Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **541,** 3993–4010 (2025) https://doi.org/10.1093/mnras/staf1206
Advance Access publication 2025 July 28

Robust direction-dependent gain-calibration of beam-modelling errors far from the target field

S. A. Brackenhoff [1]★ A. R. Offringa,[1,2] M. Mevius [2] L. V. E. Koopmans [1] J. K. Chege,[1,2] E. Ceccotti [1,3] C. Höfer,[1] L. Gao,[1,4] S. Ghosh,[1] F. G. Mertens [1,5] and S. Munshi [1]

[1]*Kapteyn Astronomical Institute, University of Groningen, PO Box 800, NL-9700 AV Groningen, the Netherlands*
[2]*Netherlands Institute for Radio Astronomy (ASTRON), PO Box 2, NL-7990 AA Dwingeloo, the Netherlands*
[3]*INAF – Istituto di Radioastronomia, Via P. Gobetti 101, I-40129 Bologna, Italy*
[4]*Liaoning Key Laboratory of Cosmology and Astrophysics, College of Sciences, Northeastern University, Shenyang 110819, China*
[5]*LUX, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Université, F-75014 Paris, France*

Accepted 2025 July 18. Received 2025 July 11; in original form 2025 April 4

ABSTRACT

Many astronomical questions require deep, wide-field observations at low radio frequencies. Phased arrays like LOFAR and SKA-Low (low band part of the Square Kilometre Array) are designed for this, but have inherently unstable element gains, leading to time, frequency, and direction-dependent gain errors. Precise direction-dependent calibration of observations is therefore key to reaching the highest possible dynamic range. Many tools for direction-dependent calibration utilize sky and beam models to infer gains.

However, these calibration tools struggle with precision calibration for relatively bright (e.g. A-team) sources far from the beam centre. Therefore, the point spread function of these sources can potentially obscure a faint signal of interest. We show that, and why, the assumption of a smooth gain solution per station fails for realistic radio interferometers, and how this affects gain-calibration results. Subsequently, we introduce an improvement for smooth spectral gain constraints for direction-dependent gain-calibration algorithms, in which the level of regularization is weighted by the expected station response to the sky model. We test this method using direction-dependent calibration method DDECAL and physically motivated beam-modelling errors for LOFAR-HBA (High-Band Antennas of the Low Frequency Array) stations. The new method outperforms the standard method for various calibration settings near nulls in the beam, and matches the standard inverse-variance-weighted method's performance for the remainder of the data. The proposed method is especially effective for short baselines, both in visibility and image space. Improved direction-dependent gain calibration is critical for future high-precision SKA-Low observations, where higher sensitivity, increased antenna beam complexity, and mutual coupling call for better off-axis source subtraction, which may not be achieved through improved beam models alone.

Key words: methods: data analysis – techniques: interferometric – software: simulations – dark ages, reionization, first stars.

1 INTRODUCTION

To accurately characterize weak astronomical signals in large, complex data sets, it is imperative to separate these signals from radiation originating from other astronomical or terrestrial sources. Such contaminants can be man-made sources of interference, or other sky sources that emit radiation in similar parts of the electromagnetic spectrum. Advanced signal separation methods are needed, because any leakage or confusion may leave the signal of interest obscured. To filter the interfering signals out of the data, they must be well characterized.

This makes identifying and characterizing known sources in the radio sky an important endeavour in sky model-based radio-interferometric calibration. Many catalogues of radio sources that can be used to build a sky model for the field of interest have been constructed, such as the Cambridge surveys (Edge et al. 1959; Pilkington & Scott 1965; Baldwin et al. 1985; Hales et al. 2007), the WEsterbork Northern Sky Survey (Rengelink et al. 1997), the NRAO VLA Sky Survey (Condon et al. 1998), the VLA Low-frequency Sky Survey (Cohen et al. 2007), the TIFR GMRT Sky Survey (Intema et al. 2017), the GaLactic and Extragalactic All-sky MWA survey (Hurley-Walker et al. 2017), the LOFAR Two-metre Sky Survey (Shimwell et al. 2017, 2022), and the LOFAR LBA Sky Survey (de Gasperin et al. 2023).

However, knowledge of the on-sky sources alone is not sufficient to properly remove them from the data, because propagation and instrumental effects affect their signature in the data. For example, ionospheric effects can cause the sources to scintillate (Cronyn 1972; Koopmans 2010). Furthermore, the chromatic components of the instrument itself, such as the bandpass filter in LOFAR, may cause imprints on all data (de Gasperin et al. 2019). Calibration

★ E-mail: brackenhoff@astro.rug.nl

attempts to correct for these effects. For wide fields of view, the most important are the temporally and spatially varying ionospheric and primary beam effects (Lonsdale 2005; Smirnov 2011b). The two main approaches to removing these are demixing (van der Tol, Jeffs & van der Veen 2007) and direction-dependent (DD) calibration. Although demixing (i.e. self-calibration on visibilities that have been phase rotated to the source that needs to be removed) can be effective for removing very bright off-axis sources, it is not as versatile as DD-gain calibration. Demixing is only able to calibrate for a single direction at a time, and the source must dominate the field around it. This places constraints on the minimum source brightness and the minimum separation from other (bright) sources in the vicinity. DD-gain calibration, on the other hand, can cover a larger range of signal-to-noise ratios and source separations, but is computationally more expensive. In DD-gain calibration, different gains are calculated in the directions of groups of sources, rather than for the full sky. This allows for the necessary flexibility to account for DD errors, but also increases the number of degrees of freedom in the calibration process, which creates a risk of overfitting in the gain solutions (Patil et al. 2016; Ewall-Wice et al. 2017; Mouri Sardarabadi & Koopmans 2019). Such an overfit can lead to suppression of the signal of interest. Therefore, additional constraints have to be introduced to regularize the calibration solutions. Because both propagation effects and individual receiver gains are physically expected to be smooth over a large range of frequencies, one of the most commonly imposed constraints is spectral smoothness. Spectral smoothness is a permissible assumption for most ionospheric effects, but can falter in areas where the instrument primary beam is highly spectrally irregular (Trott & Wayth 2016). Such spectral irregularities occur in phased array stations that consist of several antenna elements that are coherently summed to create a larger station.

Typically, instruments are designed to have a beam that is spatially smooth near the target direction, but this is not necessarily the case outside the main lobe of the primary beam. Especially for phased arrays, this beam can have a very complex spatial structure, such that even closely spaced parts of the sky experience vastly different levels of beam attenuation. This is especially the case near areas where destructive interference results in sharp rises in attenuation, called beam nulls. Because of the chromatic nature of the instrument beam, and the rotation of the Earth relative to the sky over time, these attenuations also vary spectrotemporally. As a result, the assumption of smoothness is not well satisfied in areas where the beam changes rapidly as a function of direction, time, and frequency. When strong sidelobes (areas of low beam attenuation outside the main lobe) are present and gain-calibration errors are made due to the incorrectly imposed spectral smoothness constraint, the residuals of ill-separated bright interfering sources can still drown out the signal of interest. This can occur even if these sources are spatially separated from the target direction by an angle several times greater than the width of the main lobe of the primary beam[1] (Franzen et al. 2016; Gasperin et al. 2019; Trott 2021; Munshi et al. 2024). Low-frequency instruments, which tend to have higher off-axis gains and large fields of view, run an especially high risk of having such sources cross their sidelobes.

One way to restore the adherence to the spectral smoothness assumption is to simply impose a weaker constraint (i.e. to reduce the scales on which the gains are expected to be spectrally smooth). While this can help, it also increases the number of degrees of freedom, leading to an increased risk of overfitting, which the smoothness constraint was introduced to prevent.

Another possibility is to calibrate with an improved model for the beam, such that the most rapid variations are already built into the beam model rather than absorbed into the gains (e.g. Line et al. 2018; Iheanetu et al. 2019; Nunhokee et al. 2020). This is challenging for phased arrays, however, as their beams change as a function of pointing direction, such that models with many parameters are needed to describe all possible pointings. Most interferometric beam models used in calibration therefore assume identical elements. The station beam is computed by using a full electromagnetic simulation of an element beam, and imposing the station structure through a sum of progressive phase shifts (the 'array factor'), ignoring variations in mutual coupling between antenna elements in the array (Wijnholds et al. 2019). This type of model describes the general behaviour of the beam well, but suffers from possible instability. If the station has a regular structure, a beam model created with identical elements has deep nulls and strong sidelobes, that may shift or disappear when the station symmetry is broken. This can occur when there are variations in the element beam patterns (e.g. due to mutual coupling), when there are gain variations between elements, or when elements break. Broken symmetries have a smoothing effect on the beam pattern. When the beam model used in calibration is less smooth than the true beam, calibration can imprint the structure of the model beam onto the data.

Models that do incorporate mutual coupling effects are under construction (O'Hara et al. 2025), but full numerical electromagnetic simulations are not feasible to the required accuracy, because the *in situ* behaviour of the individual antenna elements is difficult to predict (e.g. Wilensky et al. 2024). Also, failures or degradations within the station may occur. These are not straightforward to detect post-commissioning (Chokshi et al. 2024). Another option for creating improved beam models is holography using a bright source. One can introduce an artificial bright source (e.g. drone). This is often impossible, however, both because of the many holographic measurements that would be needed to cover the range of possible pointing directions, and because of the distance to the far field. This far field can start kilometres up in the air due to the station sizes, such that flying a sufficiently bright calibrator source over the instrument is not feasible (Jacobs et al. 2017). Using satellite carrier emission or reflections instead is a possibility, but these signals are typically narrow band and therefore cannot be used to create an accurate model across the full spectral band of the instrument. Finally, bright sky sources can be used, but only in a limited set of directions (Nunhokee et al. 2020). To create a model across the full band and for all pointings, most modern radio observatories opt for simplified beam models instead. However, these falter in the presence of real-world effects, such as broken antenna elements, drifting or degraded amplifiers, or electromagnetic coupling within the station. Although these models are useful to first order and sufficient in many cases, they do leave errors which are prohibitive in the most ambitious science cases.

Errors in the beam model are an issue for all interferometers, but they are most prominent for regular phased arrays, where any gain error can break the symmetry within a station and therefore change the shape of the primary beam not just quantitatively, but also qualitatively. This is the case for the stations of instruments such as the High-Band Antennas of the Low Frequency Array (LOFAR-HBA, van Haarlem et al. 2013), the New Extension in Nançay Upgrading LOFAR (NenuFAR, Zarka et al. 2012), and the Murchison Widefield Array (MWA, Bowman et al. 2013). Furthermore, it is also a prominent issue for interferometers with complex wide-field

[1]This is especially the case for the brightest radio sources sometimes referred to as the A-team, such as Cassiopeia A, Centaurus A, Cygnus A, Fornax A, Hercules A, Hydra A, Pictor A, Taurus A, and Virgo A.

antenna beams, such as the upcoming low band part of the Square Kilometre Array (SKA-Low, Dewdney et al. 2009). Beam-modelling errors affect a wide range of high dynamic range science cases. For example, transient surveys (Murphy et al. 2013), the observation of the faint large-scale structure traced out by neutral hydrogen (Hale et al. 2019), or the creation of deep extragalactic surveys (Best et al. 2023). There is also the continuing effort to detect the redshifted 21-cm signal from the Epoch of Reionization (EoR) and Cosmic Dawn with the aforementioned instruments (e.g. Koopmans et al. 2015; Trott et al. 2020; Munshi et al. 2024; Mertens et al. 2025).

We illustrate the effect of DD beam-modelling errors on the gains recovered during DD-gain calibration using the EoR science case and the LOFAR-HBA instrument, because this example is especially affected by DD errors (Patil et al. 2016; Ewall-Wice et al. 2017; Sardarabadi & Koopmans 2019). These errors are likely one of the dominant causes of unwanted residual emission ('excess variance') seen in the data after sky model subtraction with the DD-gain solutions applied (Gan et al. 2022; Mertens et al. 2025). Beam errors have also been shown to be limiting this science case on similar instruments such as the MWA (Chokshi et al. 2024) and NenuFAR (Munshi et al. 2024). Because the short baselines that are the main focus of interferometric EoR science are not strongly affected by the ionosphere (Brackenhoff et al. 2024), we can focus mainly on the beam itself.

We present the impact of realistic beam errors on LOFAR-HBA with a PYTHON code named SHIMMERR (Station Heterogeneity Impact on Multi-dimensional beam-Modelling Errors simulatoR and calibratoR).[2] We create beam errors by simulating that a number of individual antenna elements within the station are switched off, as this is expected to be a dominant source of beam-modelling errors and differences between individual LOFAR-HBA stations. We calculate the resulting beam changes, the effects they have on the visibility data, and the impact on the spectrally regularized calibration of sources far from the target direction. This is the first simulation that shows the full end-to-end effect of beam errors on LOFAR-HBA DD calibration on the scale of a full night of observations. Furthermore, we present a first potential heuristic solution to the broken spectral smoothness assumption in DD-gain calibration and show that this solution improves upon the standard method both in visibility and image space. In this more robust regularization method, the regularizer is weighted by the forward prediction of the visibility contribution of each direction, such that time- and frequency intervals with strong beam attenuation do not negatively affect the gain solutions at other times and frequencies. We use the observation and calibration settings from LOFAR EoR to illustrate the impact of beam errors, but emphasize that these results apply to many science cases and instruments, especially those needing the shorter baselines.

This paper is structured as follows: the details of DD-gain calibration and the errors introduced by incorrectly assuming spectral smoothness are explained in Section 2. We also elaborate on the heuristic to mitigate this issue here. In Section 3, the simulation and calibration setups are explained. The results of the calibration are discussed in Section 4, where we first discuss how the standard and proposed method impact the obtained gains. We then examine how errors in these gains manifest in visibility space, and finally, we assess their impact on images. We focus on the shortest baselines in particular, as this type of DD calibration is uniquely equipped to calibrate these. We conclude the paper in Section 5.

[2]https://github.com/Stefanie-B/shimmerr

Notation. Throughout this work, we indicate vectors with lowercase bold italic typesetting ($\boldsymbol{a}$), unit-length vectors with a hat ($\hat{\boldsymbol{a}}$), and matrices with uppercase bold sans-serif typesetting ($\mathbf{A}$). The complex conjugate, complex conjugate transpose, and matrix inverse are indicated by $(\cdot)^*$, $(\cdot)^H$, and $(\cdot)^{-1}$, respectively. Averages are denoted by $\langle\cdot\rangle$, the Frobenius norm (the square root of the sum of squares of the elements of a matrix) is denoted by $\|\cdot\|_F$. The inner product is denoted with $\cdot$, the Hadamard product by $\odot$, and the imaginary unit by i.

2 DIRECTION-DEPENDENT GAIN-CALIBRATION OUTSIDE THE MAIN BEAM LOBE

In radio astronomy, the sky is described by its brightness distribution, denoted by $I(l, m)$, where l and m are direction cosines. $I(l, m)$ represents the specific intensity as a function of sky direction. Under the assumptions of spatial incoherence of $I(l, m)$ and quasi-monochromatic radiation, the van Cittert–Zernike theorem provides the theoretical basis for radio interferometry. The theorem states that the spatial coherence of the electric field (i.e. a visibility) is given by the Fourier transform of the sky brightness distribution I (Thompson, Moran & Swenson 2001). A baseline (i.e. the vector connecting two receiver elements) with components (u, v, w) in units of wavelength measures a visibility $V(u, v, w)$,

$$V(u,v,w) = \iint A(l,m) I(l,m) \exp\left\{-2\pi i \left[ul + vm + w\left(\sqrt{1-l^2-m^2} - 1\right)\right]\right\} \frac{\mathrm{d}l\mathrm{d}m}{\sqrt{1-l^2-m^2}}. \tag{1}$$

Here, l, m are direction cosines on the sky, and $A(l, m)$ is the beam response of the receiver.

Whilst the van Cittert–Zernike theorem is an excellent tool for understanding the physical meaning of visibilities, it has its limitations, because it can only describe scalar fields (i.e. no polarization effects), and does not include instrumental or propagation effects. The van Cittert–Zernike theorem can be generalized to the radio interferometric measurement equation (RIME, Hamaker, Bregman & Sault 1996; Smirnov 2011a). For a single point source, the RIME is given by

$$\mathbf{V}_{p,q}(\nu, t) = \mathbf{P}_p(\nu, t)\mathbf{C}_{p,q}(\nu, t)\mathbf{P}_q^H(\nu, t) + \mathbf{N}_{p,q}(\nu, t). \tag{2}$$

In this equation, $\mathbf{V}_{p,q}$ is a 2×2 matrix that describes the visibility on the baseline between receiver elements p and q at time t and frequency ν with the Jones formalism. $\mathbf{P}_p$ is a 2×2 propagation matrix describing the signal path of incident radiation on receiver element p, as well as instrumental effects. Finally, $\mathbf{N}_{p,q}$ describes white Gaussian noise.[3] $\mathbf{C}_{p,q}$ is the 2×2 source coherency matrix, which describes the correlations between orthogonal components of the electric field originating from the source. Let $\boldsymbol{e} = \left[e_x, e_y\right]^T$ be the complex Jones vector that represent the complex electric field of the source. The source coherency matrix is then defined as

$$\mathbf{C}_{p,q}(\nu, t) = \left\langle \boldsymbol{e}\boldsymbol{e}^H \right\rangle, \tag{3}$$

where the averages $\langle\cdot\rangle$ are over time. This matrix describes both the total intensity and the polarization of incoming radiation.

For uncorrelated, discrete sources, the brightness of the full sky is additive, because the electric fields from incoherent sources sum

[3]Note that the elements of $\mathbf{N}$ describe different polarizations and not the statistical behaviour of the noise.

in power rather than amplitude. If we assume all components are sufficiently compact that the Jones matrix can be assumed constant over the extent of the source, the full-sky RIME can be broken up into a set of components denoted by d, i.e.

$$\mathbf{V}_{p,q}(\nu,t)=\sum_d \mathbf{P}_{p,d}(\nu,t)\mathbf{C}_{p,q,d}(\nu,t)\mathbf{P}_{q,d}(\nu,t)^H+\mathbf{N}_{p,q}(\nu,t). \quad (4)$$

The RIME is a very powerful tool for describing the process of calibration. In sky model-based calibration, a sky model is created to formulate $\mathbf{C}_{p,q,d}$. This model typically consists of a set of discrete sources that are forward modelled to describe their coherency as a function of baseline, space, and time to obtain the set of brightness coherency matrices. If a model of the primary beam $\mathbf{A}_{p,d}$ is known, it is applied to the sky model during forward modelling.

In standard radio interferometry, individual antennas or dishes are used as receiver elements (i.e. directly correlated to form visibilities). However, the receiver elements in many modern interferometers, such as LOFAR, SKA-Low, NenuFAR, and MWA, are phased array stations, rather than single antennas. Each station consists of a set of N_m antennas that are coherently summed to create electronically steerable beams. By introducing phase shifts based on the antenna element positions $\boldsymbol{x}_{p,m}$ in metres, the station can be steered in direction $\hat{\boldsymbol{s}}_{\mathbf{0}}$,

$$\boldsymbol{v}_p(\nu,t)=\sum_m \exp\left[-\frac{2\pi i\nu}{c}\hat{\boldsymbol{s}}_{\mathbf{0}}\cdot\boldsymbol{x}_{p,m}\right]\boldsymbol{v}_{p,m}(\nu,t). \quad (5)$$

Here, c is the speed of light in metres per second. $\boldsymbol{v}_p(\nu,t)$ is the 2×1 beam-formed voltage vector of station p, whereas $\boldsymbol{v}_{p,m}(\nu,t)$ is the voltage from the mth element in the station. The elements of the voltage vectors describe the orthogonal polarization feeds. The visibility in equation (4) is given by the correlation between two stations,

$$\mathbf{V}_{p,q}(\nu,t)=\boldsymbol{v}_p(\nu,t)\boldsymbol{v}_q^H(\nu,t). \quad (6)$$

A phased array can determine the angle of arrival from the delay by which a wavefront is received by its elements. A wavefront originating from a source in direction $\hat{\boldsymbol{s}}$ has a wave vector of $\boldsymbol{k}(\hat{\boldsymbol{s}},\nu)=-\frac{2\pi\nu}{c}\hat{\boldsymbol{s}}$ [m^{-1}]. These delays can be written as $\exp\left(-i\boldsymbol{k}\cdot\boldsymbol{x}_{p,m}\right)$. The station response to a source can be modelled as

$$\mathbf{A}_p(\hat{\boldsymbol{s}},\nu)=\mathbf{U}_p(\hat{\boldsymbol{s}},\nu)\sum_m \mathbf{G}_{p,m}\exp\left[-i\boldsymbol{k}(\hat{\boldsymbol{s}}-\hat{\boldsymbol{s}}_{\mathbf{0}},\nu)\cdot\boldsymbol{x}_{p,m}\right], \quad (7)$$

where $\mathbf{G}_{p,m}$ are per-element complex gains, and $\mathbf{U}_p(\hat{\boldsymbol{s}},\nu)$ is the antenna element beam (e.g. a simple dipole), which we assume to be identical between antennas here. The element-beam independent part of this equation,

$$\mathrm{AF}(\hat{\boldsymbol{s}},\nu)=\sum_m \mathbf{G}_{p,m}\exp\left[-i\boldsymbol{k}(\hat{\boldsymbol{s}}-\hat{\boldsymbol{s}}_{\mathbf{0}},\nu)\cdot\boldsymbol{x}_{p,m}\right], \quad (8)$$

is called the 'array factor'. It represents the far-field beam of the station based solely on the element positions and their excitations, assuming identical, isotropic elements.

The propagation matrix is then split into the beam model, and a model for all other propagation effects (both astronomical and instrumental), i.e. $\mathbf{P}_{p,d}=\mathbf{J}_{p,d}\mathbf{A}_{p,d}$. Calibration then attempts to find the matrices $\mathbf{J}_{p,d}(\nu,t)$ that make this sky model best match the data. Because the sky usually remains coherent over short time and frequency intervals, a matrix $\mathbf{J}_p$ is found over such an interval instead of per data point, both raising the signal-to-noise ratio and reducing the computational cost. We call these intervals 'solution intervals' and define a parameter l that iterates over all combinations of ν and t within a solution interval to simplify notation. For similar reasons, groups of spatially close sources (calibration clusters, denoted by k) are assigned the same Jones matrix $\mathbf{J}_{p,k}$, rather than an individual matrix per source $\mathbf{J}_{p,d}$. This is an approximation, since the clustering of sources is a heuristic procedure based on the required minimum flux level of a calibration cluster, or the total permissible number of calibration clusters. This means that one pair of closely spaced sources can be assigned the same calibration cluster, whereas another similarly spaced source could be assigned a different matrix, even though its propagation path can be very similar to that of the original pair. If the clusters are not well chosen, the gain can abruptly change on the boundaries between calibration clusters, effectively imposing a two-dimensional step function on the spatial variations in propagation. The matrix $\mathbf{J}_{p,k}$ models the typical propagation effects for sources in a calibration cluster, but is an approximation for each source individually. The simplest form of the calibration problem, for each solution interval separately, then becomes

$$\underset{\mathbf{J}_{p,k}}{\text{Minimize}}\sum_{p,q}\sum_{l}\left\|\mathbf{V}_{p,q,l}-\sum_k\mathbf{J}_{p,k}\left(\sum_{d\in k}\mathbf{A}_{p,d,l}\mathbf{C}_{p,q,d,l}\mathbf{A}_{q,d,l}^H\right)\mathbf{J}_{q,k}^H\right\|_F^2. \quad (9)$$

The aim is then to absorb the largest imperfections in the (lack of) beam model into the calculated solutions $\mathbf{J}_{p,k}$.

Typically, there are several spectral solution intervals stacked within the bandwidth of the instrument. A separate gain ($\mathbf{J}_{p,k}$) exists for each frequency solution interval $\Delta\nu_{\mathrm{sol}}$. Applying a spectral smoothness constraint means assuming that each of these $\mathbf{J}_{p,k}$ lie on a smooth function of ν. The calibration problem now becomes

$$\underset{\mathbf{J}_{p,k}}{\text{Minimize}}\sum_{\nu_{\mathrm{sol}}}\sum_{p,q}\sum_{l}\left\|\mathbf{V}_{p,q,l}-\sum_k\mathbf{J}_{p,k}(\nu_{\mathrm{sol}})\left(\sum_{d\in k}\mathbf{A}_{p,d,l}\mathbf{C}_{p,q,d,l}\mathbf{A}_{q,d,l}^H\right)\mathbf{J}_{q,k}^H(\nu_{\mathrm{sol}})\right\|_F^2$$
$$\text{subject to}\quad \text{spectral smoothness in } \mathbf{J}_{p,k}(\nu_{\mathrm{sol}}). \quad (10)$$

This is a non-convex fourth-order complex optimization problem. There is no analytical solution, which means that careful heuristic choices need to be made to obtain an numerical approximate solution.

There have been many efforts to create DD-gain calibration methods. Examples are joint calibration and imaging algorithms (Repetti et al. 2017; Arras et al. 2019; Birdi, Repetti & Wiaux 2020; Roth et al. 2023), or demixing, in which each cluster of sky-model components is subtracted separately through direction-independent (DI) self-calibration on phase-rotated visibilities (Tol et al. 2007). This latter approach suffers from a high number of degrees of freedom, such that it can only be applied for the brightest sources and at large separations. It is not able to remove several bright sources at a relatively large separation in the main beam, or faint sources in the sidelobes, for example. A modified approach, in which all other sources are subtracted with the (imperfect) calibration gains during each DI calibration step, called peeling, has also been widely used (van Weeren et al. 2016; Albert et al. 2020). In this work, however, we consider the families of calibration algorithms that simultaneously calibrate all directions during each iteration, and implement the smoothness constraint to limit the number of degrees of freedom.[4]

[4]Algorithms such as FACETCAL (Weeren et al. 2016) also depend on a smooth beam approximation and therefore may suffer from similar effects, however, our solution method is not straightforward to implement in FACETCAL.

The two most important algorithms in this context are SAGECAL-CO (Yatawatta 2015) and DDECAL (Gan et al. 2023). SAGECAL-CO uses consensus optimization to penalize gains for deviating from functions in a low-order Bernstein polynomial basis. These Bernstein polynomials are optimized during calibration as well. SAGECAL-CO therefore does not directly enforce smoothness during each iteration, but does provide a 'global collector', which contains a smooth approximation of the obtained gains in this Bernstein basis. This consensus-based approach allows the solutions to be split over frequency channels, and therefore, be solved in a highly parallel manner. DDECAL enforces spectral smoothness at each iteration through low-pass filtering the gains with a truncated Gaussian kernel in between iterations. We apply DDECAL in this work, because it makes understanding the effects of spectral regularization with erroneous modelling assumptions more intuitive, and because it is part of the widely used DP3 processing software[5] (van Diepen, Dijkema & Offringa 2018). However, we emphasize that although SAGECAL-CO and DDECAL apply the spectral smoothness constraint differently, they both suffer from the same issue; the fact that a spectrally smooth gain function $\mathbf{J}_{p,k}(\nu_{\rm sol})$ does not accurately describe the spectral variation of beam-attenuation modelling errors, especially outside the main lobe of the primary beam. Our proposed method can be implemented straightforwardly in SAGECAL-CO as well as in DDECAL. We expect our conclusions to generalize to SAGECAL-CO, and other similar DD-gain calibration algorithms that impose spectral smoothness.

2.1 DDECAL

DDECAL contains several algorithms[6] that numerically approximate the solution of equation (10). Below, we describe its `directionsolve` algorithm, which is similar to the algorithm described by Smirnov & Tasse (2015), but with the addition of a constraints framework. The solver finds the gains by solving for one station in all directions simultaneously. For all other stations, the gains of the previous iteration are assumed. In equation (10), this means that $\mathbf{J}_{q,k}$ are assumed to be known, while for each p, $\mathbf{J}_{p,k}$ are updated for all directions k simultaneously. For each solution interval, the data and gains are stacked columnwise, as,[7]

$$\widetilde{\mathbf{V}}_p = \begin{bmatrix} \mathbf{V}_{p,q=0,l=0} & \mathbf{V}_{p,q=0,l=1} & \dots & \mathbf{V}_{p,q=0,l=N_l} \\ & \mathbf{V}_{p,q=1,l=0} & \dots & \mathbf{V}_{p,q=N_s,l=N_l} \end{bmatrix}, \tag{11}$$

and

$$\widetilde{\mathbf{J}}_p = \begin{bmatrix} \mathbf{J}_{p,k=0} & \mathbf{J}_{p,k=1} & \dots & \mathbf{J}_{p,k=N_k} \end{bmatrix}. \tag{12}$$

Here, $\widetilde{\mathbf{J}}_p$ is the $2 \times 2N_k$ matrix of gains, where N_k is the number of calibration directions. $\widetilde{\mathbf{V}}_p$ is the $2 \times 2N_s N_l$ matrix of visibilities associated with station p. N_s is the total number of stations that station p shares a baseline with, and N_l are the number of time and frequency steps within a solution interval.

[5] dp3.readthedocs.io

[6] These are `directionsolve`, `directioniterative`, `hybrid`, and `LBFGS`. The `LBFGS` routine was implemented in DDECAL specifically to compare results to those obtained with SAGECAL-CO and not widely used in isolation. `directioniterative` and `hybrid` are mainly used to reduce the required computational time, which is why we focus on the more precise `directionsolve` in this work.

[7] The notation used here is a corrected version of that presented by Gan et al. (2023), but is consistent with a new version of their notation available at https://arxiv.org/abs/2209.07854v2.

Using the sky and beam models, and the DD gains from the previous iteration, a model matrix $\widetilde{\mathbf{M}}_p$ is also constructed. This is a $2N_k \times 2N_s N_l$ matrix with its rows stacked in the same order as the columns in equation (11), and elements given by

$$\left(\widetilde{\mathbf{M}}_p\right)_{k,(ql)} = \sum_{d \in k} \mathbf{A}_{p,l,d} \mathbf{C}_{p,q,l,d} \mathbf{A}^H_{q,l,d} \mathbf{J}^H_{q,k}. \tag{13}$$

Using these equations, the DD RIME reduces to

$$\widetilde{\mathbf{V}}_p = \widetilde{\mathbf{J}}_p \widetilde{\mathbf{M}}_p. \tag{14}$$

This is a least-squares fit to the RIME for each station (at each time and frequency solution interval) separately.

Because this is a non-linear least-squares problem (given that the matrices $\widetilde{\mathbf{M}}_p$ depend on $\widetilde{\mathbf{J}}_p$ and will therefore change after $\widetilde{\mathbf{J}}_p$ has been updated), a Gauss–Newton step size is used for convergence. Denoting the iteration number by $[n]$, a full iteration of DDECAL follows the procedure described below.

(i) *Prediction*. Compute matrices $\widetilde{\mathbf{M}}_p[n]$ using equation (13).

(ii) *Least-squares solution.* We denote the solution to equation (10) given the known matrices $\widetilde{\mathbf{V}}_p$ and $\widetilde{\mathbf{M}}_p[n]$ as $\widetilde{\mathbf{L}}_p[n+1]$, to differentiate between this intermediate computation step and the new gain $\widetilde{\mathbf{J}}_p[n+1]$ that is obtained after regularization. This least-squares update is computed for each spectral solution interval $\Delta\nu_{\rm sol}$.

(iii) *Update the intermediate gains.* To avoid oscillation or divergence through overshooting the global minimum, a Gauss–Newton descent with an update speed of $\alpha = 0.2$ is used. A noisy version of the new gain is found using $\widetilde{\mathbf{U}}_p[n+1] = (1-\alpha)\widetilde{\mathbf{J}}_p[n] + \alpha\widetilde{\mathbf{L}}_p[n+1]$.

(iv) *Regularization*. Finally, the new gains are obtained by regularizing the intermediate gain solutions. Although the DDECAL framework also contains other constraints, we focus on spectral regularization, in which the least-squares fits are low-pass filtered through convolution with a truncated Gaussian kernel. The standard deviation σ of this kernel is set by the user to a given spectral smoothness scale, and the kernel is truncated at three times its standard deviation. The wider this filter is, the smoother the obtained gains will be. Each component of the Jones matrix $\widetilde{\mathbf{J}}_p[n+1](\nu_{\rm sol})$ is separately regularized. For each component, a smooth update is given by

$$\widetilde{J}_{p,k}(\nu) = \frac{\sum_{\nu' \in [\nu-3\sigma,\nu+3\sigma]} K(\nu,\nu') W_{p,k}(\nu') U_{p,k}(\nu')}{\sum_{\nu' \in [\nu-3\sigma,\nu+3\sigma]} K(\nu,\nu') W_{p,k}(\nu')}, \tag{15}$$

where we have dropped the subscript 'sol' for legibility. The Gaussian kernel is given by $K(\nu,\nu') = \exp[-\frac{1}{2}(\nu-\nu')^2/\sigma^2]$, and $W_{p,k}(\nu)$ is a weight. Such weights are useful when some solution intervals contain fewer data points than others, due to interference excision, for example.[8]

2.2 Direction-dependent gain-calibration with non-smooth beams

When the beam is rapidly varying as a function of frequency, however, the assumptions underlying equation (10) break down. To explain this, we define the 'true beam', i.e. the station's effective beam with *in situ* effects (in the case of our simulations, these consist of switched-off antenna elements), and the 'model beam' or 'calibration beam', which is the expected beam model $\mathbf{A}_{p,l,d}$

[8] In this work, no interference is simulated and no such excision is performed, so these weights equal one.

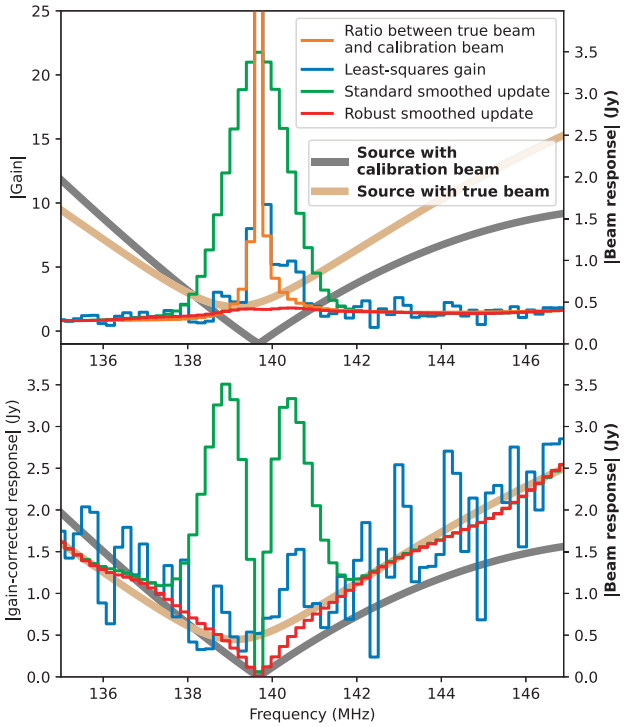

Figure 1. Illustration of the effect of an erroneous beam model on calibration. In both panels, the bold axis shows the prediction of the response of a station to a single point source, both with the beam model used in calibration (bold grey line) and the actual perturbed beam (bold tan line). Top: the regular weight axis shows the gain solutions for the same station. Here, blue is the unregularized solution $J_{p,k}(\nu_{\rm sol})$, green is the same solution after standard smoothing has been applied, and red shows the same with the proposed robust regularization. Orange illustrates the ideal gain that would perfectly match the calibration and true beam. Bottom: the regular weight axis shows the product of the source in the calibration beam and the coloured lines from the top panel. For perfect calibration, this aligns with the tan line.

(without missing elements). The most obvious issue is that there may be modelling errors between the true beam and the beam model on scales smaller than a solution interval in frequency and time. In this case, the single 2×2 Jones matrix $\mathbf{J}_{p,k}$ computed per solution interval and direction cannot describe the small-scale beam-modelling errors accurately. However, deviations from the beam model within the station typically lead to the primary beam being smoother than the ideal model beam. This is because beam shape variations, and amplitude and phase deviations between elements in a phased station break the station symmetry. Therefore, they tend to blur sharp features in the beam, but also tend to lead to a less sharp fall-off of the amplitude of the far sidelobes, as we illustrate in Sections 3.2 and 4.1. As long as a sufficiently small solution interval is chosen, the errors introduced by these intra-interval modelling errors are expected to be small, because the beam-modelling errors do not lead to more rapidly varying errors than are present in the model beam.

The second, bigger problem, and focus of this work, occurs when the beam varies strongly within a spectral smoothness interval. We illustrate this effect for an extreme case with a single polarization in Fig. 1. In both panels, the right-hand axis (corresponding to the thick lines) shows how the apparent flux of a single point source that is far from the target direction varies as a function of frequency. In this example, we use the position of Cas A from the point-source model discussed in Section 3.3, but use a flat-spectrum source with an intrinsic flux of 2.1 kJy. Cas A has a separation from the target field of $\sim 30^\circ$. The primary beam width of a LOFAR-HBA station is $\sim 4^\circ$ in the simulated frequency range, placing the source well into the sidelobes of the station beam. The thick tan-coloured line corresponds to how the source is 'seen' through the true primary beam of a single station (this is the beam in which elements have been switched off). We denote this as $A'_{p,l,s}I_s$, where $A'_{p,l,s}$ is the true primary beam in the direction of Cas A for a single polarization, and I_s is the intrinsic flux of the source. The thick grey line shows the same with an ideal model primary beam (with all elements functioning), which we denote as $A_{p,l,s}I_s$ (without a prime). The latter has an area of very high attenuation (a 'null'), that is spectrally shifted, wider, and less deep in the true beam $A'_{p,l,s}$. We simulate with a monochromatic antenna beam, such that all spectral effects here stem from beam-forming, i.e. equation (8).

The left-hand side axes (in regular weight), show mock calibration results after a single iteration. In the top panel, we illustrate gains $J_{p,s}$,[9] that would result from calibration with different types of regularization. In the bottom panel, these gains are applied to the source's model apparent flux with the calibration beam, i.e. $J_{p,s}A_{p,l,s}I_s$. Applying the gains for both stations in a baseline creates model visibilities that are corrupted in the same way as the data. These corrupted model visibilities are subtracted from the data to remove the source. Therefore, the calibrated model flux $J_{p,s}A_{p,l,s}I_s$ shown in regular weight in the bottom panel should ideally coincide with the true response $A'_{p,l,s}I_s$ shown by the thick tan-coloured line. With our assumptions, the ideal gain (that would calibrate this direction perfectly) is equal to the ratio between the true beam and the calibration beam $A'_{p,l,s}/A'_{p,l,s}$. This gain is shown as the orange line in the top panel. Because the true beam $A'_{p,l,s}$ is unknown, calibration must estimate this ratio from the noisy observations.

In a noisy observation, the first-iteration least-squares solution per spectral solution interval $\widetilde{L}_p$ looks like the blue line in the top panel. When no regularization is applied, this will be the gain after the first iteration $J_{p,s}$. There is a very large spike in the gain at the null in calibration beam. Furthermore, due to the many degrees of freedom, the observational noise is still clearly visible in the least-squares solution, as relatively large spectral channel-to-channel variations throughout the band in the blue line in the top panel. This is normally not a desirable solution, because having too many degrees of freedom leads to overfitting and an undesired interplay between the gains computed for different calibration directions. Additionally, applying noisy solutions to a bright sky model leads to increased errors in the residual data when these sources are subtracted. For unregularized calibration, the value of $J_{p,s}A_{p,l,s}I_s$ (blue line in the bottom panel) therefore does not model the actual source response $A'_{p,l,s}I_s$ (thick tan-coloured line) well.

When the low-pass filter (i.e. the truncated Gaussian kernel) is applied to constrain the solution, we obtain the green lines. Although the noise on the gain has been reduced, the channels adjacent to the null have a higher gain when regularization is applied. This does not model the true source well near the null, as is evident from comparing the green and thick tan-coloured lines in the bottom panel of Fig. 1. In this way, erroneous gains obtained in the null can spread to channels further away from that null. Therefore, although the calibration solutions will generally be improved because of the reduction in degrees of freedom through spectral regularization, areas

[9]Because the calibration direction contains only a point source, we use the same subscript for k and d.

with fast spectral beam variation will suffer from gain errors in the channels adjacent to the null.

The first solution that may come to mind is to remove the calibration beam from the algorithm, as this would remove the spike in the ratio between the true and calibration beam. However, this would mean that any beam variations smaller than a solution interval in time, frequency, or space (i.e. a calibration cluster) cannot be modelled. Therefore, a beam model that is correct to the first order is preferred over removing the beam model altogether, although erroneous solutions may need to be excised after calibration. Furthermore, we illustrate the effect here for a large beam error, but the same can occur if the primary beam model is perfect, as long as the null is present. In this case, the spike results from the low signal-to-noise ratio at the null, and the adjacent channels are still perturbed. Excising the parts of the data where these nulls occur is also not a good solution, because it leads to a high loss of data. This is because there can be multiple off-axis sources, so there can be many time and frequency intervals where one of these crosses a null. Furthermore, because this would involve excising entire stations rather than single baselines, many data can be affected.

2.3 Robust spectral regularization

As an alternative to the current form of spectral regularization, we propose a modified version of the Gaussian-smoothing algorithm that naturally fits in the current approach through reweighting the regularizer. Effectively, this reweighting scheme changes the spectral smoothness assumption; without reweighting, the assumption is that the gains are smooth in frequency. However, with the reweighted regularizer, we instead assume that the beam response (i.e. the product of the gain and the calibration beam) is a smooth function of frequency. This is a more accurate assumption, because the true beam has shallower nulls than the calibration beam, due to variations within the station that break its symmetry. In our simulations, these variations consist of switched-off elements. In general, intrastation variations occur due to a multitude of additional reasons, such as element-gain variations or variations in mutual coupling between antennas at the centre or the edge of a station (Elder & Jacobs 2024).

Concretely, we propose a new set of weights given by the old weights and the calibration model, i.e.

$$\mathbf{W}'_{p,k} = \mathbf{W}_{p,k} \odot \sum_{q,l,d\in k} \left|\mathbf{A}_{p,l,d}\mathbf{C}_{p,q,l,d}\mathbf{A}^{H}_{q,l,d}\right|, \tag{16}$$

for each frequency interval $\nu_{\rm sol}$. The sum in equation (16) acts as a model-flux-weighted estimate of the beam response in direction k. The weights are high when the beam-attenuated calibration model has a high absolute value, but low when there are nulls in the calibration beam. Therefore, the high-value gain corrections (the spike in Fig. 1) are downweighted during convolution, such that they have a lower impact on adjacent channels. This leads to a regularized gain solution that is not as sensitive to large deviations. We illustrate the reweighted gain update with the red line in Fig. 1. Outside of the channels that fall exactly on the beam null in the calibration beam, the reweighted method (red line) approximates the ideal gain solution (orange line) much better than the unregularized (blue) and standard regularized (green) lines in the top panel. As a result, it outperforms the other methods in approximating the true station response (thick tan-coloured line) in the bottom panel.

We refer to the reweighted method as the 'robust' weighting scheme (and compare it to the 'standard' weighting scheme, which is not reweighted by the model flux) and explain how it is tested on simulated data in the next sections. We do emphasize that this reweighted calibration scheme is robust against errors in the spectral smoothness assumption, but does not remedy errors on scales smaller than a solution interval or calibration cluster. A DD-gain calibration method still divides the sky in calibration clusters, effectively enforcing a two-dimensional step function on the sky, and similarly divides the time axis in discrete solution intervals. The robust calibration scheme is therefore mainly effective against beam-modelling errors that change rapidly as a function of frequency.

3 SIMULATIONS

Because beam changes resulting from gain-deviations on the individual antenna element level have not previously been investigated for full-scale LOFAR-HBA observations, we have implemented the simulation and calibration pipeline in a PYTHON code named SHIMMERR.[10] This codebase is able to simulate any hierarchical interferometer, as long as the European Terrestrial Reference System coordinates of the individual antenna elements are known. It has been highly parallelized to allow for simulations of full LOFAR data sets and can simulate beam errors stemming from broken or miscalibrated (sets of) antenna elements, and pointing errors. When these effects are disabled, the elements at each level of the beam-forming hierarchy become identical, such that their beams only need to be computed once. The computation can then be sped-up by calculating a single beam and simulating the full beam by applying the array factor (equation 8, where m can iterate over antennas or tiles, based on the level in the hierarchy).

To reduce the computational complexity of the forward prediction and because the behaviour of nulls in the beam is not expected to behave qualitatively differently between polarizations, we have chosen to implement SHIMMERR for only a single polarization, which we treat as Stokes I, currently. The pipelines for visibility prediction and calibration are illustrated in Fig. 2. In the remainder of this section, we describe how the different parts of the visibility prediction and calibration simulations are implemented and the settings we have chosen for the simulated observations and calibration.

3.1 Telescope model

The goal of the forward simulations is to generate a set of 'true' visibilities that describe the behaviour of the telescope if there are station-beam perturbations. The basic building block of a LOFAR-HBA station is a cross-bowtie antenna. Sixteen of such antennas form a 'tile' (see the first panel in the top row of Fig. 3). Tiles are analogue beam-formed (Haarlem et al. 2013). A group of 24, 48, or 96 tiles – depending on the type of station – are digitally beam-formed to create a station beam. The digitally beam-formed station responses to the sky are correlated to create visibilities, as described in equation (4).

An example of the beams for different beam-forming hierarchy levels within the station are shown in the middle row of Fig. 3. In the first column, we show a beam model for the antenna element, based on the beams presented in chapter 2 of Heald, McKean & Pizzo (2018). This beam does not need to be very accurate, because its large width in comparison to the final station-beam width results in limited impact on calibration. The second column shows the array factor of a tile (not multiplied by the element beam). This beam can be steered through analogue beam-forming. In this simulation setup, it is steered towards the North Celestial Pole (NCP) field. The third column shows

[10] https://github.com/Stefanie-B/shimmerr

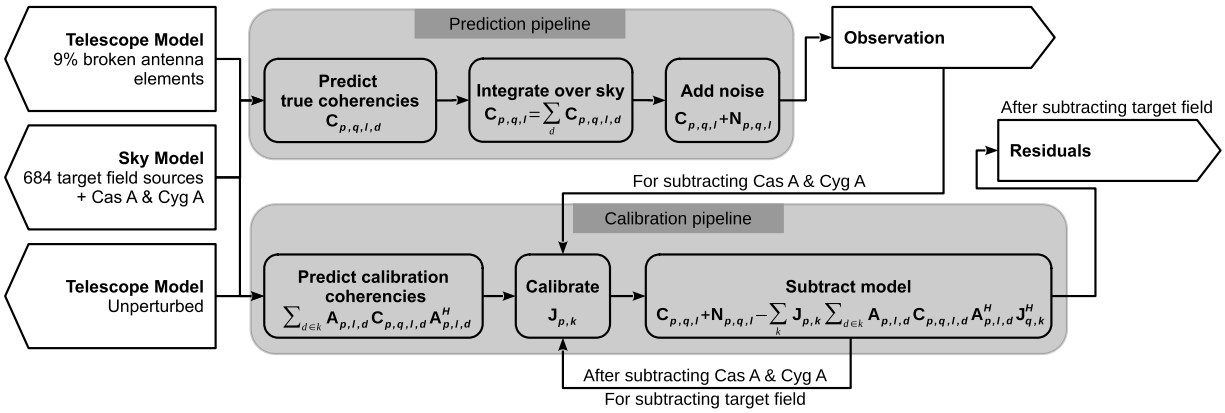

Figure 2. Pipelines used for forward simulation of the visibilities and the calibration. Boxes with an arrow on the left are inputs, boxes with an arrow on the right are outputs, and rounded boxes are processing steps.

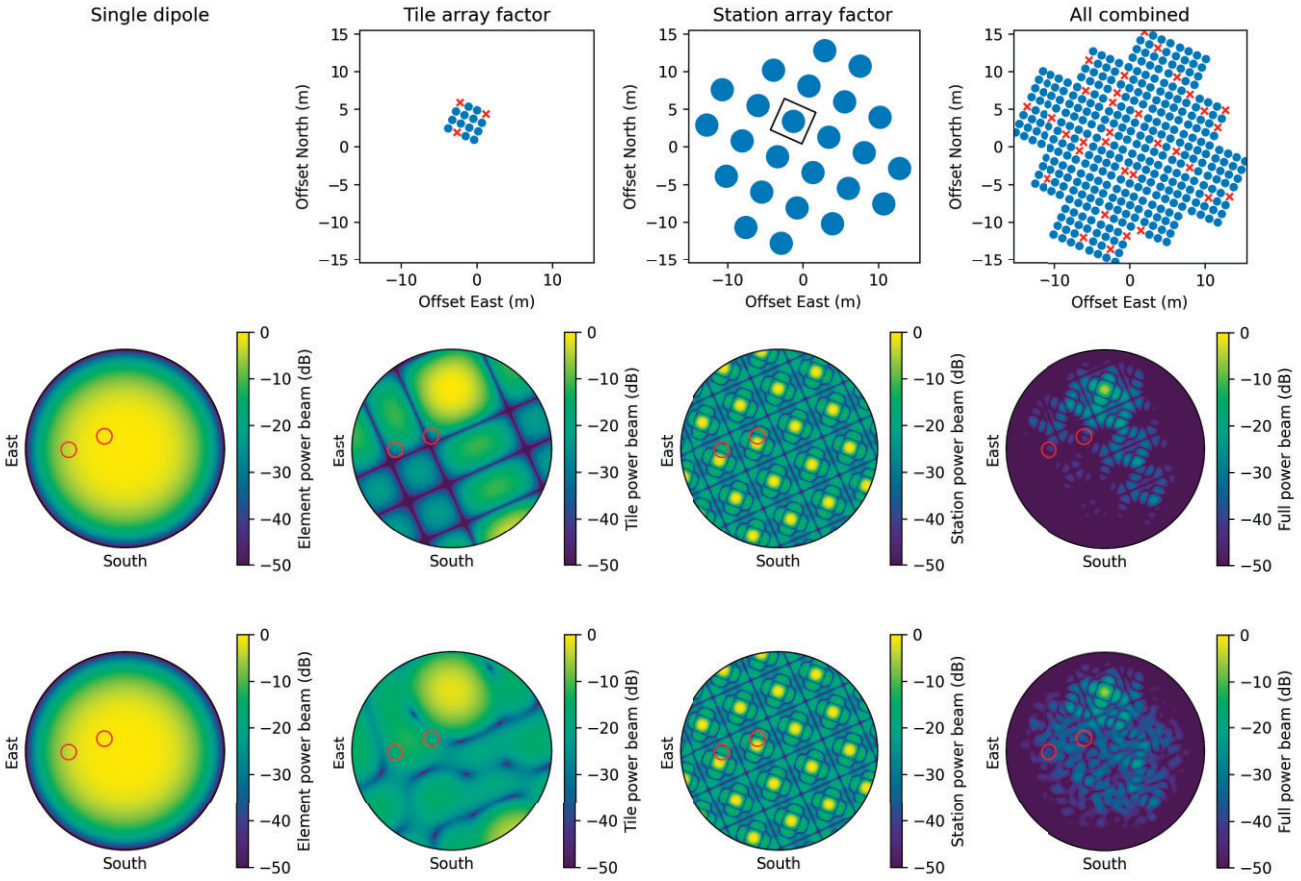

Figure 3. Illustration of the hierarchy of the LOFAR-HBA beam for station CS001HBA0 pointed to the NCP at 03:00:00.00 on 2014 June 28 UTC at 140 MHz. Top row: positions of the elements, where small circles and crosses indicate antenna elements and big circles indicate tiles, respectively. Middle row: calibration beams, where no elements are broken. Bottom row: true beams, where the antennas marked by crosses are missing. From left to right: the element beam, the per-tile array factor, the station array factor if tiles are used as elements, and the full station beam. The full station beam is the beam that is obtained if all three beams in the columns to the left are taken into account. The square in the top plot of column three shows the tile that is plotted in column two. The circles in the bottom two rows show where the two bright sources Cassiopeia A and Cygnus A are.

the array factor of the station (with tiles as elements), not multiplied by the tile beams. The final column shows the full station beam. If no perturbations to the individual elements are present, this is equal to the product of the three beams to the left of it. LOFAR-HBA stations are rotated with respect to one another to mitigate the effect of strong sidelobes and nulls (Wijnholds, Bregman & van Ardenne 2011), resulting in differing beam patterns between stations.

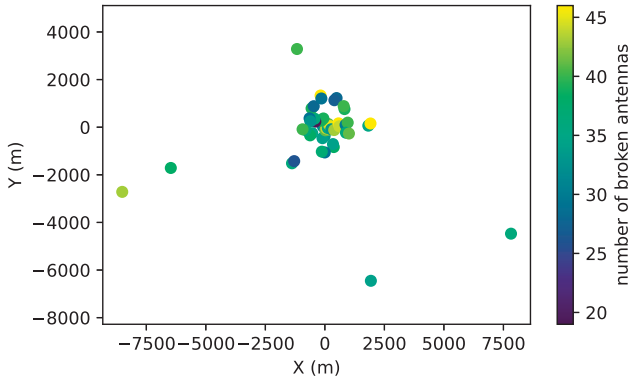

Figure 4. Layout of the stations included in the simulation. The y-axis shows the north–south offset from the centre of the interferometer, while the x-axis shows the east–west offset. The stations are colour-coded by the number of broken antenna elements in the true beam. The total number of antennas in a station is 384.

3.2 Beam perturbations

Diagnostic information on the performance of elements is only available at the tile level in LOFAR-HBA observations, because of the analogue beam former. Tiles that have degraded past a certain point will be excised upon detection, which means that they are not included in the digital beam former and not included in the simulated calibration beam. However, if an antenna within a tile breaks, this can only be detected as a different tile response. While the tile responses of LOFAR have been calibrated during commissioning, such calibration rounds can be out of date. Currently, typically up to three antenna elements are broken or malfunctioning per tile (Norden, private communication).

The exact distribution of broken elements is unknown. Therefore, we simulate beam errors by randomly switching off a number of elements, by setting their element gains in equation (7) to zero. Throughout our simulations, $N \sim \mathcal{U}(0, 3)$ elements are switched off in each tile, where $\mathcal{U}(0, 3)$ is the discrete uniform distribution between 0 and 3. This results in an average of 1.5 antennas, or 9 per cent of the tiles being simulated as 'broken'. We randomize the broken antennas for each tile in each station, such that different stations have a different configuration of broken elements. Because the number of broken antennas per tile is randomized, different stations also have varying numbers of broken elements, as shown in Fig. 4.

The effects on the beam, resulting from a random realization of removing antennas, is illustrated in the last row of Fig. 3. As is clear from the second column, switching off three antenna elements drastically alters the shape of the tile beam, and removing and altering nulls. Because the tile beams are no longer identical, the nulls and sidelobes of the full station primary beam also change. This has a limited effect near the phase centre (bright yellow spot), but does change when off-axis sources pass through nulls. We illustrate where such off-axis sources can be compared to the phase centre through red circles at the positions of Cas A and Cyg A.

It is important to emphasize that, although we focus on broken elements in this work, there are other effects that break the regular structure of a phased array station and create similar distorted beam patterns to those shown here. For example, mutual coupling (Elder & Jacobs 2024; O'Hara et al. 2025) will affect the beam by changing the patterns of individual antenna elements. Performance (e.g. gain) differences between functioning elements (Chokshi et al. 2024) will have a similar effect to setting the gains of several elements to zero. These errors can also change over time, for example, by the low-noise amplifiers being affected by on-site temperature and humidity variations. Additionally, such variations can occur at the tile (analogue-beam-formed) beam level rather than the individual antenna element level. However, because we expect broken elements to have a dominant effect, we demonstrate the resulting calibration problems arising from such beam errors.

3.3 Sky model

For sources in the main lobe, we use a model of the NCP field, which is the most studied field in the LOFAR-EoR key science project (Patil et al. 2017; Mertens et al. 2020, 2025). The model is a $10^\circ \times 10^\circ$ field of view containing 684 flat-spectrum point sources, that dominate the power spectrum of the field, distributed over 24 calibration clusters (see Brackenhoff et al. 2024).

For the off-axis sources, we use two bright radio sources: Cassiopeia A (Cas A) and Cygnus A (Cyg A). These sources are approximately 30° and 50° away from the target direction, respectively. These are important to include, because their residuals after DD-corrected model subtraction currently dominate the 21-cm signal power spectra of the NCP field (Gan et al. 2022; Munshi et al. 2024; Ceccotti et al. 2025). We base our models of Cas A and Cyg A on those provided by ASTRON,[11] but replace Gaussian components by point sources with the same total brightness. DDECAL calculates the model for a Gaussian source by simulating it at its centroid and scaling its brightness according to baseline length. This means that DD (e.g. beam) effects across an extended source are not taken into account. Therefore, we would not obtain additional DD information by modelling them as Gaussians.

Two variations of the model are considered: a single point source for the total brightness of the A-team source (the 'single-point-source model') and a version where Cas A and Cyg A consist of 9 and 5 point sources respectively (the 'multipoint-source model'). Comparing the results obtained between these models allows for comparing the extreme case of a single-point-source shifting in and out of a null to the more general case of an extended source (represented by multiple point sources), without needing to simulate Gaussians.

3.4 Prediction pipeline

Equation (7) is used to forward simulate the beam. Since the chromaticity of the element beam is limited, we model it as an achromatic beam $U(\hat{\boldsymbol{s}})$ that is identical for all antennas. The station beam can be computed with and without removing elements. Setting all element gains to unity results in the calibration beam, whereas setting some element gains to zero results in the perturbed 'true' beam.

We use the true beam (with switched-off elements) as the propagation matrix in equation (4) to simulate visibilities. The thermal noise is created by independently drawing the real and imaginary part of the noise from a Gaussian with standard deviation $\sigma = \mathrm{SEFD}/\sqrt{2\Delta\nu\Delta t}$ [Jy],[12] where $\Delta\nu$ and Δt are the spectral and temporal resolutions of the simulation in Hertz and seconds, respectively. SEFD is the system equivalent flux density in Jansky. The visibilities created in this way

[11] https://github.com/lofar-astron/prefactor/blob/master/skymodels/A-Team_lowres.skymodel

[12] The factor two is introduced by independently drawing the real and complex part of the noise.

Table 1. Prediction settings for the simulated observations.

Parameter	Setting
Telescope	LOFAR-HBA – CS (dual mode) + RS (in CS configuration)
Missing elements	9 per cent of station
Pointing	RA = $0^h00^m00^s$, Dec. = $+90°00'00''$
Start times (UTC)	27-12-2013 18:00:00 (N1)
	27-03-2014 18:00:00 (N2)
	27-06-2014 18:00:00 (N3)
Sky model	684 point sources in the central $10 \times 10°$
	Cassiopeia A (1 or 9 point sources)
	Cygnus A (1 or 5 point sources)
Bandwidth	135–147 MHz
Frequency resolution	195 kHz
Time resolution	10 s
Total duration	12 h
SEFD	4.2 kJy

(with the true beam and noise) are treated as the observation's raw uncalibrated data.

For the calibration model, we subsequently use the model beam (that does not contain broken elements). We sum the coherencies per calibration patch to create model visibilities for each direction, and do not add thermal noise.

3.5 Simulated observations

Table 1 lists the most important settings used for the simulated observations. These settings have been chosen to mimic typical LOFAR-EoR observations. Each observation consists of a 12-h rotation synthesis and has the same time resolution as as is used in LOFAR-EoR DD-gain calibration. The total bandwidth matches a LOFAR-EoR redshift bin and the spectral resolution is one channel per sub-band, equal to the calibration resolution on real data. The SEFD has also been chosen to match the noise level of these observations (Mertens et al. 2020). Finally, the telescope model has been configured to match the configuration of a typical LOFAR-EoR data set. The core stations (CS), which consist of pairs of stations of 24 tiles, are configured to act independently. This means that one CS acts as two independent but closely spaced stations, labelled by a 0 and 1, respectively (for example: CS002HBA consists of CS002HBA0 and CS002HBA1). The remote stations (RS), which are usually comprised of 48 tiles, have their outer 24 tiles switched off to match the shape of a CS. This makes the CS and RS identical in our simulations.[13]

3.6 Calibration

A telescope without missing elements is used to create model coherencies for calibration purposes. Because we have performed the forward prediction with unity gains for all except the switched-off elements, and because we do not assume bandpass errors, we only consider a DD-gain calibration step for source subtraction. If both DI and DD-gain calibration are performed, errors in the DI gains could be absorbed in the DD gains, such that the results become less interpretable. Calibration is performed in two steps, because the station beams change faster as a function of time and frequency for sources far from the target direction. Because of this, Cas A and Cyg A should be calibrated for and subtracted with a smaller time interval per solution than used for the target field[14] (Gan et al. 2023; Munshi et al. 2024). The first round of calibration is done at time interval of 2.5 min per solution and in three directions: Cas A, Cyg A, and the target field for which we assume a single direction. Cas A and Cyg A are removed with the gain-calibration solutions for their respective directions applied to their respective sky models. The residuals from this calibration step are used as an input for the second round of calibration using a solution interval of 20 min, in which the target field is divided into 24 gain-calibration directions. In the LOFAR-EoR key science project, DD-gain calibration is performed on a set of longer baselines ($250 - 5000\lambda$), to avoid absorption of the large-scale 21-cm signal in the calibration solutions (Patil et al. 2016; Mevius et al. 2021). The analysis is done on shorter baselines ($50 - 250\lambda$). Therefore, we also calibrate on these longer baselines only, but analyse all baselines between 50 λ and 5000 λ. We omit three stations, that are a part of fewer than three baselines in this range. An overview of the calibration settings is listed in Table 2.

Table 2. Calibration settings for subtracting the sky model.

Parameter	Value for Cas A and Cyg A	Value for target field
Number of calibration directions N_k	3	24
Solution time interval (min)	2.5	20
Baseline lengths calibration (λ)	250–5000	
Solution frequency interval (kHz)	195	
Number of iterations	50	
Reference station	CS002HBA0	

4 RESULTS

In this section, we analyse the effect of the beam errors on the calibration results. First, the expected and computed gains are discussed to build more intuition on the spectrotemporal signatures of the beam errors. With this in mind, the residuals after subtracting all sources with the computed DD gains are inspected for various spectral smoothness kernel widths for both the standard and the robust regularization scheme. Finally, the effect of the errors in image space is discussed.

4.1 Gain solutions

Because the Earth rotates during an observation, the spatial beam pattern shown in the last column of Fig. 3 moves across the sky and changes shape (or, in the case of the NCP, rotates across the sky). Sources that are not in the station-beam main lobe pass through multiple nulls and sidelobes. Additionally, the position and width of the station-beam sidelobes depend on the spacing between antennas in units of wavelength. This causes the beam lobes to be narrower at higher frequencies, and wider at lower frequencies. These two effects combined define the beam as a four-dimensional structure with two spatial, one spectral, and one temporal dimension. For a given point in the sky, this reduces to a two-dimensional dynamic spectrum for the gain $|\mathbf{J}_{p,k}(\nu, t)|$.

[13] In real observations, the outer tiles of the RS receive a zero-weight such that they do not directly contribute to the observation. However, because they are still present, they do electrically couple to the inner tiles, distorting the beams in a different way than the CS.

[14] In the LOFAR NCP field, these two steps are performed simultaneously with SAGECAL-CO. However, setting two different solution intervals is not possible with the `directionsolve` method of DDECAL.

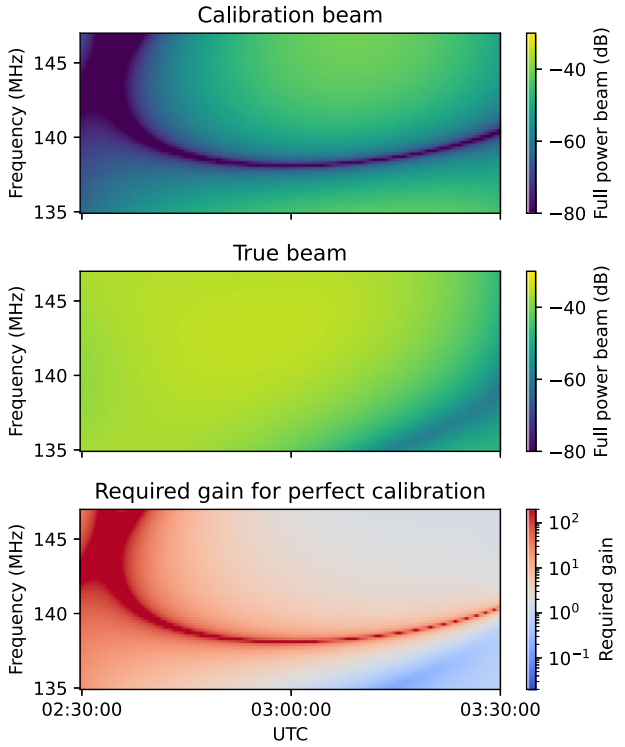

Figure 5. Effect of beam errors in frequency and time for station CS001HBA0 on 2014 June 28 UTC (N3) in the direction of Cas A. This is the same station as in Fig. 3. Top: expected beam response (calibration beam), middle: beam response with 32 broken elements ('true' beam), and bottom: gain required to fully subtract Cas A at the same resolution as the data.

Fig. 5 shows the dynamic spectrum of the station power beam[15] from Fig. 3 for the direction of Cas A. The top panel of Fig. 5 shows the calibration beam (without switched-off elements), and the middle panel shows the true beam (with switched-off elements). When calibrating, the goal is to approximate the true beam with the product of the gain and calibration beam. The gain that would perfectly do this for a single point source is equal to the ratio between the true and the calibration beams. This ideal gain is shown in the bottom panel of the figure.

Gain-calibration algorithms do not reach this ideal gain for two reasons: the finite-calibration resolution needed to reach a minimum signal-to-noise ratio for the gain solutions (gain discretization), and the non-convex nature of the optimization problem leading to solution errors. Gain discretization introduces errors because variations on scales smaller than a solution interval cannot be captured. In other words, a single matrix $\mathbf{J}_{p,k}$ cannot fully compensate for modelling errors in $\mathbf{A}_{p,l,d}$ as a function of l (time and frequency indices within a solution interval) or d (source direction within a calibration cluster). Therefore, if the variation of the station beam in space, time, or frequency is not well modelled within a solution interval, these intra-interval errors remain present after calibration. Additionally, the ideal gain may have spectral variations on scales that are suppressed by the low-pass filter. This effect is especially strong near nulls in the calibration beam, where the desired gain peak approaches infinity in case the true beam is non-zero. Around a null, the ideal gain solution has a very rapid rise and fall. However, such changes cannot be captured with a smooth gain solution. This rise and fall is visible as an increased required gain along a horizontal streak in Fig. 5. In addition to these resolution-based effects, there is a risk of non-convergence (after the maximum number of iterations) or convergence to a local minimum in the optimization problem. This occurs because the final gain solution $\mathbf{J}_{p,k}$ depends on the intermediate estimates of the gains of all stations and all directions. Therefore, an error in one direction for one station can potentially affect the estimates of all other gains.

With a noisy 12 h simulation in hand, we run DDECAL. The errors in the gain-calibration solutions in the direction of Cas A are shown in Fig. 6. In all panels, the single-point-source models are used for Cas A and Cyg A, but the spectral regularization settings vary between panels. The gain error is computed by dividing the amplitude of the calculated gain by the amplitude of the ideal gain (i.e. the ratio of the true and calibration beams shown in the bottom panel of Fig. 5). The top panel of Fig. 6 shows the result if no spectral regularization is used during calibration, and generally follows the expected gain behaviour shown in the bottom panel of Fig. 5. The unregularized solution is not useful in practice due to overfitting of the data, as is visible as the rapid variations in the lower right corner of the plot. We include the unregularized solution here purely as a reference plot, to illustrate the effect that regularization has. The discretization of the gains, discussed above, is clearly visible as steps in the gains every 2.5 min. Additionally, the gains are underestimated along the null in the calibration beam, as is evident from the orange regions in the same areas where the null was identified in Fig. 5. Because the ideal gain approaches infinity here, this underestimation is expected. However, because the null in the calibration beam is narrow, this affects few channels, that could be flagged later.

The second and third rows of Fig. 6 both use a 1 MHz wide smoothness scale. However, the second row uses the standard regularization method and the third the robust one. The fourth and fifth rows show the same, but for a wider 4 MHz spectral smoothing kernel. The standard smoothing method (the second and fourth row) shows a gain-solution artefact at 03:00:00 that is not present for the robust regularization method. The gains near the null are overestimated, with an underestimation exactly on the null. Additionally, there is an underestimation of the gain near the null that is shifted towards the centre of the band (at ∼140 and ∼143 MHz for rows two and four, respectively). The overestimation at the null surrounding an underestimation is expected from applying a low-pass filter to a peak in the gain solutions, as illustrated in Fig. 1. The location of the gain-calibration error is also in line with expectations, as the null near 03:00:00 changes more rapidly in the spectral direction than the temporal direction. We attribute the underestimation towards the centre of the band to a local minimum. The low-pass filter increases the gains here, due to the overestimated gains near the null. This can affect the gains further away in frequency as well. Therefore, in the next iteration, a least-squares solution is found where the gains towards the centre of the band (at ∼140 and ∼ 143 MHz for rows two and four, respectively) are lowered, to 'insulate' the rest of the band from the effect of the badly calibrated null during smoothing. The same does not happen on the other side of the null, away from the centre, because there are fewer channels on this side of the null. Therefore, there are not enough channels that would be affected to create the 'insulation' effect.

However, when robust regularization is applied (rows three and five of Fig. 6), the artefact disappears and the behaviour of the null at 03:00:00 is very similar to the rest of the observation. In this part of the gain dynamic spectrum, robust regularization clearly shows fewer errors than the standard method and limits

[15] As is customary in literature about beams, we use a 'power beam' in the plots. The power beam is $|\mathbf{A}_{p,k}(\nu, t)|^2$.

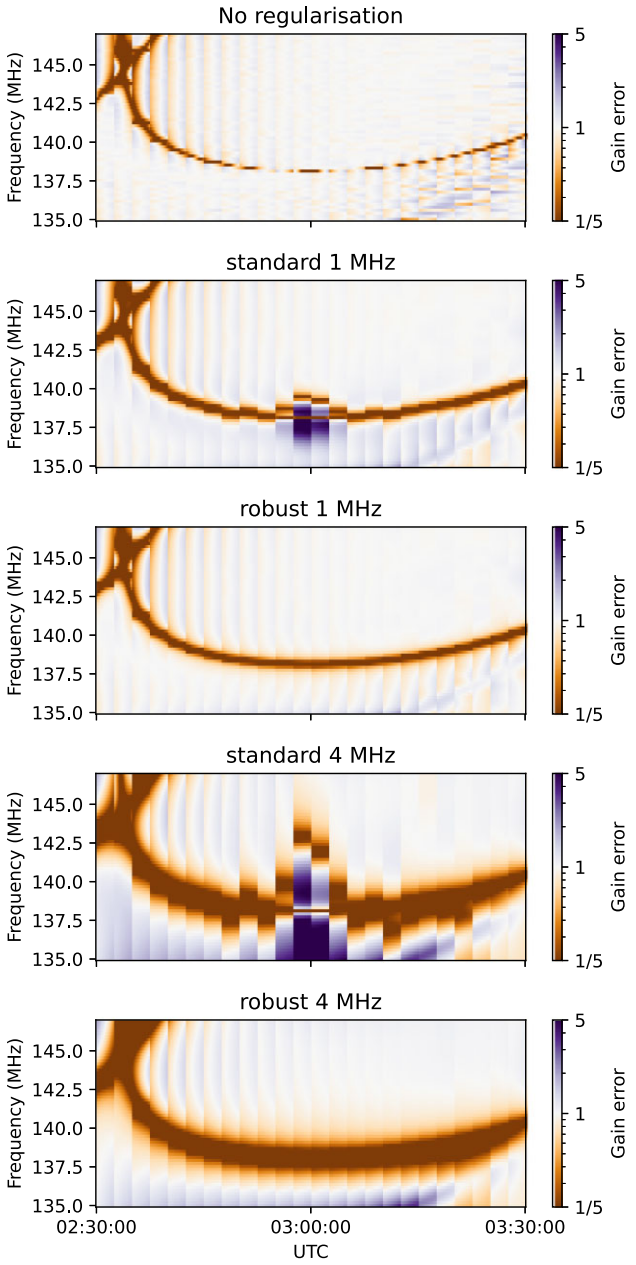

Figure 6. Ratio between the gain computed using DDECAL and the required gain in the direction of Cas A obtained with the point-source model for Cas A and Cyg A. Top to bottom: result without regularization, with a 1 MHz kernel and the standard regularization method, with a 1 MHz kernel and the robust regularization method, with a 4 MHz kernel and the standard regularization method, and with the 4 MHz kernel and the robust regularization method. White indicates an accurate gain solution, purple an overestimation of the gain and orange an underestimation. The station and time match Fig. 5.

the number of channels affected by the effect of the null in the calibration beam, as intended. There are still clear errors in the recovered gains, however. Furthermore, the wider kernel seems to both increase the number of channels with errors (wider orange streaks) and the amplitude of errors that were already present (stronger purple and orange shading near 135 MHz), increasing the bias.

4.2 Residual visibilities

We now turn our attention to the residual visibilities that are found after subtracting all sources with their respective DD-calibrated gains. Because there are no differences between the simulated and calibration sky models, a perfect calibration would result in the residual visibilities being completely noise-like.[16] The accuracy of the calibration can therefore be tested by comparing the residual visibilities to noise. In Figs 7 and 8, the dynamic spectra of the residual visibilities on single baselines are shown. One of the stations that comprise the baseline in Fig. 7 is the same station as used for Figs 3, 5, and 6, such that the nulls seen in Figs 5 and 6 can be associated with effects seen in Fig. 7. The left-hand columns of Figs 7 and 8 show the residual visibilities after subtracting the gain-calibrated sky model with a 4 MHz kernel and the standard method, and the right-hand columns show the same, but with the robust gain-calibration method. The top panels show visibility amplitudes, and the bottom panels show visibility phases.

The gain errors seen in Fig. 6 clearly have a detrimental effect on the subtraction of Cas A, because the areas that have large gain errors in Fig. 6 also exhibit an increased amplitude and structure in the phases in Fig. 7. There are also other areas with large residual visibilities, such as the streak below 140 MHz between 02:35 and 02:45. This artefact coincides with a null in the direction of Cyg A, the other bright off-axis source. Similar to what was found for the gains, robust gain regularization yields an improvement in the calibration results. It reduces the artefact at 03:00:00, both in terms of amplitude and number of frequency channels affected by it. However, even the robust regularization method does not completely remove Cas A and Cyg A. These residuals are strongly time- and baseline-dependent, however. This can be seen in Fig. 8, where robust regularization is able to subtract all sources down to the noise level.

The main improvement between the standard and the robust method is that the standard method can introduce large outliers in the gains, which can in turn result in large outliers in the visibility residuals. However, typically, such outliers in non-transient visibility data sets could be remedied with outlier-based data excision. Therefore, a more realistic comparison would be to compare the two methods after excision of the outliers. A challenge is that our simulations are solely corrupted by beam errors, whereas a real observation might have multiple different sources of outliers in the residual visibilities. Therefore, low-level artefacts specifically introduced by errors in the gain would be more difficult to identify and excise in real data. None the less, we can draw conclusions about the magnitude and number of outliers compared to the thermal noise level, as well as the residual power left after excision using these simulations. Because the outliers introduced by beam errors tend to be correlated both in time and frequency, they can be detected with traditional outlier detection methods such as the scale-invariant rank filter that flags data based on the number of outliers within a slice of data (AOFLAGGER, Offringa, van de Gronde & Roerdink 2012). A single round of AOFLAGGER is, therefore, performed on all residual visibilities.

Because perfect calibration would result in only Gaussian noise, the effectiveness of the calibration can be judged by deviations from the thermal noise level. Therefore, we compare the standard deviations of the visibilities of a complete 12 h simulation before and after outlier excision to the thermal noise level of $\sigma_n = \mathrm{SEFD}/\sqrt{\Delta\nu\Delta t} = 3.01$ Jy. The standard deviations of the

[16]Due to this being a non-linear model, a small bias is expected, but this is negligible compared to the error level.

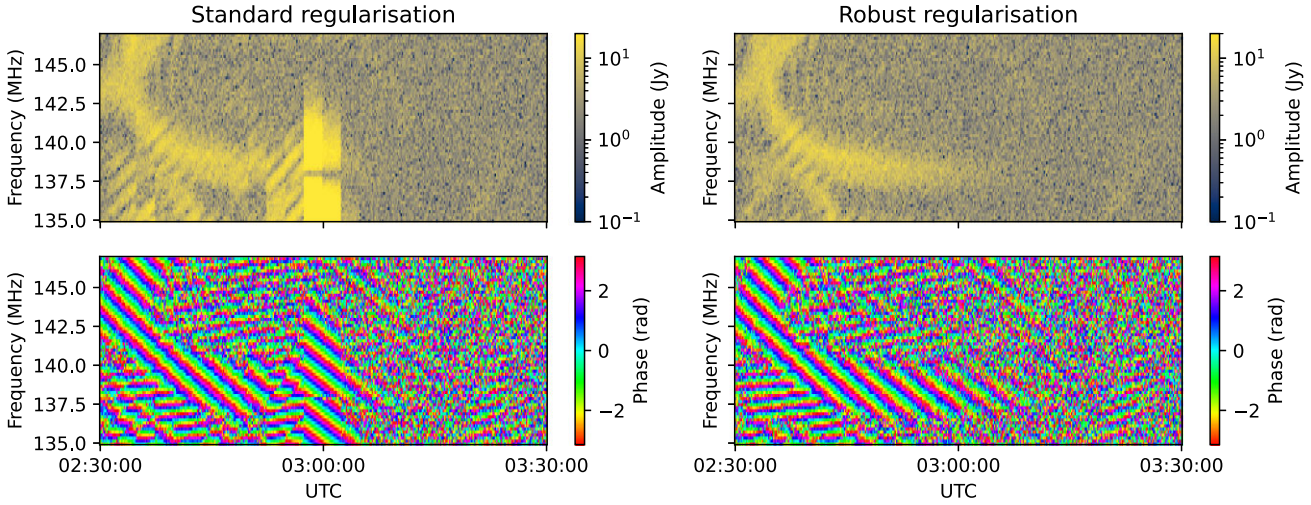

Figure 7. Residual visibilities for the baseline between CS001HBA0 and CS003HBA1 on 2014 June 28 (N3) after all sources have been subtracted with solutions created with the 4 MHz kernel. The times and frequencies match Figs 5 and 6. Left: standard DDECAL algorithm, and right: the robust method. Top: amplitudes, and bottom: phases. Ideally, these residuals should be noise-like.

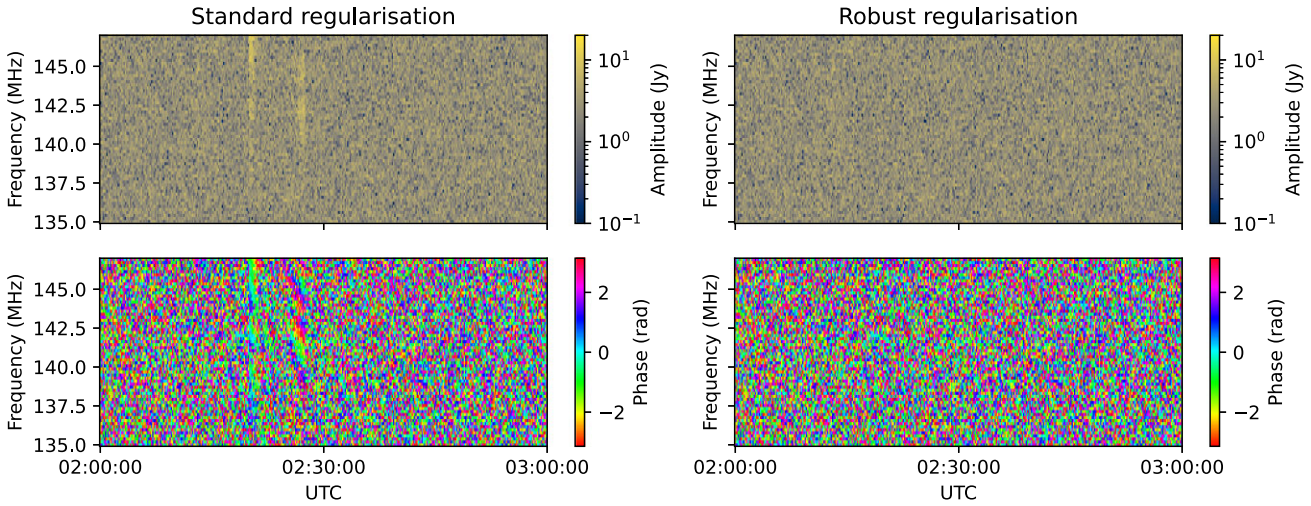

Figure 8. Similar to Fig. 7, but for the baseline between CS002HBA1 and CS007HBA1 on 2013 December 28 (N1). Here, the robust regularizer is able to subtract the sky sources down to the noise level.

simulations of N1 (that starts on 2013 December 27, see Table 1) with the single-point and multipoint models for Cas A and Cyg A are shown in Fig. 9. The figure shows the ratio between the standard deviations and noise before and after outlier flagging in the top and middle rows, respectively. The bottom row shows the percentage of data that was excised by AOFLAGGER. We consider the latter statistic, because using fewer flags leads to a lower loss of data, and therefore, a higher sensitivity. This loss of data can be significant, e.g. in LOFAR-EoR science, through various methods of flagging (see Mertens et al. 2025), up to 19 per cent of data are flagged to excise Cas A and Cyg A (Mertens, private communication). Therefore, attaining the same standard deviation with fewer flags is beneficial.

For all smoothing kernel settings and sky models, the robust regularization method matches or outperforms the standard method, both in terms of standard deviations before and after flagging, and in the number of flags that are needed to attain this. Generally, the pre-flagging standard deviation of the standard method is much larger than that of the robust method, such that flagging the data has a much larger effect.

The different panel columns show various kernel sizes for the calibration. The first and second columns show the 1 and 4 MHz kernels, respectively. Similar to what was found in Section 4.1, the 4 MHz kernel performs significantly worse than the 1 MHz kernel. This is somewhat surprising, since Gan et al. (2023) showed that the 4 MHz kernel performed better on LOFAR-EoR data. We suspect that there is a difference in which kernel width performs better on the main lobe versus off-axis sources. Because the target field does not suffer from beam-modelling errors to the same degree as the off-axis sources, the 4 MHz kernel may outperform the 1 MHz kernel for the target field sources, while this is the other way around for the far-field sources. A similar spectral smoothness difference between the main lobe and off-axis sources is also observed in NenuFAR observations (Munshi et al. 2025). For both instruments, the target field contributes more power than the off-axis sources, owing to lower beam attenuation. Because of this, the target field dominates the residuals in real data, where calibration and subtraction are less ideal than in these simulations. In our simulations, the only errors that are introduced are beam-modelling errors and thermal noise.

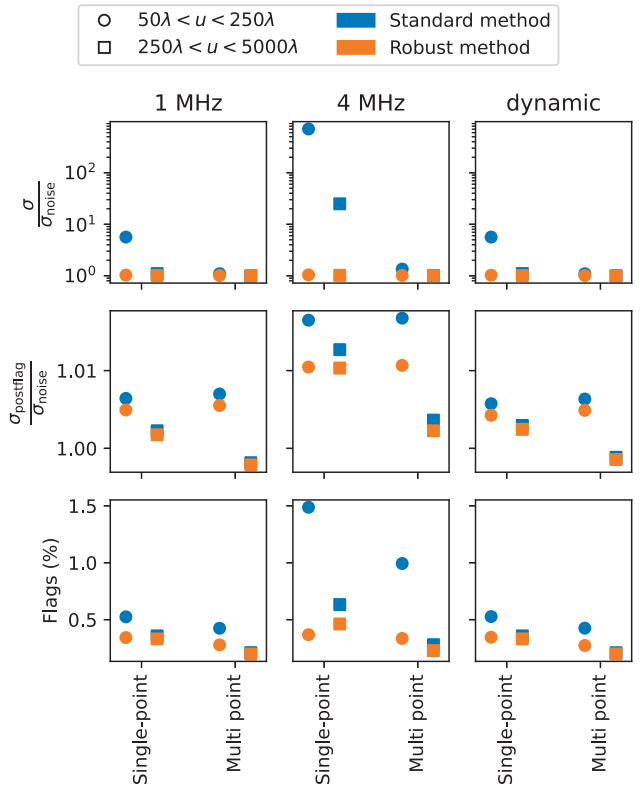

Figure 9. Effect of outlier excision on the data set, for both the single-point and multipoint off-axis models for the simulation starting on 2013 December 27. Top: standard deviation of the residual visibilities compared to that of the thermal noise, middle: the same after outlier flagging, and bottom: percentage of the data that were removed. From left to right, the residuals after calibration with a 1 MHz kernel, 4 MHz kernel, and a dynamic kernel with mixed regularization (1 MHz for Cas A and Cyg A, and 4 MHz for the target field). Blue markers indicate the standard method and orange markers the proposed method. The EoR analysis data, which are baselines $< 250\lambda$ are indicated by circles and the calibration baselines ($> 250\lambda$) by squares.

Because the beam-modelling errors are spectrally smooth within the main lobe of the primary beam, we are able to remove the target field sources to a much deeper level, revealing errors in the directions of Cas A and Cyg A.

To test this hypothesis, we also use a combined kernel on the simulation: a 1 MHz kernel is used for subtracting Cas A and Cyg A, whereas a 4 MHz kernel is used for the target field. The results for this dynamic calibration kernel are shown in the final column of Fig. 9 and are very close to the results for the 1 MHz kernel. This suggests that the outliers found in the data primarily correspond to errors made in the directions of Cas A and Cyg A, such that these dominate the statistics shown in Fig. 9.

Fig. 9 shows two sets of baselines: those between 250λ and 5000λ (i.e. the calibration baselines), and those between 50λ and 250λ (i.e. the EoR analysis baselines). Comparing the calibration and analysis baselines provides information about the level of overfitting, ideally the two sets should be as close as possible. However, there is more residual power on the analysis baselines. This is expected for a baseline cut, and corresponds to earlier findings (Mevius et al. 2021). Furthermore, the analysis baselines in the standard method require a higher flagging percentage than those in the robust method.

The horizontal axis coordinates within a panel of Fig. 9 denote two different models for the off-axis sources, describing them either with a single or multiple point sources. Because the Local Sidereal Times (LSTs) are identical between the simulations, only the off-axis contribution to the visibilities differs, whereas that of the target field stays the same. Clearly, the single-point-source model suffers more from outliers than the multipoint-source model, as is evident by the reduction in standard deviation during flagging and the number of flags. This may be due to two things: first, the multipoint-source model has a bigger spatial footprint, and secondly, each individual component of the multipoint-source model contains less power than the single component of the single-point model. To differentiate between these two, we also performed a test with a strongly attenuated single-point model (with the flux reduced by a factor four). This attenuated model suffers from large outliers in the same way as the single-point-source model does (see the complete results in Fig. A1), such that we can attribute the better calibration behaviour of the multipoint-source model to its bigger spatial footprint. As a result of the extended spatial footprint, the source is never fully inside a null, such that its flux is not as abruptly attenuated and the spectral smoothness assumption is not violated as strongly.

The scenario that most closely resembles real LOFAR-EoR observations is the multipoint-source model with the 4 MHz regularization kernel, which was previously found to be preferred over the 1 MHz regularization kernel (Gan et al. 2023). The analysis baselines show a clear improvement when using the robust regularization kernel, reducing both the residual power slightly after flagging from 3.06 to 3.04 Jy and the number of flags from 0.99 to 0.34 per cent on the analysis baselines. Furthermore, the results can potentially be improved even more by utilizing the dynamic calibration kernel, which reaches a standard deviation of 3.02 Jy after flagging 0.27 per cent of the simulated data. This is very close to the thermal noise level of 3.01 Jy. However, tests on real data are necessary to confirm that the improved performance of the dynamic calibration kernel is not hampered by real-world effects that are not present in this simulation.

4.3 Imaging

In image space, the main goal of DD-gain calibration of off-axis sources is often to remove them to a sufficient level, so that their point spread function sidelobes do not cause errors in the target field. Although Weeren et al. (2016) have shown that facetting is successful in creating images with long baselines, the power of DD-gain calibration algorithms like DDECAL and SAGECAL-CO that calibrate in multiple directions simultaneously is that they are also effective in reducing errors on shorter baselines. To measure the effectiveness of DD-gain calibration as a function of baseline length, errors of various angular scales in the image must be probed.

To this end, the residual data are continuum imaged at a resolution of 0.3 arcmin with a field of $10^\circ \times 10^\circ$ and uniform weighting using WSCLEAN (Offringa et al. 2014). WSCLEAN is an efficient wide-field imager, that computes the inverse transform of equation (1). It does so by gridding the visibilities onto a uv-grid at different levels of the w-term. Each of these grids are then inverse Fourier transformed to obtain a stack of images. These images are phase shifted with a w-term correction, and a wide-field image is formed by summing the corrected images. Offringa et al. (2019) have shown that this form of imaging is precise enough for 21-cm cosmology. All baselines between 50λ and 5000λ are used in the images. To give a realistic comparison, flags are applied to the data before imaging, such that the outliers are excised. As a result, none of the images contain very bright sources anymore, and the dirty images can be analysed directly, without a need for deconvolution. Subsequently, the images

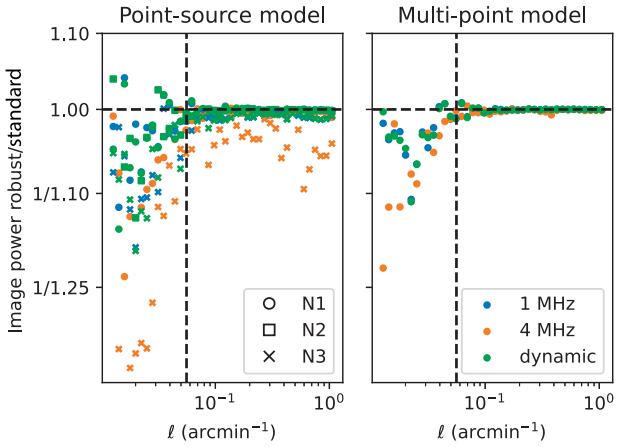

Figure 10. Ratio of power in the residual images with the standard versus proposed method as a function of logarithmically binned scales in the image. The left panel shows the point-source model for Cas A and Cyg A and the right panel the multipoint model. The colours indicate the different kernels. For the single-point model, the different observing LSTs are denoted with different markers. The horizontal dashed line denotes a ratio of unity, and the vertical dashed line the scale probed by a baseline of 250 λ.

are Fourier transformed to determine their angular power spectrum (i.e. the power at various angular scales),

$$P(\ell) = \left\langle |\mathcal{F}\{I(x,y)\}|^2 \right\rangle_\ell . \tag{17}$$

Here, $\mathcal{F}\{\cdot\}$ denotes the spatial Fourier transform, and $I(x, y)$ denotes the image intensity at pixel (x, y). The operator $\langle\cdot\rangle_\ell$ represents an azimuthal average over Fourier modes with constant angular frequency ℓ. If ℓ_x and ℓ_y are the Fourier conjugates of the image axes, $\ell = \sqrt{\ell_x^2 + \ell_y^2}$. We use the cosmological convention of using the inverse of the spatial scale as the Fourier conjugate.

If there is residual source power in the images due to gain-calibration errors, this will show as additional power above the thermal noise level. Therefore, better DD-gain calibration generally leads to less power in image space, although sources may leave more residual power on some spatial scales than others. Fig. 10 shows the ratio between the power spectrum of the images created using the robust and standard regularization methods as a function of spatial scale, i.e. $P_{\rm robust}(\ell)/P_{\rm standard}(\ell)$, where the subscripts 'robust' and 'standard' denote the calibration methods. Points below unity generally mean that the robust regularization method performs better, although the ratio also contains noise. The left panel shows the results with the point-source model (for all nights), and the right panel shows the results with the multipoint model. The different colours indicate different smoothing kernels.

Similar to what was found in visibility space, the robustly regularized DD-gain calibration method performs similarly to or better than the standard method throughout the spectrum, with only a few points exceeding unity, which we attribute to noise. In simulation N3 with the 4 MHz kernel, robust regularization performs especially well compared to standard regularization. We expect this to be because the off-axis sources have a higher elevation at the LST range of this simulation. Because of this, they are less attenuated by the antenna beam for this simulation and, therefore, have the highest impact. A similar effect is found for the standard deviation of this simulation in Appendix A. This suggests that the robust regularization method is able to mitigate the impact of strong sources in the sidelobes better on all scales in image space, even when strong spectral regularization is used.

Overall, however, the ratio remains near unity for both model types, except at the lowest ℓ-modes, especially those related to scales $\lesssim$0.06 arcmin^{-1}. These largest scales in the image correspond to the baselines with a projected length of ≤ 250 λ, at which the robust method performs significantly better than the standard method. These scales are important for many science cases, such as 21-cm signal observations of the EoR and Cosmic Dawn, and studies of the diffuse polarized foreground of the Milky Way. We attribute the improvement on the short baselines to the distinction between calibration and analysis baselines. This implies that the robust spectral regularizer results in gains that extrapolate better to shorter baselines than the standard regularizer,[17] which may occur because the gains are less biased by strong spectral variations in the beam-sidelobes.

5 DISCUSSION AND CONCLUSIONS

In this work, we present simulations of the effect of beam errors due to broken receiver elements in the LOFAR-HBA system that are unaccounted for in state-of-the-art analysis software. We analysed their impact on the calibration of off-axis sources far from the target direction. The beam errors are simulated by switching off antenna elements within LOFAR-HBA tiles at a realistic level of tile degradation (typically up to three receivers per tile of 16). We also present a software package, SHIMMERR,[18] with which similar simulations can be done for other science cases and instruments. We present the resulting beam errors, along with the effects that they have on DD-gain calibration and subtraction of far-field sources. Based on these simulations, and calibration with the state-of-the-art DD-gain calibration code DDECAL, we come to the following conclusions:

Errors caused by nulls in the calibration beam. When a bright off-axis source is modelled near a null in the calibration beam, when there is no strong null in the true beam, some of the largest errors in DD-gain calibration solutions occur. Because of the extremely low amplitude and high relative variability of the calibration beam, the required gain corrections in these nulls are typically very large. This is especially the case for arrays with stations with a regular layout. Because the nulls also vary spectrally on scales much smaller than the assumed spectral smoothness scale, the gain solutions in the nulls can cause incorrectly deviating gain solutions to spread to other channels. Nulls in the true beam that are not present in the calibration beam could theoretically cause similar problems, but the effect is smaller because the true beam is smoother than the calibration beam.

Spectral regularization heuristic. We show that applying heuristic spectral regularization, by weighting the spectral smoothing kernel for the gains by the model visibilities, is more robust against the type of errors mentioned above. It outperforms the standard spectral regularization method in terms of subtracting bright off-axis sources both in visibility and image space, especially on short baselines. Both the amplitude and number of extreme gain-solution failures introduced during calibration are reduced. This holds both before and after outlier excision, for various spectral smoothness scales and off-axis sky models. Even in the least extreme case investigated, a multipoint-source off-axis sky model, we show that robust regularization significantly outperforms the standard regularization

[17] Due to the baseline cut, short baseline gains are constructed from the gain solutions found on longer baselines.
[18] https://github.com/Stefanie-B/shimmerr

method on the shortest baselines. Although we have tested this robust weighting scheme in the DD-gain calibration method DDECAL only, it is also straightforward to implement in other DD-gain calibration codes.

We emphasize that this robust reweighting scheme is a simple heuristic that tackles only the incorrect spectral smoothness assumption used in DD-gain calibration. Now that its effectiveness has been shown and a simulation and calibration framework for quick testing is available in SHIMMERR, future work can build on this and introduce other regularization improvements. For example, a more complex basis for modelling errors in a null, that takes into account both spectral and temporal information, can be built. Furthermore, future work can investigate the spatial assumptions in DD calibration in which a constant gain is assumed for a set of sources (i.e. a calibration cluster). Because DD beam errors result from variations between elements within a station, they may be better tackled through calibrating the element gains rather than station gains. The spatial structure of the beam then follows from a physically motivated model, which does not require discontinuities.

Spectral smoothness scales. Due to large- and small-scale spectral gain changes of the beam far from the target direction, smaller spectral smoothness scales seem to be favoured for off-axis sources than for target field sources. Therefore, using dynamic spectral smoothness scales, where sources that are further from the target direction receive a lower spectral smoothness scale, perform better in the presence of DD beam-modelling errors.

The robust regularizer has not been tested in the presence of more general cases of extended emission. This is because current DDECAL evaluates the beam response of extended emission at discrete points, effectively removing the difference between point sources and extended sources in the context of beam-modelling errors. Furthermore, beam-modelling errors are expected to have a lower impact on extended emission, since extended emission does not fully fall into deep beam nulls. Therefore, the difference between the standard and robust regularizer is expected to be smaller for extended emission. Using the difference between the single-point and multipoint off-axis source models as a first-order probe for how the robust regularizer performs for extended emission, we expect extended emission to stabilize the solutions further.

Diffuse extended emission in the sidelobes of the beam has been shown to pose a challenge for 21-cm cosmology with the drift-scan instrument HERA (Hydrogen Epoch of Reionization Array, Charles et al. 2023), necessitating a spatial filtering technique to remove this emission prior to calibration. This is not expected to be needed for LOFAR, because LOFAR has more long baselines, allowing it to have a higher baseline cut, that drastically reduces the impact of diffuse emission during gain calibration. The HERA baseline cut is at 40 m, whereas that of LOFAR is at 250 $\lambda \approx 500$ m in the relevant frequency range. Furthermore, Höfer et al. (2025) have shown that the combination of beam errors and diffuse emission does not lead to a bias in the 21-cm power spectrum for LOFAR during DI calibration with a lower baseline cut.

In conclusion, beam modelling for phased-array stations such as those of LOFAR, NenuFAR, MWA (where we consider a tile to be a station), and the upcoming SKA-Low remains a challenging endeavour due to changes in the beam shapes as a function of pointing, time (due to rotation of the Earth), and observing frequency, as well as additional temporal changes due to the behaviour of the telescope due to degrading hardware and weather. The SKA-Low array, for example, will have a higher sensitivity than the current generation of instruments, and likely also have a more complex antenna beam. Therefore, beam effects such as those present for LOFAR-HBA stations will similarly affect SKA-Low. Furthermore, SKA-Low's higher sensitivity may cause far-field sources currently below the noise to appear in the data. Therefore, applying DD-gain calibration methods that are robust against often unavoidable beam-modelling errors in the distant sidelobes and nulls are necessary to use phased-array telescopes to their full capabilities.

ACKNOWLEDGEMENTS

SAB, LVEK, KC, CH, LG, SG, and SM acknowledge the financial support from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 884760, 'CoDEX'). EC (INAF) would like to acknowledge support from the Centre for Data Science and Systems Complexity (DSSC), Faculty of Science and Engineering at the University of Groningen, and from the Ministry of Universities and Research (MUR) through the PRIN project 'Optimal inference from radio images of the epoch of reionization'. FGM acknowledges support from the I-DAWN project, funded by the DIM-ORIGINS programme.

Apart from software that is mentioned in the main body of the text, in this work, we made use of the DS9 (Joye et al. 2003) FITS file image viewer, and the NUMPY (Harris et al. 2020), NUMBA (Lam, Pitrou & Seibert 2015), ASTROPY (Astropy Collaboration 2013, 2018, 2022), PYTHON-CASACORE (The CASA Team 2022), MATPLOTLIB (Hunter 2007), and PANDAS (McKinney 2010) PYTHON packages.

DATA AVAILABILITY

The simulation software underlying this article are created with SHIMMERR, available at https://doi.org/10.5281/zenodo.15114900. The simulations underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

Albert J. G., van Weeren R. J., Intema H. T., Röttgering H. J. A., 2020, A&A, 635, A147
Arras P., Frank P., Leike R., Westermann R., Enßlin T. A., 2019, A&A, 627, A134
Astropy Collaboration, 2013, A&A, 558, A33
Astropy Collaboration, 2018, AJ, 156, 123
Astropy Collaboration, 2022, ApJ, 935, 167
Baldwin J. E., Boysen R. C., Hales S. E. G., Jennings J. E., Waggett P. C., Warner P. J., Wilson D. M. A., 1985, MNRAS, 217, 717
Best P. N. et al., 2023, MNRAS, 523, 1729
Birdi J., Repetti A., Wiaux Y., 2020, MNRAS, 492, 3509
Bowman J. D. et al., 2013, Publ. Astron. Soc. Aust., 30, e031
Brackenhoff S. A. et al., 2024, MNRAS, 533, 632
Ceccotti E. et al., 2025, A&A, 696, A56
Charles N., Kern N., Bernardi G., Bester L., Smirnov O., Fagnoni N., Acedo E., 2023, MNRAS, 522, 1009
Chokshi A., Barry N., Line J. L. B., Jordan C. H., Pindor B., Webster R. L., 2024, MNRAS, 534, 2475
Cohen A. S., Lane W. M., Cotton W. D., Kassim N. E., Lazio T. J. W., Perley R. A., Condon J. J., Erickson W. C., 2007, AJ, 134, 1245
Condon J. J., Cotton W. D., Greisen E. W., Yin Q. F., Perley R. A., Taylor G. B., Broderick J. J., 1998, AJ, 115, 1693
Cronyn W. M., 1972, ApJ, 174, 181
de Gasperin F. et al., 2019, A&A, 622, A5
de Gasperin F. et al., 2023, A&A, 673, A165
Dewdney P. E., Hall P. J., Schilizzi R. T., Lazio T. J. L. W., 2009, Proc. IEEE, 97, 1482

Edge D. O., Shakeshaft J. R., McAdam W. B., Baldwin J. E., Archer S., 1959, Mem. RAS, 68, 37
Elder K., Jacobs D. C., 2024, preprint (arXiv:2411.04193)
Ewall-Wice A., Dillon J. S., Liu A., Hewitt J., 2017, MNRAS, 470, 1849
Franzen T. M. O. et al., 2016, MNRAS, 459, 3314
Gan H. et al., 2022, A&A, 663, A9
Gan H. et al., 2023, A&A, 669, A20
Hale C. L. et al., 2019, A&A, 622, A4
Hales S. E. G., Riley J. M., Waldram E. M., Warner P. J., Baldwin J. E., 2007, MNRAS, 382, 1639
Hamaker J. P., Bregman J. D., Sault R. J., 1996, A&AS, 117, 137
Harris C. R. et al., 2020, Nature, 585, 357
Heald G., McKean J., Pizzo R.eds, 2018, Low Frequency Radio Astronomy and the LOFAR Observatory: Lectures from the Third LOFAR Data Processing School. Astrophysics and Space Science Library. Vol. 426, Springer International Publishing, Cham. Available at: http://link.springer.com/10.1007/978-3-319-23434-2
Höfer C. et al., 2025, preprint (arXiv:2504.03554)
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Hurley-Walker N. et al., 2017, MNRAS, 464, 1146
Iheanetu K., Girard J. N., Smirnov O., Asad K. M. B., de Villiers M., Thorat K., Makhathini S., Perley R. A., 2019, MNRAS, 485, 4107
Intema H. T., Jagannathan P., Mooley K. P., Frail D. A., 2017, A&A, 598, A78
Jacobs D. C. et al., 2017, PASP, 129, 035002
Joye W. A., Mandel E., 2003, in Payne H. E., Jedrzejewski R. I., Hook R. N., eds, ASP Conf. Ser. Vol. 295, Astronomical Data Analysis Software and Systems XII. Astron. Soc. Pac., San Francisco, p. 489
Koopmans L. V. E. et al., 2015, Proc. Sci., The Cosmic Dawn and Epoch of Reionisation with SKA. SISSA, Trieste, Pos#001. https://arxiv.org/abs/1505.07568
Koopmans L. V. E., 2010, ApJ, 718, 963
Lam S. K., Pitrou A., Seibert S., 2015, in Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC. LLVM'15. Association for Computing Machinery, New York, p. 1. Available at: https://dl.acm.org/doi/10.1145/2833157.2833162
Line J. L. B. et al., 2018, Publ. Astron. Soc. Aust., 35, e045
Lonsdale C. J., 2005, in Kassim N. E., Perez M. R., Junor W., Henning P. A., eds, ASP Conf. Ser. Vol. 345, From Clark Lake to the Long Wavelength Array: Bill Erickson's Radio Science. Astron. Soc. Pac., San Francisco, p. 399
McKinney W., 2010, SciPy, 445, 51
Mertens F. G. et al., 2020, MNRAS, 493, 1662
Mertens F. G. et al., 2025, A&A, 698, A186
Mevius M. et al., 2021, MNRAS, 509, 3693
Mouri Sardarabadi A., Koopmans L. V. E., 2019, MNRAS, 483, 5480
Munshi S. et al., 2024, A&A, 681, A62
Munshi S. et al., 2025, A&A, 693, A276
Murphy T. et al., 2013, Publ. Astron. Soc. Aust., 30, e006
Nunhokee C. D. et al., 2020, ApJ, 897, 5
O'Hara O. S. D. et al., 2025, MNRAS, 538, 31
Offringa A. R., van de Gronde J. J., Roerdink J. B. T. M., 2012, A&A, 539, A95
Offringa A. R. et al., 2014, MNRAS, 444, 606
Offringa A. R., Mertens F., Tol S. v. d., Veenboer B., Gehlot B. K., Koopmans L. V. E., Mevius M., 2019, A&A, 631, A12
Patil A. H. et al., 2016, MNRAS, 463, 4317
Patil A. H. et al., 2017, ApJ, 838, 65
Pilkington J. D. H., Scott J. F., 1965, Mem. RAS, 69, 183
Rengelink R. B., Tang Y., de Bruyn A. G., Miley G. K., Bremer M. N., Roettgering H. J. A., Bremer M. A. R., 1997, A&AS, 124, 259
Repetti A., Birdi J., Dabbech A., Wiaux Y., 2017, MNRAS, 470 3981
Roth J., Arras P., Reinecke M., Perley R. A., Westermann R., Enßlin T. A., 2023, A&A, 678, A177
Sardarabadi A. M., Koopmans L. V. E., 2019, preprint (arXiv:1902.02482[astro-ph])
Shimwell T. W. et al., 2017, A&A, 598, A104
Shimwell T. W. et al., 2022, A&A, 659, A1
Smirnov O. M., 2011a, A&A, 527, A106
Smirnov O. M., 2011b, A&A, 527, A107
Smirnov O. M., Tasse C., 2015, MNRAS, 449, 2668
The CASA Team, 2022, PASP, 134, 114501
Thompson A. R., Moran J. M., Swenson G. W., Jr, 2001, Interferometry and Synthesis in Radio Astronomy. Springer Nature, New York
Trott C. M. et al., 2020, MNRAS, 493, 4711
Trott C. M., 2021, J. Astron. Telesc. Instrum. Syst., 8, 011011
Trott C. M., Wayth R. B., 2016, Publ. Astron. Soc. Aust., 33, e019
van der Tol S., Jeffs B. D., van der Veen A.-J., 2007, IEEE Trans. Signal Process., 55, 4497
van Diepen G., Dijkema T. J., Offringa A., 2018, Astrophysics Source Code Library, record ascl:1804.003
van Haarlem M. P. et al., 2013, A&A, 556, A2
van Weeren R. J. et al., 2016, ApJS, 223, 2
Wijnholds S. J., Arts M., Bolli P., di Ninni P., Virone G., 2019, 2019 International Conference on Electromagnetics in Advanced Applications (ICEAA), Granada, Spain, p. 437
Wijnholds S. J., Bregman J. D., van Ardenne A., 2011, Radio Sci., 46, 1
Wilensky M. J. et al., 2024, RAS Tech. Instrum., 3, 400
Yatawatta S., 2015, MNRAS, 449, 4506
Zarka P., Girard J. N., Tagger M., Denis L., 2012, in Boissier S., de Laverny P., Nardetto N., Samadi R., Valls-Gabaud D., Wozniak H., eds, SF2A-2012: Proceedings of the Annual meeting of the French Society of Astronomy and Astrophysics, p. 687

APPENDIX A: RESIDUAL VISIBILITY STATISTICS

In Fig. 9, the standard deviations and flagging fractions of two of the simulations in this work are shown. Fig. A1 in this appendix contains these statistics for all seven simulations, in all considered calibration setups.

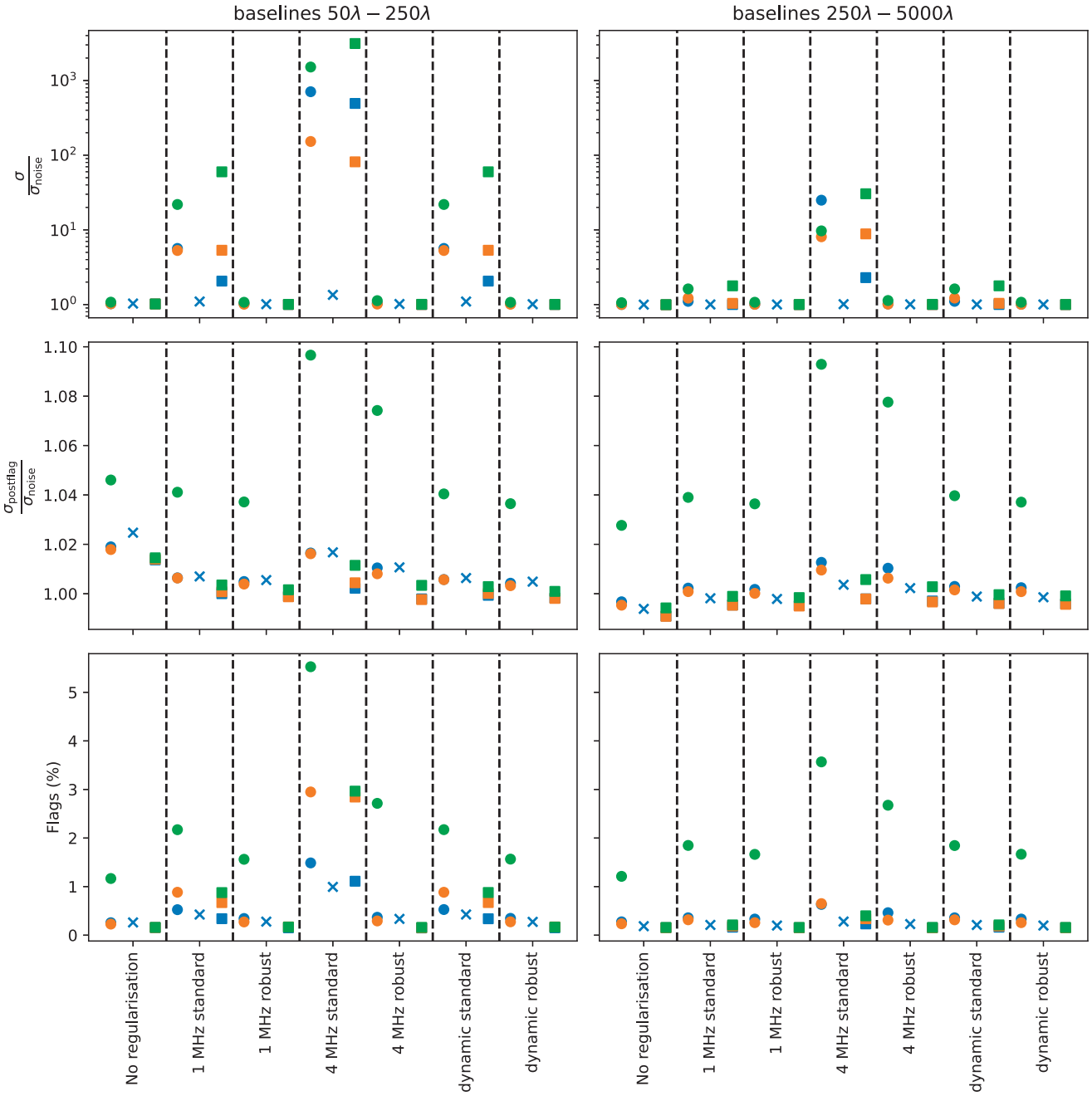

Figure A1. Standard deviations compared to the noise level before and after flagging and flagging percentages for all simulations under consideration in this work. We split the statistics into analysis baselines (< 250 λ, left-hand column) and calibration baselines (> 250 λ, right-hand column). The circles, crosses, and squares denote the single point source, multipoint source, and attenuated models for Cas A and Cyg A, respectively, and are horizontally offset for legibility. The colours denote different observing LSTs, with blue, orange, and green representing N1, N2, and N3, respectively. The bins within the panels, separated with dashed lines, show the different regularization settings using in calibration.

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