The impact of lossy data compression on the power spectrum of the high redshift 21-cm signal with LOFAR

J.K Chege, L.V.E Koopmans, A.R Offringa, B.K Gehlot, S.A Brackenhoff, E Ceccotti, S Ghosh, C Höfer, F.G Mertens, M Mevius, et al.

A&A, 692, A211 (2024)
https://doi.org/10.1051/0004-6361/202451367

Astronomy & Astrophysics

The impact of lossy data compression on the power spectrum of the high-redshift 21 cm signal with LOFAR

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

[1] Kapteyn Astronomical Institute, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands
[2] Netherlands Institute for Radio Astronomy (ASTRON), Oude Hoogeveensedijk 4, 7991 PD Dwingeloo, The Netherlands
[3] LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Université, F-75014 Paris, France

Received 3 July 2024 / Accepted 7 November 2024

ABSTRACT

Context. Current radio interferometers output multi-petabyte-scale volumes of data per year, making the storage, transfer, and processing of these data a sizeable challenge. This challenge is expected to grow with next-generation telescopes such as the Square Kilometre Array (SKA), which will produce a considerably larger data volume than current instruments. Lossy compression of interferometric data post-correlation can abate this challenge, but any drawbacks from the compression should be well understood in advance.

Aims. Lossy data compression reduces the precision of data, introducing additional noise. Since high-redshift (e.g., cosmic dawn or epoch of reionization) 21 cm studies impose strict precision requirements, the impact of this effect on the 21 cm signal power spectrum statistic is investigated in a bid to rule out unwanted systematics.

Methods. We applied DYSCO visibility compression, a technique for normalizing and quantizing specifically designed for radio interferometric data, to observed visibilities datasets from the LOFAR telescope as well as simulated ones. The power spectrum of these data was analyzed, and we establish the level of the compression noise in the power spectrum in comparison to the thermal noise. We also examined its coherency behavior by employing the cross-coherence metric. Finally, for optimal compression results, we compared the compression noise obtained from different compression settings to a nominal 21 cm signal power.

Results. From a single night of observation, we find that the noise introduced due to the compression is more than five orders of magnitude lower than the thermal noise level in the power spectrum. The noise does not affect calibration. Furthermore, the noise remains subdominant to the noise introduced by the nonlinear calibration algorithm used following random parameter initialization across different runs. The compression noise shows no correlation with the sky signal and has no measurable coherent component, therefore averaging down optimally with the integration of more data. The level of compression error in the power spectrum ultimately depends on the compression settings.

Conclusions. DYSCO visibility compression is found to be an insignificant concern for 21 cm power spectrum studies. Hence, data volumes can be safely reduced by factors of ∼4 with insignificant bias to the final power spectrum. Data from SKA-Low will likely be compressible by the same factor as data from LOFAR owing to the similarities of the two instruments. The same technique can be used to compress data from other telescopes, but a small adjustment of the compression parameters might be required.

Key words. instrumentation: interferometers – methods: data analysis – methods: observational – methods: statistical – techniques: interferometric – cosmology: observations

1. Introduction

Over the past several decades, radio interferometers have been expanding in physical dimensions, transitioning into what is referred to as the large-N regime, due to an ever-increasing number of antennas used. This trend is particularly pronounced in low-frequency instruments, where the relatively low cost of antenna components makes it economically viable to construct phased arrays consisting of hundreds of antennas. Such arrays include the Low-Frequency Array (LOFAR; van Haarlem et al. 2013), the New Extension in Nançay Upgrading LOFAR (NenuFAR; Zarka et al. 2012), the Murchison Widefield Array (MWA; Tingay et al. 2013), and the Hydrogen Epoch of Reionization Array (HERA; Deboer et al. 2017). The forthcoming Square Kilometre Array-Low (SKA-Low; Braun et al. 2019; Mellema et al. 2013) will have more than 10^5 antennas, considerably more than current instruments.

An increased number of observing antennas that are correlated (or other correlated receiver systems such as tiles or stations) results in a substantially larger data volume. This is because the data output from an interferometer – referred to as the visibility – is a short-time-integrated cross-correlation of electric fields from each antenna pair (Thompson et al. 2017), therefore scaling as $O(N^2)$, where N is the number of correlated components. Moreover, to cater to different requirements, the visibilities are often recorded and written to disk at high temporal and spectral resolutions over long observation durations and large instantaneous bandwidths. Temporal resolution is preferable for the study of highly dynamic time-domain phenomena such as transient radio-frequency interference (RFI; e.g., Gehlot et al. 2024) and ionospheric effects (e.g., Chege et al. 2022). Similarly, observations with high spectral resolution are often required, for example in the study of radio spectral lines

⋆ Corresponding author; chege@astro.rug.nl

(e.g., Asgekar et al. 2013). Some science cases such as deep all-sky radio surveys require observations of large sky areas (e.g., Shimwell et al. 2017), while others require integration of large data volumes obtained from the same sky field over a long duration. An example of the latter are the observations targeting the cosmic 21 cm signal emitted during the epoch of reionization (EoR) and the cosmic dawn (e.g., Paciga et al. 2013; Patil et al. 2017; Cheng et al. 2018; Gehlot et al. 2019; Kolopanis et al. 2019; Li et al. 2019; Mertens et al. 2020; Trott et al. 2020; HERA Collaboration 2022, 2023; Munshi et al. 2024). All these varying requirements result in considerable volumes of unprocessed observational data.

The output volumes from current interferometers have already grown to petabyte scales[1] (Sabater et al. 2015), and with the forthcoming SKA-Low telescope, the data deluge is bound to surge into unprecedented exabyte volumes. This "big data" demands large storage spaces coupled with complex network architectures for prompt retrieval and transfer, sometimes over thousands of kilometers, before any processing can be undertaken. For this reason, data archiving and management alone becomes a significantly expensive and sometimes limiting part of a science project. A viable solution capable of saving substantial storage resources and considerably mitigating the input–output bottleneck could be the application of data compression techniques.

A compression method can be said to be either lossy or lossless depending on whether or not information is lost during the data compression. Lossless compression methods (e.g., GZIP[2]), while fully preserving every single data bit, rely on structured information, which is scarce in noisy data. Hence, lossless methods achieve compression factors of only a few tens of percent on noisy data (Lindstrom 2017). Lossy methods achieve much higher compression factors on noisy data but at the cost of losing some information.

At high resolutions, radio data are very noisy, and as such lossy compression methods are usually preferred over lossless ones. However, lossy compression, as with any other data transforming step, should preferably not bias the final science output in any significant way. This is especially the case for high-redshift 21 cm studies, as they target an extremely faint signal and thus require the integration of a thousand hours (e.g., for the LOFAR telescope) before a detection can be achieved (e.g., Mesinger et al. 2016).

Furthermore, this signal is buried under strong Galactic and extragalactic foregrounds that are three to five orders of magnitude brighter (e.g., Jelić et al. 2008, 2010), exacerbating the challenge. High-redshift 21 cm studies thus aim for minimal systematic biases and errors (e.g., Barry et al. 2016). Errors resulting from lossy visibility data compression should not further complicate this already considerable challenge.

Several algorithms have already been developed for the specific purpose of compressing radio visibility data. They include BITSHUFFLE (Masui et al. 2015), implemented for lossless compression of integer data from the Canadian Hydrogen Intensity Mapping Experiment (CHIME) and reported to achieve data compression of almost a factor of 4. Several other compression methods have been developed for specific data formats such as the Flexible Image Transport System[3] (FITS; Wells et al. 1981) file format and the Astronomical Image Processing System[4] (AIPS; see, e.g., White et al. 2012). For noisy complex visibilities data in the MeasurementSet[5] data format, the Dynamical Statistical Compression (DYSCO; Offringa 2016) tool was developed to perform lossy compression and has been shown to achieve a compression factor of 4 or more on LOFAR and MWA visibilities.

In this study we investigated the impact of lossy compression on visibilities in 21 cm observations data processing as a means of tackling excessive data volume challenges. While antenna-specific recorded voltages are normally quantized before correlation in well-understood procedures, compression of data after correlation and its effect in the case of the 21 cm signal observations remain largely unexplored. Previous work by Offringa (2016) investigated the image-space effects when compressing visibility data with DYSCO. In this work we examined the impact of data compression specifically on studies of the high-redshift 21 cm signals using the power spectrum method, which is the conventional metric of 21 cm signal measurements in most current EoR studies. We quantify the compression noise added to the power spectrum and establish its behavior. Specifically, we tackle three pertinent questions: (i) What is the level of compression noise compared to thermal noise? (ii) Is compression noise incoherent? (iii) Does compression noise affect calibration?

We describe DYSCO, the compression tool used in this work, in Sect. 2 before describing the observation and simulation data used in Sect. 3. The data processing methodology is then described in Sect. 4, and the results are presented in Sect. 5. Our conclusions are outlined in Sect. 6.

2. Dynamical Statistical Compression (DYSCO)

In this section we briefly summarize the data compression tool used throughout this paper, namely DYSCO. DYSCO[6] was developed by Offringa (2016) and it consists of a visibility compression algorithm and a CASACORE[7] standard data storage manager that enables transparent storage of compressed data in the MeasurementSet format. In this way, the compressed data can be written to disk and processing can proceed normally without any additional steps. DYSCO is already integrated into both LOFAR and MWA data preprocessing pipelines.

DYSCO compression is performed in two consecutive steps: a normalization and a quantization step. The normalization ensures that the full dataset has a constant noise variance. The noise distribution in visibility data can vary across different antennas, polarizations, timesteps, and frequencies. Therefore, assumptions made during the normalization step regarding the noise distribution across the four dimensions will have an impact on the compression accuracy. For instance, the row-normalization (R) method assigns a scaling factor per "row", which contains data from the same antenna and timestep, but different polarizations and frequencies. Due to the assumption of uniform variance across multiple polarizations and frequencies, the row-normalization method has been shown to perform significantly worse in the image space by adding much higher noise in comparison to the other available methods. Alternatively, DYSCO incorporates two more normalization methods, namely the row-frequency (RF) and the antenna-frequency (AF) normalization methods. The former is similar to the row-normalization method but with an additional scaling factor per frequency channel.

[1] The LOFAR EoR Key Science Project project alone has ≈5 petabytes of archival data.
[2] https://www.gzip.org/
[3] https://fits.gsfc.nasa.gov/
[4] http://www.aips.nrao.edu
[5] https://casa.nrao.edu/Memos/229.html
[6] https://github.com/aroffringa/dysco
[7] https://casacore.github.io/casacore

Table 1. Bit size, normalization method, and quantization distribution parameters used for DYSCO compression in DP3. While many other bit sizes can be used, here we list the bit sizes as recommended from prior tests. The default DP3 values are in the second row.

Bit size	Normalization	Distribution	Expected compression factor
8	Row-Frequency	1.5σ Truncated Gaussian	6
10	Antenna-Frequency	2.5σ Truncated Gaussian	4
12	Row	3.5σ Truncated Gaussian	3.5
16		Gaussian	2.5
		Uniform	
		Student's T	

The method also stores the per-polarization normalization factors separately. The latter uses a three-term normalization factor composed of a frequency channel factor and a factor for each of the two correlated antennas. The normalization here is also done independently of each timestep and polarization.

After normalization, the data are encoded using a nonlinear quantization scheme with dithering. The encoding is optimized for complex samples using a distribution with a zero mean in such a way that probable values are more accurately compressed than less probable values. The choice of such a distribution is also optimizable by the DYSCO user. The encoded values are finally converted to binary values and bit-packed using a chosen number of bits.

The compression bit size, normalization, and quantization distribution parameters available for compression in DYSCO are listed in Table 1. The default values used by the Default Preprocessing Pipeline[8] (DP3; van Diepen et al. 2018), which is used extensively for LOFAR data analysis, are AF normalization, a Gaussian distribution that is truncated at 2.5σ (only the distribution that is used to compute the ideal encoding is truncated; actual visibilities are never truncated, as during the normalization it is made sure that all visibility values fit within the chosen distribution), and 10 bits. We studied whether these default settings are sufficient to compress LOFAR 21 cm signal data, or whether other settings are needed. We finally note that given that uncompressed data are typically stored in 32 or 64-bit format, storing them in a 10-bit format leads to the earlier-mentioned substantially smaller data volumes. We note that the metadata in the measurement sets are not compressed and some metadata (i.e., scale factors for the RF and AF normalizations) are added, hence leading to a slightly lower compression factor compared to the simple ratio of bits per visibility.

3. Data acquisition

In this section we describe the data used in the rest of the paper, which includes both real and simulated radio observations.

3.1. Real observations

The datasets used in examining the effects of lossy data compression were obtained with LOFAR (van Haarlem et al. 2013). LOFAR is a low-frequency radio interferometer and a pathfinder instrument for the SKA with a geographical footprint centered in the Netherlands and spreading out into multiple European countries. It can observe in two frequency bands using the low-band antennas (LBAs; 10−90 MHz) and high-band antennas (HBAs; 110−240 MHz), respectively. The antennas are phased-up into stations, with the core consisting of 48 stations (24 stations each split into two separate stations) densely packed within a 2-kilometer-wide area near the town of Exloo in Drenthe. An additional 14 stations are located further across the Netherlands while 14 others are located in different European countries. These are referred to as the remote and international stations, respectively. The core, remote and international stations, have maximum baselines of approximately 4, 120, and 2000 km, respectively.

In our analysis, we used a typical LOFAR HBA dataset from the Cycle 2 observing season, retrieved from the LOFAR Long Term Archive. In total, two nights of observation were used, spanning a duration of 12 hours per night and a 12 MHz bandwidth between 148 and 160 MHz. The raw data have a 12 kHz and 2 s frequency and time resolution, respectively. More information about this dataset is summarized in Table 2.

Table 2. Two nights of real LOFAR HBA observations analyzed in this work.

Parameter	L246297	L246309
Observation cycle	2	2
UTC [(a)] start date-time	2014-10-23 16:46:30	2014-10-16 17:01:41
LST [(b)] start-time [hour]	19.3	19.1
Duration [hour]	13.0	12.6
SEFD [(c)] estimate	4294	4253
Number of stations [(d)]	62	62
Frequency range (MHz)	148–160	148–160
Frequency resolution [(e)] (kHz)	12.2, 61.0	12.2, 61.0
Time resolution (s)	2.0	2.0

Notes. [(a)]Coordinated Universal Time. [(b)]Local Sidereal Time. [(c)]System Equivalent Flux density; as reported in Mertens et al. (2020). [(d)]International stations not included. [(e)]Values corresponding to data with 15 and 3 channel per sub-band, i.e., before and after frequency averaging.

3.2. Simulations

Besides using real LOFAR observation data, we complemented our study with simulated data. These simulations are based on the measurement set of the real observation L246297, listed in Table 2, with the simulated data replacing the observed data for consistent data structure and properties. We simulated two sets of data, which we refer to as simulation sets A and B. Table 3 lists the dataset properties and compression settings used for the different simulations.

Firstly, for simulation set A, we used all core and remote stations to simulate two observation datasets, both spanning the

[8] https://github.com/lofar-astron/DP3

Table 3. Simulated LOFAR HBA observation data.

Parameter	Simulation A	Simulation B
MS template	L246297	L246297
Stations	CS+RS	CS
Duration [hour]	12.0	6.0
Bit size	10	10, 12, 16
Normalization	AF	AF, RF
Distribution	2.5σ truncated Gaussian	2.5σ truncated Gaussian

same 12-hour duration. The datasets had an identical foreground emission comprising compact extragalactic radio sources but a unique and independent noise realization We modeled an area of 10° around the north celestial pole (NCP) using the brightest ∼700 sources. We also included the far-field Cygnus A and Cassiopeia A sources as they are bright enough to have a significant impact on the processing of the NCP field. More details on this simplified NCP model can be found in Brackenhoff et al. (2024).

For simulation set B, we used the same foreground model as simulation set A, but varied the compression bit size and normalization parameters. In contrast with simulation A where we used different noise realizations per dataset, in simulation B, we added the same noise realization to all datasets. Here, we also limited ourselves to including only the core LOFAR stations and a shorter observation duration of 6 hours. This reduced dataset is chosen for less memory usage and quicker computation.

All datasets generated from both simulations (A and B) included the instrumental beam attenuation effect and were carried out at the same time and frequency resolution as the real raw data before any averaging, as listed in Table 2. The simulations were done using the SAGECAL[9] algorithm (Yatawatta et al. 2013; Yatawatta 2015). For each simulated dataset, its uncompressed version was used as the reference dataset.

4. Data processing, compression, and calibration

In this section we describe the data reduction steps carried out for our analysis. The analysis closely follows the steps applied in the LOFAR EoR Key Science Project data processing pipeline (see Mertens et al. 2020). However, we introduced data compression as an additional data preprocessing step where needed.

4.1. Preprocessing and compression

In the preprocessing step, data were first run through an RFI excision step carried out using AOFlagger[10] (Offringa et al. 2012). This step serves not only to get rid of unwanted terrestrial signals but also to reduce any dynamic range contribution by RFI to the data. A reduced dynamic range improves the performance of the normalization step during compression improving the overall compression performance. While RFI flagging has no effect on simulated data, we retained the RFI flags obtained from real data in our simulation datasets to replicate a realistic observation. All data from the international stations was also flagged in this step.

The data were then compressed using DYSCO. We first used the default DP3 compression parameters recommended by Offringa (2016) for our relatively high-resolution noise-dominated data. Such noisy data are typical for 21 cm signal studies since they usually target sky fields with minimal foreground power and ease of foregrounds modeling. Later, we examined whether these default parameters are sufficient for the precision required in high-redshift 21 cm signal studies. For improved signal-to-noise ratios (S/Ns) per solution time interval during calibration and for quicker computation, the data were averaged to 61 kHz per spectral channel.

4.2. Calibration

Throughout this work, we limited ourselves to performing the first direction-independent (DI) calibration step in the LOFAR EoR pipeline. One calibration step is sufficient for answering the question of whether compression errors affect calibration and its results would apply to the other stages of calibration. This is because compression is only applied once on the highest resolution raw data as it is the most voluminous. This is typically done for archival purposes. Decompression is then done prior to any further processing and therefore any compression effects on calibration should manifest clearly in the first calibration stage. Moreover, multiple compression and decompression runs during processing are not recommended as each iteration would introduce additional noise to the visibilities. Averaging in time and frequency carried out during processing reduces the data volume by a factor of 25 (the data are averaged from 15 to 3 frequency channels per sub-band for DI and from 2 s to 10 s time integration for the direction-dependent calibration step). However, intermediate visibilities during the different calibration steps occupy additional columns in the measurement set but require relatively limited disk space due to the averaging.

All gain calibration was carried out using SAGECAL (Yatawatta 2016). For the real data, the calibration sky model was composed of two sky directions, one around the NCP and the other for the bright source 3C 61.1 with 1333 and 1545 components, respectively. A calibration solution interval of 30 s was used with a single solution being obtained for each 183.3 kHz sub-band. A minimum and maximum calibration baseline cutoff was set at 50λ and 5000λ, respectively, and the resulting gains were regularized using a third-order Bernstein polynomial. The resulting Bernstein polynomial for the central NCP direction was applied to the data to obtain the calibrated visibilities. The baseline cutoff and gains smoothing have been deemed crucial to minimize signal suppression and noise boost in the 21 cm power spectrum (e.g., Barry et al. 2016; Mevius et al. 2022). More details on these calibration parameter choices are available in Patil et al. (2017), Mouri Sardarabadi & Koopmans (2019), and Mertens et al. (2020). The simulated data were calibrated similarly, with the only difference being in the sky model, which in this case was composed of fewer components since the simulations had fewer compact foregrounds.

4.3. Generation of power spectra

Results presented in this paper are based on power spectra generated from different visibility datasets, real or simulated, raw or calibrated using the PSPIPE[11] tool. In this section we describe the process involved.

First, the visibilities from all sub-bands are gridded and transformed into image cubes using WSClean (Offringa et al. 2014). Each image cube has a field of view (FOV) of $12° \times 12°$ centered at the NCP with an angular resolution of 0.5 arcmin.

[9] https://github.com/nlesc-dirac/sagecal
[10] https://sourceforge.net/projects/aoflagger
[11] https://gitlab.com/flomertens/pspipe

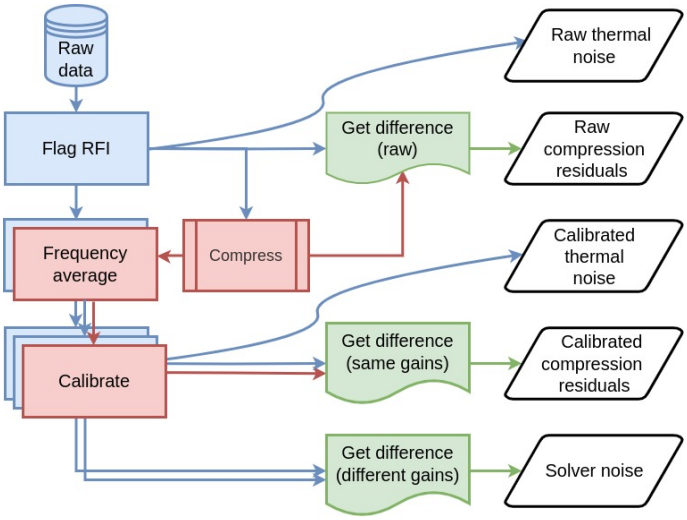

Fig. 1. Processing flow for the different visibility products used to obtain the compression residual noise, thermal noise, and solver noise power spectra displayed in Figs. 2 and 3. The flow of compressed and uncompressed data is indicated in blue and red, respectively, while different residuals are shown in green. Overlapping panels imply steps carried out multiple times on different datasets or on the same dataset). The two curved lines represent the thermal noise variance obtained as the difference of even-odd timestep image data. The exact details involved in each step are described in the text.

For the actual power spectrum, this FOV is then reduced to 4° by use of a Tukey spatial window. The image cube was then re-gridded and converted from image space units of Jy/PSF[12] into brightness temperature units of kelvins together with a spatial Fourier transform step back into visibilities space. This final step was carried out on a visibilities subset that includes only baselines between 50λ and 250λ in length. Concurrently, an estimate of the thermal noise variance is obtained by generating a new cube composed of the difference between even and odd timesteps. From such cubes, we can obtain the power spectrum by first taking a Fourier transform along the frequency direction. The coordinates are then mapped into comoving distances in the form of wavenumbers ($\boldsymbol{k}$) with the appropriate cosmological units (Morales & Hewitt 2004; McQuinn et al. 2006). We used the common cylindrically averaged (2D; $k_\perp$ and $k_\parallel$ coordinates) and the dimensionless spherically averaged (1D, k) power spectra.

5. Results

In this section we present the results of our analysis. First, we establish the scale of the compression noise in comparison to thermal noise. We then show the coherence properties of the compression noise followed by an assessment of how data compression affects calibration. Finally, we examine what are the optimal DYSCO compression settings.

5.1. Compression noise

To determine the additional noise introduced to the 21 cm power spectrum due to visibility compression, we computed the difference between the compressed and the reference (not compressed) visibilities before calibration. In the following sections, we refer to the output of this subtraction as the compression residuals or the compression noise (Fig. 1). Figure 2 shows the comparison between these compression residuals and the reference thermal noise by use of the cylindrically averaged power spectrum. The noise introduced to the power spectrum due to data compression is shown to be around 5.5 orders of magnitude lower than the raw thermal noise. Similarly, the compression noise power is shown to be 4.5 orders of magnitude lower than the uncertainty of the reference thermal noise. Hence, even if compression noise were fully coherent (e.g., the result of compression of coherent foreground emission), it would only reach the level of the error on the thermal noise by adding about 30 000 times more data. Such a huge amount of data will, most likely, never be a requirement. Additionally, the ratios are devoid of any spatial structures, implying that compression noise does not introduce any spurious or scale-dependent errors. We studied the coherence, finding that the compression noise is consistent with being incoherent.

A similar metric can be obtained for calibrated data by applying identical calibration gains solutions to both the reference and the compressed dataset before obtaining the calibrated compression residuals. The need for identical solutions is to eliminate calibration "solver noise", a term that refers to the additional noise introduced due to random initialization of parameters by SAGECAL per calibration run, which leads to slightly different gain solutions after a finite number of iterations during the optimization (e.g., Mevius et al. 2022). It is known that calibration also introduces a systematic power contribution resulting from various factors, for example, the use of incomplete sky models, transfer of gains solution from longer to shorter baseline sets, and spectrally noisy gain solutions (e.g., Barry et al. 2016; Mevius et al. 2022). This systematic power is different from the solver noise referred to here, which is indeed random and intrinsic to the calibration algorithm used. Without identical calibration gains, the compression residuals would be dominated by this solver noise, although it is well below the thermal noise, as shown below. We thus applied the DI calibration gains obtained for the reference dataset to its compressed version (we show in Sect. 5.3 that the results do not change significantly if calibration solutions obtained from the compressed dataset are applied instead).

Subsequently, we reran the power spectrum computation on the calibrated compression residuals, and the result is shown in Fig. 3. Similar to raw data, the compression of visibilities does not introduce spurious noise after calibration. The ratio between the compression residuals and the thermal noise level remains at orders lower than 10^{-5} between their 2D power spectra. A slightly higher level is seen in the ratio of the compression residuals and the thermal noise uncertainty. Additionally, solver noise bias is shown to be more dominant in comparison to the compression residuals noise, but also well below the thermal noise and the error on the thermal noise.

5.2. Compression noise coherence

Having determined that the added compression noise is far below both the thermal and solver noise for LOFAR HBA data, we investigated whether this noise has any correlation with the sky signal and therefore is coherent in nature. Partly coherent compression noise would not only average down more slowly than incoherent noise, but it could also introduce biases on the 21 cm measurements obtained from deep integrations. For this test, we obtained three pairs of compression residuals from:

1. both LOFAR nights before calibration,

[12] PSF stands for point-spread function.

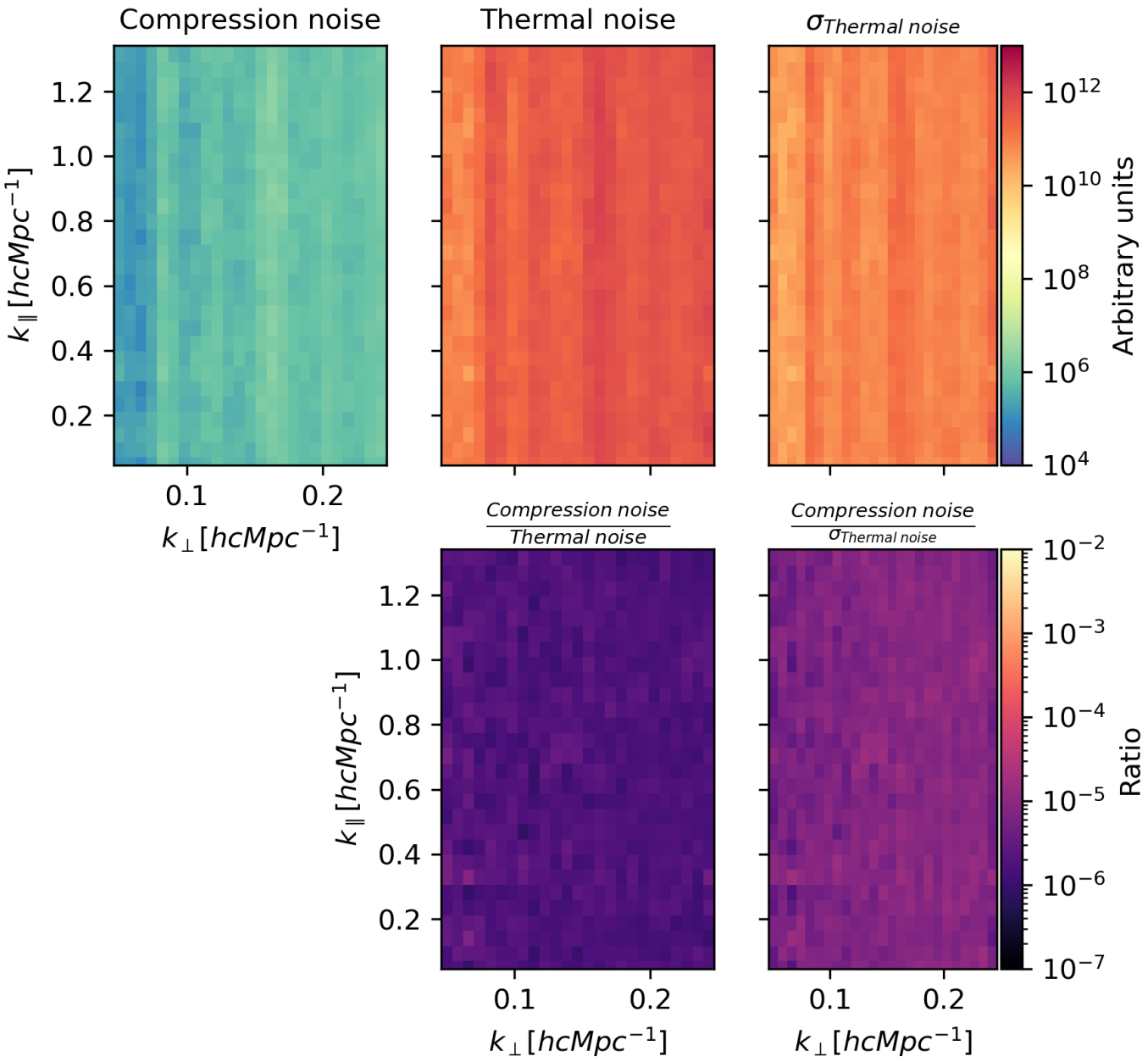

Fig. 2. Cylindrically averaged power spectra comparing the compression noise (the compressed minus uncompressed visibilities) power spectrum to the thermal noise power spectrum. In the top row, the compression noise power spectrum (left), the thermal noise (middle), and the thermal noise uncertainty (right) are shown. In the bottom row, we show the ratio between the compression noise and the thermal noise (middle) and the thermal noise uncertainty (right). These spectra are obtained from real uncalibrated data (hence the arbitrary power spectrum units).

2. both LOFAR nights after calibration, and
3. two identical simulations with different noise realizations (simulation A).

As summarized in Table 3, all the datasets used for simulation A spanned an equal 12-hour duration, with the same LST range. The simulated LOFAR HBA datasets comprising identical extragalactic foregrounds but a different thermal noise realization was added to each. We processed all the datasets similarly, again applying the gains obtained from each uncompressed dataset to its compressed version, in order to get rid of the solver noise. We then obtained the difference of the DI-calibrated visibilities for each pair of reference and compressed simulated data, which gave us a pair of DI-calibrated compression residuals. We checked for any correlation between each pair in this trio of residuals pairs.

We used the coherence metric, C, given by the real part of the normalized cross-power spectrum (Mertens et al. 2020; Gehlot et al. 2024):

$$C_{a,b}\left(k_\perp, k_\parallel\right) = \frac{\Re\left(\tilde{T}_a^*(\boldsymbol{k})\tilde{T}_b(\boldsymbol{k})\right)}{\sqrt{|\tilde{T}_a(\boldsymbol{k})|^2|\tilde{T}_b(\boldsymbol{k})|^2}}. \quad (1)$$

Taking the real values of the cross-spectrum as opposed to the absolute values used in Mertens et al. (2020), provides information on how positive and negative coherence values are distributed around zero. Therefore, this metric ranges from −1 to 1 with both extremes denoting maximum coherence while zero denotes total incoherence. Although the coherence can have an imaginary component, due to spatial shifts between modes

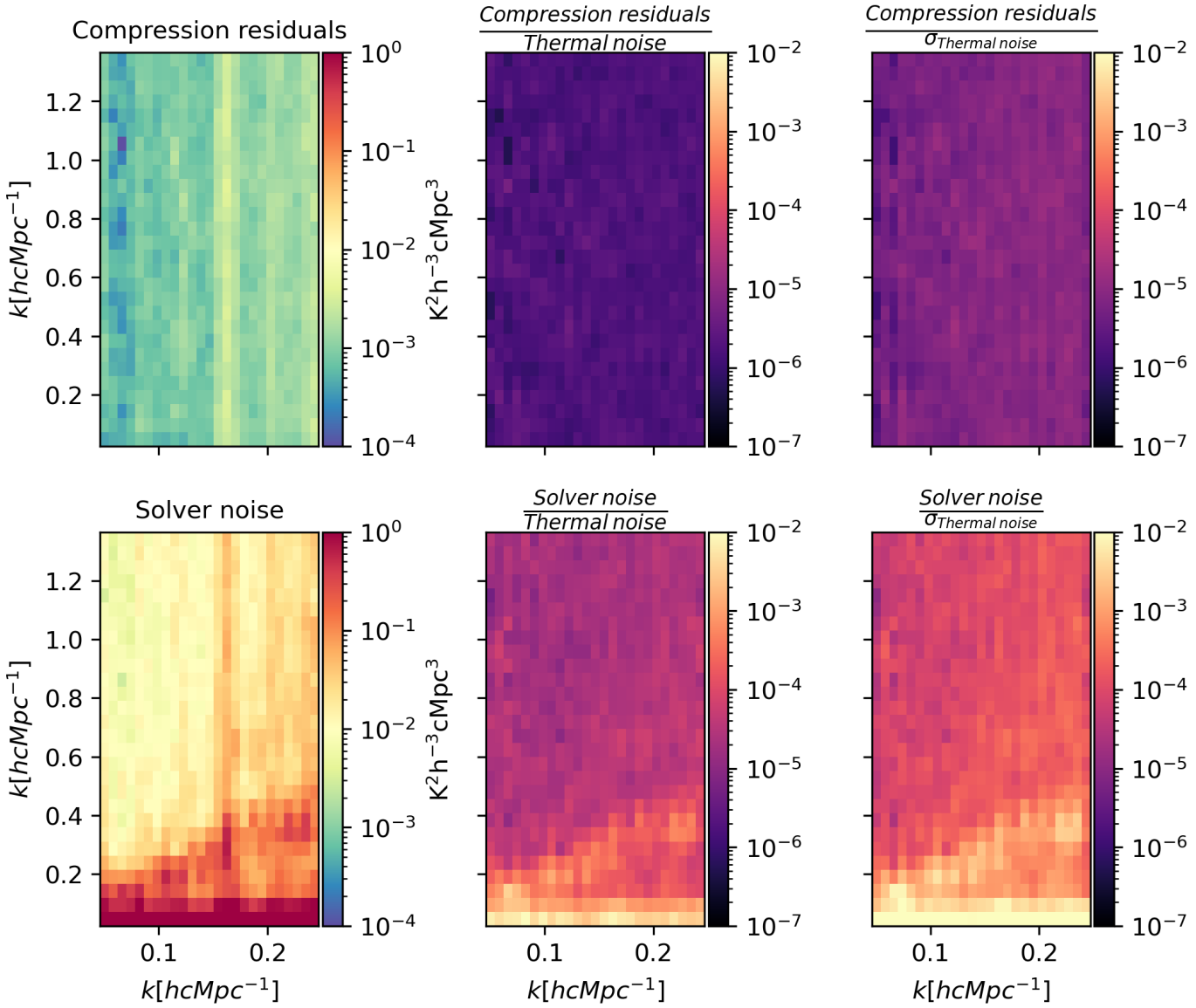

Fig. 3. Comparisons between the cylindrically averaged power spectra of both compression and solver noise to the thermal noise and the thermal noise uncertainty after calibration. The top row shows the cylindrically averaged power spectrum of the compression noise (top left), its ratio with the thermal noise power spectrum (top middle), and the thermal noise uncertainty (top right). Similarly, the bottom row shows the solver noise power spectrum (bottom left) and its ratio with both the thermal noise power spectrum (bottom middle) and the thermal noise uncertainty (bottom right). The thermal noise and compression residuals here are obtained from real calibrated data.

before and after compression, such an effect would require compression to be highly spatially correlated, which is not the case. We can therefore ignore the imaginary component. The coherences from each pair are shown in the top panels of Fig. 4. The compression noise is seen to be highly incoherent across all k-modes with a noise-like behavior around a mean of zero, devoid of any spurious coherence structures. The coherence has an rms of ∼0.14 around a mean of zero, which remains consistent across all three cases. In the bottom panel, we present the average coherence with respect to both the $k_\perp$ and $k_\parallel$ modes. The average coherence again has a noise-like behavior around zero that is within the rms and consistent for both cases.

To ascertain that this coherence level is consistent with random uncorrelated residuals data, we computed the spherically averaged power spectrum of the compression residuals from each night separately and then compared it with the power spectrum obtained from a coherent averaging of the compression residuals from both nights. Figure 5 shows the three spectra as well as a ratio of each individual night's residual power spectrum to the combined power spectrum. Both ratios show a consistent factor of ∼2 as expected from combining two equal-size datasets composed of highly incoherent noise. This verifies that compression noise will progressively average down, like normal system noise, with deeper data integrations. Any coherence in the compression noise would not rise above the error on the thermal noise for at least several hundred thousand hours of integration.

For completeness, we also examined the correlation of the solver noise as obtained from the calibration of two different observation nights. This noise is also found to have minimal coherence with a mean of −0.013 and an RMS of 0.18 as shown in Fig. A.1. This too, while not being the main subject of this paper, is a novel result. It shows that the random algorithmic solver noise, examined in this work using SAGECAL, does not introduce significant bias in the power spectrum.

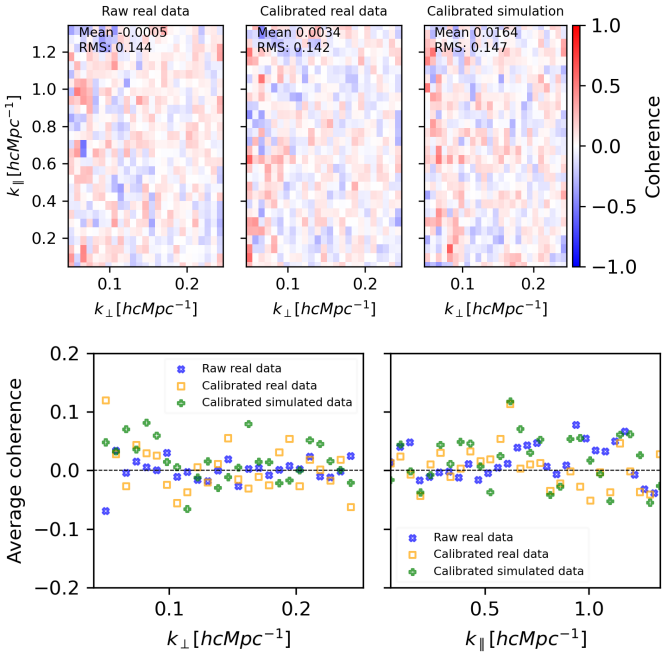

Fig. 4. Compression residuals coherence for raw real, calibrated real, and calibrated simulated data. For each panel in the top row, the coherence is computed using two compressed minus uncompressed data residuals from two separate datasets. To eliminate solver noise, the calibration gains solution from each uncompressed dataset is applied to its compressed version before obtaining the residuals. The average coherence for each top row panel is shown in the bottom row as a function of both $k_\perp$ (left) and $k_\parallel$ (right) modes.

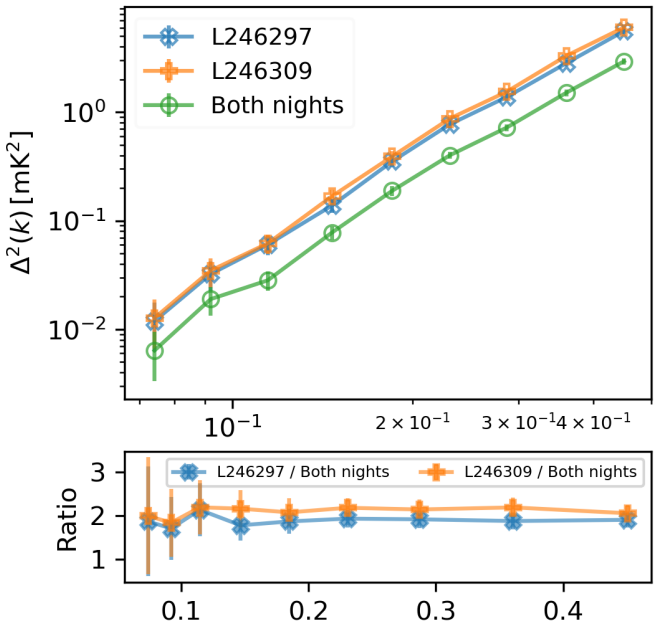

Fig. 5. Spherically averaged power spectra obtained from the compression residuals of two nights and their combined power spectrum. The ratio of each night to the combined power spectrum is shown in the bottom plot.

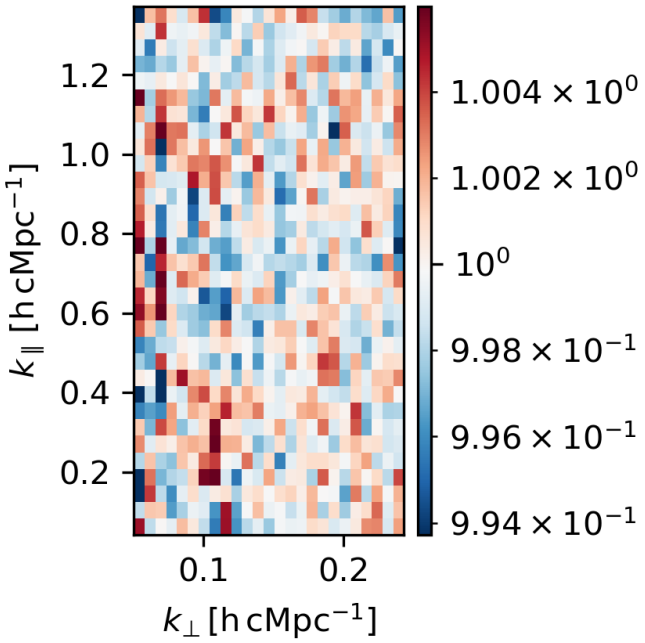

Fig. 6. Ratio of compression residuals after applying the gains solutions obtained from either uncompressed or compressed data.

5.3. Calibration on compressed data

Calibration gains solutions obtained from compressed data should not show significant discrepancies from those obtained by calibrating uncompressed data. In the tests discussed above, we used gains from the reference data for the purpose of eliminating solver noise. We show that the results remained unchanged when we used the calibration gains solutions from compressed data instead. In Fig. 6 we show the ratio of the compression residuals obtained by applying either the compressed or uncompressed data solutions. This ratio shows random fluctuations close to unity implying no significant difference.

Additionally, we examined the signal de-correlation resulting from calibrating a compressed dataset, as opposed to its original uncompressed version. A unity coherence between the calibrated reference and compressed data is expected if the calibration solutions obtained for the two datasets were identical. We wanted to measure the level of signal de-correlation resulting from the difference in the gains. The metric was obtained from the pair of reference and compressed calibrated datasets after applying the same gains solutions set to both (the gains obtained from either the reference data or the uncompressed data[13]). We used the "decoherence" metric[14] (D; Eq. (2)): an exactly zero decoherence value signals total coherence between the two calibrated datasets (i.e., no compression effect), while 1 signals maximum decoherence:

$$D^{\mathrm{abs}}_{a,b}(k_\perp, k_\parallel) = 1 - \frac{|\tilde{T}^*_a(\boldsymbol{k})\tilde{T}_b(\boldsymbol{k})|}{\sqrt{|\tilde{T}_a(\boldsymbol{k})|^2|\tilde{T}_b(\boldsymbol{k})|^2}}. \quad (2)$$

[13] We can also apply the gains solutions obtained from calibrating the compressed dataset to the reference dataset and check the decoherence as well. The two decoherence outputs should be equivalent in the case where calibration gains solution outputs are the same regardless of whether the input data were compressed or not. The two were found to be similar to the order of $\sim 10^{-7}$. Therefore, the reference solutions and DYSCO compressed solutions are almost identical. By applying either of them to both the reference and compressed data and then computing the coherence of the calibrated data pair, we obtain an equivalent estimate of the compression noise coherence.

[14] The actual de-correlation is due to a reduction in the S/N between the two calibrated datasets and not loss of the actual signal.

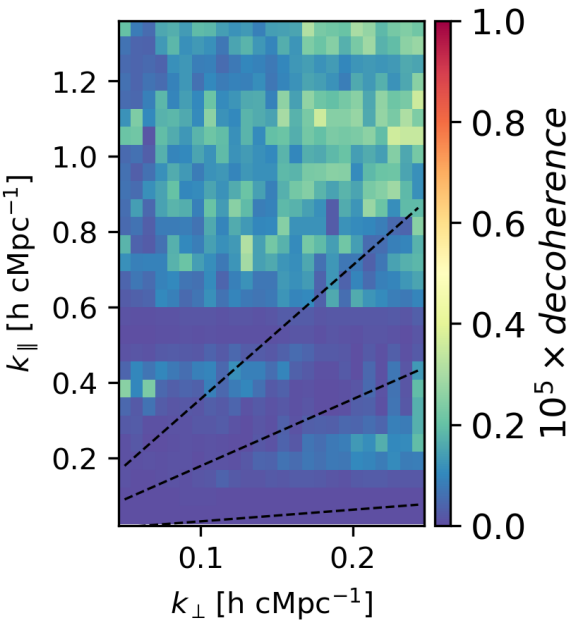

Fig. 7. Amount of signal de-correlation (due to a reduced S/N) resulting from calibration of a compressed dataset instead of its original version (without compression). The same calibration gains solutions set is applied to both the reference and compressed data before computing the coherence. The dashed lines show the 5, 30 and 90° delay lines.

Figure 7 shows the decoherence resulting from this calibration of compressed data as opposed to uncompressed data due to the reduced S/N of the data. Here we see a decoherence at the 10^{-6} level. Since we have already shown that the compression noise is far below the thermal noise level, any apparent decoherence seen in the noise-dominated regions (e.g., at higher $k_\parallel$ modes) can be attributed to the noisy nature of the data itself in those regions as opposed to being an effect of data compression. Nevertheless, this effect is overall insignificant and therefore the calibration of compressed data yet again shows no nefarious effects.

Similarly, we find that the solver noise has a decoherence on the order of $\sim 10^{-4}$, two orders of magnitude higher than the decoherence caused by data compression. This implies that the compression noise de-correlates the signal at a subdominant level in comparison to the solver noise. Compression noise is therefore not an issue of concern.

5.4. Optimal compression settings for LOFAR HBA data

The error introduced by lossy compression is expected to vary depending on the compression settings chosen for DYSCO. A higher bit-rate choice for the compressed data will result in a reduced compression error. Additionally, the choice of the data normalization and the quantization distribution will influence the final compression error. The previous sections used the default DYSCO settings (see Table 1) as implemented in the DP3 pipeline. While these default parameters might be ideal for science cases such as radio surveys and transient searches, the requirements on any resulting errors in high-redshift 21 cm signal detection experiments are much more stringent. We thus intended to show how varying these compression parameters reflects on the 21 cm signal power spectrum, specifically the bit size and normalization. We did not test parameters that are known to be worse than the default parameters, such as the row-normalization method and bit sizes of less than 10.

We also did not test the effect of different quantization distributions, firstly because any conclusions drawn from comparing different distributions would not be robust enough to apply to all datasets: and secondly, the performance of a given distribution is also coupled to other settings such as the bit size (Offringa 2016). As summarized in Table 3 (simulation B), this test was carried out using 6-hour simulations and incorporating only the core stations of the LOFAR HBAs. For reference, we also included a simulated 21 cm signal model at $z = 8.3$ from Mesinger et al. (2016).

Figure 8 shows the data compression factor and the spherically averaged power spectrum of the compression residuals obtained by varying the data compression normalization method and bit-rates. Across the three tested bit-rates (10, 12, 16), both RF and AF normalization result in similar error levels that are all far below the thermal noise level. However, the RF normalization errors are about ×1.4 higher than the AF errors. As described in Sect. 2, this error difference between AF and RF is attributable to the difference in the normalization dimensions between the two methods. Nevertheless, all compression residuals are well below the theoretical EoR level at $z = 8.3$ for both normalization methods, across all the bit sizes, even in the 6 hours of data used in this test. In 1200 hours of data, for example, these errors would be 200 times lower even.

As expected the compression noise is higher with lower bit sizes. Moving from 10 to 12 bits results in a factor of ∼4 and ∼60 (slightly lower than the expected factor of 64) from 12 to 16, respectively. The compression factor achieved varies from 3.4, 3.0, to 2.5 for 10, 12, and 16 bit-rates, respectively. The exact compression factor is dependent on the dimensions of the data, particularly on the number of channels: the more channels stored in a measurement set, the lower the relative impact of the metadata, and the higher the compression factor. Since the tests were done on a subset of the data, the compression factor on a full LOFAR observation night dataset is ∼4 when done with 10 bits and AF normalization. If done with RF normalization instead, this factor will be slightly lower; however, the RF normalization method can be used to compress both visibility cross-correlations and auto-correlations while the AF normalization is limited to the cross-correlations only.

Based on these results, for LOFAR EoR data, we recommend 10-bits DYSCO compression with RF normalization. These parameters are suitable for LOFAR and might need retuning for other instruments. However, due to the similarities between LOFAR and SKA-Low, it suffices to conclude that the optimal parameters obtained here for LOFAR are likely also applicable for SKA-Low with minimal adjustments.

6. Conclusions

Lossy data compression methods can be a means to reduce the large expected costs associated with the storage and transfer of radio interferometric data, in particular those from LOFAR and SKA. The compression should not, however, compromise the fidelity of the data, especially for high-precision studies such as 21 cm signal observations of the EoR and cosmic dawn. In this work, we have investigated the effect of lossy compression of visibilities on the 21 cm observations. Specifically, we have examined the level of compression errors and their behavior as they manifest in the 21 cm power spectrum using the DYSCO compression code (Offringa 2016).

We find that compression introduces additional noise to the power spectrum. However, this noise is around five orders of magnitude lower than the error on the thermal noise power

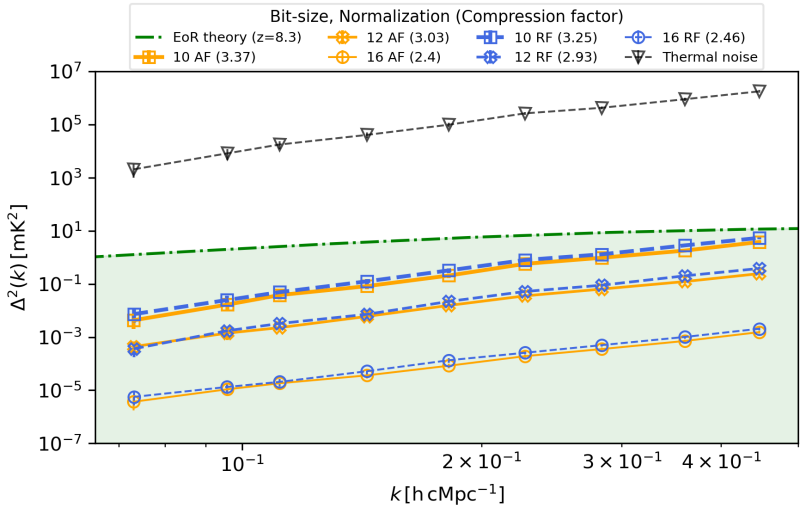

Fig. 8. Comparison of the compression error levels to the spherically averaged power spectrum and the compression factors, for different bit sizes and normalization methods. Orange and blue lines represent AF and RF normalization methods, respectively, with the three thickness levels and markers representing the bit sizes from 10 (thickest line with square markers) to 16 (thinnest line with circle markers). The green line shows a representative spherically averaged power spectrum from a simulated 21 cm model at $z = 8.3$. Compression errors from all settings are below the 21 cm signal, even from a relatively small 6-hour dataset. Since the compression noise is shown to be incoherent, a deeper data integration of, for example, 1000 hours will result in a compression noise level that is 1000 times lower for all settings. The compression factors obtained for each setting are shown in brackets in the legend.

spectrum of a single night. This noise also has been shown to not be correlated to the sky, as seen from the minimal coherence between the residuals of different datasets.

Since the compression noise is much lower than the thermal noise and is highly incoherent, its effect on calibration is found to be insignificant, as expected. The calibration solutions obtained from compressed data are highly similar to those obtained from the reference data. While this test was done using only the DI calibration step, it suffices to conclude that a similar insignificant effect applies in all calibration stages of the data. Thus, we did not delve into compression effects in direction-dependent calibration.

We examined the optimal DYSCO compression parameters for LOFAR EoR data. We find that the bit size used to store the compressed data is crucial in determining whether an error in the power spectrum remains well below the expected 21 cm signal after long integrations. Larger bit sizes result in lower compression factors on one hand, but also less compression on the other. Therefore, a balance between these two factors should be considered when choosing a suitable compression bit size. Moreover, since the compression performance depends on the instrument sensitivity, these parameters might need to be retuned for different instruments. While this paper used LOFAR HBA data, the findings reported here will likely apply to SKA-Low data since both instruments will have around the same noise per visibility.

Acknowledgements. JKC, LKV, BKG, SAB, SG, CH 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"). LVEK, ARO and EC acknowledge support from the Centre for Data Science and Systems Complexity (DSSC), Faculty of Science and Engineering at the University of Groningen. F.G.M. acknowledges support from a PSL Fellowship. LOFAR, the Low-Frequency Array designed and constructed by ASTRON, has facilities in several countries, owned by various parties (each with their own funding sources), and collectively operated by the International LOFAR Telescope (ILT) foundation under a joint scientific policy. This research made use of publicly available software developed for LOFAR telescope.

References

Asgekar, A., Oonk, J. B. R., Yatawatta, S., et al. 2013, A&A, 551, L11
Barry, N., Hazelton, B., Sullivan, I., Morales, M. F., & Pober, J. C. 2016, MNRAS, 461, 3135
Brackenhoff, S., Mevius, M., Koopmans, L., et al. 2024, MNRAS, 533, 632
Braun, R., Bonaldi, A., Bourke, T., Keane, E., & Wagg, J. 2019, Arxiv e-prints [arxiv:1912.12699]
Chege, J. K., Jordan, C. H., Lynch, C., et al. 2022, PASA, 39, e047
Cheng, C., Parsons, A. R., Kolopanis, M., et al. 2018, ApJ, 868, 26
Deboer, D. R., Parsons, A. R., Aguirre, J. E., Alexander, P., & Ali, Z. S. 2017, PASP, 129, 045001
Gehlot, B. K., Mertens, F. G., Koopmans, L. V. E., et al. 2019, MNRAS, 488, 4271
Gehlot, B. K., Koopmans, L. V. E., Brackenhoff, S. A., et al. 2024, A&A, 681, A71
HERA Collaboration 2022, ApJ, 925, 221
HERA Collaboration 2023, ApJ, 945, 124
Jelić, V., Zaroubi, S., Labropoulos, P., et al. 2008, MNRAS, 389, 1319
Jelić, V., Zaroubi, S., Labropoulos, P., et al. 2010, MNRAS, 409, 1647
Kolopanis, M., Jacobs, D. C., Cheng, C., et al. 2019, ApJ, 883, 133
Li, W., Pober, J. C., Barry, N., et al. 2019, ApJ, 887, 141
Lindstrom, P. 2017, Error Distributions of Lossy Floating-Point Compressors, 2574
Masui, K. W., Amiri, M., Connor, L., et al. 2015, Astron. Comput., 12, 181
McQuinn, M., Zahn, O., Zaldarriaga, M., Hernquist, L., & Furlanetto, S. R. 2006, ApJ, 653, 815
Mellema, G., Koopmans, L. V. E., & Abdalla, F. A. 2013, Exp. Astron., 36, 235
Mertens, F. G., Mevius, M., Koopmans, L. V. E., et al. 2020, MNRAS, 493, 1662
Mesinger, A., Greig, B., & Sobacchi, E. 2016, MNRAS, 459, 2342
Mevius, M., Mertens, F., Koopmans, L. V. E., et al. 2022, MNRAS, 509, 3693
Morales, M. F., & Hewitt, J. 2004, ApJ, 615, 7
Mouri Sardarabadi, A., & Koopmans, L. V. E. 2019, MNRAS, 483, 5480
Munshi, S., Mertens, F. G., Koopmans, L. V. E., et al. 2024, A&A, 681, A62
Offringa, A. R. 2016, A&A, 595, A99
Offringa, A. R., van de Gronde, J. J., & Roerdink, J. B. T. M. 2012, A&A, 539, A95
Offringa, A. R., McKinley, B., Hurley-Walker, F. H., et al. 2014, MNRAS, 444, 606
Paciga, G., Albert, J. G., Bandura, K., et al. 2013, MNRAS, 433, 639
Patil, A. H., Yatawatta, S., Koopmans, L. V. E., et al. 2017, ApJ, 838, 65
Sabater, J., Sánchez-Expósito, S., Garrido, J., et al. 2015, Highlights of Spanish Astrophysics VIII, 840
Shimwell, T. W., Röttgering, H. J. A., Best, P. N., et al. 2017, A&A, 598, A104
Thompson, A. R., Moran, J. M., & Swenson, G. W. 2017, Analysis of the Interferometer Response (Cham: Springer)
Tingay, S. J., Goeke, R., Bowman, J. D., Emrich, D., & Ord, S. M. 2013, PASA, 30, e007
Trott, C. M., Jordan, C. H., Midgley, S., et al. 2020, MNRAS, 493, 4711
van Diepen, G., Dijkema, T. J., & Offringa, A. 2018, Astrophysics Source Code Library [record ascl:1804.003]
van Haarlem, M. P., Wise, M. W., Gunst, A. W., Heald, G., & McKean, J. P. 2013, A&A, 556, A2
Wells, D. C., Greisen, E. W., & Harten, R. H. 1981, A&AS, 44, 363
White, R. L., Greenfield, P., Pence, W., Tody, D., & Seaman, R. 2012, ArXiv e-prints [arXiv:1201.1336]
Yatawatta, S. 2015, MNRAS, 449, 4506
Yatawatta, S. 2016, ArXiv e-prints [arXiv:1605.09219]
Yatawatta, S., de Bruyn, A. G., Brentjens, M. A., et al. 2013, A&A, 550, A136
Zarka, P., Girard, J. N., Tagger, M., & Denis, L. 2012, in SF2A-2012: Proceedings of the Annual meeting of the French Society of Astronomy andAstrophysics, eds. S. Boissier, P. de Laverny, N. Nardetto, et al., 687

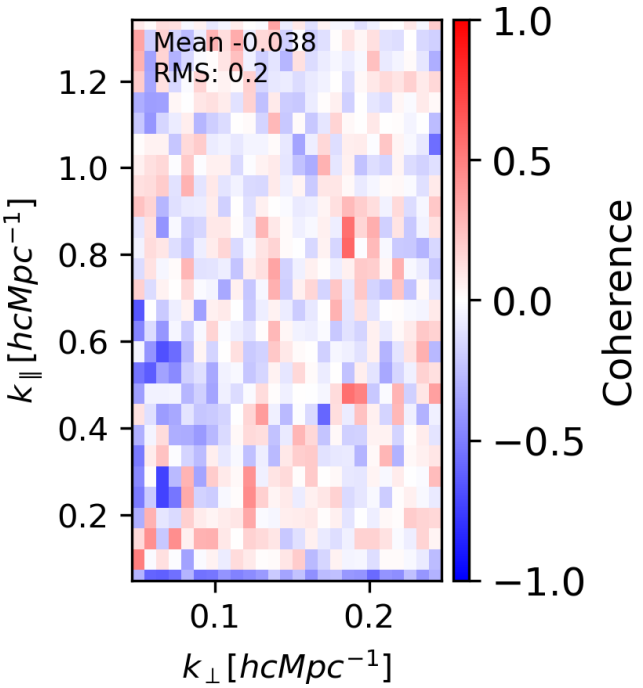

Fig. A.1. Ratio of compression residuals after applying the gains solutions obtained from either uncompressed or compressed data.

Appendix A: Coherence of the calibration solver noise

In Fig. A.1 we show the coherence between the solver noise obtained from the calibration runs of two different nights. Similar to the compression noise coherence shown in Fig. 4, solver noise shows minimal coherence. Therefore, conclusions drawn from the compression noise coherence hold for solver noise as well.