ourselves to the three well-established methods mentioned above, FOF, SPA and MFP, and not consider methods which have recently been proposed such as the watershed algorithm (Lin et al. 2016) or granulometry (Kakiichi et al. 2017). The basic question we are trying to answer is how these methods can be applied to observed 21-cm tomographic data and how well they perform in this context.

The paper is structured as follows. In the next section, we describe the quantity from which we will derive the BSDs, namely the redshifted 21-cm signal. Section 3 provides an overview of the size determination methods and the procedures to use them on 21-cm measurements. Section 4 gives a description of the simulations used in this study as well as how we generate the mock observations. Section 5 presents the results of our study. The sixth section contains the discussion along with the summary.

2 REDSHIFTED 21-CM SIGNAL

The neutral hydrogen 21-cm line will be a very powerful tool to study the EoR (Madau, Meiksin & Rees 1997). It is a hyper-fine line of wavelength 21 cm, caused by the ground state spin-flip transition in the atom's electron–proton configuration. The neutral hydrogen atoms in the IGM during reionization could be observed through the redshifted 21-cm signal using low-frequency radio telescopes. The signal is seen against the CMB and the measured differential brightness temperature can be written as (e.g. Mellema et al. 2013):

$$\delta T_b \approx 27 x_{HI}(1+\delta) \left(\frac{1+z}{10} \right)^{\frac{1}{2}} \left(1 - \frac{T_{CMB}(z)}{T_s} \right)$$
$$\times \left(\frac{\Omega_b}{0.044} \frac{h}{0.7} \right) \left(\frac{\Omega_m}{0.27} \right)^{\frac{1}{2}} \left(\frac{1-Y_p}{1-0.248} \right) \text{mK}, \qquad (1)$$

where x_{HI} and δ are the neutral hydrogen fraction and the density fluctuation, respectively, T_s is the spin temperature or the excitation temperature of the distribution of the two states of hydrogen. $T_{CMB}(z)$ is CMB temperature at redshift z and Y_p is the primordial helium abundance.

The above equation shows that the 21-cm signal would not be observed when T_s is fully coupled to the CMB temperature T_{CMB}. During EoR, the spin temperature is expected to approach the gas temperature due to the Wouthuysen–Field effect (Madau et al. 1997) and the 21-cm signal would be visible with CMB as the background. When the gas temperature is below T_{CMB}, the signal is seen in absorption and when it is above, it is seen in emission. For the case $T_s \gg T_{CMB}$, typical for the later stages of reionization, $1 - \frac{T_{CMB}(z)}{T_s} \to 1$ and the signal becomes independent of the value of T_s. Throughout this paper, we will use this high spin temperature limit.

The signal is observed at a frequency given by

$$\nu_{obs} = \frac{\nu_0}{1+z_{obs}} = \frac{\nu_0(1-v_\parallel/c)}{1+z}, \qquad (2)$$

where $\nu_0 = 1.42$ GHz is the frequency of the transition and z_{obs} is the observed redshift which is different from the cosmological redshift z due to line-of-sight component of the peculiar motions in the intergalactic gas, $v_\parallel$. This causes a distortion of the signal when observed in redshift space. See Mao et al. (2012) for a comprehensive description of this RSD effect.

Observations will produce a 3D data set $\delta T_b(\boldsymbol{\theta}, \nu_{obs})$, where $\boldsymbol{\theta}$ indicates a position in the sky. Since ν_{obs} depends on the cosmological redshift z, this data set will cover a range of look back times. The fact that the signal at different frequencies originates from different look back times is known as the LC effect and the data set $\delta T_b(\boldsymbol{\theta}, \nu_{obs})$ is referred to as an LC.

3 BUBBLE SIZE STATISTICS METHODS

In this section, we first provide a brief description of all the size determination methods which we explore in this paper. For a more detailed description of the methods, we encourage the readers to consult the original papers as well as Friedrich et al. (2011) and Lin et al. (2016). After this, we describe the method we use to identify ionized regions in 21-cm observations.

3.1 Methods

3.1.1 Mean-free-path (MFP)

This method was introduced in Mesinger & Furlanetto (2007) and is based on Monte Carlo inference. It selects a random ionized location and casts a ray in a random direction. The ray is followed until a stopping criteria is met at which point the length of the ray is recorded. This process is repeated numerous times, in our case 10^7 times. The final result is a histogram of ray length values. Different stopping criteria can be defined. When applied to a binary ionization fraction field in which cells are labelled as either ionized or neutral, the ray is stopped when it reaches the first neutral cell. This is how we will use this method.[1] The MFP method derives its name from the fact that the ray traced corresponds to the 'mean free path' of an ionizing photon, given that the ionized region is typically highly ionized and neutral region nearly completely neutral.

In the simplest implementation, seminumerical methods, such as 21cmFAST (Mesinger et al. 2011), directly produce binary fields. Fully numerical methods such as C^2-Ray produce continuous values between 0 and 1 for the ionization fraction x_{HII}. This means that a certain threshold value has to be chosen to convert this continuous field to a binary field. Often $x_{HII} = 0.5$ is chosen. As shown in Friedrich et al. (2011) the MFP–BSD is sensitive to the precise choice of this threshold value.

A BSD method is diffusive if applying it to a collection of non-overlapping bubbles of radius R_0 will not yield the correct BSD which is $\delta(R - R_0)$, but rather a distribution stretching over a range of bubble sizes. For the MFP method, ray lengths can vary from 0 to $2R_0$ and therefore it is diffusive. A BSD method is biased if the peak of the distribution is not at R_0 but at a different size. Lin et al. (2016) showed that the MFP–BSD peaks very close to R_0 and classified the method as unbiased.

3.1.2 Spherical-average (SPA)

The SPA method was proposed in Zahn et al. (2007) and used to compare the analytic calculation of the BSD from excursion set theory with results from radiative transfer simulations. For each ionized location, this method finds the largest sphere around it for which the average ionization fraction is above some chosen threshold value. The final result point is a histogram of radius values which is meant to represent the BSD. Since this method evaluates the average ionization fraction in a certain region, it requires a threshold value even in the case of a binary ionization fraction field. Since we want our bubbles to be mostly ionized, threshold values close to 1 are chosen. We will use 0.9.

As pointed out by Friedrich et al. (2011) and Lin et al. (2016), the SPA method is both diffusive and strongly biased. For a collection of non-overlapping bubbles of radius R_0, the bubble sizes range from 0 to R_0 and the peak of the distribution is found close to $R_0/3$. Note that Lin et al. (2016) refer to this method as the distance transform (DT).

3.1.3 Friends-of-friends (FOF)

The FOF method is based on the idea of hierarchical clustering used in the field of data mining and statistics (Ivezić et al. 2014). The linkages between data points are used to find clusters. If a data point is within a (chosen) linking length of any of the points in a cluster, it becomes part of that cluster. This method is extensively used for cluster analysis in *N*-body simulations (e.g. Press & Davis 1982; Davis et al. 1985). Iliev et al. (2006a) first used FOF to analyse gridded x_{HII} data from an EoR simulation by linking any neighbouring cells which have been labelled as ionized. For each cluster, the volume is calculated and the final result is a histogram of cluster volumes. Furlanetto & Oh (2016) pointed out that this approach is the same as the Hoshen–Kopelman algorithm (Hoshen & Kopelman 1976). The typical histogram, particularly in the middle and late stages of reionization shows a bimodal distribution in which one large, percolated cluster contains most of the ionized points and a collection of much smaller isolated clusters contain the remainder. Friedrich et al. (2011) and Furlanetto & Oh (2016) analysed this property further. The latter authors showed how this dominant cluster grows as reionization progresses. They also found this feature to be universal and due to the fact that reionization is percolation process.

Just as the MFP method, FOF does not require a threshold when applied to a binary field. However, as explained above the generation of a binary field from a continuous field does require the choice of a threshold. The FOF method is neither diffusive nor biased, but it measures volumes of topologically connected ionized regions and not radii.

3.2 Binary field creation

All the methods discussed above have been previously applied to ionization fraction fields. Depending on the origin of the field, these were either binary fields or continuous fields transformed into binary ones through the choice of an appropriate threshold value, for example $x_{HII} = 0.5$. However, we now want to apply these methods to data sets containing the observed value of δT_b. We therefore require a method to transform δT_b into a binary field of neutral and ionized regions. It is difficult to define a fixed threshold value to achieve this as δT_b depends on both density and ionization fraction variations, as well as on redshift. Fully ionized regions will of

---

[1] If the MFP method is applied to an H I density field, one can also calculate the optical depth for hydrogen ionizing photons τ along the ray and use a limit on τ as a stopping criterion, see Iliev et al. (2008) for an example of this approach.
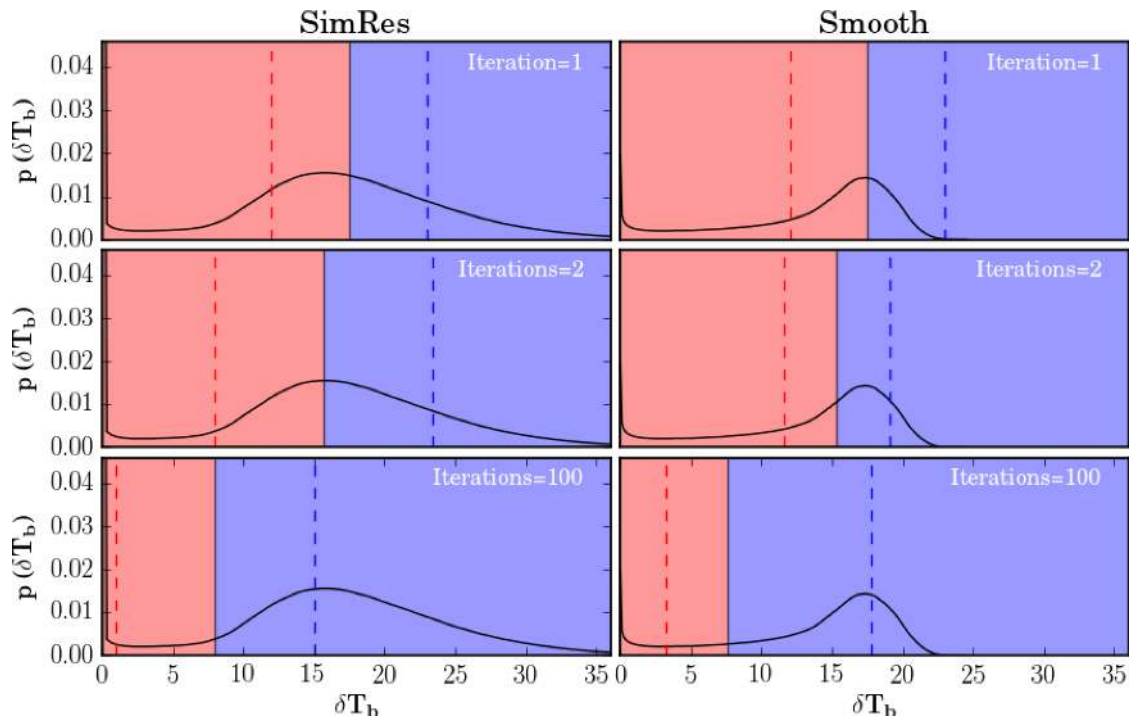
Figure 1. Illustration of the steps followed by K-Means to determine the threshold from the PDF has been shown. Row 1: two random points (cluster centres) are chosen and all the other points are linked to these values based on their distance. The dashed lines show the cluster centres and the colour of the shaded region shows linkage. Row 2: the mean of all the points present in each shaded region of 'row 1' gives the new cluster centres and the same linking process is repeated. Row 3: after many iterations of the same process, the cluster centres cease to shift, and the algorithm has converged.

course have $\delta T_b = 0$. However, *interferometric* tomographic data, due to the lack of baselines of length zero, do not allow the measurement of the absolute value of δT_b. Furthermore, the observations will not resolve scales below several comoving megaparsec and will therefore most likely not resolve the ionization fronts, thus further complicating the choice of a threshold.

The problem of dividing an image or 3D data set into different regions is called segmentation and in the field of image processing, many methods exist which use the data itself to achieve this. One obvious way for the case of the 21-cm signal is to consider its PDF. Previous works (see e.g. Ichikawa et al. 2010) have shown this PDF to be bimodal. This property allows automatic selection of a threshold value in the δT_b data to label regions as either ionized or neutral and hence create the binary field.

To automatically select an appropriate threshold value, we use the K-Means clustering algorithm (e.g. Hartigan & Wong 1979; Kanungo et al. 2002). K-Means is an unsupervised clustering method that finds clusters in large data sets (e.g. Sánchez Almeida & Allende Prieto 2013). A bimodal PDF has the data points clustered at the two peak values of the modes. K-Means then finds these two clusters and puts a threshold in between them. The method starts with choosing two random values in the range of values present in the data cube. These are the initial guess for the cluster centres. All other values are connected to one of these points to form two clusters. Next the centroids of these two clusters are found. Using these calculated centroids as the new centre of the clusters, the clustering of the points is recalculated. This process is repeated until the calculated centroids overlap with the cluster centres.

Fig. 1 illustrates our threshold selection process pictorially. Here, we start with guesses for the initial cluster centres that are far away from their final position. The cluster centres converge to the required positions when the K-Means algorithm is allowed to run a sufficient

number of iterations. It has been shown that the algorithm always converges to the solution in a finite time (e.g. Bottou & Bengio 1995; Kanungo et al. 2002), which makes it an apt choice for our case. We note that unlike some global thresholding methods, like Otsu's method (Otsu 1979), K-Means does not actually construct a PDF but works directly on the data points. However, for the 1D case considered here (values of the 21-cm signal) both methods produce similar results (Liu & Yu 2009). Both the 1D version of K-Means and Otsu's method become unreliable when the PDF does not possess a clear bimodality.

We will evaluate the performance of K-Means below in Section 5.1. As the continuous ionization fraction field also displays a bimodal PDF (see e.g. fig. 36 in Iliev et al. 2006b), K-Means can also be used when analysing such simulation results, as we will show below. We like to point out that there exist other algorithms to produce binary ionization fields from 21-cm data, which do not necessarily rely on the PDF. We will explore such other segmentation approaches in a future paper (Giri et al., in preparation).

3.3 BSD curves

After running the various BSD algorithms, we obtain the number of bubbles within a given size range. For both MFP and SPA, we will plot the following quantity against size R,

$$P(R) = R \frac{\mathrm{d}n}{\mathrm{d}R} = \frac{\mathrm{d}n}{\mathrm{d}\log(R)}, \tag{3}$$

where the latter expression shows that $P(R)$ is equivalent to using logarithmic binning of R. This quantity gives the fraction of bubbles that fall in a given size range, which means that it does not provide information on *how many* bubbles have been identified.

Table 1. Parameters of the reionization simulations that are used in this study. The labels used are the same as the ones used in Dixon et al. (2016), where these simulations were first introduced.

Label	Box size (Mpc)	g_γ (HMACH)	g_γ (LMACH)	g_γ (LMACH)$_{supp}$	Mesh	τ
LB1 (fiducial)	349	1.7	0	0	250^3	0.049
LB3	349	1.7	7.1	1.7	250^3	0.068
LB4	349	1.7	1.7	Equation 4 of Dixon et al. (2016)	250^3	0.057

Plotting the equivalent $P(V)$ curves for the volumes determined from the FOF method has the undesirable effect that the single largest cluster, which contains most of the ionized volume, becomes unnoticeable compared to the many small regions. To make the contribution of this largest cluster more prominent, we instead plot $V/V_{ionized}P(V)$ for the sizes determined from FOF, where $V_{ionized}$ is the total volume of ionized regions. This quantity therefore shows what fraction of the total ionized volume is contained in regions of a given volume.

4 SIMULATED 21-CM SIGNAL

4.1 Numerical simulation

To investigate the use of the different BSD algorithms on 21-cm signal data cubes, we use the results from large-scale fully numerical reionization simulations. The details of our simulation methodology have been discussed in previous papers (Mellema et al. 2006b; Iliev et al. 2006a; Datta et al. 2012). In short, we follow the evolution of matter with an *N*-body simulation using the CUBEP^{3}M code (Harnois-Déraps et al. 2013). We then postprocess the results with a radiative transfer simulation using the C^2-RAY code (Mellema et al. 2006a), where we assign an ionizing luminosity based on physically motivated models to the haloes found in the *N*-body simulation.

The specific simulations we have used follow reionization in a comoving volume of (349Mpc)^3. We assume a ΛCDM universe with cosmological parameters $\Omega_m = 0.27$, $\Omega_k = 0$, $\Omega_b = 0.044$, $h = 0.7$, $n = 0.96$, $\sigma_8 = 0.8$ and $Y_p = 0.248$, consistent with the *Wilkinson Microwave Anisotropy Probe* (*WMAP*) (Komatsu et al. 2011) and Planck (Planck Collaboration et al. 2016a) results. We will use three different assumptions for the source properties, labelled LB1, LB3 and LB4. These simulations were described and studied in detail in Dixon et al. (2016). A summary of the source parameters used for those simulations is given in Table 1. LB1 is our fiducial case. In this case, the only active sources are located in dark matter haloes of masses larger than 10^9 M$_\odot$ (high-mass atomically cooling haloes or HMACHs). These haloes release ionizing photons at a rate of 1.7 photons per baryon every 10^7 yr. Simulation LB3 uses additional low-mass sources with halo masses between 10^8 and 10^9 M$_\odot$ (low-mass atomically cooling haloes or LMACHs) with an ionizing photon rate of 7.1 per baryon every 10^7 yr. These haloes are assumed to be subject to radiative feedback and their ionizing photon rates drop to 1.7 photons per baryon every 10^7 yr once they are located inside an ionized region. In the LB4 case, the same low-mass sources are used, but the radiative feedback is implemented by a mass-dependent suppression factor in ionized regions, as described in Dixon et al. (2016). Apart from the simulation parameters Table 1 also lists the value for the electron scattering optical depth derived from these simulations. The values are all consistent within 1σ with the measurements by the *Planck* satellite (Planck Collaboration et al. 2016b,c).

We construct $\delta T_b(\boldsymbol{x}, z)$, the differential brightness temperature, originating from a given location at a given time corresponding to cosmological redshift z using equation (1). The ionization fractions $x_{H II}$ are produced by the radiative transfer simulation, while the density fluctuations δ are taken from the results of the *N*-body simulation. Those density fluctuations have been smoothed and gridded into the radiative transfer mesh. Since all values in this 3D data set correspond to the same cosmological redshift z, we refer to it as coeval cubes (CCs) to distinguish it from the LC data set discussed next.

4.2 Light-cone construction

As explained in Section 2, the data set observed by a radio interferometer is a nLC in which the images at different frequencies correspond to different signals originating from different redshifts and which are in addition distorted due to peculiar motions in the intergalactic gas (i.e. the RSDs). We construct LC data from the coeval simulation data using the procedure described in Mellema et al. (2006b) and Datta et al. (2012). The neutral fraction LC and the density LC are constructed separately and then are used to construct a 21-cm signal LC using equation (1). To be able to study the impact of the RSD we produce two types of LC data, one without RSD for which $z_{obs} = z$ and one with RSD, using equation (2). We account for the RSD in the LC using the MM-RRM scheme explained in Mao et al. (2012). The top two panels of Fig. 2 show cuts along a non-distorted and a redshift space distorted LC from our fiducial simulation.

4.3 Telescope resolution

The typical full width at half-maximum (FWHM) of the point spread function of an interferometer is given by (e.g. Rohlfs & Wilson 2013),

$$\Delta\theta = \frac{\lambda}{B}. \tag{4}$$

In the above equation, λ is redshifted value of λ_{21} (i.e. $21.1(1 + z)$ cm) and B represents the maximum baseline length used for producing the images. Unless otherwise specified, we use the planned maximum baseline of the core of SKA1-Low, which is 2 km. We will refer to this as SKA1-Low resolution although the actual resolution may be slightly different.

To mimic the response of SKA1-Low, we smooth the LC with a Gaussian kernel of FWHM $\Delta\theta$ in the angular direction. This FWHM is frequency dependent as the resolution of the radio telescope decreases as we go to higher redshift. In the frequency direction of the LC, we smooth the data with a top-hat kernel of the same width as the FWHM of the corresponding angular smoothing kernel. The two lower panels of Fig. 2 illustrate how this smoothing affects the simulated signals for both the non-distortied and the redshift space distorted case.
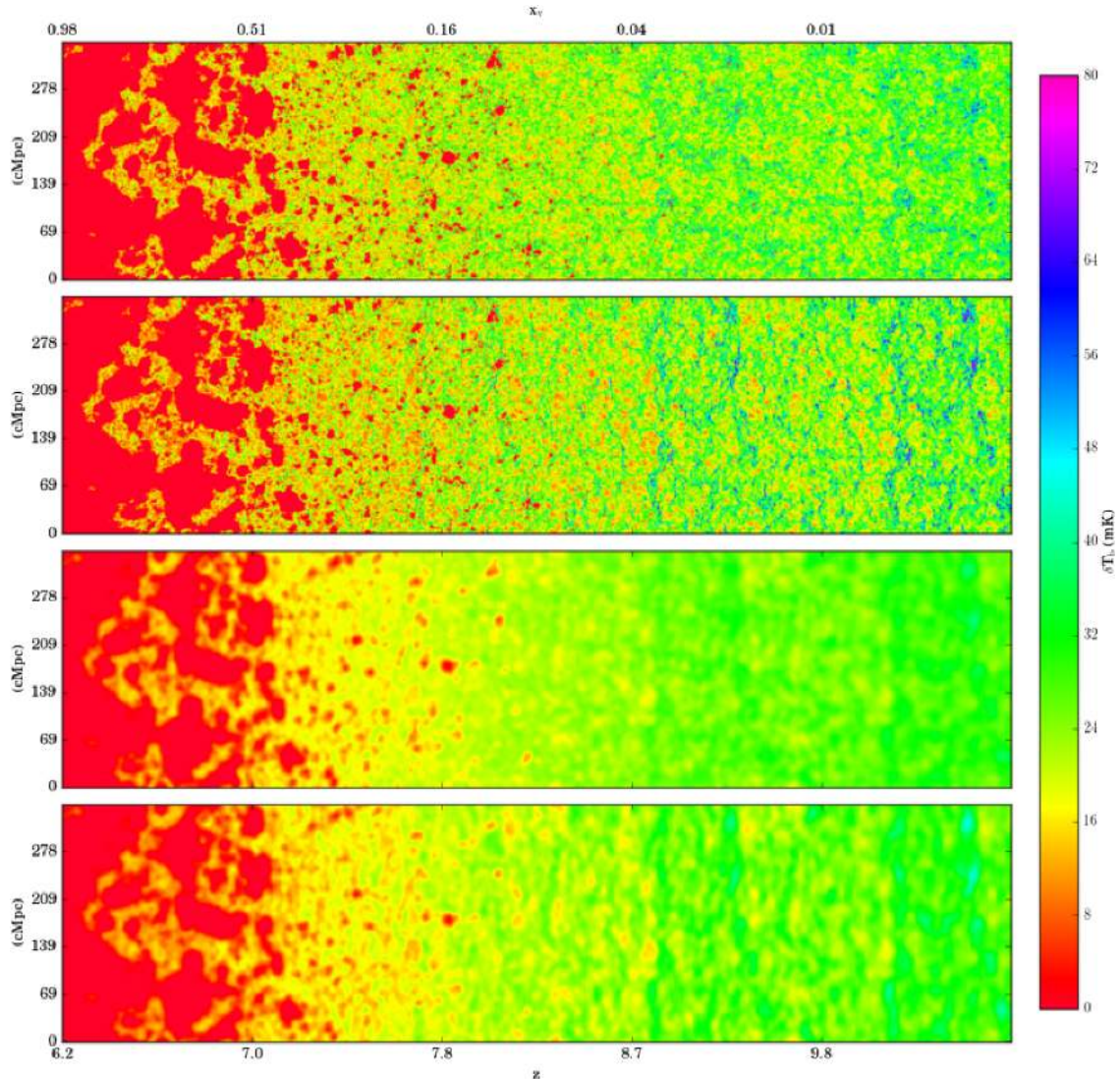
Figure 2. A cut along the LC from our fiducial simulation. The top two panels show the results at the resolution of the simulation without and with RSD effects, respectively. Similarly, the bottom two panels show those LCs after smoothing with a Gaussian beam corresponding to the resolution of a radio interferometer with a maximum baseline of 2 km. In the frequency direction, the LC data have been smoothed with a top-hat kernel of the same width as the FWHM of the corresponding angular smoothing kernel. The colour bar shows the absolute value of the differential brightness temperature δT_b. The vertical axes of the LC slice gives the length in comoving units. The horizontal axis in the bottom panel shows the redshift values whereas the top panel indicates the corresponding mean ionization fraction.

5 RESULTS

5.1 Global ionization fraction

After segmenting a tomographic 21-cm data set into a binary ionization fraction field, the first quantity to consider is the global ionized fraction by volume, x_v, or in other words what fraction of space is contained in ionized regions. This quantity is easily calculated from simulation results but for the observations will depend on the chosen segmentation as well as the resolution. In Fig. 3, we show the measured global ionization fraction $\hat{x}_v$ against the actual one x_v for our fiducial simulation for the entire reionization history. We consider four different binary fields. The first two were generated from the ionized fraction and δT_b fields at the resolution of the simulation. The other two were obtained from δT_b fields where we reduced the resolution to the SKA1-Low case and twice worse, the latter implying maximum baselines of 1 km. The binary

fields were produced with the K-Means algorithm as described in Section 3.2.

We see that the segmentation of the 21-cm signal and the ionization fraction data at the resolution of the simulation give the same values for $\hat{x}_v$, hence K-Means recovers the ionized regions well. Even at this resolution the measured value is always lower than the actual one, with differences reaching ∼20 per cent. When creating the binary field, partially ionized cells with ionization fractions below the threshold value will be classified as neutral and do not contribute to the measured global ionization fraction. On the other hand, partially ionized cells above the threshold will contribute 100 per cent to the measured global ionization fraction. The results show that the missing contribution of the former group dominates over the additional contribution of the latter group.

The measured global ionization fraction deviates even more from the actual one after reducing the resolution (dashed and dash–dotted
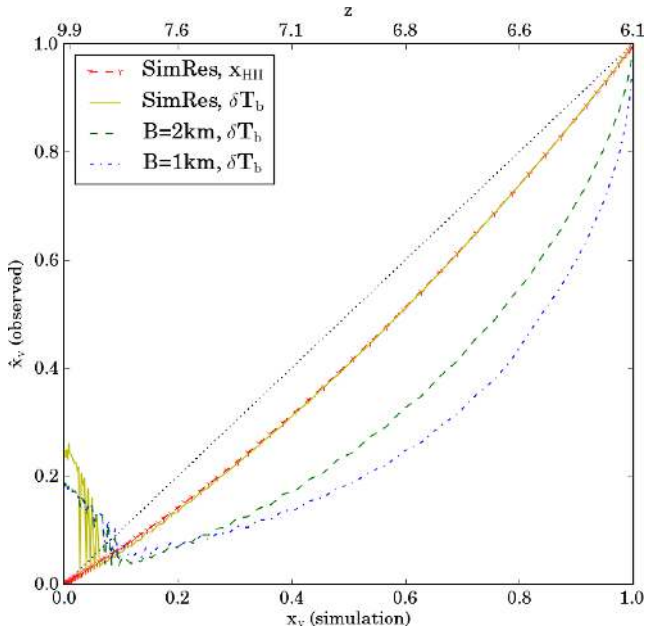
Figure 3. The fraction of ionized cells $\hat{x}_v$ identified by the K-Means algorithm against the mean volume-weighted ionization fraction x_v from the simulation. K-Means was used to produce binary fields from following data sets: the ionized fraction x_{HII} at the resolution of the simulation and the differential brightness temperature at three different resolutions: that of the simulation, and those corresponding to maximum baselines of 2 km and 1 km.

Table 2. A list of the global ionization fractions at different redshifts for our fiducial simulation LB1. x_v gives the average value of the ionization fraction data cube. $\hat{x}_{v,\text{sim}}$ gives the fraction of 21-cm cells labelled as ionized after segmentation. $\hat{x}_{v,\text{smooth}}$ gives the same fraction for 21-cm data cubes smoothed to SKA1-Low resolution.

z	x_v	$\hat{x}_{v,\text{sim}}$	$\hat{x}_{v,\text{smooth}}$
6.4	0.90	0.88	0.83
6.7	0.70	0.64	0.49
6.8	0.60	0.54	0.40
6.9	0.50	0.45	0.28
7.3	0.30	0.24	0.12
7.4	0.25	0.20	0.09
7.8	0.15	0.11	0.05

lines in Fig. 3). While smoothing the data one can on the one hand expect ionized bubbles below a certain size to be no longer visible while on the other hand larger bubbles may appear even larger due to apparent joining. From the reduced values of $\hat{x}_v$ for lower resolution we infer that the first effect dominates. For the lowest resolution considered here, the measured global ionization fraction can be less than half of the actual value.

We conclude that it will be hard to obtain an accurate determination of the actual global ionization fraction from tomographic images. Values of x_v and $\hat{x}_v$ for a number of representative redshifts are given in Table 2. These are the redshifts for which BSDs are presented in the following section.

Below a global ionization fraction of $x_v \approx 0.10$, the $\hat{x}_v$ derived from the 21-cm signals become very noisy. Here, the K-Means method has difficulty in dividing the PDF of the signal into ionized and neutral values since the number of ionized data points is
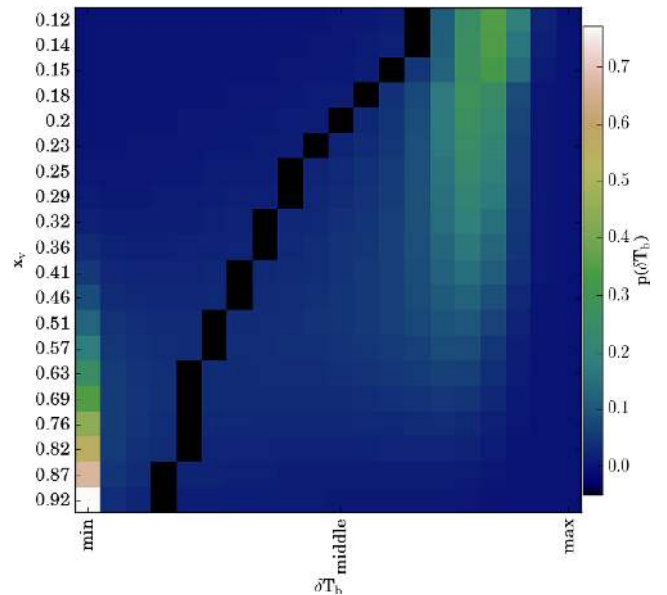

Figure 4. The PDF of observed 21-cm signal at different redshifts shown as a colour map. Each horizontal slice represents a PDF at a particular global ionization fraction (vertical axis). The horizontal axis shows the binned values of δT_b which have been rescaled between their minimum and maximum. The bin labelled as 'middle' is average value of minimum and maximum. The colours represent the number density of the PDF which is normalized to unity. Along each of the PDFs a black spot indicates the threshold value found by the K-Means algorithm.

very low during the early times. Therefore, K-Means is not a good classifier for the 21-cm signal from the early stages of reionization.

To better appreciate the performance of the K-Means algorithm we show in Fig. 4, where it places the threshold value with respect to the PDFs of δT_b. The colours show the value of its PDF from early (top) to late stages (bottom) of reionization against the values of δT_b (scaled from minimum to maximum). The column at the minimum value correspond to the highly ionized cells and the bright areas near δT_{max} correspond to the neutral cells. The spread in the latter is due to the density fluctuations. The threshold should be such that it separates these two modes. The black spots indicate where K-Means puts the threshold value. We can infer that the method works quite well for most of reionization epochs. However, for the earliest stages K-Means places the threshold value very close to the density fluctuation mode. As a consequence it identifies points in the tail of the density mode as ionized. The ionized cluster that K-Means looks for is so small that the method cannot define a prominent centroid and as a result both centroids are found inside the density fluctuations cluster. This behaviour explains the noisy results in Fig. 3. For early stages of reionization, a different threshold algorithm should be considered. We postpone a further discussion of this to a future paper (Giri et al., in preparation).

5.2 Effect of limited resolution on bubble size distributions

Now that we have established the performance of the K-Means algorithm and the effect of smoothing and segmentation on the ionization fractions, we can address the performance of the different BSDs introduced in Section 3.1. In this section, we address the effects of using the 21-cm signal at limited resolution.

In left-hand panels of Figs 5 and 6, we compare the SPA–BSDs and MFP–BSDs at the resolution of the simulation (solid lines) to
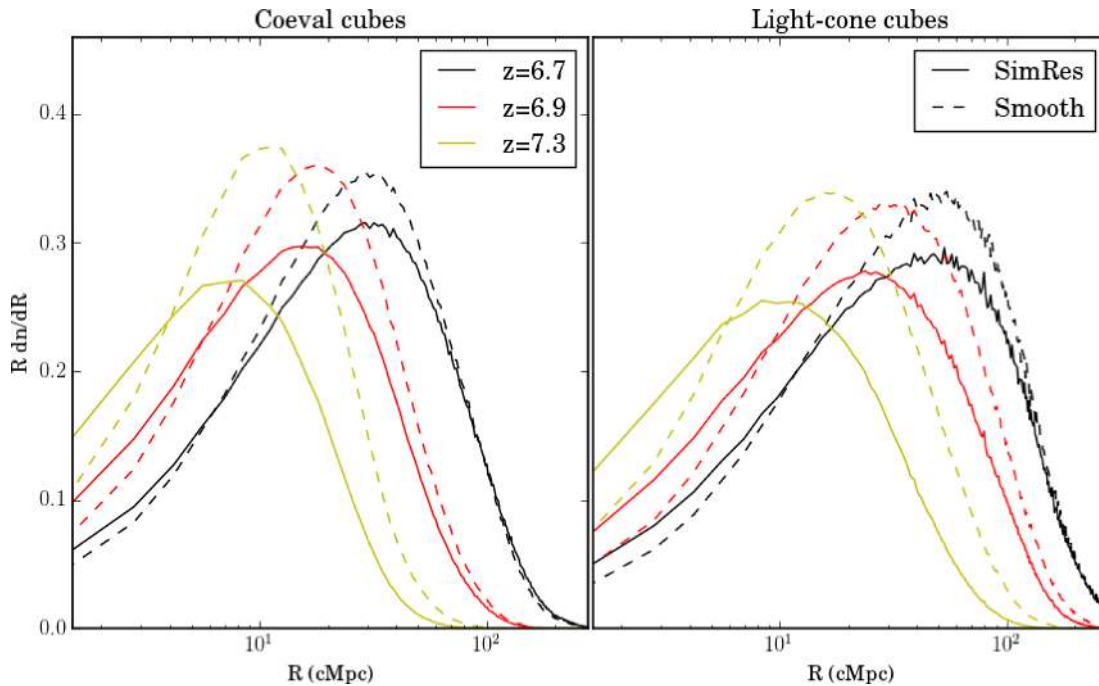
Figure 5. MFP results: the left-hand panel displays the BSDs for CCs at three different redshifts at the resolution of the simulation (dashed) and at SKA1-Low resolution (solid). The right-hand panel display the size distribution of the LC with the indicated central redshift at the resolution of the simulation (dashed) and at SKA1-Low resolution (solid). The BSDs for $z = 7.3$, 6.9 and 6.7 are shown as the curves from left to right, respectively. The corresponding ionization fractions are given in Table 2.
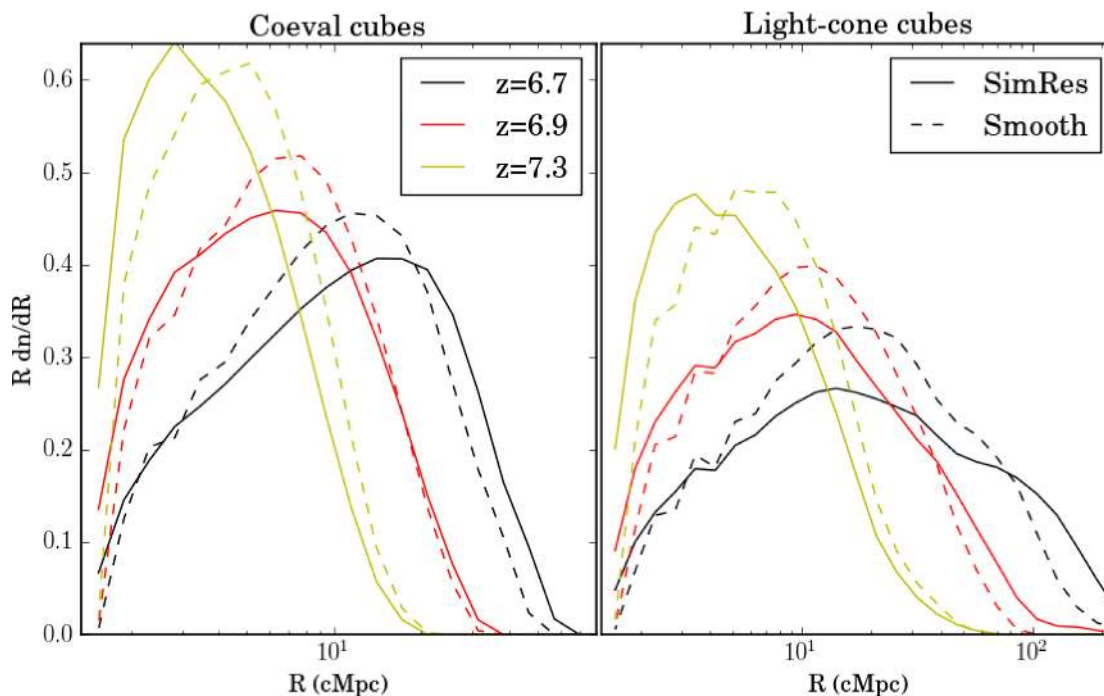

Figure 6. The same as Fig. 5 but for the SPA–BSD.

the ones at a SKA1-Low resolution (dashed lines). We have picked three redshift values for which the intrinsic and measured global ionization fractions are listed in Table 2. Both methods show that during the early stages of reionization the peaks of the BSDs shift to larger sizes after reducing the resolution, which is a consequence of

smaller bubbles being smoothed out and larger bubbles thus taking up a larger fraction of the distribution. For the later stages ($z = 6.7$), we notice that the relative frequency of both the smallest and largest regions is reduced in the smoothed data making the distributions more narrow. This is seen in both the MFP and SPA distributions

but is more clear in the latter. As a result, the peak value falls at a smaller radius in the smoothed case. As the largest regions display quite complex morphologies with tunnels, bridges and islands, we attribute this behaviour to the smoothing removing some of the connections which exist in the non-smoothed case.

Kakiichi et al. (2017) also noted a shift to larger sizes in the BSDs found by the granulometry method and attributed this to smoothing causing an apparent joining of ionized regions, labelling this effect as the 'smoothing bias'. However, normalized curves, as the ones we are using here and also used in Kakiichi et al. (2017), display the *fraction* of bubbles at a particular size. Therefore, a reduction in the absolute number of smaller regions can shift the peak of the distribution to larger values without the need of increasing the absolute number of larger regions. The impact of apparent joining of ionized regions can therefore not be determined from these BSDs.

Due to their change of shape, the BSDs at lower resolution show less evolution than those for the full resolution case. As a result, it may be hard to distinguish between two different stages of reionization based solely on these measured curves. However, these normalized BSDs are not sensitive to the total number of ionized regions or the global ionized fraction. To track the progress of reionization, we therefore need to analyse these BSDs jointly with the measured global ionization fractions $\hat{x}_{\mathrm{v,smooth}}$. As can be seen from Table 2, $\hat{x}_{\mathrm{v,smooth}}$ does evolve substantially, from 0.12 to 0.49, for the low-resolution case.

The MFP–BSDs (Fig. 5) and SPA–BSDs (Fig. 6) show similar behaviour and the relative shifts of the curves are very similar in both cases. However, it should be noted that the radius at which the BSD peaks is always lower for the SPA method than for the MFP method. For example, the peak distribution values of the SPA–BSD for $\hat{x}_{\mathrm{v}} = 0.1$ and $\hat{x}_{\mathrm{v}} = 0.4$ differ by about 6 Mpc, whereas for the MFP–BSD we see a difference of about 20 Mpc. Hence, we confirm the results from Friedrich et al. (2011) and Lin et al. (2016) that the peak values of the SPA–BSDs are around three times smaller than the peak values of the MFP–BSDs. Lin et al. (2016) have shown that this is due to the strong bias of the SPA method.

Lin et al. (2016) have also shown that shape of the MFP–BSD is closer to the intrinsic BSD leading to the MFP method being preferred over the SPA method. Since the MFP method uses random positions and directions for the rays, it can be sensitive to sampling noise, as can be seen in the results for the later stages of reionization in Fig. 5. This noise can be reduced by increasing the number of rays being traced.

Fig. 7 (left-hand panel) shows the FOF–BSD at the resolution of the simulation (solid) and at SKA1-Low resolution (dashed). As usual, the distribution is bimodal with one large ionized region that dominates the total ionized volume and a population of much smaller regions making up the rest. The large ionized region has been referred to as the 'percolation cluster' (see Furlanetto & Oh 2016), and appears at $x_{\mathrm{v}} \approx 0.10$. It forms when almost all the ionized regions connect through small bridges and is a distinct feature of any percolation process. The reduction in the amplitude of the mode containing the smaller bubbles illustrates that this population becomes less important as reionization progresses.

The FOF results clearly show the impact of limited resolution on the population of small regions, as the smallest bubbles are suppressed by a factor of more than 1000. However, we also see a joining effect since in the low-resolution data the largest smaller regions tend to be larger than in the high-resolution case, which suggests that joining does affect the measured BSDs, also for the MFP and SPA methods. Note however, that the percolation cluster actually has a somewhat smaller volume in the low-resolution case.

For $z = 7.3$, the percolation cluster has not yet fully developed in the smoothed data whereas a series of smaller regions in the volume range 10^5–10^6 Mpc^3 is detected. This is actually consistent with the value of $\hat{x}_{\mathrm{v}}$ which is 0.12 for this redshift. The percolation cluster typically emerges around this value (Furlanetto & Oh 2016). Hence, the reduction in resolution makes the observable smaller bubbles larger and percolation cluster smaller. As the percolation cluster dominates the entire ionized volume, the measured ionization fraction is always lower at lower resolution, consistent with the results in the previous section.

Furlanetto & Oh (2016) have predicted that the $V^2 \mathrm{d}n/\mathrm{d}V$ curve for the population of smaller regions determined by the FOF method should be flat due to the nature of reionization as a percolation process. We indeed observe this behaviour at simulation resolution. However, after smoothing the slope becomes positive. Interestingly for all cases, the slope is such that $V^2 \mathrm{d}n/\mathrm{d}V \propto V$ or $\mathrm{d}n/\mathrm{d}V \propto V^{-1}$. If this transition to a positive slope is a universal result, FOF–BSDs from observations could still be used to confirm the percolative behaviour of the reionization process.

5.3 Line-of-sight evolution

The previous subsection described the results for CC, but the observations will of course deliver LC image cubes instead, where the frequency axis covers the signals from a range of redshifts. In this section, we consider the impact of the LC effect.

The right-hand panels in Figs 5–7 show the different BSDs for LC data of which the central redshift coincides with the redshift values indicated in the figure. The width of the LC corresponds to a distance of 349 Mpc which is roughly $\Delta z \approx 0.80$ and is the same as our simulation volume. We see that the LC effect affects all BSDs, pushing them to larger sizes than found in the CCs. The smoothing affects the LC data in a similar way as it does for coeval data. The largest difference is seen for the FOF distribution at early times ($z = 7.3$), where the population of larger bubbles that appeared in the coeval data after smoothing is absent in the LC data and the percolation cluster is again apparent. It should however be noted that conclusions for large regions are sensitive to sample variance effects as they are based on only one or two regions.

Datta et al. (2012) showed that the BSD determined from LC data can be approximated by the one from the coeval cube at the central redshift of the LC data. They used SPA for their analysis and only considered one redshift value. In Fig. 8, we compare the MFP–BSD for LC data with the distributions from coeval data corresponding to the highest, central and lowest redshift contained in the LC. The left-hand panels show the MFP curves for the signal at the resolution of the simulation and the right-hand panels the same for SKA1-Low resolution. We see that the MFP–BSD is bracketed between those for the higher and lower redshift CCs, also for the smoothed case. The plot also indicates that the BSD for the central redshift is not a good representation of the one from the LC data. The LC data reveal the presence of larger bubble sizes and its BSD appears to fall in between those from the central and lowest redshifts. The SPA–BSDs (not shown) exhibit a similar behaviour.

Fig. 9 shows the same analysis for the FOF–BSD. These results present a mixed message. On the one hand, the size of the percolation cluster for the LC data is larger than at the central redshift. On the other hand, the distribution of smaller regions in LC case appears to fall in between that seen in the central and highest redshifts, although it is much closer to that of the central redshift.

As studied in more detail in Datta et al. (2014), the impact of the LC effect depends on the width of the LC data set. If the evolution
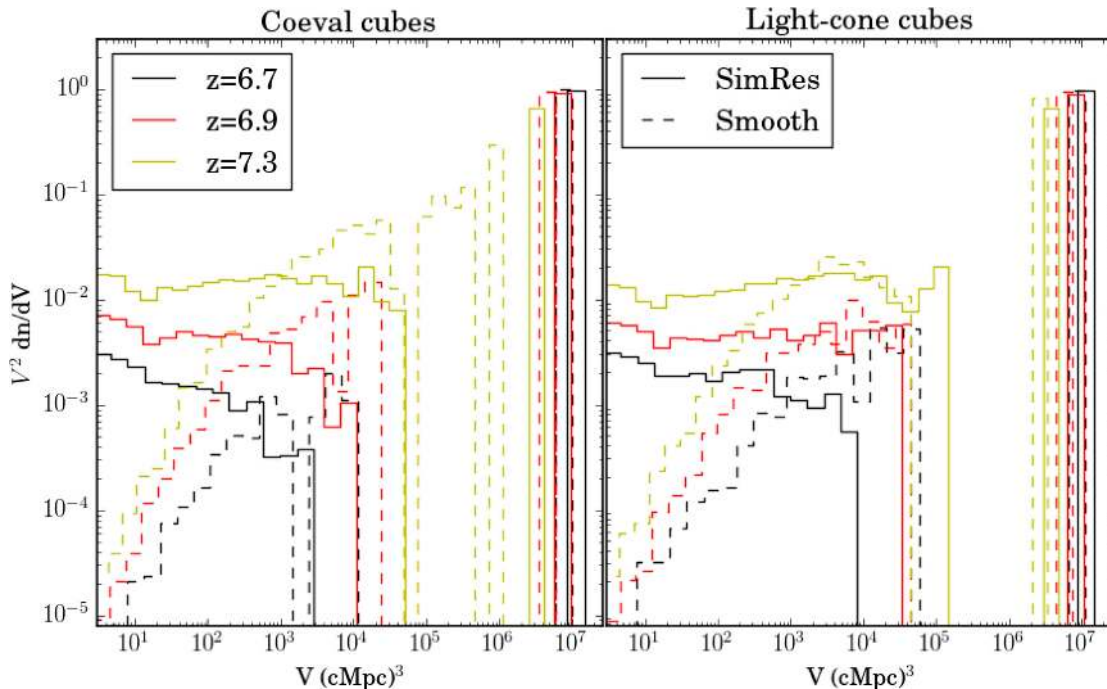
Figure 7. FOF curves: the left-hand panel shows the volume distribution $V^2 \mathrm{d}P/\mathrm{d}V$ of bubbles in CCs at different redshifts at the resolution of the simulation (solid) and at SKA-Low resolution (dashed). The right-hand panel displays the volume distribution of the LC with the indicated central redshift at the resolution of the simulation (dashed) and at SKA1-Low resolution (solid). The BSDs for $z = 7.3$, 6.9 and 6.7 are shown as the curves from left to right, respectively. The corresponding ionization fraction is given in table 2.

of the signal over the extent of the LC is weak or linear, statistical measurements will be similar to those at its central redshift. However, if there is substantial evolution, this will no longer be the case. Datta et al. (2014) recommended that LC data should not have an extension larger than $\Delta z \approx 0.50$ (which during reionization corresponds to a frequency width of $\sim$10 MHz). The LC data presented above have a width $\Delta z \approx 0.80$.

To explore the effect of the width of the LC, we show BSDs for different LC widths in Fig. 10. To keep the data sets cubic in the sense that they have the same comoving size in all directions, we select smaller cubes from the large LC cube. We tested that the reduced volumes affect the BSDs in a marginal way. These results confirm the conclusion from Datta et al. (2014) that the LC effect becomes a minor nuisance for widths $\Delta z \lesssim 0.50$.

5.4 Effect of RSD

Early studies have shown that RSDs have appreciable impact on the 21-cm power spectrum (Bharadwaj, Nath & Sethi 2001; Bharadwaj & Ali 2004). The matter on average moves towards the high-density regions, therefore in redshift space RSDs tend to compress high-density regions and to expand low-density regions. This is known as the Kaiser effect (Kaiser 1987). Clearly, the sizes of the bubbles observed using the redshifted 21-cm signal could also be affected by the gas peculiar velocities. As the ionized regions are typically associated with the high-density, source rich regions, we expect that RSDs will decrease the observed bubble sizes. As shown by Jensen et al. (2013), the effect of RSDs is largest during early reionization and becomes progressively weaker as reionization progresses. Close inspection of Fig. 2 indeed reveals small but observable differences in the shapes of ionized regions between the non-distorted and distorted cases, even at SKA1-Low resolution.

To study the effect of RSDs on the BSDs, we first consider the volume of the largest connected region in the cube as found by the FOF method at different ionization fractions x_v. Fig. 11 shows the ratio of this volume from an LC cube (width 349 Mpc) without RSD to one with RSD. We consider both the simulation resolution (solid curve) and SKA1-Low resolution (dashed curve). We see that this largest region is larger without RSD and that the size ratio approaches unity as reionization progresses. This result confirms our expectation that the RSD effect decreases the measured bubble sizes. However, the results also show that the magnitude of this effect is at most 20 per cent and for most of reionization even lower.

In Fig. 12, we show the effect of RSDs on the MFP–BSD at $z = 7.8$. The redshift choice is owing to the previous inference that RSDs have a larger effect earlier during reionization. We see a shift to the smaller R in the BSDs for the RSD case. This again supports the idea that RSDs decreases the observed sizes. However, the two BSDs at SKA1-Low resolution (dashed lines) are almost identical, which indicates that even though RSDs have an effect on the sizes, this may not be detectable in low-resolution data.

Both of these results show that RSDs do not have a major impact on the size distribution of the ionized regions from the observed low-resolution data. This can be understood from the realization that RSD affect the sizes of the ionized region in only one direction (along the frequency direction). Since the MFP, SPA and FOF methods all consider 3D structures, the small change in size along one dimension caused by the RSDs is mostly averaged away.

5.5 Comparing different source models

One of the most important reasons to study the 21-cm signal from the EoR is to understand the nature of the sources of reionization. Hence, the variation in the BSDs for different source models and
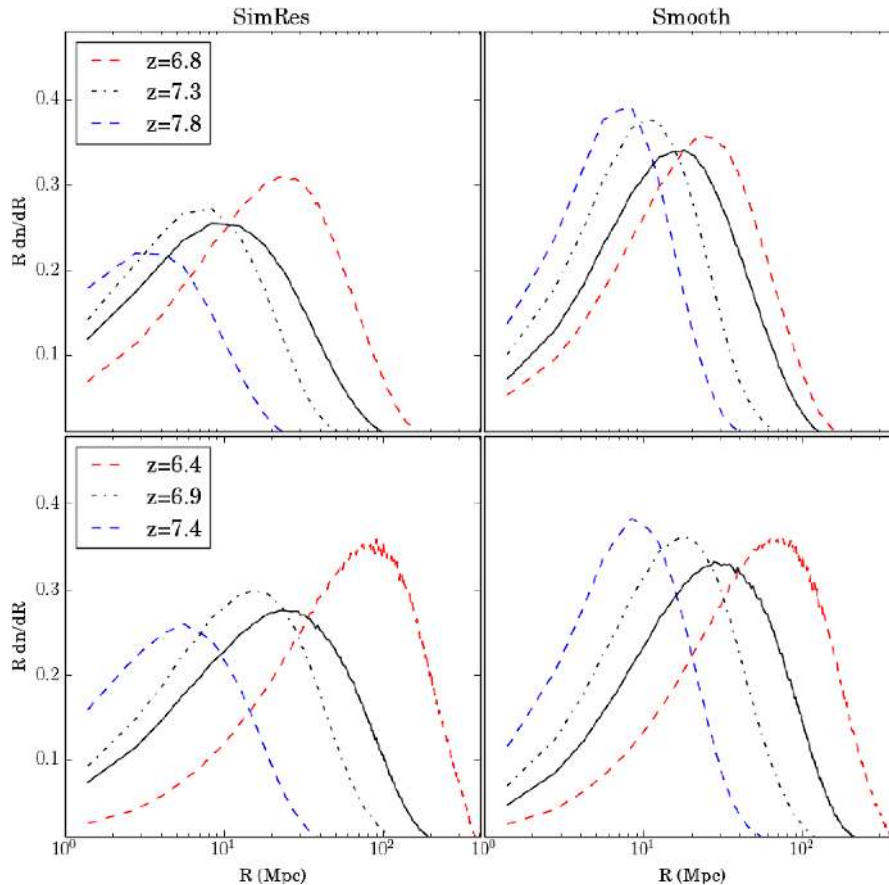
Figure 8. The LC effect in MFP measurements. The BSDs from the two LC data sets (black solid) are compared to the coeval BSDs of the central (black dot–dashed), lowest (blue dashed) and highest (red dashed) redshifts in the LC. The left-hand panels show the results at the resolution of the simulation and the right-hand panels at SKA1-Low resolution.

whether BSDs can be used to differentiate them are relevant to study. Although a full exploration is beyond the scope of this paper, we compare here BSDs for three different source models taken from Dixon et al. (2016), namely simulations LB1 (only massive sources), LB3 (partial suppression of low-mass sources) and LB4 (gradual, mass-dependent suppression of low-mass sources).

In Fig. 13, we show the MFP–BSDs for these three source models at different epochs ($x_v = 0.3, 0.5, 0.7$) at the resolution of the simulation (upper panels) and at the SKA1-Low resolution (bottom). We see that initially case LB1 is quite different from LB3 and LB4, since it does not have low-mass sources, resulting in later reionization with larger ionized bubbles. However, as reionization progresses the MFP–BSDs of LB1 and LB3 become more similar since low-mass sources become partially suppressed over time, resulting in late-time reionization being dominated by the same high-mass sources in both cases (although the timing of reionization is quite different in the two cases). In the LB3 case, the suppression of low-mass sources is mass-dependent, so the lowest mass ones are most suppressed, while larger ones remain less affected. This yields a different BSD, shifted towards somewhat smaller scales at any given stage of the reionization history.

At SKA1-Low resolution, the MFP–BSDs for cases LB3 and LB4 are initially more different than in the unsmoothed case, due to the difference in their suppression mechanisms and the different timing of reionization. Otherwise, we see the same behaviour, except that, as already noted above, the evolution of the curves spans a smaller

range in R values. Given that the horizontal axes in these panels are logarithmic, these curves should be clearly distinguishable when observed at high enough signal to noise to identify the ionized regions.

We performed a similar analysis as in Fig. 13, but now based on the FOF–BSDs (figure not shown). The largest differences in the results are in the volume of the largest connected region and in how large the largest of the population of smaller regions are. However, the differences appear to be small and the FOF–BSDs do not appear to be a good tool to distinguish different source models. The most likely cause is that the form of FOF–BSD is dominated by the nature of reionization as a percolation process (Furlanetto & Oh 2016), with modest dependence on the details of the different models, although a confirmation of this conclusion requires analysing a larger set of simulations.

Dixon et al. (2016) compared the 21-cm power spectra of these same three models and found that they also differ. This implies that for these three cases the BSDs do not break a degeneracy which could be present in a power spectrum analysis. However, the differences in the power spectra are only in the amplitude of the curve when we consider the observable range of k values ($k \lesssim 0.5 \, \mathrm{Mpc^{-1}}$). The shapes and slopes of the curves are quite similar and the power spectra do not show clear features at a specific scale. This makes the power spectra sensitive to calibration errors in the absolute flux scale or to foreground of instrumental residuals which add power to the signal (Cathryn Trott, private communication).
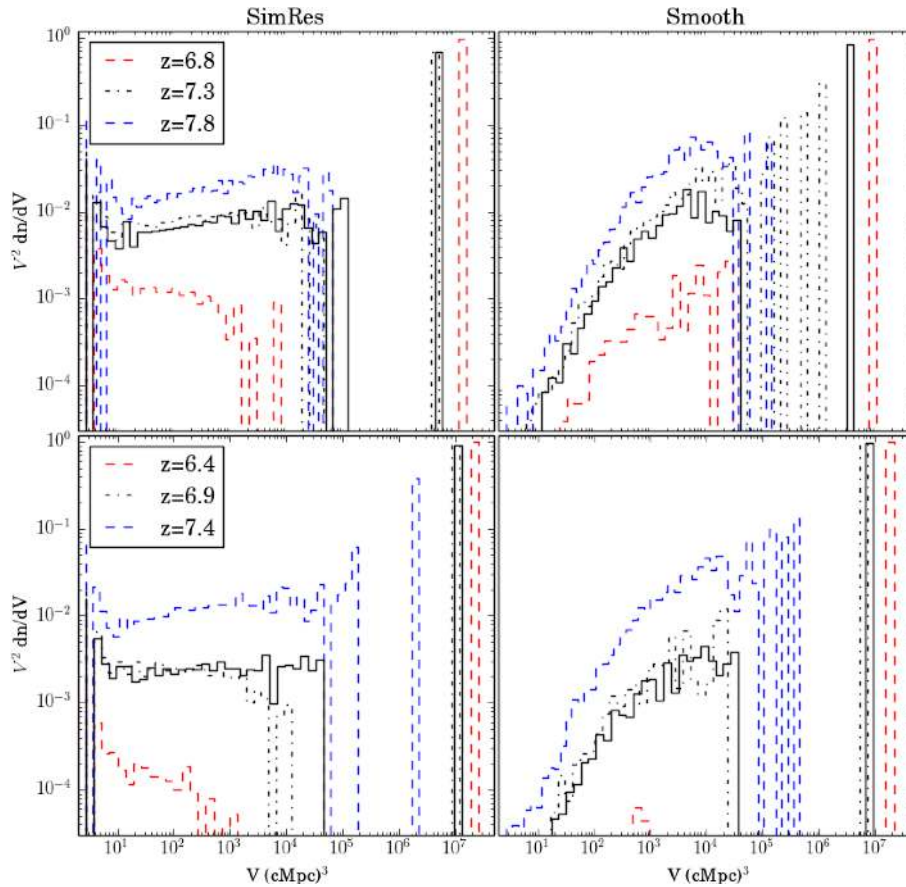
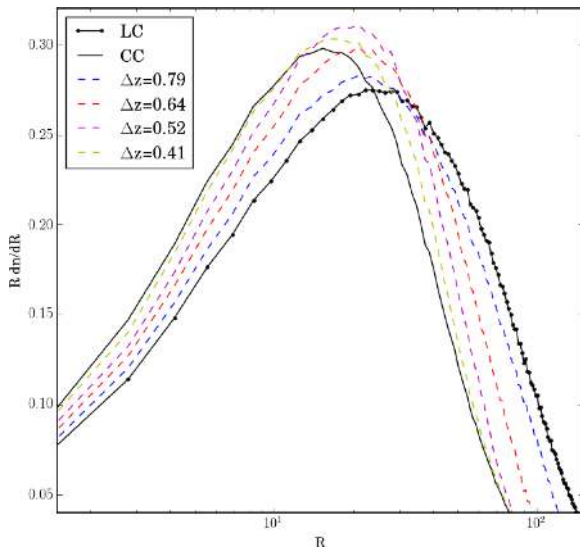**Figure 9.** As Fig. 8 but for FOF–BSDs.

Figure 10. The dependence of the LC effect on the bandwidth used. The black solid curve represents the MFP–BSD for the CC at the central redshift whereas the solid black curve with dots shows the results from the LC from Fig. 8, bottom left-hand panel ($z = 6.9$, $\Delta z = 0.82$). The coloured dashed curves show the size distributions for LCs with the same central redshift but narrower bandwidths. The BSDs from the LC data converge to the CC one as the bandwidth is being reduced.

The BSDs are insensitive to deviations in the flux scale and could therefore be used to reduce the uncertainties while comparing the observations to models.

5.6 Percolation

We mentioned above that the FOF–BSDs do not distinguish clearly between the three source models. However, there is an another aspect to the FOF method, which is the emergence of the dominating largest ionized region. The growth curves for this percolation cluster were studied by Furlanetto & Oh (2016). They found that for a range of models the rapid rise in the growth curve happens around $x_v \approx 0.1$ and that this behaviour is expected from percolation theory.

In Fig. 14, we show growth curves of the largest region found by the FOF method for the three source models considered. The fraction of the ionized volume contained in the largest connected region is plotted against the measured ionization fraction. As above, we consider both the resolution of the simulation (solid curves) and SKA1-Low resolution (dashed curves).

These percolation curves show the same behaviour as noted by Furlanetto & Oh (2016); around $\hat{x}_v \approx 0.1$, the largest cluster starts a rapid growth and contains most of the ionized volume before $\hat{x}_v \approx 0.2$ is reached. We see some differences between the curves of the different models with LB3 showing the earliest and most rapid growth. This can be understood from the presence of a large population of low-mass sources in that model, which also shifts the evolution to earlier times.
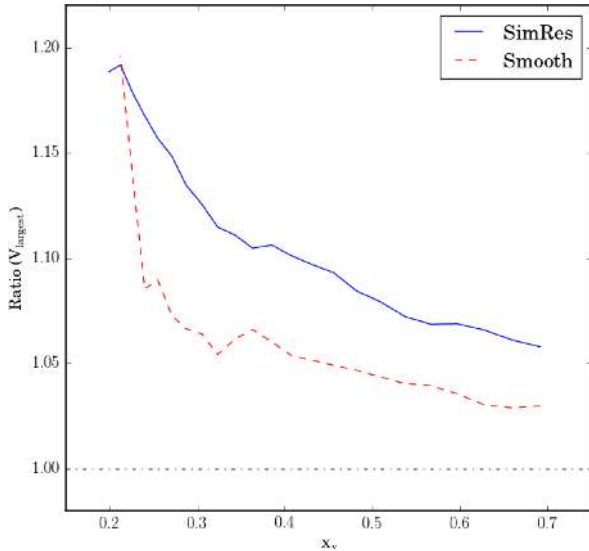
Figure 11. The ratio of the volume of the largest ionized region found in the sub-volume from the LC without RSD to the ones with RSD versus the global ionization fraction (x_v). The largest region is found using the FOF algorithm.
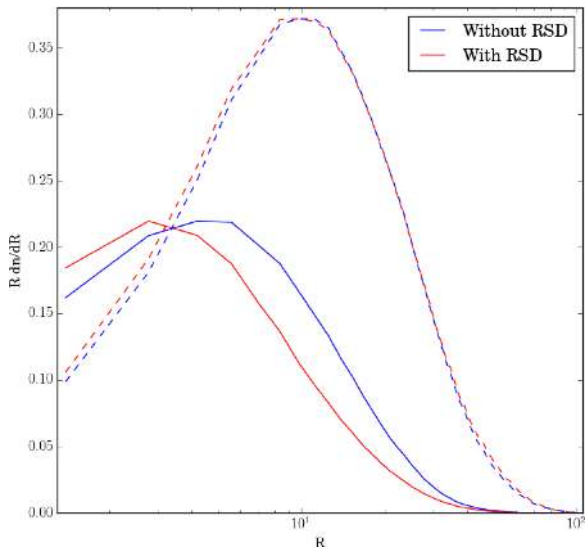

Figure 12. The MFP–BSD of sub-volume from LC with RSD and without RSD are given. The solid curve gives ones at the simulation resolution and the dashed ones are for the lower resolution case. All the plots are for the epochs with $x_v = 0.15$.

For the lower resolution results (dashed curves), the results are more similar between the three models and the rise starts earlier, around $\hat{x}_v \approx 0.06$. It is also much more rapid, reaching a fraction of 80 per cent around $\hat{x}_v \approx 0.1$. The reduced resolution thus leads to increased connectivity and a larger relative size for the percolation cluster at a given observed mean ionized fraction.

In Section 5.2, we saw that smoothing decreases the size of the observed percolation cluster. However, lower resolution causes the $\hat{x}_v$ to decrease as well. As the slope of the curve is greater than unity, the decrease in both numerator and denominator causes the fractional volume to increase. As shown in Section 5.1 and Table 2, the observed mean ionized fraction underestimates the mean ionized

fraction from the simulation, implying that the observed transition takes place around $x_v \approx 0.2$.

The shape of these percolation curves is sensitive to the chosen threshold value for segmenting the data into ionized and neutral clusters. Furlanetto & Oh (2016) used a threshold value for the ionized fraction that made the total fractional volume of ionized regions equal to the mean ionized fraction at a given redshift. Since this quantity is unknown for real observations, we used thresholds values for the 21-cm signal determined by the K-Means method, which means we cannot compare in detail to the results in Furlanetto & Oh (2016). We should also point out that for the low mean ionization fractions at which the transition happens ($\hat{x}_v \lesssim 0.1$), K-Means has difficulty finding good values for the threshold, leading to some noisiness in the results.

6 SUMMARY AND DISCUSSION

In this work, we propose a new approach to analyse tomographic 21-cm data sets from reionization. It consists of two steps: first, the 21-cm data is segmented into a binary field of ionized and neutral regions. After that, one of the existing BSD methods can be applied to this field. For the first step, we introduced a new method known as K-Means clustering. For the second step, we have investigated three different BSD methods – MFP, SPA and FOF, each with its own strengths and weaknesses. In particular, we are interested in how they perform when applied to 21-cm data cubes generated by future radio interferometers such as SKA1-Low.

We considered a number of effects which will be present in the 21-cm data cubes:

(i) Finite resolution (corresponding to maximum baselines of 2 km)
(ii) Absence of zero base lines (causing the average signal in an image to be zero)
(iii) Light-cone effect
(iv) Redshift space distortions

We did not consider the effects of noise and telescope calibration.

The K-Means algorithm can be described as a self-adjusting thresholding technique. Use of such a technique is important in view of the reduced resolution and lack of an absolute zero-point in the interferometric observation. We find K-Means to work well if a sufficient number of ionized resolution elements are present in the data. In terms of the measured volume-averaged ionization fraction of the IGM, we find this criterion to imply $\hat{x}_v \gtrsim 0.1$. For lower values of $\hat{x}_v$, other methods may perform better. However, it is also possible that at these early times it is fundamentally difficult to distinguish between small, partly resolved ionized regions and density fluctuations.

The results from the different BSD methods show some shared trends, while also describing different aspects of the ionization field. Of the three BSD methods we considered, MFP proves to be somewhat more useful than SPA, largely because it is less diffusive. We confirm the result of Friedrich et al. (2011) and Lin et al. (2016) that the bubble sizes from the SPA–BSDs and MFP–BSDs are not directly comparable due to their different biases; the SPA method gives roughly three times smaller bubbles than the MFP method. The FOF results cannot be compared to either the MFP or SPA methods, since unlike those methods it is finding the volumes of topologically connected regions. Based on our limited exploration, FOF appears to be mostly useful to confirm the nature of reionization as a percolation process, but does not clearly distinguish between models with different properties for the sources of reionization.
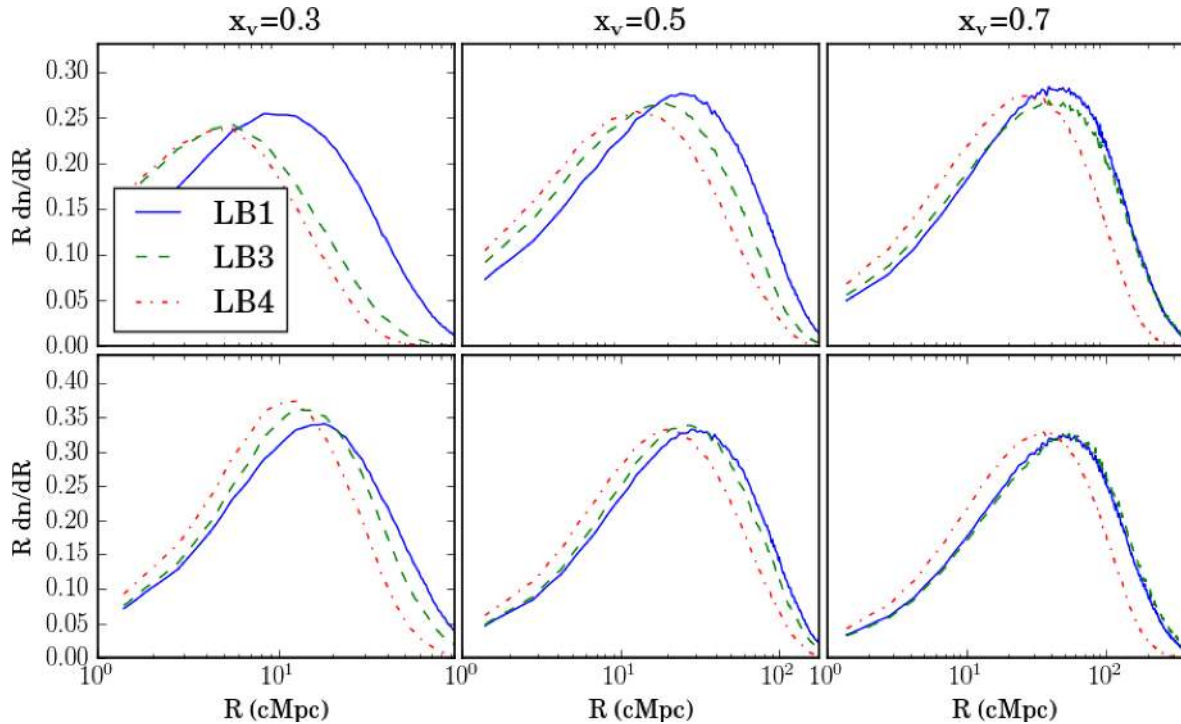
Figure 13. MFP–BSDs for the three simulations considered. We compare them at epochs corresponding to $x_v = 0.3, 0.5, 0.7$. The upper panel shows the comparison at the resolution of the simulation whereas the lower panel shows the study of smoothed signal.
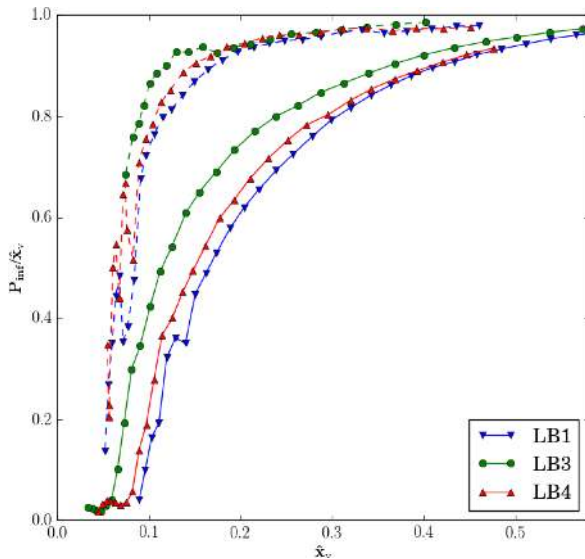

Figure 14. The relative volume of the largest cluster identified by the FOF method plotted against the measured global ionization fraction $\hat{x}_v$. The solid lines show the results at the resolution of the simulation and the dashed curves at SKA1-Low resolution. The three colours show the results for the three different simulations.

When the resolution is decreased, the BSD curves from the SPA and MFP methods at early times show a shift to larger sizes. This shift diminishes with the progress of reionization and at the later stages of reionization it becomes marginal. Consequently, the MFP- and SPA–BSD curves at the typical resolution of SKA1-Low show less evolution than the intrinsic ones at simulation resolution. In the presence of errors and sample variance, the derivation of the

reionization history solely from these BSDs may therefore be difficult. However, taking into account the global ionized fraction measured from the binary data set may be able to alleviate this difficulty.

The FOF–BSDs show that the largest (or percolation) cluster is always smaller for the lower resolution data, which may explain why the shift seen in the MFP- and SPA–BSDs decreases as reionization progresses. Late in reionization the size measurements by the MFP and SPA algorithms will mostly take place inside the large percolation cluster. In fact, both the SPA and MFP results show a hint of having a lower fraction of regions at the largest sizes.

Real 21-cm data sets will be affected by both the LC effect and RSDs. We found that the impact of the LC effect is only significant if the data extend for more than ∼10 MHz in frequency or about 0.5 in redshift. The RSDs typically change the BSDs by at most 10 per cent at the simulation resolution and much less at SKA1-Low resolution and can therefore be safely ignored when measuring and interpreting BSDs.

Due to the strongly non-Gaussian PDF of the 21-cm signal, a power spectrum analysis in principle may suffer from degeneracy as it does not provide a complete statistical description of the results. BSDs obtained from tomographic imaging data should be able to break these degeneracies. However, for the three models studied in this paper, both the power spectra and BSDs are different and therefore it remains to be shown that BSDs can distinguish scenarios which show very similar power spectra. Still, even if such scenarios never occur in reality, the measured BSDs will be affected differently by measurement and calibration errors and will therefore improve the reliability of the astrophysical and cosmological parameters derived from the 21-cm data.

In this study, we have assumed that the noise level in the tomographic data is low enough not to affect the segmentation and the BSD measurements. However, as for example shown in

Kakiichi et al. (2017) this assumption is optimistic since root-mean-square noise levels as low as ∼2 mK can impact the results. A full evaluation of the impact of noise is beyond the scope of the current paper. However, some general considerations can be made. The presence of noise in the signal would first of all affect the segmentation step. If the segmentation procedure labels noisy pixels in the wrong way, erroneous neutral spots might for example appear inside the ionized regions, or vice versa. The SPA method can be expected to be less affected by such erroneous spots, as it determines the size of an ionized region by its average volume filling factor. However, the appearance of erroneous ionized spots may boost the number of small bubbles found. The MFP method may be more sensitive to the presence of erroneous neutral spots, as hitting such a spot with a ray will always reduce the length of the ray compared to what it should be. On the other hand, the MFP method can be expected to be less sensitive to the appearance of erroneous ionized spots, as the random selection process makes the selection of these points as starting points for rays very unlikely. The FOF method would determine volumes that are reduced by the number of neutral spots and also show an increase in the number of small ionized volumes due to the erroneous ionized spots. Kakiichi et al. (2017) showed that for the granulometric method noise introduces a 'splitting bias' meaning that it shifts the BSDs to smaller sizes by splitting connected regions into separate ones. The same effect can be expected for the FOF method.

The root-mean-square noise level depends not only on the integration time, but also on the resolution chosen. The analysis of the tomographic 21-cm data will require a careful balance between a low enough resolution to achieve acceptable noise levels and a high enough resolution to extract useful BSDs. Although we have not presented a detailed resolution study, the lower panel of Fig. 13 indicates that reducing the resolution reduces the differences between the BSDs at different phases of reionization and the differences between different models (see also Kakiichi et al. 2017). If the telescope data require too low resolution to obtain meaningful BSDs, other statistical measures for the tomographic data should be considered.

As explained in Section 2, we assumed the high spin temperature limit when constructing the 21-cm signal. In this case, the lowest value for the signal is zero which corresponds directly to the ionized regions. This allows us to identify ionized regions from the 21-cm signal. Although this limit is generally thought to be valid during most of the EoR, it is possible that spin temperature fluctuations exist during reionization. If regions exist with $T_S < T_{CMB}$, the lowest value for the 21-cm signal will be less than zero. It immediately follows that in this case it will be hard or even impossible to identify ionized regions, especially without a calibration of the absolute value of the 21-cm signal. Even if we would know which regions have $\delta T_b \approx 0$, we would still not be able to tell whether they correspond to regions with $x_{HI} \approx 0$ or $T_S \approx T_{CMB}$. Furthermore, the 21-cm PDFs from models with spin temperature fluctuations have smooth shapes which means it will be hard to define physically motivated threshold values for any type of size analysis.

However, it has also been shown that during the period of spin temperature fluctuations the signal also is significantly non-Gaussian (Watkinson & Pritchard 2015; Ross et al. 2017), so it should be worthwhile to explore tomographic techniques for this regime. However, at this time it is difficult to say whether BSDs, with some other definition of what constitutes a bubble than we have used, are a useful tool in this context or whether other techniques are preferable. This analysis of tomographic data with spin temperature fluctuations needs to be considered carefully.

In this paper, we considered the three classical methods developed to characterize BSDs in simulation data and applied them to mock 21-cm observations. Recently, two alternative methods for deriving BSDs have been proposed. The first is the watershed method, which was proposed for reionization studies by Lin et al. (2016), who used it on simulated x_{HII} data and compared to the SPA, MFP and FOF methods. The method has a marker based algorithm (Barnes, Lehman & Mulla 2014). The markers are the points from which 'flooding' starts until the watershed lines are reached. Lin et al. (2016) use the local minima in the field of distance to the nearest neutral resolution element to determine the markers, which leads to an oversegmentation. The oversegmentation can be controlled by carefully choosing a smoothing parameter. In the comparison, watershed performs well although the authors did 'tune' the method to optimize its performance. Its application to 21-cm tomographic data merits further exploration. However, given the results of Lin et al. (2016), we expect the method to show a similar behaviour as we found for the MFP method.

The other new method is granulometry, described in detail in Kakiichi et al. (2017). These authors not only introduced the method, but also considered many of the issues related to applying this method to finite resolution and noisy 21-cm data. In our paper, we have referred in several places to those results. The method shows good promise, but a comparison to the methods used in this paper as well as the watershed method would be useful.

Both the watershed and granulometry method require a segmentation of the data into neutral and ionized elements. Kakiichi et al. (2017) chose a very simple approach, namely labelling all regions lower than the mean 21-cm signal as ionized. As explained in more detail in that paper, this choice only properly identifies ionized regions during part of reionization and may erroneously label low-density regions as ionized. The granulometry method would clearly benefit from a more robust segmentation method, for example the one used in the current paper.

We considered BSDs of ionized regions. As reionization approaches completion, the concept of discrete ionized regions becomes less and less applicable. We therefore did not consider BSDs beyond global ionization fractions of 0.7. Beyond that a study of the BSDs of neutral regions will make more sense. We did not address this, but the methodology would be completely equivalent and we postpone an investigation of this to a future paper.

BSDs, irrespective of what method or component is chosen, are of course not the only metric that can be applied to tomographic data. Other possible metrics are the Minkowski functionals that describe the global topological characteristics of ionized or neutral regions (see e.g. Friedrich et al. 2011), metrics which describe the shapes of ionized/neutral regions or metrics which depend on the (relative) positions of ionized/neutral regions. Some of these metrics may be relatively insensitive to the parameters of reionization, but others may be able to break degeneracies in an analysis based solely on power spectra. Metrics for shapes and positions have been previously applied to voids in the galaxy surveys, see e.g. Foster & Nelson (2009).

This paper presents the first exploration of the application of BSDs on 21-cm tomographic data. As discussed above, several directions for future investigations present themselves. The most important one is the presence of noise, which will mostly complicate the segmentation of the data into ionized and neutral regions. The choice of segmentation method may actually be important to minimize the impact of noise. We will address these issues in a follow up paper. A comparison between the MFP, granulometry and watershed BSDs would be another useful step in finding an optimal

size distribution tool for 21-cm tomography. However, we expect the qualitative conclusions from our current work to hold also for these other BSD methods.

ACKNOWLEDGEMENTS

REFERENCES

Barnes R., Lehman C., Mulla D., 2014, Comput. Geosci., 62, 117
Bharadwaj S., Ali S. S., 2004, MNRAS, 352, 142
Bharadwaj S., Nath B. B., Sethi S. K., 2001, J. Astrophys. Astron., 22, 21
Bottou L., Bengio Y., 1995, in Advances in Neural Information Processing Systems 7. MIT Press, pp. 585–592
Datta K. K., Mellema G., Mao Y., Iliev I. T., Shapiro P. R., Ahn K., 2012, MNRAS, 424, 1877
Datta K. K., Jensen H., Majumdar S., Mellema G., Iliev I. T., Mao Y., Shapiro P. R., Ahn K., 2014, MNRAS, 442, 1491
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, 371
Dillon J. S., Liu A., Tegmark M., 2013, Phys. Rev. D, 87, 043005
Dixon K. L., Iliev I. T., Mellema G., Ahn K., Shapiro P. R., 2016, MNRAS, 456, 3011
Foster C., Nelson L. A., 2009, ApJ, 699, 1252
Friedrich M. M., Mellema G., Alvarez M. A., Shapiro P. R., Iliev I. T., 2011, MNRAS, 413, 1353
Furlanetto S. R., Oh S. P., 2016, MNRAS, 457, 1813
Furlanetto S. R., Zaldarriaga M., Hernquist L., 2004a, ApJ, 613, 1
Furlanetto S. R., Zaldarriaga M., Hernquist L., 2004b, ApJ, 613, 16
Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Phys. Rep., 433, 181
Greig B., Mesinger A., 2015, MNRAS, 449, 4246
Harker G. et al., 2010, MNRAS, 405, 2492
Harnois-Déraps J., Pen U.-L., Iliev I. T., Merz H., Emberson J. D., Desjacques V., 2013, MNRAS, 436, 540
Hartigan J. A., Wong M. A., 1979, J. R. Stat. Soc. C, 28, 100
Hoshen J., Kopelman R., 1976, Phys. Rev. B, 14, 3438
Ichikawa K., Barkana R., Iliev I. T., Mellema G., Shapiro P. R., 2010, MNRAS, 406, 2521
Iliev I. T., Mellema G., Pen U.-L., Merz H., Shapiro P. R., Alvarez M. A., 2006a, MNRAS, 369, 1625
Iliev I. T. et al., 2006b, MNRAS, 371, 1057
Iliev I. T., Shapiro P. R., McDonald P., Mellema G., Pen U.-L., 2008, MNRAS, 391, 63
Iliev I. T., Mellema G., Shapiro P. R., Pen U.-L., Mao Y., Koda J., Ahn K., 2012, MNRAS, 423, 2222
Ivezić Ž., Connolly A. J., VanderPlas J. T., Gray A., 2014, Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data. Princeton Univ. Press, Princeton, NJ
Jacobs D. C. et al., 2015, ApJ, 801, 51
Jensen H. et al., 2013, MNRAS, 435, 460
Jensen H., Majumdar S., Mellema G., Lidz A., Iliev I. T., Dixon K. L., 2016, MNRAS, 456, 66
Kaiser N., 1987, MNRAS, 227, 1
Kakiichi K. et al., 2017, MNRAS, 471, 1936
Kanungo T., Mount D. M., Netanyahu N. S., Piatko C. D., Silverman R., Wu A. Y., 2002, IEEE Trans. Pattern Anal. Mach. Intell., 24, 881
Komatsu E. et al., 2011, ApJS, 192, 18
Lidz A., Zahn O., McQuinn M., Zaldarriaga M., Hernquist L., 2008, ApJ, 680, 962
Lin Y., Oh S. P., Furlanetto S. R., Sutter P. M., 2016, MNRAS, 461, 3361
Liu A., Tegmark M., 2011, Phys. Rev. D, 83, 103006
Liu D., Yu J., 2009, in Hybrid Intelligent Systems, 2009. HIS'09. Ninth International Conference on. p. 344
Lonsdale C. J. et al., 2009, Proc. IEEE, 97, 1497
Madau P., Meiksin A., Rees M. J., 1997, ApJ, 475, 429
Mao Y., Shapiro P. R., Mellema G., Iliev I. T., Koda J., Ahn K., 2012, MNRAS, 422, 926
McQuinn M., Lidz A., Zahn O., Dutta S., Hernquist L., Zaldarriaga M., 2007, MNRAS, 377, 1043
Mellema G., Iliev I. T., Alvarez M. A., Shapiro P. R., 2006a, New A, 11, 374
Mellema G., Iliev I. T., Pen U.-L., Shapiro P. R., 2006b, MNRAS, 372, 679
Mellema G. et al., 2013, Exp. Astron., 36, 235
Mesinger A., Furlanetto S., 2007, ApJ, 669, 663
Mesinger A., Furlanetto S., Cen R., 2011, MNRAS, 411, 955
Nadathur S., Hotchkiss S., 2015, MNRAS, 454, 2228
Otsu N., 1979, IEEE Trans. Syst. Man Cybern., 9, 62
Paciga G. et al., 2011, MNRAS, 413, 1174
Parsons A. R. et al., 2010, AJ, 139, 1468
Patil A. H. et al., 2017, ApJ, 838, 65
Planck Collaboration et al., 2016a, A&A, 594, A13
Planck Collaboration et al., 2016b, A&A, 596, A107
Planck Collaboration et al., 2016c, A&A, 596, A108
Press W. H., Davis M., 1982, ApJ, 259, 449
Rohlfs K., Wilson T., 2013, Tools of Radio Astronomy. Springer Science & Business Media
Ross H. E., Dixon K. L., Iliev I. T., Mellema G., 2017, MNRAS, 468, 3785
Sánchez Almeida J., Allende Prieto C., 2013, ApJ, 763, 50
Schaap W. E., van de Weygaert R., 2000, A&A, 363, L29
Shimabukuro H., Yoshiura S., Takahashi K., Yokoyama S., Ichiki K., 2016, MNRAS, 458, 3003
Trott C. M. et al., 2016, ApJ, 818, 139
Watkinson C. A., Pritchard J. R., 2014, MNRAS, 443, 3090
Watkinson C. A., Pritchard J. R., 2015, MNRAS, 454, 1416
Watkinson C. A., Majumdar S., Pritchard J. R., Mondal R., 2017, MNRAS, 472, 2436
Zahn O., Lidz A., McQuinn M., Dutta S., Hernquist L., Zaldarriaga M., Furlanetto S. R., 2007, ApJ, 654, 12

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