

The brightness and spatial distributions of terrestrial radio sources

Journal:	<i>Monthly Notices of the Royal Astronomical Society</i>
Manuscript ID:	MN-13-0509-MJ
Wiley - Manuscript type:	Main Journal
Date Submitted by the Author:	01-Mar-2013
Complete List of Authors:	<p>Offringa, André; ARC Centre of Excellence for All-sky Astrophysics (CAASTRO), de Bruyn, Ger; ASTRON, Astronomy Group Zaroubi, Saleem; University of Groningen, Kapteyn Astronomical Institute Koopmans, Leon; Kapteyn Astronomical Institute, University of Groningen Wijnholds, Stefan; ASTRON, Astronomy Group Abdalla, Filipe; UCL, Physics and Astronomy Brouw, Wim; Kapteyn Astronomical Institute, University of Groningen Ciardi, Benedetta; Max Planck Institute for Astrophysics, Iliev, Ilian; University of Sussex, Physics and Astronomy; Harker, Geraint; University of Groningen, Kapteyn Institute; University of Durham, Department of Physics Mellema, Garrelt; Stockholm University, Stockholm Observatory Bernardi, Gianni; Harvard-Smithsonian Center for Astrophysics, Zarka, Philippe; Observatoire de Paris - CNRS, LESIA Ghosh, Abhik; University of Groningen, Kapteyn Astronomical Institute Alexov, A. Anderson, J.; Max-Planck-Institut für Radioastronomie, Asgekar, Ashish; Netherlands Institute for Radio Astronomy (ASTRON), Radio Observatory Avruch, I.; University of Groningen, Kapteyn Astronomical Institute Beck, Rainer; Max-Planck-Institut fuer Radioastronomie, Bell, Martin; ARC Centre of Excellence for All-sky Astrophysics (CAASTRO), Bentum, M.; ASTRON, Astronomy Group</p>
Keywords:	atmospheric effects < Astronomical instrumentation, methods, and techniques, instrumentation: interferometers < Astronomical instrumentation, methods, and techniques, methods: observational < Astronomical instrumentation, methods, and techniques, techniques: interferometric < Astronomical instrumentation, methods, and techniques, radio continuum: general < Sources as a function of wavelength, (cosmology:) early Universe < Cosmology

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



The brightness and spatial distributions of terrestrial radio sources

A. R. Offringa^{1,2,3,*}, A. G. de Bruyn^{4,3}, S. Zaroubi³, L. V. E. Koopmans³,
 S. J. Wijnholds⁴, F. B. Abdalla⁵, W. N. Brouw^{3,4}, B. Ciardi⁶, I. T. Iliev⁷,
 G. J. A. Harker⁸, G. Mellema⁹, G. Bernardi¹⁰, P. Zarka¹¹, A. Ghosh³,
 A. Alexov¹², J. Anderson¹³, A. Asgekar⁴, I. M. Avruch^{14,3}, R. Beck¹³,
 M. E. Bell^{2,15}, M. R. Bell⁶, M. J. Bentum⁴, P. Best¹⁶, L. Bîrzan¹⁷, F. Breitling¹⁸,
 J. Broderick¹⁹, M. Brüggen²⁰, H. R. Butcher^{4,1}, F. de Gasperin²⁰, E. de Geus⁴,
 M. de Vos⁴, S. Duscha⁴, J. Eislöffel²¹, R. A. Fallows⁴, C. Ferrari²²,
 W. Frieswijk⁴, M. A. Garrett^{4,17}, J. Griebmeier²³, T. E. Hassall¹⁹, A. Horneffer¹³,
 M. Iacobelli¹⁷, E. Juette²⁴, A. Karastergiou²⁵, W. Klijn⁴, V. I. Kondratiev^{4,26},
 M. Kuniyoshi¹³, G. Kuper⁴, J. van Leeuwen^{4,27}, M. Loose⁴, P. Maat⁴,
 G. Macario²², G. Mann¹⁸, J. P. McKean⁴, H. Meulman⁴, M. J. Norden⁴,
 E. Orru⁴, H. Paas²⁸, M. Pandey-Pommier²⁹, R. Pizzo⁴, A. G. Polatidis⁴,
 D. Rafferty¹⁷, W. Reich¹³, R. van Nieuwpoort⁴, H. Röttgering¹⁷,
 A. M. M. Scaife¹⁹, J. Sluman⁴, O. Smirnov^{30,31}, C. Sobey¹³, M. Tagger²³,
 Y. Tang⁴, C. Tasse¹¹, S. ter Veen³², C. Toribio⁴, R. Vermeulen⁴, C. Vocks¹⁸,
 R. J. van Weeren¹⁰, M. W. Wise^{4,27}, O. Wucknitz^{33,13}

¹RSAA, Australian National University, Mt Stromlo Observatory, via Cotter Road, Weston, ACT 2611, Australia

²ARC Centre of Excellence for All-sky Astrophysics (CAASTRO)

³Kapteyn Astronomical Institute, PO Box 800, 9700 AV Groningen, The Netherlands

⁴Netherlands Institute for Radio Astronomy (ASTRON), Postbus 2, 7990 AA Dwingeloo, The Netherlands

⁵UCL Department of Physics and Astronomy, London WC1E 6BT, United Kingdom

⁶Max Planck Institute for Astrophysics, Karl Schwarzschild Str. 1, 85741 Garching, Germany

⁷University of Sussex, Falmer, Brighton BN1 9QH, UK

⁸Center for Astrophysics and Space Astronomy, University of Colorado Boulder, CO 80309, USA

⁹Stockholm University, AlbaNova University Center, Stockholm Observatory, SE-106 91 Stockholm, Sweden

¹⁰Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA

¹¹LESIA, UMR CNRS 8109, Observatoire de Paris, 92195 Meudon, France

¹²Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA

¹³Max-Planck-Institut für Radioastronomie, Auf dem Hügel 69, 53121 Bonn, Germany

¹⁴SRON Netherlands Institute for Space Research, Sorbonnelaan 2, 3584 CA, Utrecht, The Netherlands

¹⁵Sydney Institute for Astronomy, School of Physics, The University of Sydney, NSW 2006, Australia

¹⁶Institute for Astronomy, University of Edinburgh, Royal Observatory of Edinburgh, Blackford Hill, Edinburgh EH9 3HJ, UK

¹⁷Leiden Observatory, Leiden University, PO Box 9513, 2300 RA Leiden, The Netherlands

¹⁸Leibniz-Institut für Astrophysik Potsdam (AIP), An der Sternwarte 16, 14482 Potsdam, Germany

¹⁹School of Physics and Astronomy, University of Southampton, Southampton, SO17 1BJ, UK

²⁰University of Hamburg, Gojenbergsweg 112, 21029 Hamburg, Germany

²¹Thüringer Landessternwarte, Sternwarte 5, D-07778 Tautenburg, Germany

²²Laboratoire Lagrange, UMR7293, Université de Nice Sophia-Antipolis, CNRS, Observatoire de la Côte d'Azur, 06300 Nice, France

²³Laboratoire de Physique et Chimie de l'Environnement et de l'Espace, LPC2E UMR 7328 CNRS, 45071 Orléans Cedex 02, France

²⁴Astronomisches Institut der Ruhr-Universität Bochum, Universitätsstrasse 150, 44780 Bochum, Germany

²⁵Astrophysics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH

²⁶Astro Space Center of the Lebedev Physical Institute, Profsoyuznaya str. 84/32, Moscow 117997, Russia

²⁷Astronomical Institute 'Anton Pannekoek', University of Amsterdam, Postbus 94249, 1090 GE Amsterdam, The Netherlands

²⁸Center for Information Technology (CIT), University of Groningen, The Netherlands

²⁹Centre de Recherche Astrophysique de Lyon, Observatoire de Lyon, 9 av Charles André, 69561 Saint Genis Laval Cedex, France

³⁰Centre for Radio Astronomy Techniques & Technologies (RATT), Department of Physics and Electronics, Rhodes University, PO Box 94, Grahamstown 6140, South Africa

³¹SKA South Africa, 3rd Floor, The Park, Park Road, Pinelands, 7405, South Africa

³²Department of Astrophysics/IMAPP, Radboud University Nijmegen, P.O. Box 9010, 6500 GL Nijmegen, The Netherlands

³³Argelander-Institut für Astronomie, University of Bonn, Auf dem Hügel 71, 53121, Bonn, Germany

ABSTRACT

Faint undetected sources of radio-frequency interference (RFI) might become visible after long time integration of radio observations if they are stationary and consistently present over time. Thereby, they might obstruct the detection of the weak astronomical signals of interest. This issue is especially important for Epoch of Reionisation (EoR) projects that try to detect the faint redshifted HI signals from the time of the earliest structures in the Universe. We explore the low-frequency RFI situation by studying brightness histograms of visibility data observed with LOFAR, similar to radio-source-count analyses that are used in cosmology. The histograms show an RFI distribution that follows a power-law distribution with an estimated exponent around -1.5 . With several assumptions, this can be explained with a uniform distribution of terrestrial radio sources whose radiation follows existing propagation models. Extrapolation of the power law would imply that the current LOFAR EoR observations should be severely RFI limited if these sources add up consistently. This is in contrast with actual observations, which almost reach the thermal noise and are thought not to be limited by RFI. Therefore, we conclude that it is likely that most RFI sources do not linearly increase the noise. We believe that RFI will not prevent an EoR detection with LOFAR.

Key words: atmospheric effects – instrumentation: interferometers – methods: observational – techniques: interferometric – radio continuum: general – dark ages, reionisation, first stars

1 INTRODUCTION

Radio astronomy concerns itself with the observation of radiation from celestial sources at radio wavelengths. However, astronomical radio observations can be affected by radio-frequency interference (RFI), which might make it difficult to calibrate the instrument and achieve high sensitivities (Pankonin & Price 1981; Thompson et al. 1991; Lemmon 1997; Fridman & Baan 2001). Many techniques have been designed to mitigate this effect, such as detection and flagging of data (Weber et al. 1997; Leshem et al. 2000; Ryabov et al. 2004; Baan et al. 2004; Niamsuwan et al. 2005; Flöer et al. 2010; Offringa et al. 2010a), adaptive cancellation techniques (Barnbaum & Bradley 1998; Briggs et al. 2000) and spatial filtering (Leshem et al. 2000; Ellingson & Hampson 2002; Smolders & Hampson 2002; Boonstra 2005; Kocz et al. 2012; Offringa et al. 2012b). In most cases, and with the help of careful spectrum allocation management and radio-quiet zones, these methods excise enough of the interference to allow calibration and imaging of the data.

Typical radio observations record a few hours of data, and the results are integrated. In these cases, excising only the interference that is apparent and thus above the noise often suffices, i.e., the observation can still reach the thermal noise limit of the instrument. A new challenge arises, however, when one desires much deeper observations, and hundreds of hours of observations need to be integrated. In such a case, weak interference caused by stationary sources might not manifest itself above the noise in individual observations, but might be coherently present in the data. Subsequently, when averaging these data, the interference might become apparent and occlude the signal of interest. This is very relevant for the 21-cm Epoch of Reionisation (EoR) experiments, because they involve long integration times. Several such experiments are underway, to either measure the angular power spectrum (Paciga et al. 2011; de Bruyn et al. 2011; Jacobs et al. 2011; Williams et al. 2012) or the global signal (Bowman & Rogers 2010). Ground-based Cosmic Microwave Background (CMB) experiments are another class of experiments involving long integration times (e.g.,

Subrahmanyan & Ekers 2002). For these experiments, it is important to know the possible effect of low-level interference on the data, as these might overshadow or alter the signal of interest. In this article, we will connect new insights about RFI to the angular Epoch of Reionisation experiment that is using the Low-Frequency Array (LOFAR) (de Bruyn et al. 2011, van Haarlem et al. 2013, *submitted*).

This work explores the information that is present in interference distributions, in order to analyse possible low-level interference that is not detectable by standard detection methods. Our approach is similar to the radio-source-count analyses that are used in cosmology (Condon 1984), also named $\log N - \log S$ analyses, where N and S refer to the celestial source count and brightness respectively. The slope in such a plot contains information about source populations, their luminosity functions and the geometry of the Universe. We analyse such a double-logarithmic plot for the case of terrestrial sources, with the ultimate goal of estimating their full spatial and brightness distributions. This results in a better insight into the effects of low-level interference and allows one to simulate the effects of interference more accurately. A distribution can be expressed either as a differential or cumulative function. This work uses differential distributions, because they are better suited for visualizing the parameters we are interested in, although the two distributions basically convey the same information.

This paper is organised as follows: in Sect. 2, we calculate a model for terrestrial interfering source distribution based on various assumptions. Sect. 3 presents the methods that we use to generate and analyse the histograms. This is followed in Sect. 4 by a short description of the two LOFAR data sets that have been used to perform the experiment. The results of analysing the sets are presented in Sect. 5. Finally, in Sect. 6 the results are discussed and conclusions are drawn.

2 MODELLING THE BRIGHTNESS DISTRIBUTION

Interference is generated by many different kinds of transmitters, and these will have different spatial and brightness distributions (“spatial” refers here to the distribution on the Earth). For example, aeroplanes and satellites have widely different heights, while

* E-mail: offringa@mso.anu.edu.au

Brightness and spatial distributions of terrestrial radio sources 3

other sources are ground-based. Even ground-based sources might be spread differently. For example, it can be expected that citizens' band (CB) devices, that are often used in cars, are distributed differently from broadcasting transmitters. For deliberate transmitters, the frequency at which interference occurs can identify the involved class of devices, because devices are constrained by the bands that have been allocated for the given class.

In a time-frequency plot, interfering sources can have complex structures. They can also be intermittent and different sources might overlap in time-frequency space. An example of interfering sources can be seen in Fig. 1, which shows raw visibility data of one baseline of a LOFAR observation in a time-frequency diagram (dynamic spectrum). Because many sources change over time, are repetitive or affect multiple channels, many sources produce multiple unconnected features in the time-frequency diagram. It is often not clear what constitutes a single interfering source, hence it is hard to count individual sources. Instead, we will count the number of times a given brightness occurs in time-frequency space. This — as well as many other effects — will of course influence the distribution. If sources overlap in the time-frequency space, the situation is somewhat similar to the case where multiple unresolved celestial radio sources in the reception pattern of a telescope only allow observation of a sum of sources. However, in that case it is still possible to validate radio source models by comparing $\log N - \log S$ histograms (Scheuer 1957).

First we will derive the expected intrinsic source distribution for interfering radio sources. After that, we will analyse the issues that arise when measuring the distribution by counting samples.

In every dynamic spectrum we can measure the number of times that the flux density is within a particular range. Dividing this quantity by the total number of samples yields the relative number of events as a function of intensity. We will refer to this quantity with the statistical term “frequency density” (not to be confused with the physical radiation frequency). We will now start by estimating the frequency density function of ground-based interfering sources. Consider an interfering point source of strength I measured in flux density. This source is observed by an interferometer that consists of two antennas or stations with gains g_1, g_2 , which include all instrumental effects. The antennas are located at distances r_1, r_2 from the source. The interferometer will record an apparent instantaneous strength S of

$$S(r_1, r_2) = I \frac{g_1 g_2}{r_1 r_2}, \quad (1)$$

with (real-valued amplitude) gains $g_1, g_2 > 0$ and $r_L > r_1, r_2 > 0$. Here, r_L is a limiting distance, which will be well below the diameter of the Earth. The formula represents a spherically propagating wave in free space. We will limit our analysis to cross-correlated antennas; the auto-correlations will be ignored.

We assume that the source observed is fully coherent, but a possible de-coherence factor can be absorbed in the gains. Due to the small bandwidth of most interfering sources, most RFI will be received coherently, because of the narrow-band condition. For example, if the bandwidth of the signal $\Delta\nu = 1$ kHz, the narrow-band condition $\Delta\nu \ll (2\pi\tau)^{-1}$ with correlation delay τ will hold for baselines up to a few km, because it holds as long as the baseline length is significantly less than $\Delta x = c(2\pi\Delta\nu)^{-1} \approx 50$ km.

Although two antennas do not necessarily observe the same RFI sources, for source count analysis we can treat the interferometer geometrically as a single point, as both antennas will see the same distribution. Then, we can express the received amplitude S

for a given distance r and interferometric gain $g = g_1 g_2$ as

$$S(r) = \frac{Ig}{r^2}. \quad (2)$$

Next, we assume that all RFI sources have equal constant strength I and follow a uniform spatial distribution in the local two-dimensional horizontal plane. This is obviously a simplification, as the RFI sources are actually distributed on the surface of a sphere with different heights. Beyond some distance, the Earth will partly block the path between transmitter and receiver. Therefore, the assumption of uniformly distributed sources on a two-dimensional plane is only approximately valid. Using these assumptions, we can express the expected cumulative frequency density of sources $F_{\text{distance} \leq r}(r)$ at distance r as

$$F_{\text{distance} \leq r}(r) = \rho \pi r^2, \quad (3)$$

for some constant ρ that represents the number of sources per unit area. In other words, we will observe F sources that are at most a distance r away.

We need the reverse of Eq. (3) to calculate the amplitude distribution, because sources with an amplitude of *at most* a given strength will have a distance of *at least* some distance. Sources with at least a distance of r are given by $F_{\text{distance} \geq r}(r) = N - \rho \pi r^2$, with N the total number of sources. Because r is restricted, N is finite.

The cumulative number of sources $F_{\text{amplitude} \leq S}$ that have an amplitude of at most S can now be calculated with

$$\begin{aligned} F_{\text{amplitude} \leq S}(S) &= F_{\text{distance} \geq r}(\mathcal{R}(S)) \\ &= F_{\text{distance} \geq r}(\pm \sqrt{\frac{Ig}{S}}) \\ &= N - \frac{\rho \pi Ig}{S} \end{aligned} \quad (4)$$

where $\mathcal{R}(S) = S^{-1}$, the inverse of S , i.e., the function that returns the distance r for a given amplitude S . Finally, the differential frequency density can be calculated by taking the derivative,

$$f_S(S) = \frac{dF_{\text{amplitude} \leq S}}{dS} = \frac{\rho \pi Ig}{S^2}. \quad (5)$$

Therefore, if we plot the histogram of the RFI amplitudes in a log-log plot, we expect to see a power law with a slope of -2 over the interval in which the RFI sources are spread like uniform sources on a two-dimensional plane. For this, we have assumed that the variables I, g and ρ are constants. In reality, these variables have a stochastic nature. The effect of this will be discussed in §2.4.

2.1 Spherical case

If we do take into account the fact that the sources lie on a sphere instead of a plane, Eq. (3) changes. In this case, the RFI sources are distributed over the surface of a sphere. If we assume their difference in height is negligible, we can use the formula for the surface area of a spherical cap,

$$A_{\text{cap}} = \pi (a^2 + h^2) \quad \text{with } (a \leq R) \wedge (h \leq 2R), \quad (6)$$

with R the radius of the Earth, a the radius of the base of the cap and h the height of the cap. Because the Euclidean distance r between the source and the antenna is $r^2 = a^2 + h^2$, the resulting number of sources within Euclidean distance r becomes

$$\tilde{F}_{\text{distance} \leq r}(r) = \rho A_{\text{cap}} = \rho \pi r^2 \quad \text{with } r \leq 2R, \quad (7)$$

which equals the formula for a planar source distribution given in Eq. (3). Therefore, the spherical and two-dimensional cases lead to

4 *A. R. Offringa et al.*

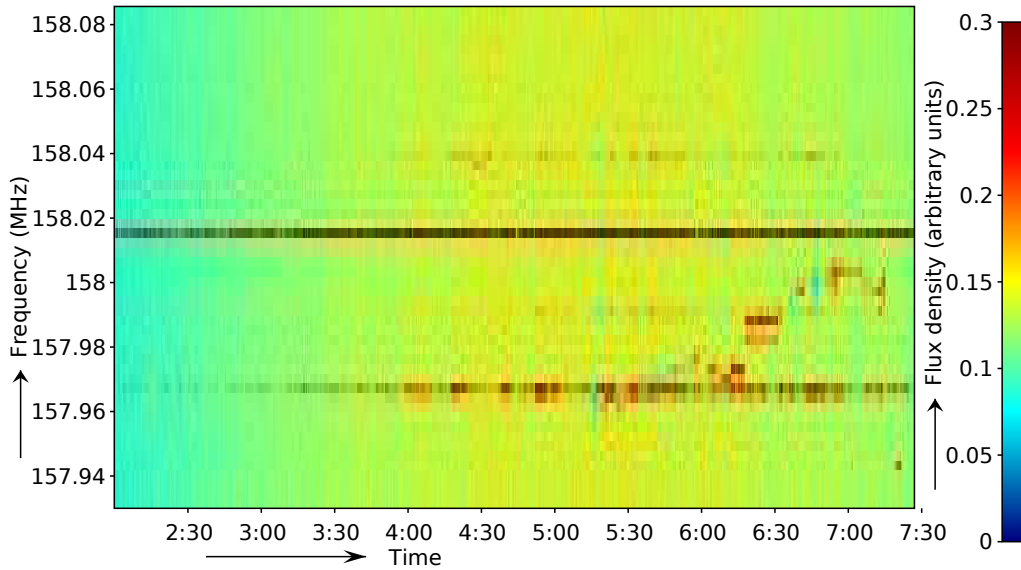


Figure 1. A small part of an observation displayed in a time-frequency plot. The features with significantly higher values are caused by interference. Some of these have a constant frequency, while others are more erratic.

a distribution with equal power law. Eq. (7) is valid for $r \leq 2R$, but in practice the lowest amplitude of observed sources is further constrained because of the curvature of the Earth. This causes distant sources to settle below the horizon. Nevertheless, some of those sources might still be visible due to reflection of the ionosphere and other (semi-)spherical propagation effects.

2.2 Propagation effects

So far, we have assumed that the electromagnetic radiation propagates through free space, resulting in an r^{-2} fall-off. In reality, the radiation will be affected by complicated propagational effects. Because of the irregular surface of the Earth (including urban areas) and the absence of a direct line of sight between transmitter and receiver, the propagation mode might be indirect. Multiple indirect paths might be formed that include reflection, refraction, diffraction and absorption of the electromagnetic wave.

A commonly used propagation model is the empirical model determined by Okumura et al. (1968), which was further developed by Hata (1980). The original model is based on urban areas, but corrections are given by Hata for sub-urban areas with a lower population density and for open areas. Hata gives the following analytical estimate for L_p , the electromagnetic propagation loss over land:

$$L_p = 69.55 + 26.16 \log_{10} f_c - 13.82 \log_{10} h_b - a(h_m) + (44.9 - 6.55 \log_{10} h_b) \log_{10} r, \quad (8)$$

where L_p the loss in dB; f_c the radiation frequency in MHz; h_b the height of the transmitting antenna in meters; h_m the height of the receiving antenna in meters; r the distance between the antennas in meters; and $a(h_m)$ a correction factor in dB that corrects for the height of the receiving antenna and the urban density. Hata found this model to be representative for frequencies $f_c \sim 150$ – 1500 MHz, with transmitter heights $h_b \sim 30$ – 200 m, receiver heights $h_m \sim 1$ – 10 m and over distances $r \sim 1$ – 20 km.

Converting from a subtracted term in decibels to a flux density factor L_S , and collecting the terms of Eq. (8) that are not depending

on r in a single variable ζ , results in

$$L_S = \zeta r^{4.49 - 0.655 \log_{10} h_b}, \quad (9)$$

with

$$\zeta = \frac{f_c^{2.616}}{h_b^{1.382}} - 10^{6.955 - \frac{1}{10} a(h_m)}. \quad (10)$$

Note that according to Hata's model, the exponent of the power law in Eq. (9) depends only on the height of the transmitting antenna, i.e., it is independent of frequency, receiver height and urban density. Now, if in Eqs. (4) and (5) one replaces the definition of $S(r)$ from Eq. (2) with one that includes the propagation effects,

$$S(r) = \frac{I_g}{\zeta r^\eta}, \quad (11)$$

with $\eta = 4.49 - 0.655 \log_{10} h_b$, one finds the frequency density function f_p that considers propagation effects,

$$f_p(S) = \frac{d}{dS} \left[N - \rho \pi \left(\frac{I_g}{\zeta S} \right)^{2/\eta} \right] = \frac{\rho 2\pi}{\eta S} \left(\frac{I_g}{\zeta S} \right)^{2/\eta}. \quad (12)$$

Consequently, due to non-free-space propagation effects, the observed log-log histogram is predicted to have a $-(\frac{2}{\eta} + 1)$ slope. By substituting η , one finds

$$\text{slope}(h_b) = \frac{1}{0.3275 \log_{10} h_b - 2.245} - 1. \quad (13)$$

This is valid for transmitters that have a height of 30–200 m, the range over which Hata's model was defined. This yields estimated distribution slopes of -1.57 and -1.67 for 30 m and 200 m high transmitters respectively. In Figure 2, the slope value is plotted as a function of the transmitter height, including extrapolated values for transmitter heights down to 1 m.

2.3 Thermal noise contribution

The full measured distribution will consist of the power-law distribution combined with that of the thermal noise and the celest-

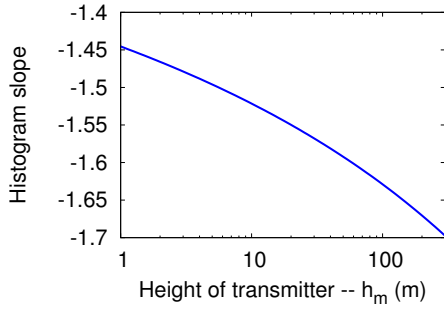


Figure 2. Effect of transmitter height on the slope of a log-log histogram. According to Hata's model, this is valid for the range 30–200 m. The trend of the slope will not continue indefinitely when increasing the height further. Instead it will converge to a -2 slope, which corresponds to free-space propagation.

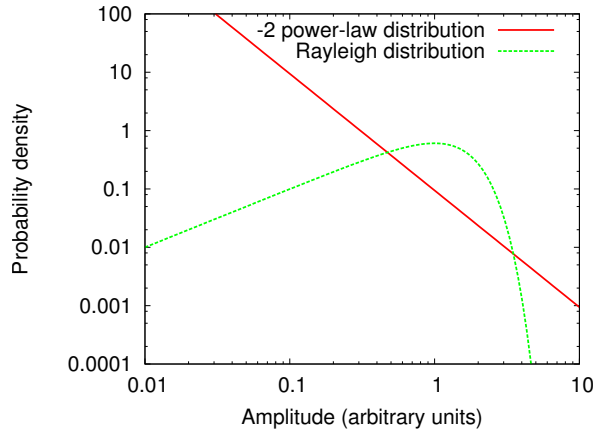


Figure 3. The Rayleigh and power-law distributions in a log-log plot. The power-law distribution (Eq. (5)) has a constant slope of -2 . The slope of the Rayleigh distribution in the limit of the origin is 1. Its maximum occurs where the amplitude value equals its mode σ , which is 1 in this example. For higher amplitudes, its slope decreases exponentially.

tial signal. For now, we will ignore the contribution of the celestial signal, as its contribution to the amplitude distribution will be minimal when observing fields without strong sources. For example, the strongest apparent source in the NCP EoR field is around 5 Jy (Yat-tawata et al. 2013, *in press*). The standard deviation of the noise, however, is around 100 Jy on highest LOFAR resolutions, and will have a larger contribution on the histogram.

The real and imaginary components of the noise in the cross-correlations are independent and identically Gaussian distributed with zero mean and equal variance. Consequently, an amplitude x will be Rayleigh distributed (Papoulis & Pillai 2001, §6-2):

$$f_{\text{noise}}(x) = \begin{cases} \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}} & x > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (14)$$

Because most of the samples will be unaffected by RFI, this will be the dominating distribution. The Rayleigh distribution is plotted together with the -2 power-law distribution of Eq. (5) in Fig. 3.

So far, these are the expected histograms for pure noise and pure RFI that propagates through free space. However, the measured distribution is a mixture of the two. A perfect RFI detector would classify the samples in two groups; one group of samples

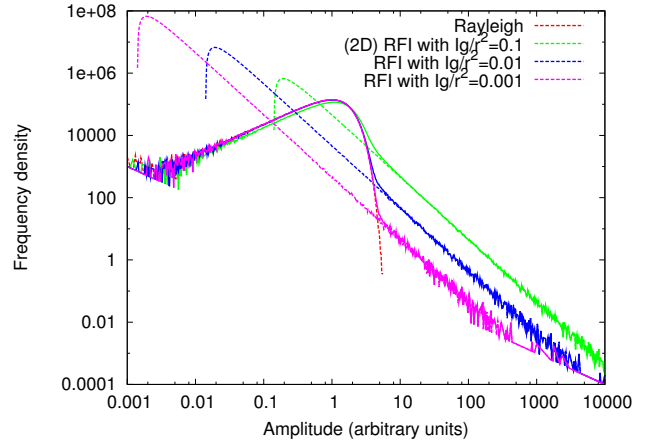


Figure 4. Histograms of simulated samples that all have a contribution of noise and RFI. Various settings of the parameters were used, and samples were drawn as described in Eq. (17). Solid lines: the combined distributions, dashed lines: the power-law distributions before mixing.

that are not affected by RFI, and therefore contain noise only, and one group of samples that are contaminated, and are the sum of RFI and noise. Real RFI detectors can classify samples into these groups to some extent, but due to false positives and false negatives, the groups will get mixed nevertheless. Even more, we still need to take into account that the samples classified as RFI are also affected by noise. Although this effect is moderate, because ‘most’ RFI samples are of much higher amplitude than what is added because of the noise, we are interested in the low-level RFI samples, and the relative effect of noise on those samples is large.

To understand the distribution of samples consisting of RFI + noise, one needs to consider the individual contributions of the real and imaginary components. The real and imaginary components of samples that are affected by RFI, will follow a multi-variate distribution: each complex component is the sum of a complex component of RFI and a noise sample. The noise comes from a Gaussian distribution with zero mean. The RFI distribution is assumed to be a power-law distribution with yet unknown exponent.

To calculate the corresponding amplitude distribution that is formed by the combination of the real and imaginary component distributions, it is easier to perform numerical simulations, as analytic derivation is not trivial. The histogram can be numerically estimated by drawing complex samples from the two distributions and calculating and counting the amplitudes. A sample can be drawn from the RFI distribution by scaling and inverting the cumulative frequency function in Eq. (4) and evaluating it for a uniformly distributed variable. Note however that $F_{\text{amplitude} \leq S}(S)$ is not limited; when decreasing the amplitude S towards zero, the number of sources will go to infinity. Therefore, one needs to assume that there are no samples beyond some limiting distance r_L to invert the cumulative function. With this assumption, a single complex RFI contaminated sample S_{RFI} can be sampled with

$$S_{\text{RFI}} \leftarrow \frac{Ig}{x_u r_L^2} e^{i2\pi y_u}, \quad (15)$$

with $x_u, y_u \sim U(0, 1)$. The result is an amplitude sample S_{RFI} , with $S_{\text{RFI}} \geq Ig/r_L^2$, that follows a power-law distribution. The corresponding formula for drawing a sample S that is contaminated

6 *A. R. Offringa et al.*

by both RFI and noise is

$$S \leftarrow v_n + w_n i + \frac{Ig}{x_u r_L^2} e^{i2\pi y_u}, \quad (16)$$

with $v_n, w_n \sim N(\mu = 0; \sigma)$. The final amplitude sample S_A can be calculated with

$$S_A \leftarrow |S| = \sqrt{\text{real}^2(S) + \text{imag}^2(S)}. \quad (17)$$

An example of distribution curves for various settings of Ig/r_L^2 is given in Fig. 4.

If we assume that I , the average source strength; g , the average instrumental gain; and r_L , the distance over which those sources are visible, are all unknown, it can be seen from Eq. (16) that given the histogram we can solve neither r_L , I or g , as they all have the same effect of scaling the -2 power-law part of the histogram. Although calibration could in theory solve g , almost all sources will come in through the far-sidelobes of the phased-array beam, and finding the expected values for the gains is therefore hard. The effects of these parameters on the histogram will be further discussed in §2.4.

Fig. 4 shows that the left side of the graph corresponds with the original Rayleigh distribution. From the right side of the histogram, we are able to estimate Ig/r_L^2 , as the placing of the power-law curve is independent of the Rayleigh distribution for large S . Methods to constrain the parameters of the RFI distribution will be discussed in §3.3. The curves in Fig. 4 are from histograms of samples that are all contaminated by both noise and RFI. In practice, we can not make this distinction accurately, as RFI detectors have a limited accuracy. Consequently, the observed histogram will be the sum of two types of histograms: the first histogram being contaminated by both noise and RFI, the second one only by noise.

For completeness we show the formula for sampling apparent source strengths that include propagation effects according to Hata's model, which can be used for sampling realistic source amplitudes in simulations. This can be derived by integration, scaling and inversion of the frequency density function in Eq. (12), resulting in:

$$\tilde{S}_{\text{RFI}} \leftarrow \frac{Ig}{\zeta x_u^{\eta/2} r_L^\eta} e^{i2\pi y_u}. \quad (18)$$

Here, \tilde{S}_{RFI} is a new complex RFI sample; ζ and η are as defined below Eq. (10); I is the average intrinsic strength; g is the gain of the instrument; $0 \leq x_u, y_u < 1$ are two independently drawn uniformly distributed samples; and r_L is the maximum distance of visible sources.

2.4 Parameter variability

In reality, the parameters ρ , I and g , which are the source density per unit area, source strength and instrumental gain respectively, will not be constant, but can change over time and frequency. Therefore, they act like stochastic variables. However, since each specific value for these parameters produces a power law, the combined distribution will still show a power law, as long as the parameters follow a distribution that is steep at high amplitudes (in log-log space), such as a Gaussian or uniform distribution.

One instrumental effect that is absorbed in g is the frequency response of the instrument, i.e., the antenna response in combination with the band-pass of the analogue and digital filters. Because the data that are analysed in Sect. 5 have initially not been band-pass calibrated, the instrumental response is not uniform over frequency. We determined that the variation due to the band-pass is

about one order of magnitude for the low-band antennas (LBA, 30.1–77.5 MHz) and about a factor of two for the high-band antennas (HBA, 115.0–163.3 MHz).

The effect of the band-pass on the data distribution is consequently limited to one order of magnitude or less. If the RFI sources have an apparent power-law distribution with a certain exponent, the distribution will have a feature at low amplitudes due to the frequency response, but a power law with equal exponent will still appear when the distribution is observed over a wide enough range.

Another effect that is absorbed in g , is the beam of the instrument. At the point of writing, LOFAR beam models are still being developed and are not yet well parametrized near the horizon. It is likely that most RFI sources are observed at the edges of the beam. Nevertheless, most sources will be observed with similar gains (within one order of magnitude), and it can be expected that the beam will have a limited effect on the histogram properties of an observation. It is therefore comparable with the effect of the frequency response.

The stochastic nature of I , that is caused by the spread of transmitters with different intrinsic strengths, might also have an effect on the $\log N - \log S$ histograms. It is unlikely that I follows a uniform or Gaussian distribution, because the distribution will contain few strong transmitters (such as radio stations) and many weak transmitters (such as remote controls). Therefore, variable I might follow a power-law distribution by itself. It is likely that strong transmitters transmit more on average, and therefore contaminate more samples as well. High-power transmitters, such as radio stations, have a typical transmitting power in the order of 100 kW. Low-power transmitters, such as remote controls, transmit with an order of 100 mW or even less. Therefore, these devices have a spread of around 6 orders of magnitudes in power. As long as the exponent of the power law of I is less steep (i.e., less negative) compared to the power law caused by the spatial distribution, the spatial distribution will dominate the histogram at high amplitudes. With a spatial -1.5 power law and the given transmitting powers, the low-power transmitters should contaminate a factor of 10^9 more samples compared to the high-power transmitters to dominate the high-amplitude distribution, which is unlikely. Therefore, it is likely that the spatial distribution will dominate the power law in the histogram. Otherwise, a turn-over point should be visible in the histogram.

3 METHODS

In this section we will briefly discuss how the histograms are created, how the slope of the underlying RFI distribution is estimated and show how to constrain some of the intrinsic RFI parameters.

3.1 Creating a histogram

While creating a histogram is trivial, it is important to note that it is necessary to have a variable bin size. This is mandated by the large dynamic range of the histogram that we are interested in. Therefore, we chose to have a bin size that increases linearly with the amplitude S , and the frequency counts are divided by the bin size after counting. Consequently, in parts of the histogram that have a sparse number of samples, the outlying samples will follow a $1/S$ curve, or a -1 slope in a log-log plot. This can be seen in the tails of the distributions of Fig. 4. This, however, is not an intrinsic feature of the data but a consequence of the binning method.

3.2 Estimating σ and slope parameters

The mode σ of the Rayleigh distribution is estimated by finding the amplitude with the maximum occurrences, i.e., the amplitude corresponding to the peak of the histogram. The slope is estimated using linear regression over a visually selected interval.

One should note that fitting straight lines to the distribution curve in a log-log plot is not the most accurate way of estimating the exponent of a power-law distribution (Clauset et al. 2009). However, because of our enormous sample size, which allows fitting the line over a large interval, the estimator will be sufficiently accurate for our purpose. Nevertheless, we will additionally calculate a maximum likelihood estimator for comparison. The maximum-likelihood estimator for the exponent in a power-law distribution is given by the Hill estimator $\hat{\alpha}_H$ (Hill 1975; Clauset et al. 2009), defined as:

$$\hat{\alpha}_H = 1 + N \left(\sum_{i=1}^N \ln \frac{x_i}{x_{\min}} \right)^{-1}, \quad (19)$$

with N the number of samples and x_i for $0 < i \leq N$ the samples that follow a power law with lower bound x_{\min} . However, this estimator assumes the distribution is not cut-off at a high point. In our case, the distribution is cut-off at both sides, for example because of the limit of the analogue-digital converter (see §3.3). Therefore, using this estimator will result in an estimate that is steeper (i.e., more negative) than the actual distribution. Moreover, the estimator requires reading and processing all data points, and because of the size of the data sets this is somewhat impractical. However, we can calculate Eq. (19) by adding the bin centre values N times to the calculation, where N is the bin frequency. Because our bin size is small and the distribution is well sampled, the resulting estimate will be close to the normal Hill estimator. This avoids the need to iterate over all data.

3.3 Determining RFI distribution limits

In this section we will show methods to find S_U and S_L , the upper and lower flux limits of the power-law distribution at which the power law breaks down. Once the exponent of the power-law part of the distribution is estimated with the previously discussed techniques, the distribution can be extrapolated to find a lower flux limit. Assume that we have found a power law with exponent α and factor β over an amplitude region $[S_1; S_2]$, resulting in the following frequency density function h :

$$h(S) = \beta S^\alpha. \quad (20)$$

S_1 and S_2 are selected by visually inspection of the histogram. Assume that the histogram contains N_{part} (RFI) samples with amplitude $> S_1$, as sketched in Fig. 5, and that the effect of the Rayleigh component on the histogram $> S_1$ is negligible. The hypothetical upper limit S_U of the distribution can now be found, i.e., the highest amplitude that would be observed when the distribution follows the power law up to that point, by solving

$$\int_{S_1}^{S_U} h(S) dS = N_{\text{part}}. \quad (21)$$

In practice, the observed histogram will break down beyond some amplitude because of several reasons. First and most importantly, the samples themselves are digitized with an analogue-to-digital

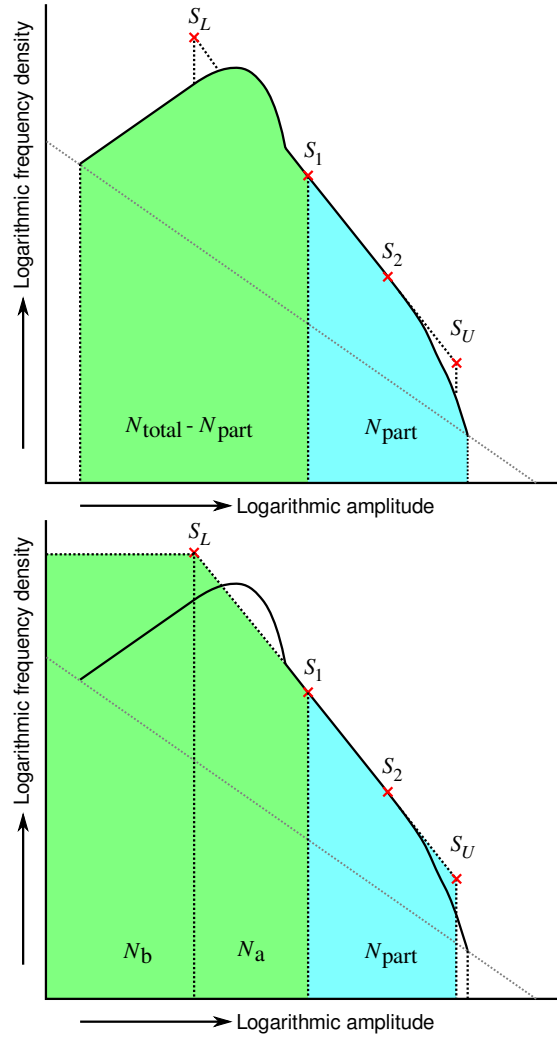


Figure 5. Cartoon of how a constraint on the lower fall-off point of the power-law distribution can be determined. Note that the labelled areas are areas as occupied in a linear plot, i.e., the integration of the density function. Areas in a log-log plot are not linearly related to the integral. There are two ways to estimate the lower constraint S_L : (i) the areas N_a and $N_{\text{total}} - N_{\text{part}}$ are equal if Ig/r^2 is constant during the observation, and (ii) if one assumes $Ig/r^2 \sim \text{uniform}$, then $N_a + N_b = N_{\text{total}} - N_{\text{part}}$.

converter (ADC) with limited range. Second, we observe for a limited time and the frequency count is discrete. Because the chance of finding a sample with a very high intensity is low, it is unlikely to observe samples beyond some amplitude within the finite observing time. Finally, under the assumption of a uniform spatial distribution of RFI transmitters, samples with very high amplitude would have to be produced by transmitters that are very close to the telescope. However, it is likely that the uniform spatial distribution of transmitters will break down at closer distances.

Solving Eq. (21) results in

$$S_U = \alpha+1 \sqrt[\alpha+1]{\frac{\alpha+1}{\beta} N_{\text{part}} + S_1^{\alpha+1}}. \quad (22)$$

In some cases there will be no solution for S_U . This situation arises because the integral of the distribution function is finite even when $S_U \rightarrow \infty$. If N_{part} exceeds the integrated value of the fitted dis-

8 A. R. Offringa et al.

tribution function, there is no solution. This can happen when the empirical distribution is not limited at the high end, or when it contains features that are not in the model.

Similar to the calculation of S_U and by using N_{total} as an upper limit to the total number of samples that are affected by the power-law distribution, one can estimate the lower limit S_L . For Fig. 5 this means that the areas labelled N_a and $N_{\text{total}} - N_{\text{part}}$ are equal. The area labelled N_b is assumed to be zero for now, which assumes the power law has a sharp cut-off on the left side, e.g., because of the curvature of the Earth. Solving $N_a = N_{\text{total}} - N_{\text{part}}$ results in

$$S_L = {}^{\alpha+1}\sqrt{\frac{\alpha+1}{\beta} (N_{\text{part}} - N_{\text{total}}) + S_1^{\alpha+1}}. \quad (23)$$

With the assumption that $Ig/r_L^2 \sim$ a uniform distribution, the area labelled in Fig. 5 as N_b is also part of the RFI distribution, and a stronger constraint \tilde{S}_L can be found by solving

$$\int_{\tilde{S}_L}^{S_U} h(S) dS + \tilde{S}_L h(\tilde{S}_L) = N_{\text{total}} - N_{\text{part}}, \quad (24)$$

which yields

$$\tilde{S}_L = {}^{\alpha+1}\sqrt{-\frac{1}{\alpha} \left(\frac{\alpha+1}{\beta} (N_{\text{part}} - N_{\text{total}}) + S_1^{\alpha+1} \right)}. \quad (25)$$

With estimates of α , β , S_L and S_U , one has obtained a parametrization of the RFI distribution. As was shown in §2.3, the left-most point where the power-law distribution falls off is Ig/r^2 – assuming free space propagation for the moment – and therefore $S_L = Ig/r^2$. For Hata's propagation model, the more generic solution $S_L = Ig/(\zeta r^\alpha)$ is found. This value represents the apparent brightness of the sources that are the furthest away from the telescope. With a fully parametrized distribution of the effect of RFI sources, an empirical model for RFI effects can be made. Moreover, one can calculate $E(S_R)$, the expected apparent strength of RFI:

$$E(S_R) = \frac{1}{N_{LU}} \int_{S_L}^{S_U} \beta S^\alpha dS = \frac{\beta}{N_{LU}} \left[\frac{1}{\alpha+2} S^{\alpha+2} \right]_{S_L}^{S_U} \quad (26)$$

$$= \frac{\beta (S_U^{\alpha+2} - S_L^{\alpha+2})}{N_{LU} (\alpha+2)} \quad (27)$$

Here, N_{LU} is the number of samples between S_L and S_U after normalizing for the bin size:

$$N_{LU} = \int_{S_L}^{S_U} \beta S^\alpha dS. \quad (28)$$

Substitution and simplification of these two results in

$$E(S_R) = \frac{(S_U^{\alpha+2} - S_L^{\alpha+2}) (\alpha+1)}{(S_U^{\alpha+1} - S_L^{\alpha+1}) (\alpha+2)}. \quad (29)$$

This is essentially the average flux density that is caused by RFI without using RFI detection or excision algorithms. $E(S_R)$ has the same units as S_L and S_U , thus after calibration (see §3.4) could be given in Jy. In practice, the increase of data noise after correlation is much less severe because of RFI flagging, which excises a part of the RFI. One can assume that all RFI above a certain power level is found by the detector. Since modern RFI detection algorithms can find all RFI that is detectable “by eye” (Offringa et al. 2010a), this power level will be near the level of the noise mode. In

Offringa et al. (2013) the false-positives rate for the AOFlagger is estimated to be 0.5%. An estimate of S_d , the average lower limit of detected RFI, can be calculated by finding the point on the distribution where the area under the distribution to the right of S_d equals the real number (true positives) of RFI samples. Therefore, the limit is calculated similar to Eq. (23), where the term $N_{\text{part}} - N_{\text{total}}$ needs to be replaced with N_{RFI} , which equals the total number of samples detected as RFI minus the 0.5% false positives.

Finally, $E(S_{\text{leak}})$, which is the expected value of leaked RFI not detected by the flagger, can be calculated by replacing S_U with S_d in the numerator of Eq. (29) and subtracting the removed number of samples from the total of number of samples. Assume that a fraction of κ samples are not detected as RFI and $1 - \kappa$ have been detected as RFI, then

$$E(S_{\text{leak}}) = \frac{1}{\kappa N_{LU}} \int_{S_L}^{S_d} \beta S^\alpha dS = \frac{(S_d^{\alpha+2} - S_L^{\alpha+2}) (\alpha+1)}{\kappa (S_U^{\alpha+1} - S_L^{\alpha+1}) (\alpha+2)}. \quad (30)$$

This is the average contribution that leaked RFI will have on a single sample. It has the same units as the parameters S_L , S_U and S_d . Typical values for κ are 0.95–0.99.

3.4 Calibration

We can assign flux densities to the horizontal axis of the histogram by using the system equivalent flux density (SEFD) of a single station. The current LOFAR SEFD is found to be approximately 3000 Jy for the HBA core stations and 1500 Jy for the remote stations in the frequency range from 125–175 MHz. For all Dutch LBA stations, in the frequency range 40–70 MHz the SEFD is approximately 30,000 Jy. The standard deviation σ in the real and imaginary values is related to the SEFD with

$$\sigma = \frac{\text{SEFD}}{\sqrt{2\Delta\nu\Delta t}}, \quad (31)$$

where $\Delta\nu$ is the bandwidth and Δt is the correlator integration time. The standard deviation will appear as the mode of the Rayleigh distribution. By fitting a Rayleigh function with fitting parameter σ to the distribution, one finds the corresponding flux density scale.

RFI sources will enter through the distant sidelobes of the station beams from many unknown directions. Moreover, models for the full beam are often hard to construct. Therefore, we will not try to calibrate the beam, and the flux densities in the histogram are apparent quantities. Consequently, we will not be able to say something about the true intrinsic power levels of RFI sources.

3.5 Error analysis

An estimate for the standard deviation of the slope estimator $\hat{\alpha}$ can be found by calculating $\text{SE}(\hat{\alpha})$, the *standard error* of $\hat{\alpha}$. The standard error of the slope of a straight line (Acton 1966, pp. 32–35) is given by

$$\text{SE}(\hat{\alpha}) = \sqrt{\frac{SS_{yy} - \hat{\alpha} SS_{xy}}{(n-2) SS_{xx}}}, \quad (32)$$

where SS_{xx} , SS_{xy} and SS_{yy} are the sums of squares, e.g., $SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ and n is the number of samples. However, we found that this is not a representative error in our case, because the errors in the slope are not normally distributed.

The noise in the part of the histogram over which the slope is calculated, is very small due to the large number of samples. Consequently, the estimated standard deviation of $\hat{\alpha}$ is very low. Nevertheless, when the slope is calculated over subsections of the original range over which the slope is calculated, one finds the line is not completely straight and the slope is changing more than is predicted by the standard error. Therefore, we introduce an error estimate $\epsilon_{\hat{\alpha}}$ that quantifies a normalized standard deviation of the slope over the range. This error is formed by calculating the slope over n_{α} smaller sub-ranges in the histogram, creating n_{α} estimates α_i . If the errors in α_i are normally distributed with mean zero, the standard deviation over the full range will be $\sqrt{n_{\alpha}}$ times smaller. Therefore, an estimate of the variance of $\hat{\alpha}$ can be calculated with

$$\epsilon_{\hat{\alpha}} = \sqrt{\frac{\sum (\alpha_i - \hat{\alpha})^2}{n_{\alpha}^2 - n_{\alpha}}}. \quad (33)$$

This estimate is slightly depending on the number of sub-ranges that is used, n_{α} , because the errors are not Gaussian distributed, but we found that $\epsilon_{\hat{\alpha}}$ is more representative than the standard error of $\hat{\alpha}$.

The Hill estimator of Eq. (19) is a different estimation method for the exponent in the power-law distribution, and yields therefore also a different standard error. The standard error of the Hill estimator is (Clauset et al. 2009)

$$\text{SE}(\hat{\alpha}_H) = \frac{-\alpha - 1}{\sqrt{n}} + \mathcal{O}\left(\frac{1}{n}\right). \quad (34)$$

Because the number of samples is very large ($> 10^{11}$), the \mathcal{O} -term will be very small. Therefore, we will calculate the quantity without this term. As with the standard error for the slope of a straight line in Eq. (32), the standard error for the Hill estimator yields very small quantities. Again, this is because it assumes the underlying power law has a fixed exponential, while in our case the power law seems to vary. Therefore, this value is given only for completeness.

Because our distributions are huge, we decided not to do goodness-of-fit tests, because these require many similar distributions to be simulated in order to reach accurate decision. Instead, we will try to evaluate the distributions visually.

4 DATA DESCRIPTION

We have analysed the distributions of two data sets. Both data sets are 24-h LOFAR RFI surveys and are extensively analysed in Offringa et al. (2013). We refer to van Haarlem et al. (2013, *submitted*) for a full description of the capabilities of LOFAR. In one set, the low-band antennas (LBA) were used and the frequency range 30.1–77.5 MHz was recorded, while in the other the high-band antennas (HBA) were used to record the frequency range 115.0–163.3 MHz. More stations were used in the LBA set. The specifications of the two sets are listed in Table 1. The stations that have been used are geometrically spread over an area of about 80 km and 30 km in diameter at maximum for the LBA and HBA sets respectively. For EoR detection experiments, the HBA are more important than the LBA, because they cover the frequency range of the redshifted EoR signal.

Although we have used Hata's model to estimate the RFI log-log histogram slope, our frequency range falls partly outside the frequency range over which Hata's model has been verified. However, according to Hata's model the observing frequency does not influence the power-law exponent in the frequency range 150–1500

Table 1. Data set specifications

	LBA set	HBA set
Observation date	2011-10-09	2010-12-27
Start time	06:50 UTC	0:00 UTC
Length	24 h	24 h
Time resolution	1 s	1 s
Frequency range	30.1–77.5 MHz	115.0–163.3 MHz
Frequency resolution	0.76 kHz	0.76 kHz
Number of stations	33	13
Total size	96.3 TB	18.6 TB
Field	NCP	NCP
Amount RFI detected	1.77%	3.18%

MHz, thus it can be assumed the exponent will at least not significantly differ over the HBA range.

To detect RFI, the AOFlagger (Offringa et al. 2010b, 2012a) was used. However, since this flagger is partly amplitude-based, it is likely that low-level RFI will leak through the detector. Since it is also low-level RFI we are interested in, we will analyse unflagged data and the RFI classified data.

5 RESULTS

In this section we present the histograms of the LBA and HBA sets and the results that were obtained by applying the methodology discussed in Sect. 3.

5.1 Histogram analysis

Fig. 6 shows the histograms with logarithmic axes for the LBA and HBA sets. In both sets, it is clear that at least one component with a Rayleigh and one with a power-law distribution have been observed. The left part of the histogram matches the Rayleigh distribution well up to the mode of the distribution. The bulge around the mode of the LBA histogram is wider due to the larger effect of the antenna response, i.e., variability of g as discussed in §2.4. As can be seen in Fig. 7, the Rayleigh-bulges of individual sub-bands are not that wide, but they are not aligned because of the differing noise levels. This effect is not so strong in histograms of the HBA sub-bands in Fig. 8, because the HBA antenna response changes less over frequency.

It is to be expected that the RFI-dominated part of the distributions at different frequencies will reflect the underlying source populations. Both Figs. 7 and 8 show that the power-law part of the distributions are very different for different sub-bands. Nevertheless, combining the data of all the sub-bands results in reasonably stable power-law distributions. The variation could be caused by the different power-law exponents that source populations at different frequencies might have. It could also be caused by a differing number of transmitters. In that case, the underlying power law might not always be apparent, because not enough samples are combined.

To make sure that the antenna response does not influence the result of the slope, we have also analysed the curves after a simple band-pass calibration. This was performed by dividing each sub-band by its standard deviation after RFI excision. Because the standard deviation of the distribution might be affected by the RFI tail of the distribution, we compare the two histograms to make sure the power-law distribution is not significantly changed. The resulting histograms are shown in Fig. 9. This procedure makes the bulge

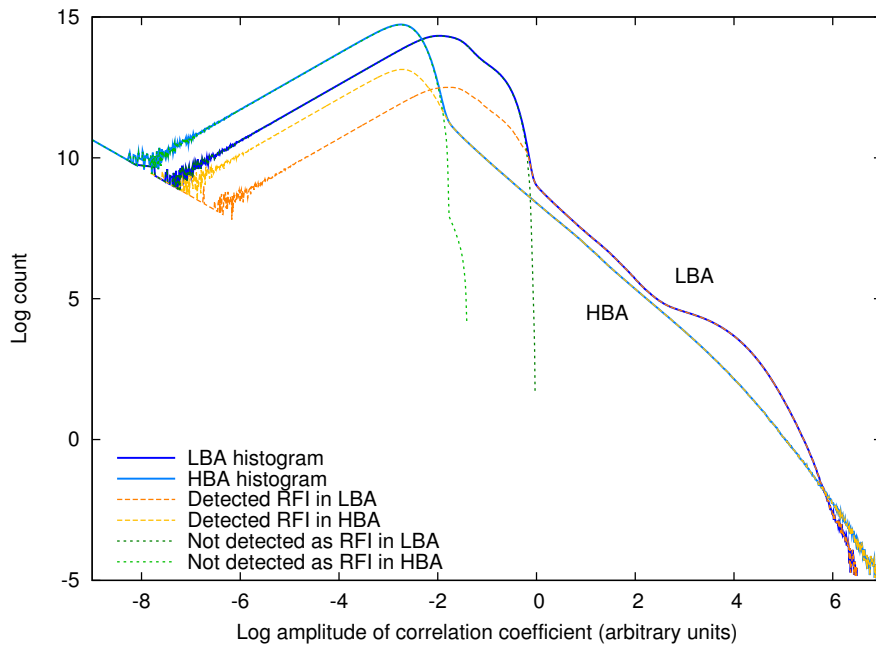


Figure 6. The histograms of the two data sets before band-pass correction and flux calibration.

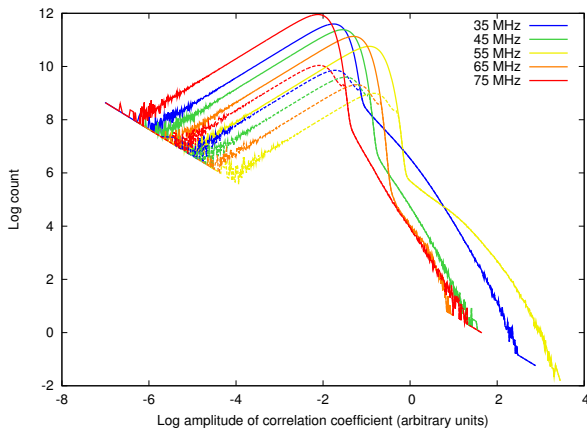


Figure 7. Histograms for 5 different 0.2 MHz LBA sub-bands without band-pass correction and flux calibration. The continuous lines represent the data before RFI flagging. The dashed lines are the histograms of the samples that have been classified as RFI.

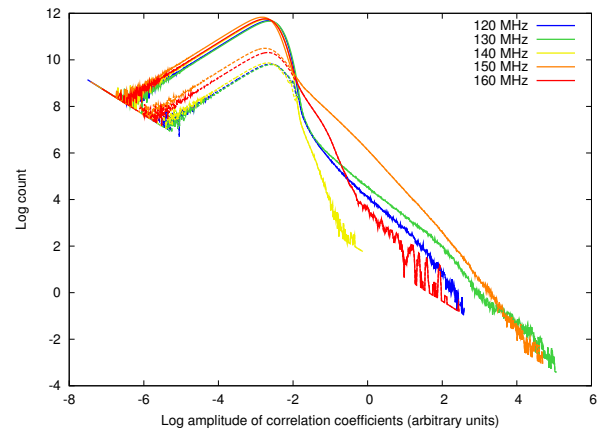


Figure 8. Histograms for 5 different 0.2 MHz HBA sub-bands without band-pass correction and flux calibration. The continuous lines represent the data before RFI flagging. The dashed lines are the histograms of the samples that have been classified as RFI.

of the LBA histogram similar to the bulge of a Rayleigh curve and extends the power-law part. Nevertheless, it does not change the log-log slope of the power law in either histograms. This validates that the variable gain that is caused by the antenna response does not change the observed power law. Consequently, it can be expected that other stochastic effects, such as the intrinsic source strength and the beam gain due to a differing direction of arrival, will similarly not affect the power law. Because the band-pass corrected histograms should provide a more accurate analysis, we will use the corrected histograms for further analysis.

The Rayleigh parts of the distributions are plotted in Fig. 10, along with a least-squares fit and its residuals. Both histograms follow the Rayleigh distribution for about five orders of magnitude,

which is validated by the residuals that show only noise. It breaks down about one order of magnitude before the mode of the distributions. This is because of the multi-component nature of the distributions, as was described in §2.3.

If we go back to Fig. 9, we see that in the LBA the power law is stable for about three orders of magnitude, and one order more in the HBA. Fig. 11 shows the slope of the log-log plot as a function of amplitude, which was constructed by performing linear regression in a sliding window, with a window size of 1 decade. The HBA shows very little structure in the slope, but the LBA is less stable and shows some features in its power-law part. Linear regression on the power-law part of the log-log plot results in a slope of -1.62 for the LBA and -1.53 for the HBA. These and the other

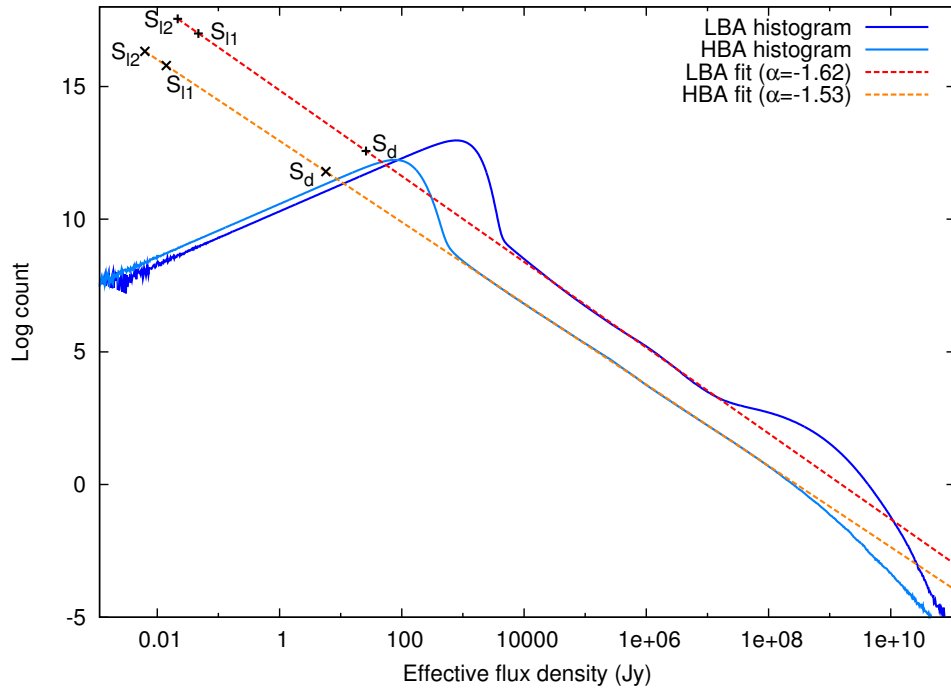


Figure 9. Observed LBA distribution after band-pass correction and flux calibration. S_{l1} and S_{l2} denote the limits of the distribution with a sharp lower cut-off (Eq. (23)) and uniform lower limit (Eq. (25)), S_d is the average lower limit of RFI that is detected.

derived quantities have been summarized in Table 2. Although the HBA slope does not show any other significant features besides the Rayleigh and power-law curves, the LBA power law ends with a bulge around an amplitude of 10^6 . This bulge is caused by a very strong source affecting lots of samples, and is a single outlier in the spatial distribution. We found this is caused by RFI observed for about an hour in the late afternoon in the lower LBA frequency regime, around 30–40 MHz. Leaving this frequency range out flattens the bulge significantly, but does not completely eliminate it, because the source put the receivers in a non-linear state, causing leakage at lower intensity levels in the other sub-bands. Unlike linear regression, the fitting region of the Hill estimator is not limited at the high end. Consequently, because of the bulge, the Hill estimator evaluates for the LBA into a slope that is less steep, with a value of -1.53 . For the HBA set, the Hill estimator is equal to the -1.53 value found by linear regression.

On the assumption that the histogram is zero below amplitude S_L , we find that $S_L = 21$ mJy for the LBA and $S_L = 6.2$ mJy for the HBA (see Table 2). If instead it is assumed that the histogram has a uniform distribution below some amplitude \tilde{S}_L , we find that the amplitude at which the power-law distribution breaks down is approximately a factor two higher. The two different assumptions on how the power-law distribution breaks down have a small effect on $E(S_{\text{leak}})$, the expected value of the leaked RFI. By using \tilde{S}_L instead of S_L , it is a few percent lower. By assuming a 100% RFI occupancy, we find that the expected value of leaked RFI is 484–496 mJy for the LBA and 167–171 mJy for the HBA. By assuming 10% occupancy, the value for $E(S_{\text{leak}})$ is about 25% reduced. The RFI occupancy only starts to have a significant effect on $E(S_{\text{leak}})$ if it is well below 10%.

6 CONCLUSIONS AND DISCUSSION

We have analysed the histogram of visibility amplitudes of LOFAR observations and found that, within a significant range of the histogram, the contribution of RFI sources follows a power-law distribution. The found power-law exponents of -1.62 and -1.53 for the 30–78 MHz LBA and 115–163 MHz HBA observations respectively, can be explained by a uniform spatial distribution of RFI sources, affected by propagation described surprisingly well by Hata’s electromagnetic propagation model. Taken at face value these exponents imply in Hata’s model that the average transmitting heights for sources affecting the LBA and HBA are 79 and 13 m respectively. There are no 79 m high transmitters nearby LOFAR stations in the LBA frequency range. Hata’s model only goes down to 150 MHz, and it is possible that the electromagnetic fall-off due to propagation will be different for lower frequencies, e.g. because the effect of the ionosphere becomes stronger. Intervals for the exponents with representative 3σ boundaries are $[-1.80; -1.44]$ for the LBA and $[-1.53; -1.49]$ for the HBA. The estimate of the HBA is thus more accurate, because its histogram deviates less from the power law.

On the assumption that the power-law distribution for RFI sources will continue down into the noise, we have constructed a full parametrization of the RFI apparent flux distribution. By assuming that all samples contain some contribution of RFI, we find that the average flux density of RFI after excision by automated flagging is 484–496 mJy for the LBA and 167–171 mJy for the HBA. These values should be compared to the noise in individual samples of 770 Jy (LBA) and 77 Jy (HBA) (see Table 2), and are upper limits for what can be expected. If in fact not all samples are affected by RFI, the leaked RFI flux will be smaller, and will of course be zero in the extreme case that the detector has found and removed all RFI.

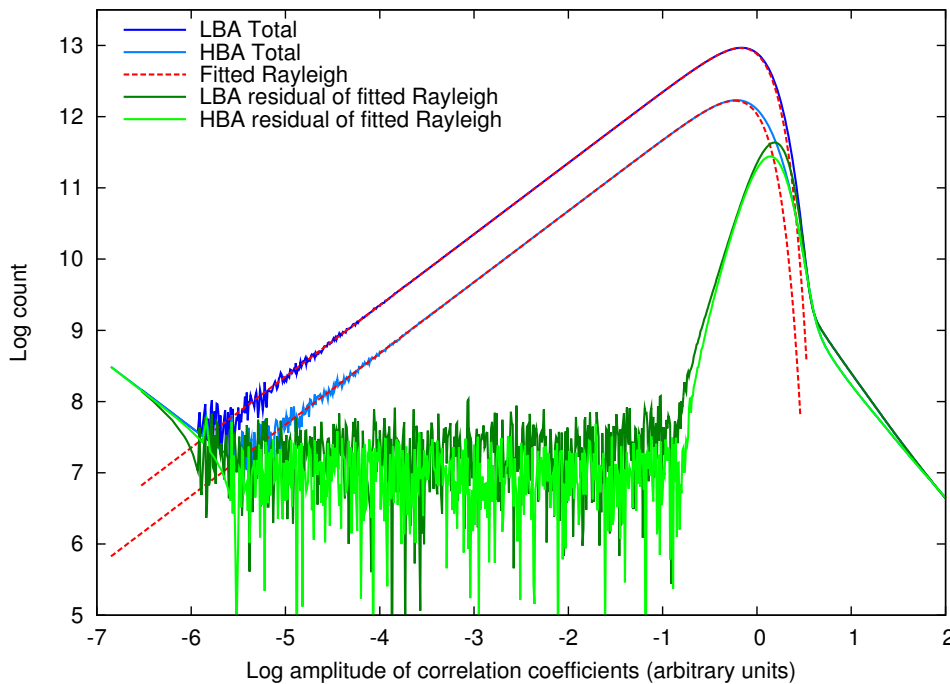


Figure 10. Least-squares fits of Rayleigh distributions to the observed LBA and HBA histograms, after band-pass correction but without flux calibration.

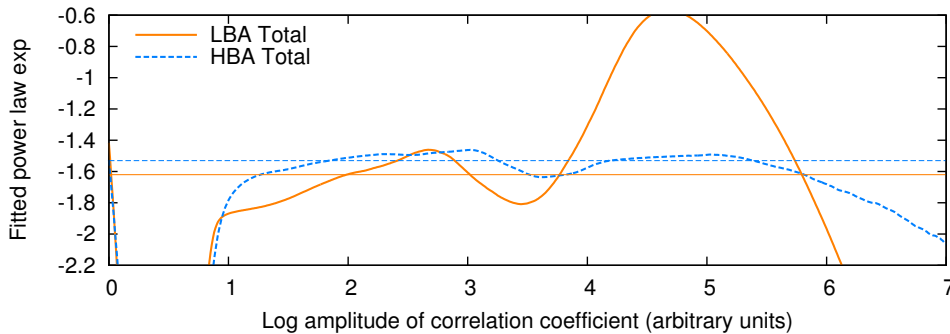


Figure 11. The slope of the band-pass corrected log-log histogram as a function of the brightness. The horizontal lines indicate the fitted slope over the full (semi-) stable region. The horizontal axis is not calibrated.

In projects such as the EoR detection experiment with LOFAR, a simulation pipeline is used to create a realistic estimate of the signal that can be expected. Currently, these simulations do not include the effects of RFI. With the construction of empirical models for the RFI source distributions, we are one step closer to including these effects in the simulation. Using Eq. (18), one can sample a realistic strength of a single RFI source, add the feature to the data and run the AOFlagger. What is still needed for accurate simulation, is to obtain a likely distribution for the duration that one such source affects the data. For example, it is neither realistic that all RFI sources are continuously transmitting nor that they affect only one sample. The RFI detector is highly depending on the morphology of the feature in the time-frequency domain. Finally, the coherency properties of the RFI might be even more important to simulate correctly, but these have been not been explored. However, these have large implications for observations with high sensitivity. This will be discussed in the next section.

The derived values for the average lower level of detected RFI, S_d , show that the AOFlagger has detected a large part of the RFI that is well below the sample noise. In both sets, S_d is more than one order of magnitude below the Rayleigh mode. This can be explained with two of the algorithms it implements. The first one is the `SumThreshold` method (Offringa et al. 2010a), that thresholds on combinations of samples, and is thus able to detect RFI that is weaker than the sample noise. The second one is the scale-invariant rank (SIR) operator (Offringa et al. 2012a). This operator is not dependent on the sample amplitude, but flags based on morphology.

6.1 Implications for very long integrations

In theory, faint RFI could impose a fundamental limit on the attainable noise limit of long integrations. As an example, we will analyse the situation for the LOFAR EoR project. This project will

Brightness and spatial distributions of terrestrial radio sources 13

Table 2. Estimated distribution quantities per data set.

Symbol	Description	LBA set	HBA set
N_{total}	Total number of samples in histogram	8.0×10^{11}	5.4×10^{11}
σ	Rayleigh mode (assumed to be $\text{SEFD}/\sqrt{2\Delta t\Delta\nu}$, where SEFD is the System Equivalent Flux Density)	770 Jy	77 Jy
<i>Estimators for power-law distribution parameters</i>			
α	Exponent of power law in RFI distribution	−1.62	−1.53
$SE(\alpha)$	Standard error of α	2.8×10^{-3}	6.9×10^{-4}
α_H	Hill estimator for power-law exponent	−1.53	−1.53
$SE(\alpha_H)$	Standard error of α_H	8.9×10^{-6}	1.0×10^{-5}
ϵ_α	Sampled estimate of standard deviation of α	6.1×10^{-2}	1.2×10^{-2}
β	Scaling factor of power law with exponent α	4.0×10^{17}	3.4×10^{15}
η	Radiation fall-off speed for α ($\eta = 2$ is free space)	3.23	3.77
<i>Limits</i>			
S_L	Constraint on lower fall-off point of power law	21 mJy	6.2 mJy
\tilde{S}_L	As S_L , but assuming $I_g/r^2 \sim \text{uniform}$	47 mJy	14 mJy
S_d	Expected lowest apparent level of RFI detected	26 Jy	5.7 Jy
$E(S_R)$	Apparent RFI flux density	2700 Jy	140 Jy
$E(S_{\text{leak}})$	Residual apparent RFI flux density after excision	484–496 mJy	167–171 mJy
	Same as above, but by assuming 10% occupancy	384 mJy	120 mJy
REFD	RFI equivalent flux density	18.9–19.3 Jy	6.5–6.7 Jy

use the LOFAR high-band antennas to collect on the order of 50–100 night-time observations of 6 h for a few target fields. The final resolution required for signal extraction will be about 1 MHz. The project will use about 60 stations, each of which provides two polarized feeds. This will bring the noise level in a single 6 h observation in 1 MHz bandwidth to

$$\sigma_{\text{eor-night}} = \text{SEFD} (2\Delta t \Delta\nu N_{\text{feed}} N_{\text{interferometers}})^{-\frac{1}{2}} \approx 250 \mu\text{Jy}. \quad (35)$$

Therefore, after 100 nights the thermal noise level will be 25 μJy .

Because some RFI sources might be stationary, the signals from these sources will add coherently over time. Therefore, the amount that time integration can decrease the flux density of RFI might be limited. Additionally, some RFI sources are received by multiple stations of the array, and by multiple feeds of the individual antennas. Therefore, integrating data from different interferometers and data from the two polarized feeds might not lower the noise level that is caused by RFI at some point. In summary, RFI is unlike normal noise and might be coherent over time, interferometer and feed.

On the other hand, many RFI signals observed in the LOFAR bands have a limited bandwidth. Indeed, the majority of the detected RFI sources affect only one or a few LOFAR channels of 0.76 kHz. Therefore, frequency averaging will lower the flux density of the RFI signal. If the frequency range contains only one stationary RFI source, the strength of this RFI source will go down linearly with the total bandwidth. If we assume that all channels are affected by RFI sources and all these sources transmit in approximately one channel, then the noise addition that is produced by RFI will go down with the square root of the number of averaged channels. This is a consequence of the random phase that different RFI sources have.

In summary, some class of stationary RFI sources are expected to be coherent over time, polarization and interferometer, but not over frequency. Therefore, in this case the noise level at which RFI leakage approximately becomes relevant is given by

$$\sigma_{\text{RFI}} = \frac{\text{REFD}}{\sqrt{2\Delta\nu}}, \quad (36)$$

where REFD is the RFI equivalent flux density at 1 Hz and 1 s resolution for a single station, in analogue to how the SEFD is de-

fined. This only holds when the observational integrated bandwidth $\Delta\nu$ is substantially higher than the average bandwidth of a single RFI source. Otherwise, if the $\Delta\nu$ is small relative to the average bandwidth of RFI sources, some RFI might show up earlier. The empirically found upper limits in this work are $\text{REFD}_{\text{LBA}} = 18.9\text{--}19.3$ Jy and $\text{REFD}_{\text{HBA}} = 6.5\text{--}6.7$ Jy (see Table 2).

For the EoR project with 1 MHz resolution, Eq. 36 results in $\sigma_{\text{RFI}} \approx 4.7$ mJy. However, the first EoR results of observations of one day have approximately reached the thermal noise of about 1.7 mJy per 0.2 MHz sub-band (Yatawatta et al. 2013), and the resulting images show no signs of RFI. This implies that either the upper limit is far from the actual RFI situation, or Eq. 36 is not applicable to most of the RFI that is observed with LOFAR. In the following section we will discuss effects that could cause a reduced contribution of RFI.

6.2 Interference-reducing effects

When integrating data, it is likely that the actual noise limit from low-level RFI will be significantly lower than the given upper limit, which was determined at highest LOFAR resolution, before further averaging. There are several reasons for this, which we will summarize one by one.

There are several effects related to the origin of the interference, which can reduce their effect. One such effect, is when the RFI sources have a variable geometric phase. Many RFI sources have a variable geometric phase, because they move or because their path of propagation changes. This would cause them to sum (partly) incoherently over time, and thus go approximately down with the noise. Moreover, many RFI sources will be seen by only a few stations or are not constant over time, which further attenuates their effect. Finally, for the shortest baselines at 150 MHz, the far field starts around 1 km. Some RFI sources will therefore be in the near field, especially for the longer baselines. In that case, a source will not add up coherently over the interferometers that see the particular source, as the interferometers see it with different phases.

A large number of RFI sources together can also behave like normal incoherent noise. Stationary RFI sources in a uniform spa-

14 *A. R. Offringa et al.*

tial distribution will interfere both constructively and destructively with each other. Individually, the sources will have fixed geometric phases, but because the baselines are much longer than the wavelength, their phases will become uniformly spread. Therefore, they will add incoherently.

Some standard data processing actions can also attenuate RFI, especially fringe stopping and the imaging of data. Fringe stopping interferometers can partly average out RFI sources. Nevertheless, stationary RFI that is averaged out by fringe stopping will leave artefacts in the field centre (Offringa et al. 2012b). This is not relevant when observing the North Celestial Pole — which is one of the LOFAR EoR fields — because no fringe stopping is applied when observing the NCP. Imaging sometimes attenuates RFI, because the Fourier transform that is involved in data imaging will localize the contribution from stationary RFI near the NCP. If RFI artefacts would show in the image of the NCP field, they can easily be detected and possibly be removed, or processing could ignore data near the pole.

Because of the above two arguments, when considering RFI it is a risk to use the NCP as one of the EoR target fields. At the same time, this field is useful for analysing the RFI coherency properties. It is also a simple field to observe with LOFAR, because it is always at high declination in the Netherlands and it does not contain bright foreground sources. Preliminary analysis of EoR NCP observations of a single night have reached the thermal noise (Yatawatta et al. 2013), but do not show leaked RFI at the pole.

On top of all these possible RFI reducing effects, our given RFI constraints could be too pessimistic because we have assumed 100% of the spectrum is occupied by RFI over the entire duration of the observation. If, say, only 10% of the samples are affected by RFI, the expected value of the leaked RFI level decreases by about 25%, and if the detected 2.68% true-positives contain all RFI, there is no leaked RFI at all. With current data, one can only speculate how much of the electromagnetic spectrum is truly occupied.

Finally, future RFI excision strategies will further enhance detection accuracy. Currently, RFI excision is applied only on the raw data from single interferometers at high resolution. Once data from a large number of nights are collected, it will be possible to detect and excise RFI more accurately, by looking at the averaged data from multiple nights and/or multiple interferometers. We have shown that the current detection algorithms can detect RFI well below the noise. Therefore, if data from different nights or interferometers are summed, and there is stationary RFI in the set, it will become detectable if it is consistently present. If it is still below the noise and not detectable, but behaves similar to the RFI we are already seeing, it will act like normal noise and will therefore be harmless.

With the current strategy, it is likely that the LOFAR EoR project will encounter some RFI on some frequencies when averaging lots of observing nights, although this still remains to be seen. To mitigate this leaked RFI, the detection can be executed at higher signal-to-noise levels. The current results indicate that a lot of RFI is not coherent, and the situation is promising. Considering the current RFI results, and the availability of further mitigation steps, we conclude that RFI will likely not be problematic for the detection of the Epoch of Reionisation with LOFAR.

ACKNOWLEDGMENTS

LOFAR, the Low-Frequency Array designed and constructed by ASTRON, has facilities in several countries, that are owned by var-

ious parties (each with their own funding sources), and that are collectively operated by the International LOFAR Telescope (ILT) foundation under a joint scientific policy. Parts of this research were conducted by the Australian Research Council Centre of Excellence for All-sky Astrophysics (CAASTRO), through project number CE110001020. C. Ferrari and G. Maccario acknowledge financial support by the “Agence Nationale de la Recherche” through grant ANR-09-JCJC-0001-01.

REFERENCES

- Acton F. S., 1966, *Analysis of Straight-Line Data*. New York: Dover
- Baan W. A., Fridman P. A., Millenaar R. P., 2004, *AJ*, 128, 933
- Barnbaum C., Bradley R. F., 1998, *AJ*, 115, 2598
- Boonstra A.-J., 2005, PhD thesis
- Bowman J. D., Rogers A. E. E., 2010, *Nature*, 468, 796
- Briggs F. H., Bell J. F., Kesteven M. J., 2000, *AJ*, 120, 3351
- Clauset A., Shalizi C. R., Newman M. E. J., 2009, *SIAM Review*, 51, 661
- Condon J. J., 1984, *ApJ*, 287, 461
- de Bruyn A. G., Brentjens M. A., Koopmans L. V. E., Zaroubi S., Labropoulos P., Yatawatta S. B., 2011, in *Proc. of General Assembly and Scientific Symposium, 2011 XXXth URSI Detecting the EoR with LOFAR: Steps along the road*. IEEE, pp 1–4
- Ellingson S. W., Hampson G. A., 2002, *IEEE Trans. on Antennas & Propagation*, 50, 25
- Flöer L., Winkel B., Kerp J., 2010, in *Proc. of Science, RFI2010, RFI mitigation for the Effelsberg Bonn HI Survey (EBHIS)*
- Fridman P. A., Baan W. A., 2001, *A&A*, 378, 327
- Hata M., 1980, “*IEEE Trans. on Vehicular Technology*”, VT-29
- Hill B. M., 1975, *Ann. Statist.*, 3, 1163
- Jacobs D. C., Aguirre J. E., Parsons A. R., Pober J. C., Bradley R. F., Carilli C., Gugliucci N. E., Manley J. R., van der Merwe C., Moore D. F., Parashare C., 2011, *ApJ Letters*, 734, L34
- Kocz J., Bailes M., Barnes D., Burke-Spolaor S., Levin L., 2012, *MNRAS*, 420, 271
- Lemmon J. J., 1997, *Radio Science*, 32, 525
- Leshem A., van der Veen A.-J., Boonstra A.-J., 2000, *ApJS*, 131, 355
- Niamsuwan N., Johnson J. T., Ellingson S. W., 2005, *Radio Science*, 40
- Offringa A. R., de Bruyn A. G., Biehl M., Zaroubi S., 2010b, in *Proc. of Science, RFI2010, A LOFAR RFI detection pipeline and its first results*
- Offringa A. R., de Bruyn A. G., Biehl M., Zaroubi S., Bernardi G., Pandey V. N., 2010a, *MNRAS*, 405, 155
- Offringa A. R., de Bruyn A. G., Zaroubi S., 2012b, *MNRAS*, 422, 563
- Offringa A. R., de Bruyn A. G., Zaroubi S., et al., 2013, *A&A*, 549
- Offringa A. R., van de Gronde J. J., Roerdink J. B. T. M., 2012a, *A&A*, 539
- Okumura Y., et al., 1968, *Radio Service Rev. Elec. Comm. Lab.*, pp 825–873
- Paciga G., Chang T.-C., Gupta Y., Nityanada R., Odegova J., Pen U.-L., Peterson J. B., Roy J., Sigurdson K., 2011, *MNRAS*, 413, 1174
- Pankonin V., Price R. M., 1981, *IEEE Trans. on Electromagnetic Compatibility, EMC-23*, 308

Brightness and spatial distributions of terrestrial radio sources 15

- Papoulis A., Pillai S., 2001, Probability, Random Variables and Stochastic Processes, 4 edn. McGraw-Hill
- Ryabov V., Zarka P., Ryabov B., 2004, Planetary and Space Science, 52, 1479
- Scheuer P. A. G., 1957, in Math. Proc. of the Cambridge Phil. Soc. 53, A statistical method for analysing observations of faint radio stars. pp 764–773
- Smolders B., Hampson G., 2002, IEEE Antennas & Propagation magazine, 44, 13
- Subrahmanyam R., Ekers R. D., 2002, in Proc. of XXVIIth General Assembly. CMB observations using the SKA. URSI, Maastricht, The Netherlands, p. 710
- Thompson A. R., Gergely T. E., Vanden Bout P. A., 1991, Physics today, 44, 41
- Weber R., Faye C., Biraud F., Dansou J., 1997, A&AS, 126, 161
- Williams C. L., Hewitt J. N., Levine A. M., de Oliveira-Costa A., Bowman J. D., et al., 2012, ApJ, 755, 47
- Yatawatta S., de Bruyn A. G., Brentjens M. A., et al., 2013, A&A, 550