# LOFAR: Calibration and Imaging on JURECA

M. Hoeft, A. Horneffer, A. Drabent, S. Fröhlich

published in

**NIC Symposium 2016**
K. Binder, M. Müller, M. Kremer, A. Schnurpfeil (Editors)

Forschungszentrum Jülich GmbH,
John von Neumann Institute for Computing (NIC),
Schriften des Forschungszentrums Jülich, NIC Series, Vol. 48,
ISBN 978-3-95806-109-5, pp. 69.
http://hdl.handle.net/2128/9842

# LOFAR: Calibration and Imaging on JURECA

**Matthias Hoeft[1], Andreas Horneffer[2], Alexander Drabent[1], and Stefan Fröhlich[3]**

[1] Thüringer Landessternwarte, Sternwarte 5, 07778 Tautenburg, Germany
*E-mail: hoeft@tls-tautenburg.de*

[2] Max-Planck-Institut für Radioastronomy, Auf dem Hügel 69, 53121 Bonn, Germany

[3] Forschungszentrum Jülich, Jülich Supercomputing Centre, 52425 Jülich, Germany

The Low Frequency Array (LOFAR) is a novel radio telescope, operating at very low radio frequencies. It uses a large frequency bandwidth and many stations comprising a multitude of simple dipole antennas. LOFAR stations are relatively small, giving them a large field-of-view (FoV). Moreover, the stations are distributed over many European countries, which gives LOFAR a superb image resolution. With its large bandwidth, large collecting area, large FoV, and high spatial resolution LOFAR achieves unprecedented sensitivity, resolution and in particular high survey speed at this little explored frequency range. The LOFAR data processing is realised via digital electronics starting from the signals of individual dipoles. Each station produces a tremendous amount of data to be sent to a central correlator via fast internet connections, which again leads lo huge amounts of data to be processed further. For LOFAR the analysis is complicated by the ionosphere which can severely disturb the observations. Moreover, the simplicity of the antenna fields introduces distortions that need to be corrected when processing the data. Traditionally, the analysis of radio interferometer data has been carried out interactively on work stations. The enormous amounts of data produced by LOFAR and the complex data reduction schemes demand much larger computing resources and automated processes. The data reduction, thus, needs to be carried out on supercomputers. Our aim is to adapt the LOFAR software for analysing data on JURECA, to develop a framework which allows astronomers to set up a data reduction including observation specifics. Moreover, with selected observations we demonstrate the feasibility of reducing LOFAR data on general-purpose, multi-user supercomputers such as JURECA. The algorithms developed for LOFAR pave the way to a new generation of powerful radio telescopes at low frequencies.

## 1 The Low Frequency Array

The *Low Frequency Array* (LOFAR) is a novel radio telescope[3], mainly developed and operated by the Dutch Institute for Radio Astronomy, ASTRON. LOFAR observes in the largely unexplored frequency range from 10 to 240 MHz. Instead of having large dishes it consists of simple dipole antennas grouped into stations. With several new stations becoming operational in 2015 (one close to Hamburg, three in Poland) LOFAR now comprises more than 100.000 dipoles distributed over Dutch and international stations. Each station operates as a phased array, i.e. the beam and the direction of the beam is set by delays when combining the signals of the individual dipoles (*beam-forming*). LOFAR allows us to observe in many directions at the same time, since the delays are set when processing the data: the processes can run in parallel as often as the processing and data transport resources permit. The beam-formed data are sent from each station to a central correlator via fast internet connections, e.g. international stations are equipped with 10 Gbit/s data links. Via cross-correlations and averaging the correlator combines the incoming data to *visibility* data.

The collecting area and the frequency coverage is maximised to increase the telescope sensitivity. LOFAR samples the signals from each antenna with a rate of 200 MHz, thus the maximum frequency window allowed by the Nyquist sampling theorem is 100 MHz. The maximum bandwidth a LOFAR station can transmit is 96 MHz, which can be spread over several directions. On the other hand, a very high frequency resolution is necessary to mitigate Radio Frequency Interferences (RFI) and to avoid smearing out signals on long baselines provided by international LOFAR stations. A typical frequency resolution used in LOFAR observations is 3 kHz. The number of frequency channels, $N_k$, is consequently of the order of 30,000. The collecting area is determined by the number of dipoles used. The diameter of the station determines also the size of the station beam, i.e. the area on the sky visible with one pointing (field-of-view, FoV). The larger the FoV the less pointings are necessary to survey the entire sky. The diameter of each station is chosen as a compromise between survey speed and sensitivity. LOFAR currently uses stations with a diameter of about 50 and 70 metres and 50 stations in total (38 stations in the Netherlands, six stations in Germany, three in Poland, and one in France, UK and Sweden each). The large frequency coverage and the large collecting area allows LOFAR to increase in sensitivity with respect to earlier telescopes by a factor of about 100.

The resolution of the instrument is given by the largest distances between the stations. With stations in UK, France, Sweden and Poland LOFAR has baselines of thousands of kilometres, permitting sub-arcsecond resolution in this frequency regime. Within one observation a FoV of several degrees can be imaged, see Fig. 1, with a resolution even below one arcsecond. The image fidelity also depends on baselines with intermediate distances. Covering baselines from several tens of metres to thousands of kilometres LOFAR allows us to reconstruct the sky brightness distribution with unprecedented imaging capabilities.

The fundamental operation of a radio interferometer is to cross-correlate and average the beam-formed signals of all stations. More precisely, the signals of all pairs of stations are multiplied and averaged by the central correlator. As noted above, the sampling rate for LOFAR is 200 MHz, the typical averaging time is one or two second(s). For each antenna pair, for each frequency channel, and for each averaging time step the correlator computes one complex number (in fact four complex numbers since we have to deal with two polarisations), $\mathcal{V}_{ij}, (t_n, \nu_k)$, where $ij$ refers to the baseline between the stations $i$ and $j$, $t_n$ refers to the time step $n$, and $\nu_k$ refers to the frequency channel $k$. Initially, a BlueGene/L supercomputer, ranking number six in the TOP 500[a] list of the fastest international supercomputers when built, was used as correlator. In 2008 it was replaced with a BlueGene/P. The correlation process itself is excellently suited for Graphics Processing Units (GPUs). So when support from IBM for the BlueGene/P ran out, a NVIDIA Tesla K10 cluster, named COBALT, has taken over the task of correlating the LOFAR station signals. The maximum input rate is 240 Gbit/s and the maximum output rate is 80 Gbit/s. The total number of visibilities produced for a typical 12 hours observation is

$$N_{\rm pol} \times \frac{N_{\rm st}\,(N_{\rm st}-1)}{2} \times N_n \times N_k, \tag{1}$$

with $N_{\rm pol}$ the number of polarisations, $N_{\rm st}$ the number of stations, $N_n$ the number of integration time steps, and $N_k$ the number of frequency channels. In total this implies $10^{13}$ complex visibilities. Data of the order of 100 TByte have to be stored and further processed for one typical LOFAR observation that includes all available stations.

[a] http://www.top500.org/lists/2005/06/

## 2 Reconstructing the Sky Brightness Distribution

LOFAR is a very flexible instrument, allowing quite different modes of operation. The instrument is primarily operated in the 'interferometer mode' as described above with the aim to reconstruct the sky brightness distribution in the frequency range used by LOFAR. In an ideal case the visibility data, as generated by the correlator, depend on the sky brightness distribution, $I(l, m, n)$, according to:

$$\mathcal{V}_{ij}(u, v, w) = \int \frac{\mathrm{d}l\, \mathrm{d}m}{\sqrt{1 - l^2 - m^2}} \sqrt{A_i(l, m) A_j(l, m)}\, I(l, m) \times e^{-2\pi i \left[ul + vm + w(\sqrt{1 - l^2 - m^2} - 1)\right]}, \tag{2}$$

where $l, m$ gives the sky direction, $u, v, w$ encodes the separation and direction from station $i$ to station $j$, and $A_i(l, m)$ the station sensitivity into direction $l, m$, (*station beam*). In case of $l, m \ll 1$ (or $w = 0$) the relation between sky brightness distribution and visibility data becomes very close to a Fourier transform. Therefore, with a planar array, a small FoV, and no distortions we could simply apply a Fourier transform to reconstruct the sky brightness

$$\text{Visibility} \overset{\mathcal{FT}}{\longleftrightarrow} \text{Sky Brightness} \tag{3}$$

However, there are many aspects which need to be taken into account when reconstructing the sky brightness from LOFAR data.

### 2.1 Data Preparation

Man-made radio signals may affect or dominate measured visibilities. This RFI typically occurs either at certain times and affect the entire spectrum (broad band RFI), or may be present for long times but only in a small frequency range (narrow band RFI, e.g., transmitters for communication). Often RFI also shows a very complex time-frequency behaviour. It is essential to remove RFI (*flagging*) before further processing the data since it may significantly distort the sky brightness reconstruction. The size of the LOFAR data prohibit a manual inspection, so RFI needs to be identified automatically with using the high frequency and time resolution as produced by the correlator. For LOFAR the special software (AOFlagger), which allows to automatically flag the data affected by RFI[7], was developed.

One undesired effect of the phased array technique at LOFAR is that the station beams have strong side lobes. Therefore, bright radio sources many degrees away from the actual target may significantly affect the visibilities when located at a side lobe. This causes rapid visibility variation in frequency and time direction. For LOFAR an algorithm called `demixing` has been developed which allows us to subtract those sources before averaging[10].

### 2.2 Nobody Is Perfect: Calibration

Signal modifications may be caused by the amplifiers and the processing of the analog signal before it gets digitised. Since each antenna has its own electronics which might be affected by the environment, e.g., variations of temperature or humidity, these distortions are impossible to predict perfectly. In addition to changing the signal as such, the variations between the dipoles in a phased array like a LOFAR station also affect the shape of the station beam.

At the low radio frequencies used by LOFAR the ionosphere has a significant impact. The electron column density along the line of sight determines the "refraction index" of the ionosphere, with two dominant effects on the waves propagating through the ionosphere: introducing a delay and rotating the polarisation direction. Since the column density varies with time and position, the disturbances of the visibilities are extremely difficult to determine. So correction factors have to be derived from the observations, since the ionosphere cannot be measured directly with sufficient accuracy. These correction factors vary from station to station, with time, and across the field of view.

The signal propagation from the source to the telescope can be represented as a series of $2 \times 2$ complex matrices[4]. Typical parameters are amplitude and phase for each of the two polarisations and a polarisation rotation angle, all for each station, resulting in hundreds of parameters to solve for in each time and frequency interval. Calibrating on the measured data itself requires a reasonable model of the sky brightness distribution. One calibration step is done by solving for parameters in these matrices that can convert this model to the measured values and then inverting the matrices and multiplying them to the measured data.

Calibration of interferometer data usually is direction independent. It uses only one set of parameters for the full field of view. In contrast to many other radio telescopes LOFAR has a large field of view –which increases the survey speed– but in particular the ionosphere introduces disturbances which vary across the field of view. Dealing with these direction dependent effects is done by solving for solutions in the direction of several bright sources within the field of view and then interpolating between those directions[6] or imaging only small facets around the bright sources. Correcting for these direction dependent effects is a major task when processing LOFAR data on JURECA, and is a field of ongoing research.

### 2.3 Imaging and Deconvolution

As any other radio interferometer, LOFAR measures only visibilities at $u, v, w$ coordinates determined by the antenna positions. Hence, the $uv$-plane is only sparsely sampled. This can be denoted with a function $B(u, v)$, which is 1 for $(u, v)$ pairs for which a visibility is actually measured and 0 otherwise. In other words, the 'ideal' visibilities are masked by the $uv$-coverage $B(u, v)$,

$$\text{Visibility} \times B(u, v) \quad \overset{\mathcal{FT}}{\longleftrightarrow} \quad \text{Sky Brightness} \otimes \text{PSF}, \tag{4}$$

where the Fourier Transform of the $uv$-coverage results in the point spread function (PSF), which can be quite complicated. Hence, the brightness distribution obtained after the Fourier Transform can be understood as a convolution of the true sky brightness distribution and the PSF. In order to restore the true sky brightness distribution we 'deconvolve' the initially obtained image. A standard *deconvolution* scheme is the *CLEAN* algorithm. In its simple version[5] the position of the maximum brightness in the image is successively determined, assumed to represent a point of true sky brightness distribution. The PSF is then subtracted from the residual image at this position. A more precise version[8] interleaves *minor* cycles of simple CLEAN with *major* cycles in which the new sky brightness

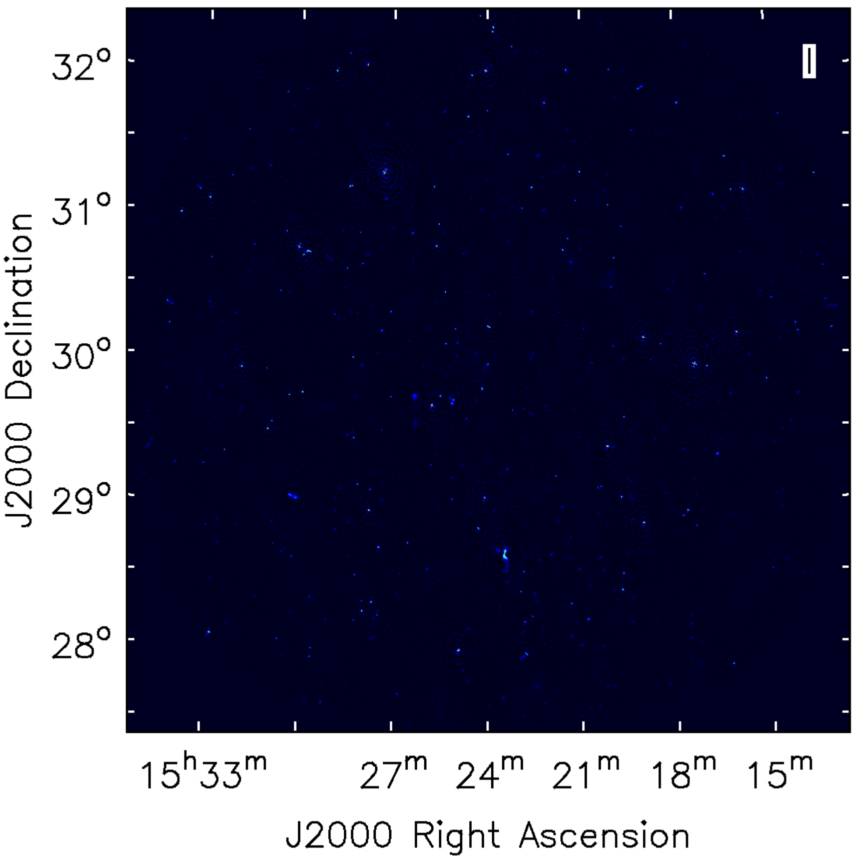


Figure 1. The field of the galaxy cluster Abell 2069 and the Corona Borealis Supercluster observed with LOFAR HBA in May 2014. This map was created by imaging frequency-limited chunks of the whole data and stacking the images afterwards. 100 out of 370 subbands was used. Image resolution is $106'' \, times \, 103''$.

components are transformed to Fourier space and subtracted from the visibilities. The new residual image is then generated by transforming the residual visibilities to image space. Hence, the deconvolution is already a complex process in which the image needs to be computed from the visibilities many times.

Since we work with large amounts of data in all imaging steps a Fast Fourier Transform needs to be applied. The field of view of the LOFAR stations is as large as several degrees. On the other hand the resolution, determined by the longest baselines, can be of the order of arcseconds or even below. Even if one is only interested in a comparatively small object the full field of view needs to be imaged in order to be able to deconvolve all sources, which would otherwise generate image artefacts all over the image. The image size must therefore be of the order of several ten thousand pixels in both directions.

The LOFAR stations have a large field of view, which prohibits to ignore the $w$-term in Eq. 2. Also the beam shape of the stations $A_i(l, m)$ and the polarisation properties of the LOFAR dipoles depend on the pointing direction relative to the station, so they vary

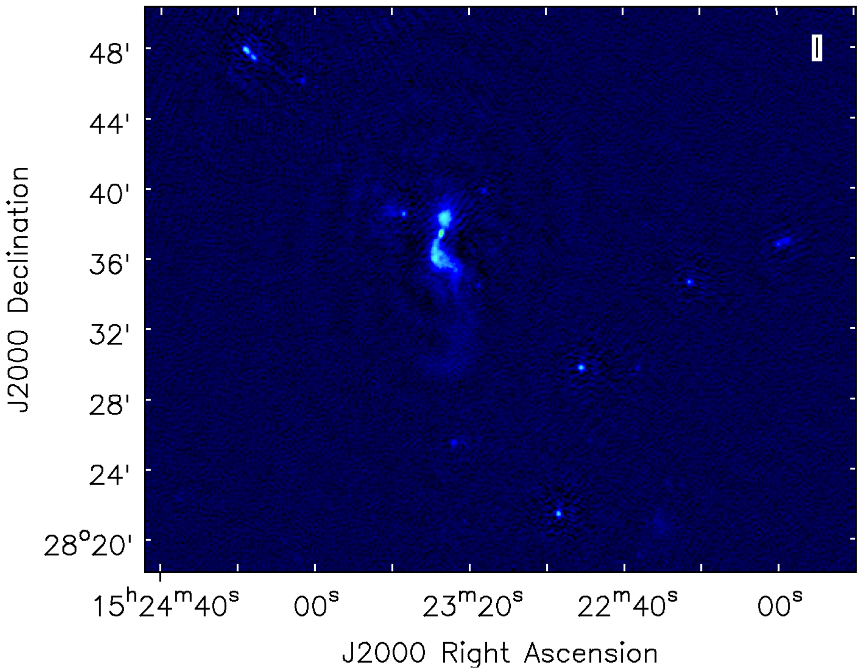


Figure 2. The radio galaxy 4C+28.39 in the field of the galaxy cluster Abell 2069 and the Corona Borealis Supercluster.

with time when tracking a source on the sky. Therefore, the sky brightness for LOFAR cannot be obtained via simply Fourier transforming the visibility data. The visibility data cannot be corrected for direction dependent effects, they have to be corrected in the images, hence a dedicated imager for LOFAR has been developed, the `awimager`[9]. The direction dependent effects are multiplicative in image space but they vary in time and frequency so correcting in image space would require Fourier transforming many time and frequency bins separately. They can be corrected for by a convolution in Fourier space, called the A-projection, and because the effects vary only on large scales in direction this keeps the support of the convolution kernel small, making it more efficient than many Fourier transforms[1]. Eq. 2 shows that the $w$-term can be seen as a multiplicative factor in image space, thus it can also be corrected for by a convolution in Fourier space[2]. Those two corrections together give the aw-projection.

## 2.4 The Processing Flow: Selfcal

Usually the sky brightness distribution of the target is not fully known. The calibration factors for the instrument can be generated by observing and calibrating on a well known and bright calibrator source, but ionospheric corrections are different for different directions. Fortunately even calibrating on an imperfect model of the sky (e.g. with low resolution) results in an improved image. Therefore several calibration–imaging loops are usually necessary before the sky model converges to a final result. This process is called *selfcal*.

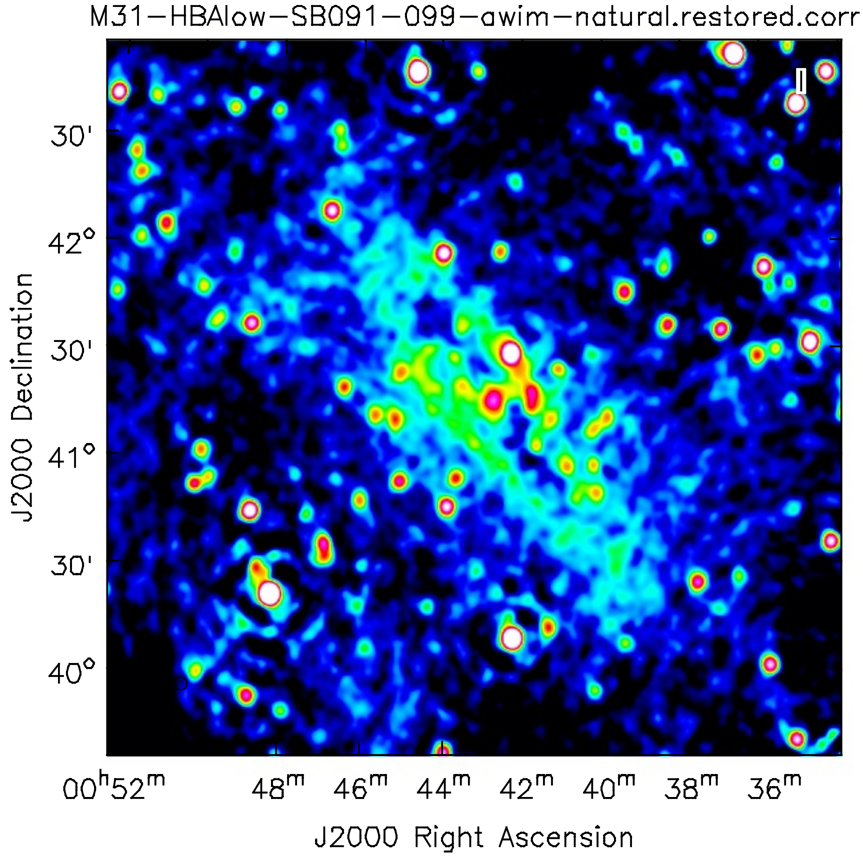


Figure 3. Image of the Andromeda galaxy M31 with LOFAR at 130 MHz. The data for this image was calibrated on JUROPA and imaged on JURECA.

The exact setup of these loops depends on the data and the scientific objective. E.g., at low frequencies the calibration needs to be done on high time resolution in order to track the changes in the ionosphere, while for the imaging the data can be averaged much more before the image is significantly distorted, so it is useful to add an averaging step to speed up processing.

The major steps of creating a sky image out of the measured visibility data are therefore: (i) flagging of RFI, (ii) demixing of bright sources in the vicinity of the target, (iii) averaging the data, (iv) using calibrator sources to calibrate the absolute amplitude of the measured visibilities, (v) image the data and successively determine variations in time, frequency, and across the field of view of the calibration matrices, and (vi) create the final image of the scientific target. For normal users the first three steps are usually done by the observatory right after observation. This minimises the amount of data to be transferred and stored to typically a few ten TByte for each observation. The computing power at JURECA allows us to do those steps on our own, freeing up computing resources at the observatory for more observations and allowing us to perform a more involved processing.

## 3 Pipeline Framework and Image Examples

In our project we aim for enabling processing of LOFAR data on JURECA. To get good results it is still necessary to adopt the calibration strategy to a specific observation. Many parameters depend on brightness of the sources in the field and the properties of the actual target, e.g., size, morphology and brightness. Figuring out the best sequence of processing steps is still manual work, but once they are known the steps can be added together into a pipeline that runs without human interaction. To run this pipeline a pipeline framework was developed, that executes several consecutive processing steps and takes care of distributing the workload to the processing nodes and generating checkpoints. The original framework is too cumbersome to adopt to different processing strategies and thus was only used for standard processing by the observatory but not by the astronomers. On the other hand simple scripts do not make efficient use of a supercomputer like JURECA. To ease pipeline development we developed an extension to the framework that allows to specify several processing steps in one configuration file in a user-friendly manner. This makes it possible to rapidly specify a processing strategy and have it executed in an efficient manner.

Fig. 3 shows the Andromeda galaxy at 130 MHz, after only direction independent calibration. The direction independent calibration is insufficient, which can be seen in the ring-like artefacts around strong sources. Currently, within the LOFAR community we are developing a strategy which allows to do direction dependent calibration in an automated fashion. This pipeline determines calibration factors in the direction of many bright sources by subtracting all other sources and working on only a small part of the field, the facet calibration. Due to its large size with extended emission of more than $3^\circ$ adopting this strategy to the observation of the Andromeda galaxy will be a challenge.


## Acknowledgements

The authors gratefully acknowledge the computing time granted by the John von Neumann Institute for Computing (NIC) and provided on the supercomputer JURECA at Jülich Supercomputing Centre (JSC).


## References


1. S. Bhatnagar, T. J. Cornwell, K. Golap, and J. M. Uson, *A&A* **487**, 419, 2008.
2. T. J. Cornwell, K. Golap, and S. Bhatnagar, *IEEE J. Select. Top. Signal Proc.* **2**, 647, 2008.
3. M. P. van Haarlem, M. W. Wise, A. W. Gunst, et al., *A&A* **556**, 2, 2013.
4. J. P. Hamaker, J. D. Bregman, and R. J. Sault, *A&AS* **117**, 137, 1996.
5. J. A. Högbom, *A&AS* **15**, 417, 1974.
6. H. T. Intema, arXiv:1402.4889, 2014.
7. A. R. Offringa, J. J. van de Gronde, and J. B. T. M. Roerdink, *A&A* **539**, A95, 2012.
8. F. R. Schwab, *AJ* **89**, 1076, 1984.
9. C. Tasse, S. van der Tol, J. van Zwieten, G. van Diepen, and S. Bhatnagar, *A&A* **553**, 105, 2013.
10. S. van der Tol, B. D. Jeffs, and A.-J. van der Veen, *IEEE Transactions on Signal Processing* **55**, 4497, 2007.