Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **514,** 4655–4668 (2022)
Advance Access publication 2022 June 15

https://doi.org/10.1093/mnras/stac1588

Characterizing beam errors for radio interferometric observations of reionization

Ainulnabilah Nasirudin [1]★ David Prelogovic [1] Steven G. Murray,[2] Andrei Mesinger [1] and Gianni Bernardi[3,4,5]

[1]*Scuola Normale Superiore, Piazza dei Cavalieri 7, I-56126 Pisa, Italy*
[2]*School of Earth and Space Exploration, Arizona State University, Tempe, AZ 85281, USA*
[3]*INAF-Istituto di Radioastronomia, via Gobetti 101, I-40129 Bologna, Italy*
[4]*Department of Physics and Electronics, Rhodes University, PO Box 94, Grahamstown 6140, South Africa*
[5]*South African Radio Observatory (SARAO), 2 Fir Street, Observatory, Cape Town 7925, South Africa*

Accepted 2022 June 1. Received 2022 May 31; in original form 2022 January 20

ABSTRACT

A limiting systematic effect in 21-cm interferometric experiments is the chromaticity due to the coupling between the sky and the instrument. This coupling is sourced by the instrument primary beam; therefore it is important to know the beam to extremely high precision. Here, we demonstrate how *known* beam uncertainties can be characterized using data bases of beam models. In this introductory work, we focus on beam errors arising from physically offset and/or broken antennas within a station. We use the public code OSKAR to generate an 'ideal' SKA beam formed from 256 antennas regularly spaced in a 35-m circle, as well as a large data base of 'perturbed' beams sampling distributions of broken/offset antennas. We decompose the beam errors ('ideal' minus 'perturbed') using principal component analysis (PCA) and Kernel PCA (KPCA). Using 20 components, we find that PCA/KPCA can reduce the residual of the beam in our data sets by 60–90 per cent compared with the assumption of an ideal beam. Using a simulated observation of the cosmic signal plus foregrounds, we find that assuming the ideal beam can result in 1 per cent error in the epoch of reionization (EoR) window and 10 per cent in the wedge of the 2D power spectrum. When PCA/KPCA is used to characterize the beam uncertainties, the error in the power spectrum shrinks to below 0.01 per cent in the EoR window and $\leq$ 1 per cent in the wedge. Our framework can be used to characterize and then marginalize over uncertainties in the beam for robust next-generation 21-cm parameter estimation.

Key words: dark ages, reionization, first stars – interferometric – statistical.

1 INTRODUCTION

Measuring the epoch of reionization (EoR) 21-cm signal is one of the key science goals of current and upcoming low-frequency interferometers such as the Murchison Widefield Array (MWA; Tingay et al. 2013; Wayth et al. 2018), the Hydrogen Epoch of Reionization Experiment (HERA; DeBoer et al. 2017), the Low Frequency Array (LOFAR; van Haarlem et al. 2013), the Giant Metrewave Radio Telescope (GMRT; Swarup et al. 1991), and the upcoming Square Kilometre Array (SKA; Dewdney et al. 2009; Mellema et al. 2013). To date, several upper limits of the EoR power spectrum have been published (see e.g. Barry et al. 2019; Li et al. 2019; Trott et al. 2020; Mertens et al. 2020; Abdurashidova et al. 2022). For an actual detection of the EoR signal, an unprecedented level of precision is required. This is because foregrounds and instrumental systematics dominate over the reionization signal by several orders of magnitude.

Possibly the most complicated and pronounced instrumental systematic comes from the uncertainty in the model of the primary beam (Sutinjo et al. 2015; Jacobs et al. 2017; Line et al. 2018). For the wide field-of-view (FoV) instruments common to 21-cm cosmology, this beam must be well-characterized over essentially the entire sky, including in side-lobes close to the horizon. The beam itself is a highly multidimensional quantity, changing over direction, frequency, pointing, and polarization. Furthermore, accurate measurements in the far-field regime are incredibly difficult; individual elements are far too large to be characterized with anechoic chambers. While novel techniques such as mapping with pulsars (Newburgh et al. 2014) and drones (Jacobs et al. 2017) show some promise, it is unclear if they will achieve the necessary angular resolution and coverage required.

In principle, simulations of the beam via electromagnetic modelling (EM) can be highly accurate, although to generate the most accurate models requires significant computational investment. Even so, there are multiple potential sources of error in such models. Considering the primary beam of an SKA 'station', which consists of 256 individual dipoles phased together to generate a single primary beam, there are two potential avenues towards modeling: (i) EM-modelling of an individual dipole, followed by synthesis of the primary beam under some assumptions (e.g. all dipoles are identical, we know to high precision the location and rotation of each dipole, and we can neglect the effect of surrounding dipoles on the response of the individual dipole), or (ii) direct EM-modelling

★ E-mail: ainulnabilah.nasirudin@sns.it

of an entire station. Most work to date has taken the first approach, for reasons of computational feasibility and the fact that different stations can easily be modelled without re-performing the expensive EM modelling of the dipole, simply by rearranging the individual dipole models. In particular, this is the approach taken by OSKAR, a core instrumental simulation code for the SKA. In this approach, there are two kinds of errors one expects to be present; first, there may be errors in the EM model of the individual dipole, arising from e.g. small *in situ* physical defects of the dipole, uncertainties in physical parameters of the simulation (e.g. soil permittivity and conductivity) and discretization of the model itself. Secondly, there are errors associated with how the dipole models are synthesized, including deviations of different dipoles from identity, small positional and/or rotational errors, the potential for a small fraction of the dipoles in the station to be offline for a particular measurement, and perhaps most importantly, 'mutual coupling' effects that modify an individual dipole's response based on the presence of surrounding electrically-conducting dipoles (Sutinjo et al. 2020; Bolli et al. 2021; Fagnoni et al. 2021). Fractional errors in the modelled beam with respect to the true *in situ* beam, from all of these potential sources, tend to be small close to zenith (for a good model), but can grow quite large towards the horizon, where mutual coupling, ground reflections and other imperfections are most active. While this potentially large fractional error close to the horizon is attenuated in its effect on model observations in proportion to the amplitude of the beam in that region, the presence of side-lobes close to the horizon can up-weight these imperfections strongly enough that they become a concern for the extreme precision requirements of 21-cm cosmology.

In this paper, we focus on characterizing the primary beam of SKA stations. As already mentioned, current state-of-the-art characterizations involve EM modelling of an individual dipole, followed by synthesis into a station of 256 dipoles given their location within the array. Here, we adopt this same general approach, which has many benefits in terms of computational efficiency, but supplement it with a characterization of the inevitable error that must be present in the resulting primary beam model. Our 'error characterization' framework, based on principal component analysis (PCA) decomposition, is highly flexible, and is not limited to a particular *kind* of modelling error. Instead, we present a framework whose purpose is to be able to characterize any *given* uncertainty in the physical modeling, and compress the resulting errors from their native extremely high-dimensional space into a few parameters that capture the spread of potential models. As a worked example of this framework, in this paper we focus on two particular sources of potential error in the primary beam model of an SKA station: (i) dipole location offsets, and (ii) the presence of offline dipoles.

We note that, in practice, these sources of error are not likely to be the dominant source of model error for the SKA. The SKA is expected to have a system that reports offline dipoles with each station alongside each observed timestamp. This would in principle allow one to accurately account for these missing dipoles in one's primary beam model, without uncertainty. Nevertheless, in practice, doing so individually for each station at each observation time might be prohibitively expensive, or this information may not be available to a particular analyst, and a framework which is able to simply capture the resulting errors may be more feasible. Furthermore, location errors of the dipoles are expected to be limited to $\sim$1 mm using current technology, as is the case for LOFAR. For offsets this small, the induced errors will likely be subdominant to the more complicated effects of mutual coupling and EM modeling errors. Nevertheless, since these two sources of error are conceptually and computationally simple to model, we use them in this paper to showcase our framework, which itself is principle able to cover more complicated sources of error, so long as they can be physically modeled to produce a training set.

As mentioned, our framework is based on a PCA compression of the beam model residuals, which we shall describe in more detail in Section 2. In general, we'd like to have an unbiased estimate of the true beam, over its high-dimensional space, with an accurate assessment of our uncertainty of that estimate. This model should propagate through instrumental calibration and through to power spectrum estimation. There are two possible approaches for modelling the beam: physical and empirical, both of which have drawbacks. A physical approach would involve identifying a set of physical parameters that are uncertain in detail, and fitting for those parameters (or, rather, marginalizing over them). Problems with this approach are (i) it is extraordinarily computationally demanding, as it requires performing an EM simulation for each dipole for every posterior sample, (ii) the number of unknown physical parameters is potentially very large, and (iii) it is likely that not all relevant physical parameters are identified, which means the posterior sample does not include the true beam. An empirical approach avoids problems (i) and (iii) by adopting a flexible description of the *output* beam model, which is intended to be flexible enough to include the true beam in its posterior space. However, in general such an approach has the drawback that the native dimensionality of the output model is extraordinarily large (including spatial angles, frequencies, polarizations, and pointings), and furthermore the flexibility of the model can lead to negative consequences for the posterior spread of the desired parameters due to high uncertainty on the priors. Our framework adopts this empirical approach, but circumvents the problem of high dimensionality via PCA compression, resulting in a limited set of parameters able to cover the majority of the posterior sample space, and also an approximate prior based on the physical training set.

The paper is organized as follows. We first describe the beam simulation and the basis set that we use in Section 2. In Section 3, we present the results of our analysis that motivates our choice for the perturbation model and the impacts on the reconstructed beam errors. In Section 4, we study the impact of using the ideal beam and the reconstructed beam errors on the power spectrum of a mock sky consisting of the 21-cm signal and foregrounds in an interferometric framework. Finally, we discuss our findings and conclude the paper in Section 5.

2 METHOD FOR CHARACTERIZING BEAM UNCERTAINTIES

In this section, we explain the method used to generate the beam and the basis set used to characterize the beam errors. We generate electromagnetic simulations of an idealized beam, as well as thousands of realizations of non-ideal beams. This training set is used in two PCA-based approaches to identify the primary modes of variation amongst the perturbed beams, and to determine the number of such modes that adequately reconstruct any perturbed beam. The linear coefficients of these modes thus represent a compressed parameter space which can be marginalized in parameter estimation. We first create a data base of perturbed beam realizations and sample the errors in Section 2.1. We then introduce our PCA and kPCA methods for characterizing the beam errors in Section 2.2.

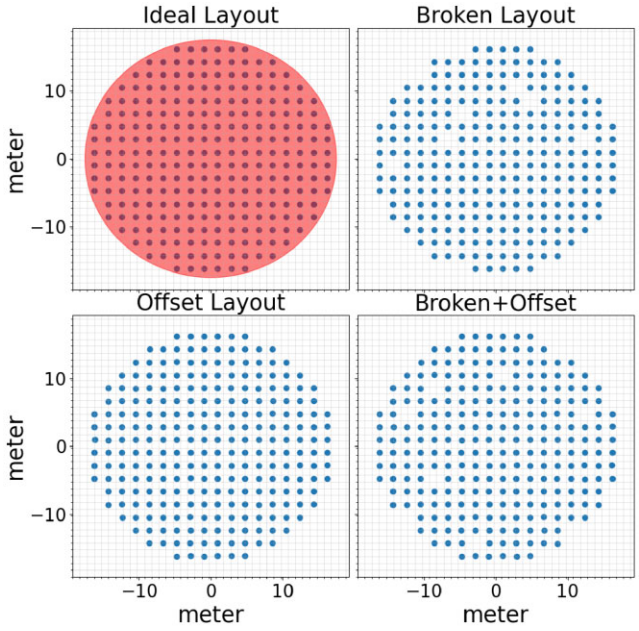

Figure 1. The ideal antenna layout (top left panel), along with the same layout but with broken (top right panel), offset (bottom left panel), and both broken + offset antennas (bottom right panel).

2.1 Data base of beam models

We use OSKAR (Dulwich et al. 2009) version 2.7.6 to simulate the primary beam response of a station based on the antenna layout. The station has 256 antennas positioned within a circle of diameter 35 m, motivated by the design of upcoming SKA stations. We define the ideal station as one in which all the antennas are regularly spaced within the circle, with a spacing of ∼1.8 m between each antenna. Because the actual SKA stations are expected to have pseudo-randomly distributed antennas, we have included a brief investigation of having this configuration in Appendix A. For computation purposes, we only generate the beam at 150, 170, and 190 MHz for a zenith-pointing observation. The exact input parameters we use in OSKAR are presented in Table C1 in Appendix C.

Our perturbation model is based on two scenarios:

(i) broken (i.e. offline) antennas, in which some antennas are excluded from the beam synthesis process. The number of broken antennas in each realization is sampled from a uniform distribution between $N_{\rm broken} = 1$–12 (corresponding to $\lesssim 5$ per cent of the total 256 antennas), with their positions assigned randomly.

(ii) offset antennas, in which *all* antennas are displaced from the ideal position following a zero-mean normal distribution with $\sigma_x = \sigma_y = 3$ cm.

We create three separate data sets, each comprised of 10 000 perturbed beam realizations: (i) broken + offset, (ii) broken only, and (iii) offset only. Fig. 1 shows the ideal antenna layout, along with three examples of the same layout but with broken-only, offset-only, and both broken + offset antennas, respectively. The 10 000 realizations in each data set are divided into 7000 training and 3000 test realizations.

Throughout, we refer to the simulated power beam of the 'ideal' station as $B_{\rm ideal}(\nu, \theta, \phi)$, where ν is the frequency, θ is the zenith angle, and ϕ the angle around the zenith pole, and the the simulated power beam of a particular realization of a perturbed station as $B_{\rm perturbed}(\nu, \theta, \phi)$. Instead of modelling the perturbed beams themselves, we model their *residuals*, i.e:

$$\Delta B(\nu, \theta, \phi) = B_{\rm perturbed}(\nu, \theta, \phi) - B_{\rm ideal}(\nu, \theta, \phi). \quad (1)$$

A particular model of the residuals will be represented as $\widehat{\Delta B}$, where we use subscripts to denote the basis used (see next subsection). Note that the residual of residuals is equivalent to the residual in the modelled beam, i.e. $\Delta B - \widehat{\Delta B} \equiv B_{\rm perturbed} - (B_{\rm ideal} + \widehat{\Delta B})$. For visualization purposes, in Fig. 2 we show the ideal beam, $B_{\rm ideal}$ (left-hand panel), an example realization of $B_{\rm perturbed}$ (middle panel) from the broken + offset data, and the corresponding ΔB (right-hand panel) at $\nu = 150$ MHz within angular difference, $(\theta_x, \theta_y) = \pm 10^\circ$ from the pointing direction. By comparing Figs 2 to A2, we have shown that having either a regularly spaced or a pseudo-random station does not affect the main results of this paper because the levels of error in the beam are comparable in both cases, both in the mainlobe and sidelobe. In any case, as we highlighted previously, our purpose here is to illustrate the approach of empirically characterizing systematics to be used in forward models, and not to have ultrarealistic examples of any specific systematic.

To quantify the beam errors we calculate the fractional residual, $X = |\Delta B|/B_{\rm ideal}$ for each model in the data sets, and then compute the mean and standard deviation of X over all realizations at each frequency, ν. We present the sample mean (top panels) and sample standard deviation (bottom panels) at 150 MHz for the three sets in Fig. 3. For the broken + offset (left panels) data sets, the mainlobe areas are mostly unaffected by the different antenna configuration as shown by $\sigma(X)$ that is consistent with 0, although the sidelobes can differ by as much as 100 per cent near the nulls. In contrast, the impact of the antenna offsets (middle panels) is very large, with errors exceeding 100 per cent around the nulls and in the side-lobes. Interestingly, the residuals in the broken-only (right-hand panels) data set are intermediate between the ones in the offset-only and broken + offset. This could imply that the impact of 'breaking' and 'offsetting' can partially compensate for one another. Indeed, one might consider that the positional offsets partially act to 'fill the gaps' left by the broken antennas, resulting in a net lower error when both effects are at play. Nevertheless, to quantitatively understand such effects, a proper investigation with electromagnetic simulations needs to be conducted.

2.2 Perturbation basis set

We utilize PCA and Kernel PCA to model the beam residuals, ΔB, arising from the different antenna configurations. We describe each in turn. These are applied to the 7000 samples in each of our three beam error training sets, and then tested on the remaining 3000 samples.

2.2.1 *Principal component analysis (PCA)*

The goal of a traditional PCA is to reduce the dimensionality of a data set by performing a linear change of basis and determining the extent to which each eigenvector captures the variation within the data set. The most significant basis vectors are termed the 'principal components'. Typically, a data set $Y = (y_1, y_2, ..., y_{n-1}, y_n)$ is first standardized with respect to its mean and standard deviation,

$$Z = \frac{Y - \mu(Y)}{\sigma(Y)}, \quad (2)$$

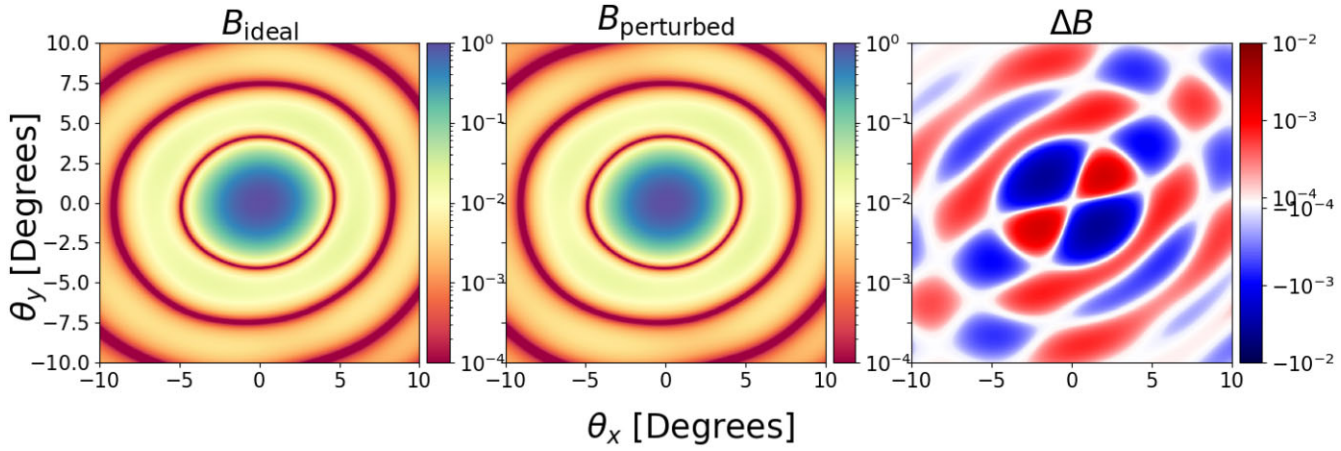

Figure 2. The ideal beam, $B_{\rm ideal}(\nu)$ (left-hand panel), an example realization of $B_{\rm perturbed}(\nu)$ (middle panel) broken + offset data, and the consequent $\Delta B(\nu)$ (right-hand panel) at $\nu = 150$ MHz within angular difference $(\theta_x, \theta_y) = \pm 10^\circ$ from the pointing direction. The fractional error is of the same general magnitude as the fractional error using a pseudo-random station layout in Fig. A2.

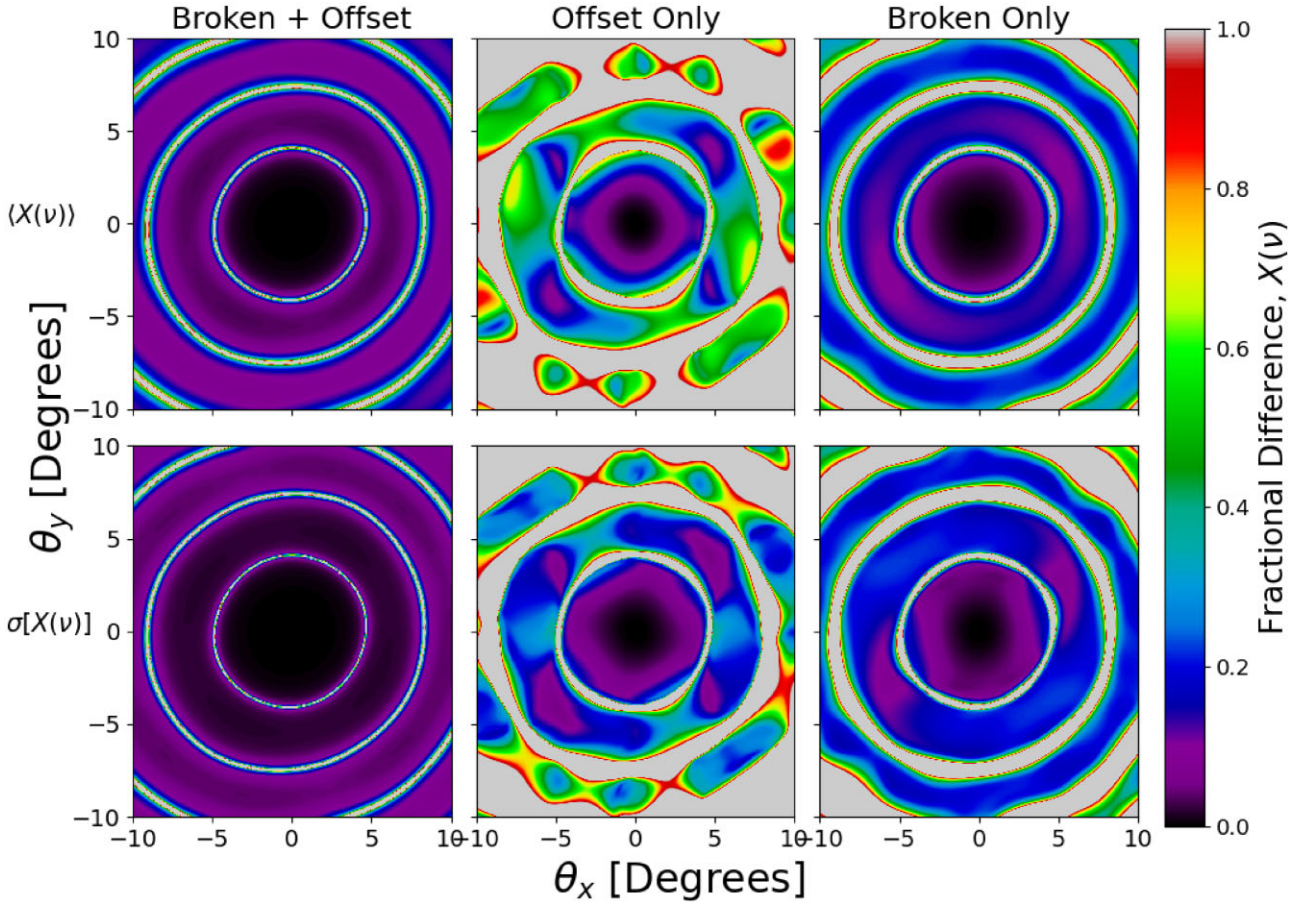

Figure 3. Mean of the fractional residual, <*X*> (top panels), and their standard deviations, $\sigma(X)$ (bottom panels), over the 10 000 realizations for the broken + offset (left), offset-only (middle), and broken-only (right) data sets at 150 MHz, where the maximum value for the colour-scale has been limited to 100 per cent.

and its covariance matrix, $C(Z)$, is computed. The eigenvectors, $\boldsymbol{v}$, and eigenvalues, a, are then calculated following

$$C\boldsymbol{v} = a\boldsymbol{v}. \tag{3}$$

Finally, a are arranged in descending order, yielding principal components in order of significance in which a feature vector can be formed with some number of features or components, N, and the reconstructed beam, $\widehat{\Delta B}_{N;\rm PCA} = \sum_i^N a_i \boldsymbol{v}_i$. The principal components have the limitation that they are *linear* transformations

of the input data set; non-linear transformations that require fewer terms to adequately describe the data may exist.

2.2.2 *Kernel principal component analysis (KPCA)*

KPCA extends the PCA method via non-linear transformations of the data set. The data is first mapped to an arbitrary higher dimension, often referred to as the *feature space*, and then linear PCA is performed on this feature space. The feature space, however, does not need to be explicitly computed. Instead, it is sufficient to compute the kernel,

$$K(y_i, y_j) = \psi(y_i)^T \psi(y_j), \tag{4}$$

where $\psi(y_i)$ is the non-linear transformation from real to feature space (Schölkopf, Smola & Müller 1997). One downside of KPCA is that a unique, one-to-one inverse relation that transforms $\psi(y_i)$ back to y_i does non exist. However, other methods such as ridge regression (Hoerl & Kennard 1970a,b) can be used for this purpose, which is what is being used here. For a simple introduction to KPCA, we refer the reader to Appendix D.

To model our data set comprised of $\Delta B(\nu)$, we developed SPAX,[1] an efficient PCA and KPCA code that is GPU and CPU-optimized. The following kernels are available within SPAX for KPCA:

$$K(y_i, y_j) = y_i^T y_j \text{ [linear]} \tag{5}$$

$$= \tanh\left(\kappa y_i^T y_j\right) \text{ [tanh]} \tag{6}$$

$$= \left(\kappa y_i^T y_j + r\right)^d \text{ [polynomial]} \tag{7}$$

$$= \exp\left(-\kappa[|y_i + y_j|]^2\right) \text{ [radial basis]} \tag{8}$$

$$= y_i^T y_j / (y_i^2 \cdot y_j^2) \text{ [cosine]}. \tag{9}$$

The kernels, K, regularization parameter, and hyperparameters κ, r, and d are flexible; different kernels (and/or hyperparameter values) can be used for the transform and inverse transform, respectively, in order to improve the fit to the training data set. We perform hyperparameter optimization using a simple, coarse grid search, selecting the parameter combination that minimizes the mean square error (MSE) between ΔB and the reconstructed residual using N features, $\widehat{\Delta B}_N$ for all three frequencies. We note that reconstructing the beam error/residual is obtained by finding the best-fitting set of eigenvalues using the given eigenvector basis. Hyperparameter calibration is performed separately for all three data sets, and we allow the inverse kernel to be different from the transform kernel. Our course grid assumes integer values, with $N = 10$, $d = 2$, and $r = 1$ where applicable to reduce computation. The best set of parameters that gives the lowest MSE for each data set are presented in Table 1, in which the subscript 'inv' refers to the parameter for the inverse transform. We highlight that using more sophisticated hyperparameter Bayesian optimization should yield even better results; we defer this to future work when we apply our method to mock data.

The relative performance of KPCA versus (linear) PCA can depend strongly on the processes which generate the data itself. In simplest terms, if the data itself is a linear combination of effects, then PCA is optimal. However, if the data is inherently a non-linear transformation from a more compact basis, then KPCA may be better in compressing the information content. In practice, if the data are inherently most compact in a non-linear basis, we may expect KPCA to outperform (i.e. have a smaller MSE) linear PCA when reconstructing with a 'small' number of components. However, linear PCA is guaranteed to achieve perfect reconstruction if using $N \to N_{\rm dim}$ (i.e. an MSE of close to zero), whereas KPCA is not,[2] and therefore there may be a crossover at some N.

[1] https://github.com/dprelogo/SPax

Table 1. The best set of parameters and kernels that gives the lowest MSE for each data set based on a simple, coarse Monte Carlo 'grid-search'. The subscript 'inv' refers to the parameter and kernel for the inverse transform.

	Broken + Offset	Broken only	Offset only
κ	54	2	79
$\kappa_{\rm inv}$	1	1	23
K	tanh	tanh	tanh
$K_{\rm inv}$	poly	rbf	poly

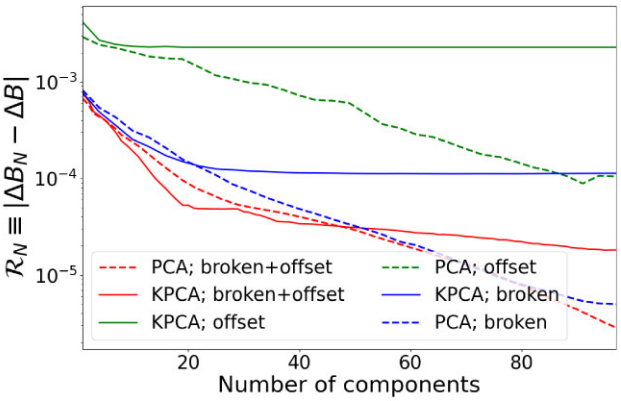

Figure 4. The mean of $\mathcal{R}_N$ for the broken + offset (red), broken (blue), and offset (green) data sets with varying N (number of components) using PCA (dash lines) and KPCA (solid lines).

3 RESULTS: HOW WELL IS THE BEAM ERROR RECOVERED?

To decide how many components to include in the final reconstruction of the beam across all frequencies, we vary N and evaluate the reconstruction error $\mathcal{R}_N \equiv |\widehat{\Delta B}_N - \Delta B|$. We present the mean, $\langle \mathcal{R}_N \rangle$ (coloured lines) across the 3000 realizations of broken + offset (red lines), offset (green lines), and broken (blue lines) test data at only 150 MHz for simplicity for PCA (dash lines) and KPCA (solid lines) in Fig. 4. Although PCA yields lower values of the mean reconstruction error $\langle \mathcal{R}_N \rangle$ at $N \geq 50$ for the broken + offset data set (red dash line), its decrease with the number of components is slow. In contrast, with KPCA, $\langle \mathcal{R}_N \rangle$ decreases rapidly by $N = 20$ for both the broken + offset and broken data sets and then plateaus somewhat as N increases. This relative performance is in qualitative agreement with our expectations from the previous section. On the other hand, the offset only data set plateaus quicker with KPCA at $N \sim 10$ with only a small improvement in $\langle \mathcal{R}_N \rangle$.

Since we want to model the beam error with the least number of components possible, below we limit ourselves to the first $N = 20$

[2] KPCA essentially performs standard linear PCA in a non-linearly transformed space. While it is guaranteed to minimize the MSE in this space, it is not guaranteed to minimize MSE in the space of the data. It is difficult to judge whether this is better or worse without understanding the natural basis of the data.

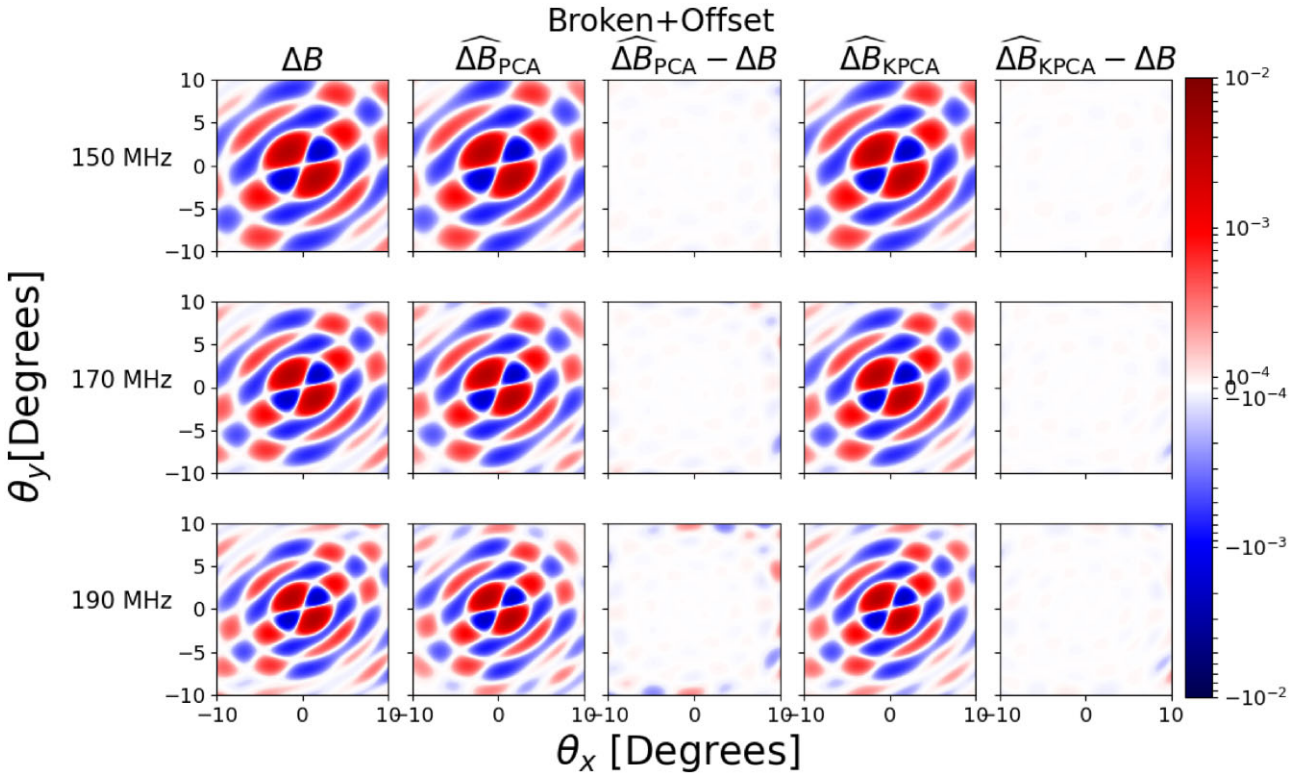

Figure 5. From left to right, the actual beam error (ΔB), the PCA-reconstructed beam error ($\widehat{\Delta B}_{\rm PCA}$), the difference between the PCA-reconstructed and actual error ($\widehat{\Delta B}_{\rm PCA} - \Delta B$), the KPCA-reconstructed beam error ($\widehat{\Delta B}_{\rm KPCA}$), and the difference between the KPCA-reconstructed and actual error ($\widehat{\Delta B}_{\rm KPCA} - \Delta B$) from one sample realization of the broken + offset data for all three frequencies. In this example, the standard deviation of the fractional difference between the reconstructed and actual error over all θ, ϕ and ν for PCA and KPCA is 344.1 and 114.68, respectively.

components for both PCA and KPCA for all data sets. We note that even though the reconstruction with KPCA for the offset only data gives minimal improvement with $N = 20$ compared to $N = 10$, we have still chosen to use the former for simplicity. Hereafter, $\widehat{\Delta B}_{N=20}$ is referred to simply as $\widehat{\Delta B}$ and the PCA/KPCA reconstruction is done on the full data set with all three frequencies.

We illustrate the recovery of the beam error in Figs 5–7, for a randomly chosen sample from each of our test sets. From left to right, we show the actual beam error (ΔB), the PCA-reconstructed beam error ($\widehat{\Delta B}_{\rm PCA}$), the difference between the PCA-reconstructed and actual error ($\widehat{\Delta B}_{\rm PCA} - \Delta B$), the KPCA-reconstructed beam error ($\widehat{\Delta B}_{\rm KPCA}$), and the difference between the KPCA-reconstructed and actual error ($\widehat{\Delta B}_{\rm KPCA} - \Delta B$). Rows correspond to our three frequency bins.

With 20 features, both PCA and KPCA can effectively reconstruct ΔB of our sample from the broken + offset example shown in Fig. 5, including the frequency evolution of the features, the size of the mainlobe, and the magnitude of the perturbation. All these result in average difference of $\leq|10^{-4}|$. At 190 MHz, however, the model seems to be slightly less sensitive to structures in the sidelobe region, as is apparent in the middle and right-most panel on the bottom row, especially with PCA. For reference, the distribution of the KPCA eigenvalues (which, as expected, follow a Gaussian distribution), are presented in Fig. C2 in the Appendix.

For the broken-only sample in Fig. 6, both PCA and KPCA are able to capture the overall details of ΔB, including the evolution of the features and the size of the mainlobe, as shown in the second column from the left and fourth column from left in Fig. 6. However, the error in the reconstruction can be up to an order of magnitude higher with PCA and there are more small-scale features compared to the reconstruction with KPCA.

For the offset-only example shown in Fig. 7, both PCA and KPCA perform worse than seen in the previous two examples. The reconstructions (second and fourth columns from left) somewhat resemble the large-scale structures of ΔB, but instead of having two large 'half-ring' structures in the sidelobe, both PCA and KPCA model them as multiple radial features. Moreover, the reconstructed error can be up to two orders of magnitude higher than in the previous examples, as is evident in the third and right-most columns.

To summarize, we present the mean and standard deviation of $|\Delta B|$ and $\mathcal{R}_{20}$ over the 3000 test realizations and all three frequencies for the three test data sets with PCA and KPCA in Table 2. Using PCA/KPCA, there is up to a factor of 10 reduction in beam error compared with the assumption of an ideal beam. The mean and standard deviation of $\mathcal{R}_{20}$ vary from 10 per cent to almost 50 per cent for the broken + offset and offset data sets, respectively.

4 IMPACT OF BEAM RECONSTRUCTION ON THE POWER SPECTRUM

Although we have established that PCA and KPCA do a good job in capturing the beam error from our data sets, the reconstruction is not perfect. Hence in this section, we investigate the impact of these residual errors on the recovery of the power spectrum, using a realistic sky composed of the EoR signal and point-source foregrounds.

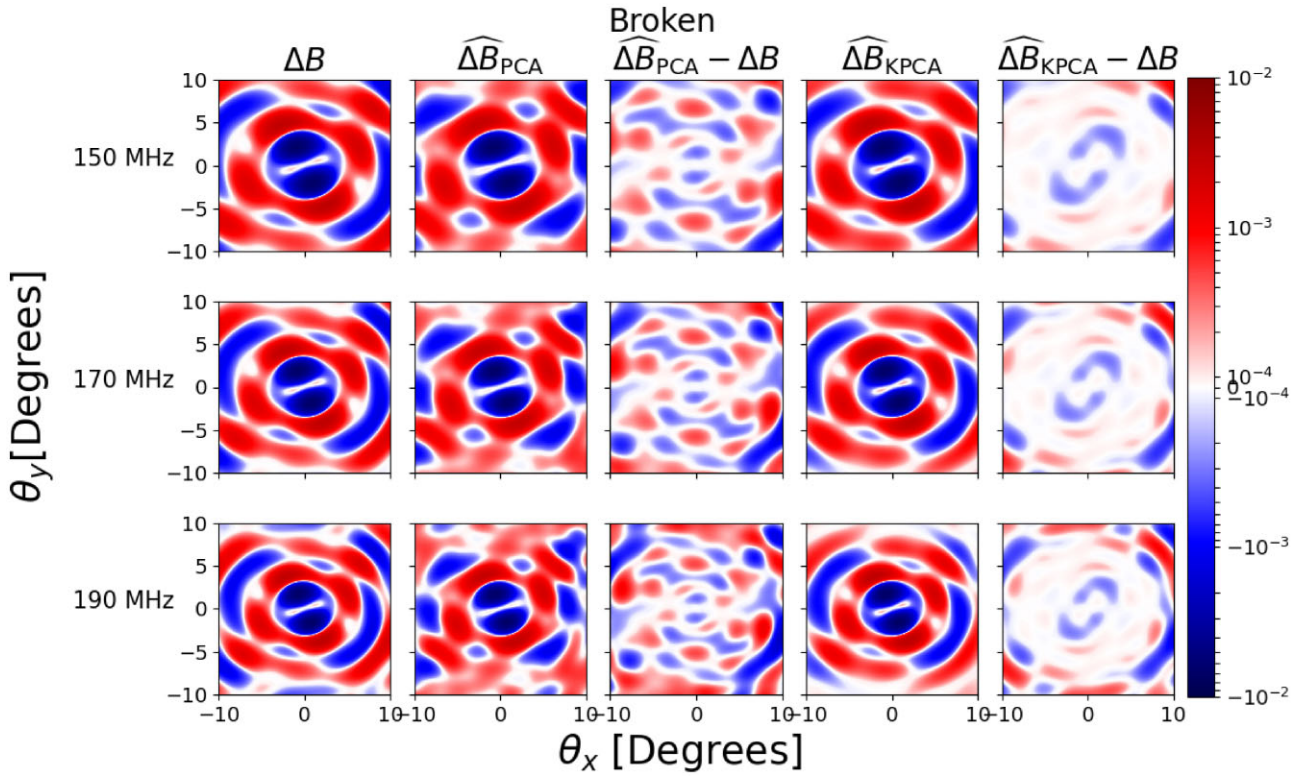

Figure 6. From left to right, ΔB, $\widehat{\Delta B}_{\rm PCA}$, $\widehat{\Delta B}_{\rm PCA} - \Delta B$, $\widehat{\Delta B}_{\rm KPCA}$, and $\widehat{\Delta B}_{\rm KPCA} - \Delta B$ from one sample realization of the broken-only set for all three frequencies. In this example, the standard deviation of the fractional difference between the reconstructed and actual error over all θ, ϕ and ν for PCA and KPCA is 347.70 and 59.49, respectively.

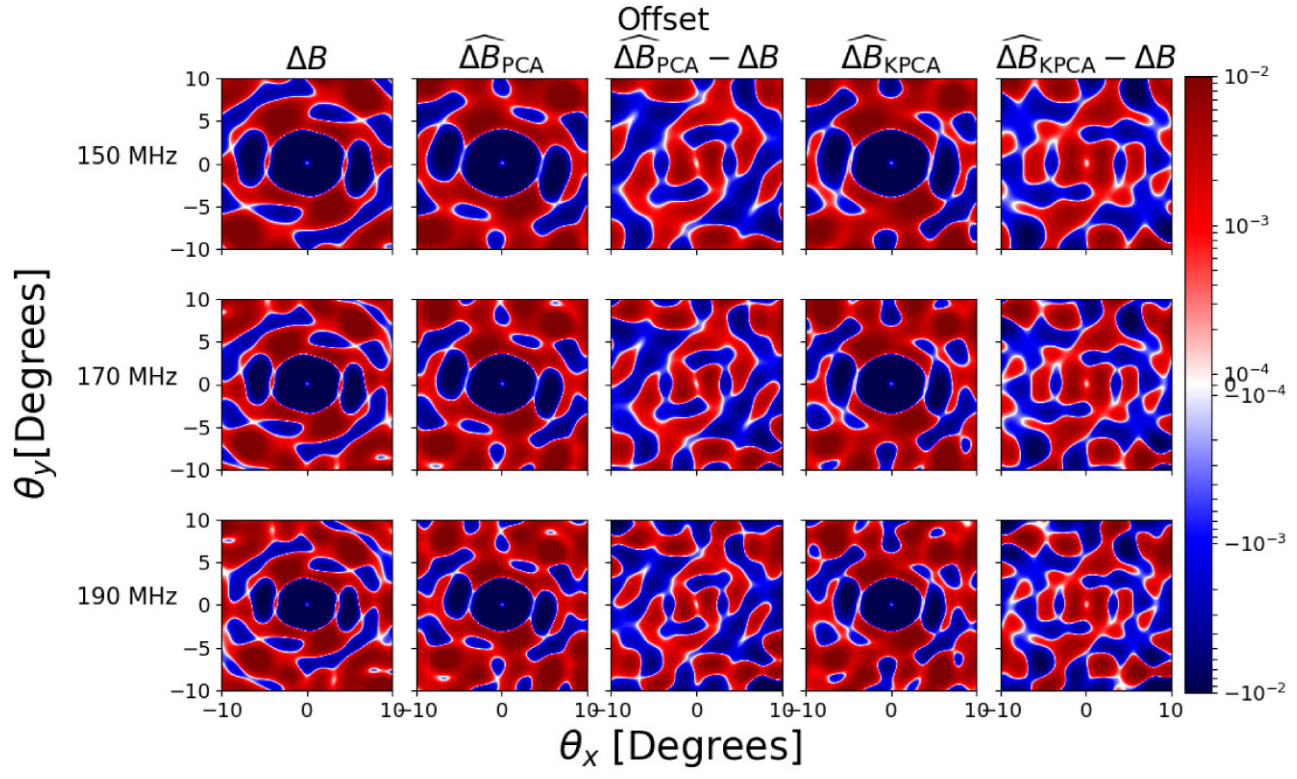

Figure 7. From left to right, ΔB, $\widehat{\Delta B}_{\rm PCA}$, $\widehat{\Delta B}_{\rm PCA} - \Delta B$, $\widehat{\Delta B}_{\rm KPCA}$, and $\widehat{\Delta B}_{\rm KPCA} - \Delta B$ from one sample realization of the offset-only set for all three frequencies. In this example, the standard deviation of the fractional difference between the reconstructed and actual error over all θ, ϕ and ν for PCA and KPCA is 247.78 and 280.07, respectively.

Table 2. The MSE and standard deviation of MSE of ΔB and $|\widehat{\Delta B} - \Delta B|$ with PCA and KPCA for the three test data sets across all three frequencies and pixels.

	Broken + Offset	Broken only	Offset only
$\langle\lvert\Delta B\rvert\rangle$	3.7×10^{-4}	8.4×10^{-4}	5.8×10^{-3}
$\sigma(\lvert\Delta B\rvert)$	4.5×10^{-4}	2.3×10^{-3}	8.5×10^{-3}
$\langle\mathcal{R}_{\rm PCA}\rangle$	4.4×10^{-5}	1.9×10^{-4}	2.0×10^{-3}
$\sigma(\mathcal{R}_{\rm PCA})$	5.7×10^{-5}	1.9×10^{-4}	1.8×10^{-3}
$\langle\mathcal{R}_{\rm KPCA}\rangle$	3.9×10^{-5}	1.7×10^{-4}	2.2×10^{-3}
$\sigma(\mathcal{R}_{\rm KPCA})$	4.9×10^{-5}	3.9×10^{-4}	2.5×10^{-3}

4.1 Foreground model

Following Nasirudin et al. (2020), we simulate extragalactic point-source foregrounds with a flux-density source count distribution with the power-law relation

$$\frac{\mathrm{d}N}{\mathrm{d}S}(S,\nu) = \alpha S_\nu^{-\beta}\left(\frac{\nu}{\nu_0}\right)^{-\gamma\beta} (\mathrm{Jy}^{-1}\mathrm{sr}^{-1}), \tag{10}$$

where $\mathrm{d}N/\mathrm{d}S$ is the source spatial density per unit flux density, S_ν is the flux at a specific frequency ν, β is the slope of the source-count function, and γ is the mean spectral-index of point sources. Based on an observational result from Intema et al. (2011), we set $\alpha = 4100\ \mathrm{Jy}^{-1}\ \mathrm{sr}^{-1}$, $\beta = 1.59$, and $\gamma = 0.8$ at $\nu_0 = 150$ MHz. Having drawn source fluxes from the above distribution, we situate them uniformly randomly across the sky and bin them into a regular grid that matches the beam output from OSKAR. We sample the point sources between $S_{\rm min} = 50\ \mu$Jy and $S_{\rm max} = 50$ mJy[3] at 150 MHz. The observation consists of 128 linearly spaced frequency channels between 150 to 165 MHz.

4.2 Reionization model

The differential brightness temperature, δT_B, during the EoR can be approximated as

$$\delta T_b(z) \approx 27 x_{\rm HI}(1+\delta_{\rm nl})\left(\frac{H(z)}{\mathrm{d}v/\mathrm{d}r + H(z)}\right)\left(1-\frac{T_\gamma}{T_s}\right) \times \left(\frac{1+z}{10}\frac{0.15}{\Omega_m h^2}\right)^{\frac{1}{2}}\left(\frac{\Omega_b h^2}{0.023}\right) (\mathrm{mK}), \tag{11}$$

where $x_{\rm HI}$ is the neutral fraction, $\delta_{\rm nl}$ is the evolved Eulerian over-density, H is the evolving Hubble constant, $\mathrm{d}v/\mathrm{d}r$ is the gradient of the line-of-sight velocity component, T_γ is the temperature of the CMB, $T_{\rm s}$ is the spin temperature of neutral hydrogen (H I), z is the redshift, $\Omega_{\rm m}$ is the dimensionless matter density parameter, $\Omega_{\rm b}$ is the dimensionless baryonic density parameter and h is the normalized Hubble constant (Furlanetto, Oh & Briggs 2006).

We use the efficient seminumerical EoR modelling tool, 21CMFASTv3 (Mesinger, Furlanetto & Cen 2011; Park et al. 2019; Murray et al. 2020), to generate the light-cone of $\delta T_{\rm b}$ during the EoR. For a detailed description of the code and astrophysical model, we refer readers to Mesinger et al. (2011), Park et al. (2019), and Murray et al. (2020). In our research, we use the default parameter values of 21CMFASTV3, which are shown to reproduce current high-z observations (Park et al. 2019) and simulate the light-cone of a 512 Mpc h^{-1} box. This choice of parameters corresponds to a neutral fraction of 0.5 at $z \sim 6.5$. Because the light-cone covers only $\sim$3.3° at 150 MHz, we tile it across the 20° mock sky and coarsen the grids to match with the resolution of the beam.

[3]Choudhuri, Bull & Garsden (2021) found that the brightest sources to be particularly important in the presence of beam variations/non-redundancy, but because we assume that brighter sources have been perfectly peeled from the observation, hence we only model those below the peeling threshold.

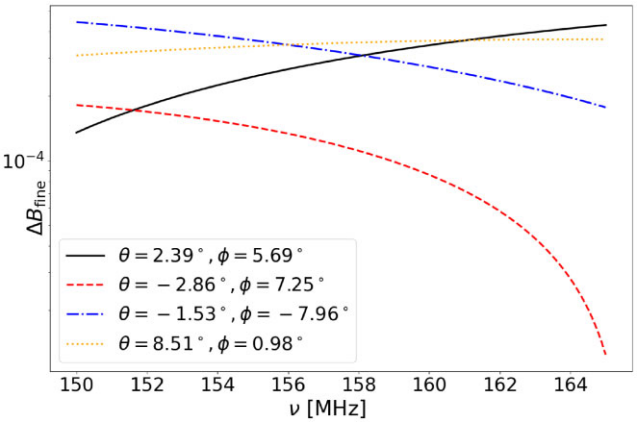

Figure 8. The spectral behaviour of the residual between the ideal and perturbed beam generated by OSKAR with 128 frequency channels, $\Delta B_{\rm fine}$ for some values of (θ, ϕ). Because $\Delta B_{\rm fine}$ is spectrally smooth, this justifies the interpolation of ΔB, $\widehat{\Delta B}_{\rm KPCA}$, and $\widehat{\Delta B}_{\rm KPCA}$.

4.3 Interferometric framework

At wavelength, λ, the baseline displacement, $\boldsymbol{u} = (u, v)$, is defined as $\boldsymbol{u} = \boldsymbol{x}/\lambda$, where $\boldsymbol{x}$ is the physical displacement between the stations in meters hence it is frequency dependent. The sky coordinate, $\boldsymbol{l}$, is defined as $\boldsymbol{l} = (l, m) = (\sin\theta\cos\phi, \sin\theta\sin\phi)$.

Using the flat-sky approximation, the visibility at frequency ν, $V(\boldsymbol{u}_j, \nu)$, for each baseline j is defined as

$$V(\boldsymbol{u}_j, \nu) = \int S(\boldsymbol{l}, \nu) B(\boldsymbol{l}, \nu) \exp(-2\pi i \boldsymbol{u}_j \cdot \boldsymbol{l}) \mathrm{d}\boldsymbol{l} \quad (\mathrm{Jy}), \tag{12}$$

where $S(\boldsymbol{l}, \nu)$ and $B(\boldsymbol{l}, \nu)$ are the flux density of each point-source and the beam attenuation at $\boldsymbol{l}$ and ν. The observed interferometric visibility is identical to the Fourier transform of the product of signal and the beam model under the flat-sky approximation. Here, we assume all stations have the exact same layout hence the same beam and that the beam databases from Section 2 with three frequencies spanning 150, 170, and 190 MHz have been linearly interpolated to 128 channels between 150 and 165 MHz. The interpolation of the reconstructed beam using only the three frequencies is motivated by our finding that the spectral behaviour of the residual between the ideal and perturbed beam generated by OSKAR with 128 frequency channels, $\Delta B_{\rm fine}$ is smooth in frequency, as shown for some (θ, ϕ) in Fig. 8.

For computation purposes, we Fast Fourier Transform over the 2D image to a regular-spaced 2D grid $\boldsymbol{u}_k$, and then interpolate $V(\boldsymbol{u}_k, \nu)$ from the regular 2D grid to the baselines $\boldsymbol{u}_j$. We then apply a Blackman–Harris taper $H(\nu)$ over the frequency axis, and calculate the delay transform (i.e. Fourier transform of un-gridded visibilities along the frequency axis),

$$V(\boldsymbol{u}_j, \tau) = \int V(\boldsymbol{u}_j, \nu) H(\nu) \exp(-2\pi i \tau \cdot \nu) \mathrm{d}\nu \quad (\mathrm{Jy\ Hz}). \tag{13}$$

The delay power spectrum is then calculated by cylindrically averaging the power of the visibilities within radial bin $\boldsymbol{r} = \sqrt{u^2 + v^2}$, which is proportional to the angular mode, $k_\perp$. We approximate the delay power spectrum as the power spectrum, in which τ is

Table 3. The visibility simulation parameters used in this research.

Parameter	Value
$N_{\rm sources}$	288 812
$N_{\rm antennas;\ perturbed}$	255
$\nu_{\rm range}$	150–165 MHz
$N_{\rm channel}$	128
$\Delta\nu_{\rm channel}$	78.74 kHz
$N_{\rm timestamp}$	1
Pointing	Zenith

proportional to the line-of-sight mode, $k_\parallel$. The conversion of the power spectrum, $\boldsymbol{r}$ and τ to cosmological units are outlined in Appendix B. For reference, we calculate the wedge region given by

$$k_\parallel \leq k_\perp \frac{\sin(\theta_{\rm FoV})E(z)\int_0^z {\rm d}z'/E(z')}{(1+z)} \quad (h\ {\rm Mpc}^{-1}), \tag{14}$$

where $\theta_{\rm FoV}$ is the angular radius of the FoV (Thyagarajan et al. 2013; Dillon et al. 2014).

Table 3 provides a summary of the visibility simulation parameters, mainly the number of foreground sources, $N_{\rm sources}$, the number of antennas in the perturbed beam synthesis, $N_{\rm antennas;\ perturbed}$, the frequency range, $\nu_{\rm range}$, the number of frequency channels, N_ν, the channel width, $\Delta\nu_{\rm channel}$, and the number of timestamp, $N_{\rm timestamp}$. We note that because we have not included any thermal noise in this work, the total integration time is not relevant.

4.4 Impact of different beams on the power spectrum

To understand the impact of beam errors on the recovery of the cylindrical power spectrum (hereafter PS), we convolve our sky described in Sections 4.1 and 4.2 with the sample realization of $B_{\rm perturbed}$ shown in Fig. 5. Following the steps outlined in Section 4.3, we then simulate the effects of interferometric observation and calculate the respective PS. From left to right, the panels in Fig. 9 show the PS of a sky consisting of the cosmic signal, foregrounds, and both cosmic signal plus foregrounds, respectively, that has been convolved with $B_{\rm ideal}$. In the foreground PS (middle panel), the well known key features shown in e.g. Dillon et al. (2014) and Barry et al. (2016) are clearly visible, mainly the foreground dominated region in dark blue where $k_\parallel \leq 0.11\,h$ Mpc^{-1}, the yellow-blue region of the wedge at $k_\perp \geq 0.1\,h$ Mpc^{-1}, and the mostly red EoR window.

Finally in Fig. 10 we present the fractional error in the PS of the beam-convolved total signal (EoR + FG) with respect to the ideal PS from the third panel of Fig. 9, i.e. $\Delta P/P_{\rm ideal} = [P(B{*}[{\rm FG} + {\rm EoR}]) - P(B_{\rm ideal}{*}[{\rm FG} + {\rm EoR}])]/P(B_{\rm ideal}{*}[{\rm FG} + {\rm EoR}])$ for $B = [B_{\rm perturbed}, B_{\rm perturbed} - \widehat{\Delta B}_{\rm PCA}, B_{\rm perturbed} - \widehat{\Delta B}_{\rm KPCA}]$ (left-hand to right-hand panels, respectively). From the left-hand panel we see that not accounting for beam errors mis-estimates the power spectrum throughout k-space, with errors peaking at $\sim$10 per cent in the wedge region. Instead, modeling $\widehat{\Delta B}$ using either PCA or KPCA reduces the error in the recovered power spectrum by over a factor of a hundred in the EoR window and a factor of ten in the wedge (compare middle and right to the left-hand panel).

Because the visibilities in the wedge are highly correlated, any deficit or surplus of beam attenuation with respect to $B_{\rm perturbed}$ is reflected in the entire wedge region. Indeed, as expected, beam errors affect foregrounds more than the cosmic signal, even in the PS space. However, there is some excess power close to the horizon line on long baselines, in which the spur-like structure above the horizon line seems to be caused by slight variations in the true beam for different antennas. This could be caused by our small FOV coupled with the type of perturbations on these scales. Because we are less likely to capture the perturbations perfectly at the edge of the FOV as shown in Figs 5, 6, and 7, most of the differences are clustered in the horizon thus potentially exacerbating the effect.

Finally, we stress that this exercise is highly idealized, providing only a minimum estimate of the PS recovery error. In practice, we will not be able to fit the PCA and KPCA coefficients to the perturbed

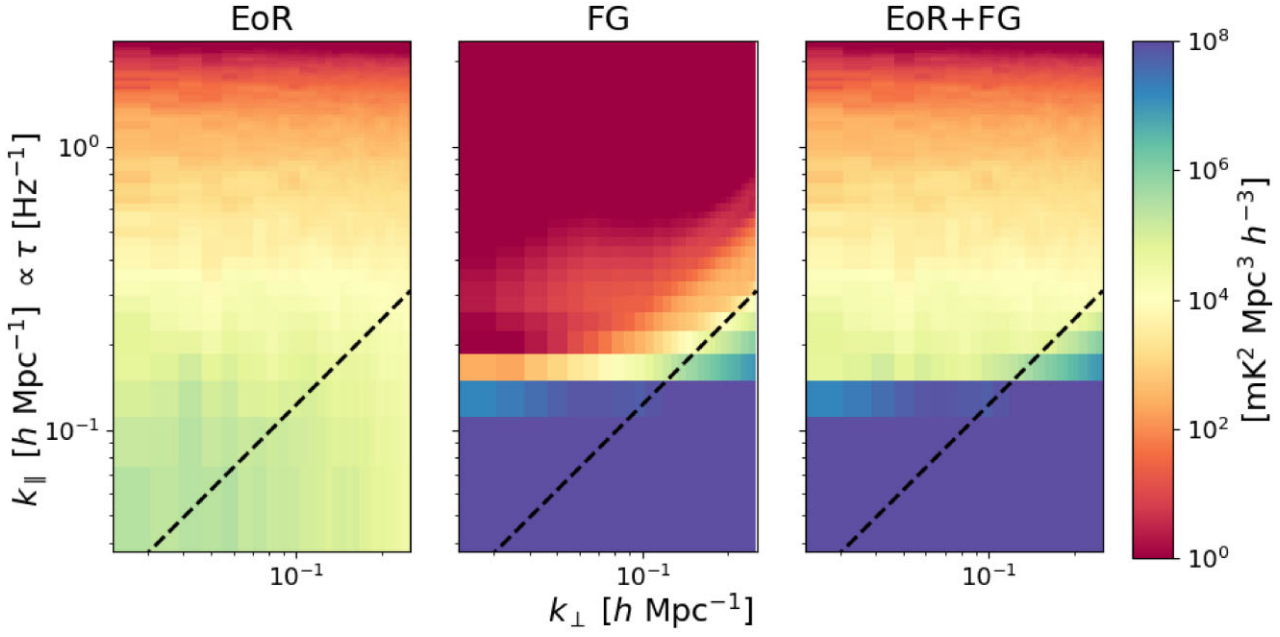

Figure 9. From left to right, the PS of cosmic signal, foregrounds, and both cosmic signal and foregrounds, respectively, for a mock sky convolved with $B_{\rm ideal}$. The black dash line shows the extent of the wedge which is calculated following equation (14).

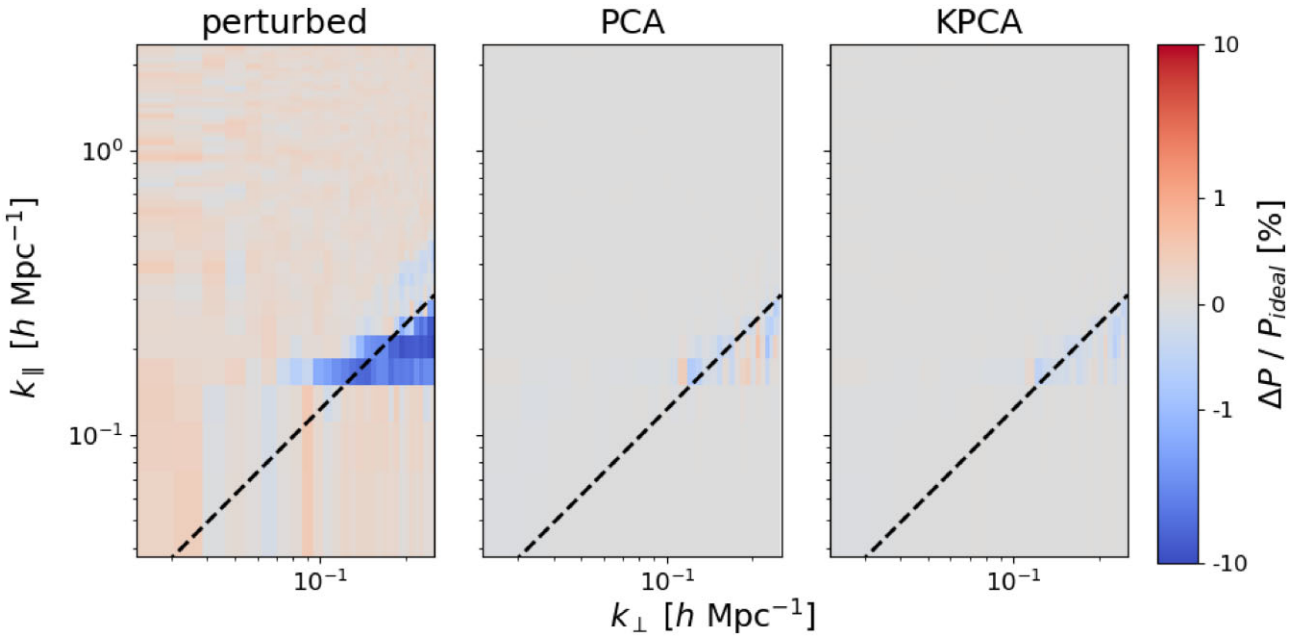

Figure 10. From left to right, the fractional difference between the PS of both cosmic signal and foregrounds convolved with $B_{\rm perturbed}$, $B_{\rm perturbed} - \widehat{\Delta B}_{\rm PCA}$, and $B_{\rm perturbed} - \widehat{\Delta B}_{\rm KPCA}$ with respect to the PS of the same sky convolved with $B_{\rm ideal}$.

beam directly, as we have done here. Instead, eigenvalues would need to be co-varied when performing calibration and inference. We defer this analysis to future work.

5 CONCLUSIONS

Some of the most important systematics in radio interferometry arise from imperfect knowledge of the telescope beam. In this work, we demonstrate an empirical approach to characterizing known sources of beam errors. Focusing on offline and offset antennas for an SKA-like beam, we generate thousands of realizations of beam errors. We use these realizations to define a beam error basis using PCA and KPCA.

We demonstrate that both PCA and KPCA perform well in recovering beam errors from offline and offset antennae. Compared with assuming an 'ideal' beam, using the top 20 components in either basis can reduce the MSE by ∼tens–100 per cent over our test data sets.

We demonstrate how this beam error characterization translates to improved power spectrum recovery. We generate a mock sky comprised of point source foregrounds and the cosmic signal, and recover the cylindrical power spectra assuming different beam models. For a random realization of beam error, we find that assuming the 'ideal' beam results in PS errors that peak at ∼10 per cent around the wedge region. Instead if either PCA or KPCA is used to characterize the perturbed beam with 20 components, the PS error is reduced by a factor of ∼10–100 throughout k-space.

We stress that we did not include additional errors from, e.g. calibration, in this work. We expect that fractional errors in sky-based calibration will be much more sensitive to errors in the assumed beam model, and these are further squared when propagated to power spectrum space. Depending on the spectral structure of these calibration errors, inaccuracies as small as 10^{-5} can be crippling to a power spectrum estimation (Barry et al. 2016; Patil et al. 2016). Therefore, we expect improved beam characterization to be even more important when calibration is also included; we defer this to future work.

Our general framework of using an empirical basis to characterize systematics should prove useful for an end-to-end inference pipeline for 21-cm interferometry. The principal eigenvectors from PCA and KPCA can provide an optimal basis for systematics, with the corresponding eigenvalues being co-varied together with cosmological parameters when performing Bayesian inference. We will demonstrate this in a follow-up work.

ACKNOWLEDGEMENTS

We thank P. Bull for helpful comments on a draft version of this work. This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 638809 - An Illumination to the Dark Ages: modelling reionization and interpreting observations (AIDA) - PI: Mesinger). The results presented here reflect the authors' views; the ERC is not responsible for their use. We gratefully acknowledge computational resources of the Center for High Performance Computing (CHPC) at Scuola Normale Superiore (SNS).

DATA AVAILABILITY

The data from this study will be shared on reasonable request to the corresponding author.

REFERENCES

Abdurashidova Z. et al., 2022, ApJ, 925, 221
Barry N., Hazelton B., Sullivan I., Morales M. F., Pober J. C., 2016, MNRAS, 461, 3135
Barry N. et al., 2019, ApJ, 884, 1
Bolli P., Di Ninni P., Bercigli M., Labate M. G., Virone G., 2021, in 15th European Conference on Antennas and Propagation (EuCAP). IEEE, Dusseldorf, Germany, p. 1
Choudhuri S., Bull P., Garsden H., 2021, MNRAS, 506, 2066
DeBoer D. R. et al., 2017, PASP, 129, 045001
Dewdney P. E., Hall P. J., Schilizzi R. T., Lazio T. J. L., 2009, Proc. IEEE, 97, 1482

Dillon J. S. et al., 2014, Phys. Rev. D, 89, 023002
Dulwich F., Mort B. J., Salvini S., Zarb Adami K., Jones M. E., 2009, Proceedings of Wide Field Astronomy & Technology for the Square Kilometre Array (SKADS 2009). Chateau de Limelette, Belgium, p. 31
Fagnoni N. et al., 2021, MNRAS, 500, 1232
Furlanetto S. R., Oh S. P., Briggs F. H., 2006, Phys. Rep., 433, 181
Hoerl A. E., Kennard R. W., 1970a, Technometrics, 12, 55
Hoerl A. E., Kennard R. W., 1970b, Technometrics, 12, 69
Intema H., Van Weeren R., Röttgering H., Lal D., 2011, A&A, 535, A38
Jacobs D. C. et al., 2017, PASP, 129, 035002
Li W. et al., 2019, ApJ, 887, 141
Line J. L. B. et al., 2018, PASA, 35, e045
Mellema G. et al., 2013, Exp. Astron., 36, 235
Mertens F. et al., 2020, MNRAS, 493, 1662
Mesinger A., Furlanetto S., Cen R., 2011, MNRAS, 411, 955
Morales M. F., Wyithe J. S. B., 2010, ARA&A, 48, 127
Murray S. G., Greig B., Mesinger A., Muñoz J. B., Qin Y., Park J., Watkinson C. A., 2020, J. Open Source Softw., 5, 2582
Nasirudin A., Murray S., Trott C., Greig B., Joseph R., Power C., 2020, ApJ, 893, 118
Newburgh L. B. et al., 2014, in Stepp L. M., Gilmozzi R., Hall H. J., eds, Proc. SPIE Conf. Ser. Vol. 9145, Ground-based and Airborne Telescopes V. SPIE, Bellingham, p. 91454V
Park J., Mesinger A., Greig B., Gillet N., 2019, MNRAS, 484, 933
Patil A. H. et al., 2016, MNRAS, 463, 4317
Schölkopf B., Smola A., Müller K.-R., 1997, International Conference on Artificial Neural Networks. Springer, Berlin, Heidelberg, p. 583
Sutinjo A. T. et al., 2015, IEEE Trans. Antennas Propag., 63, 5433
Sutinjo A. T., McKinley B., Belostotski L., Ung D. C., Thekkeppattu J. N., 2020, Union Radio-Sci. Int., 2
Swarup G., Ananthakrishnan S., Kapahi V., Rao A., Subrahmanya C., Kulkarni V., 1991, Curr. Sci., 60, 95
Thyagarajan N. et al., 2013, ApJ, 776, 6
Tingay S. et al., 2013, PASA, 30, 21
Trott C. M. et al., 2020, MNRAS, 493, 4711
van Haarlem M. P. et al., 2013, A&A, 556, A2
Wayth R. B. et al., 2018, Publ. Astron. Soc. Aust., 35, e033

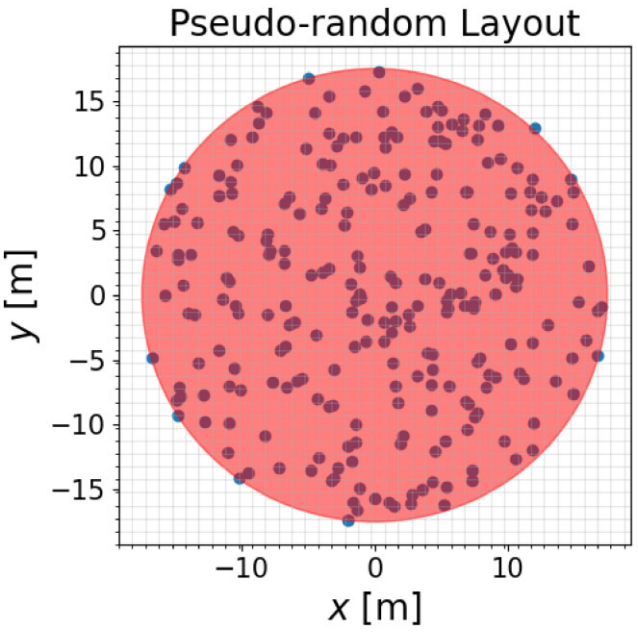

Figure A1. An example of a pseudo-random station layout.

APPENDIX A: PSEUDO-RANDOM STATION LAYOUT

In this section, we present some extra materials regarding a pseudo-random station layout that we have investigated in this work. Fig. A2 shows the OSKAR-generated beam using the layout shown in Fig. A1, along with a broken + offset perturbation on the same beam and the difference between the two.

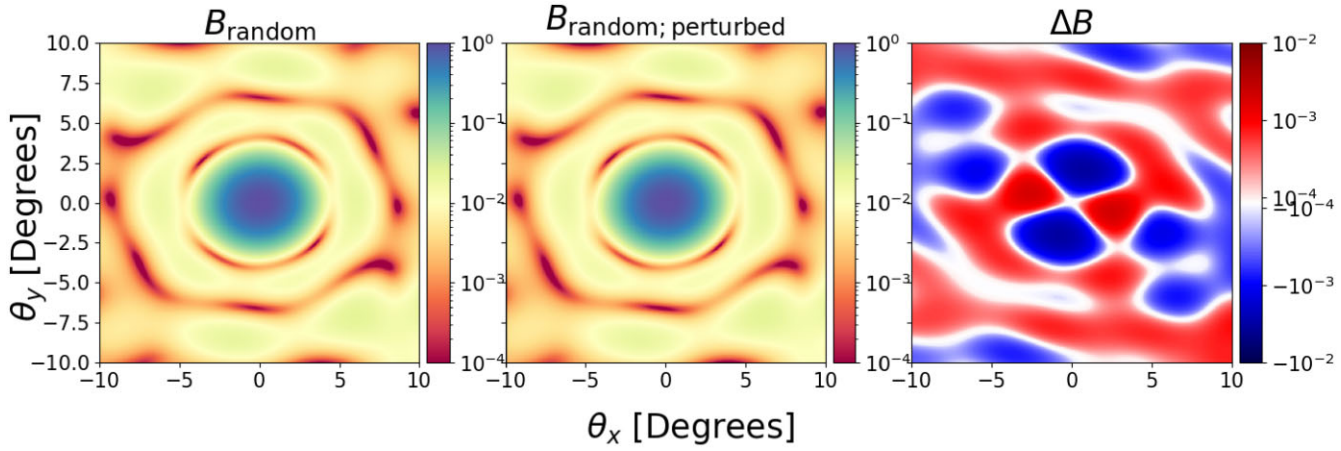

Figure A2. The beam using a pseudo-random layout, $B_{\text{random}}(\nu)$ (left-hand panel), an example realization of broken + offset perturbation on the same beam $B_{\text{random; perturbed}}(\nu)$ (middle panel), and the consequent $\Delta B(\nu)$ (right-hand panel) at $\nu = 150$ MHz. The fractional error is of the same general magnitude as the fractional error using a regularly spaced station layout in Fig. 2.

APPENDIX B: UNIT CONVERSION

The conversion of δT_B to $S(\nu)$ (and vice versa) follows the Rayleigh–Jeans law

$$S(\nu) = \left(\frac{2k_B\nu^2\delta T_B}{c^2}\right) \times 10^{26} \quad (\text{Jy sr}^{-1}), \tag{B1}$$

where k_B is the Boltzmann constant.

Under the assumption that τ is equivalent to the Fourier counterpart of the line-of-sight mode, η, both $k_\perp$ and $k_\parallel$ are converted from $\mathbf{r}$ and τ in Fourier dimensions following:

$$k_\perp = \frac{2\pi|\boldsymbol{r}|}{D_M(z)} \quad (\text{Mpc}^{-1}\,\text{h}), \tag{B2}$$

and

$$k_\parallel = \frac{2\pi H_0 f_{21} E(z)}{c(1+z)^2}\tau \quad (\text{Mpc}^{-1}\,\text{h}) \tag{B3}$$

from Morales & Wyithe (2010). Here, z is the observation redshift, $D_M(z)$ is the transverse comoving distance, H_0 is the Hubble constant, f_{21} is the rest frequency of the 21-cm hydrogen hyperfine transition and $E(z)$ is defined as

$$E(z) = \sqrt{\Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda}, \tag{B4}$$

where Ω_Λ, and Ω_k are the dimensionless density parameters for dark energy and the curvature of space.

APPENDIX C: EXTRA MATERIALS

In this section, we present some extra materials concerning the research for interested readers. The parameter input used for OSKAR is presented in Table C1 and the SKA-like station layout is shown in Fig. C1. In addition, Fig. C2 shows the probability density functions (PDFs) of the first 20 components in the higher dimension space from the KPCA, ordered from largest (top left panel) to smallest (bottom right panel) variance.

Table C1. The OSKAR parameter input used in this research.

Parameter	Values
FoV (°)	20
RA of Observation (°)	0
Dec of Observation (°)	−27
Latitude of Telescope (°)	−27
Longitude of Telescope (°)	117
Observation Time (UTC)	17:00:00
Observation Date	1 August 2020
Frequencies (MHz)	[150, 170, 190]

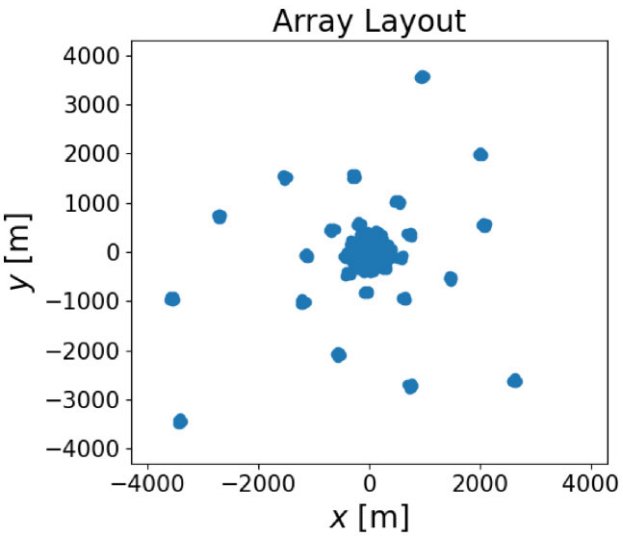

Figure C1. The SKA-like array layout.

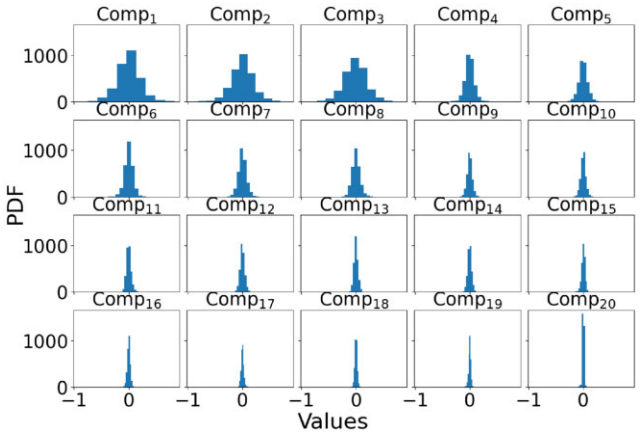

Figure C2. The PDFs of the first 20 components in the higher dimension space from the KPCA, ordered from largest (top left panel) to smallest (bottom right panel) variance.

APPENDIX D: KERNEL PCA EXAMPLE

The purpose of this simple example is to qualitatively describe the main ingredients of the KPCA algorithm – in particular: data space – y, feature space – $\psi(y)$, and two kernels (κ and $\widetilde{\kappa}$) defining mappings from one to another. Moreover, we would like to show very different roles the two kernels have in the process.

In Fig. D1, we show a bi-modal distribution (with the two modes labeled '1' and '2'). The horizontal axis represents data space. In this simple example, we would like to use KPCA to accentuate the bi-modality of this distribution.

We first pull samples from the distribution and use the kernel $\kappa(y_i, y_j) = \psi(y_i) \cdot \psi(y_j)$ as a part of the KPCA algorithm. Here $\psi(y)$ is in feature space, which is not directly accessible and can be infinite dimensional. Performing KPCA in this space and selecting the first N components amounts to selecting basis vectors in the feature space following the largest variance of the samples. After fitting the data, this subset of a feature space $\boldsymbol{\psi}_N$ is accessible and the mapping $\boldsymbol{\psi}_N(y)$ is known.

D1 Mapping to and from feature space

For the example above, one can show the first component of the feature space $\psi_1(y)$ (see Fig. D2). Samples from the modes in the distribution are distinguished by different colours, for better visualization. Starting from the distribution on the top, we pass it through the (learned) transformation shown in the middle, getting the distribution on the right. As expected, the two modes (blue versus orange) are much better separated in the first component of the feature space, ψ_1, than they were in data space, y.

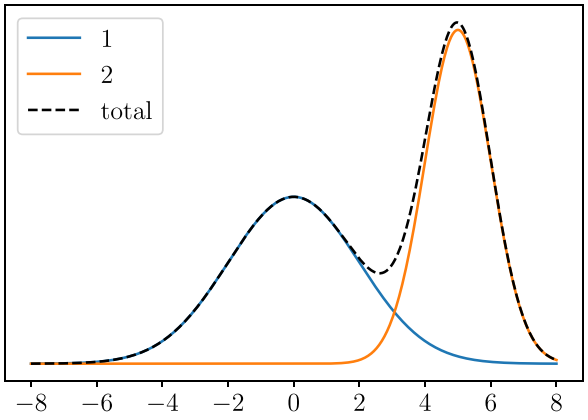

Figure D1. Initial distribution on which we would like to run a KPCA algorithm.

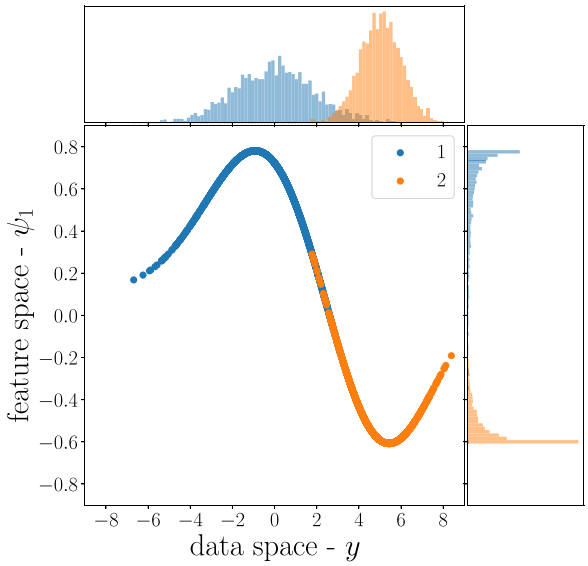

Figure D2. Mapping from data space to the KPCA feature space $\boldsymbol{\psi}_N$. Starting from the top distribution in data space, points are transformed through the learned function in the middle into the first component of the feature space, shown on the right. Samples from two underlying distributions are denoted with different colours for better visualization.

Contrary to linear PCA, however, an exact inverse transform to return from ψ_1 back to data space generally does not exist. Therefore, we define the inverse transform using kernel ridge regression. In linear ridge regression from $\boldsymbol{\psi}_N$ back to y, we would minimize the mean square error over the data:

$$\sum_i \left(y_i - \mathbf{w}^T \boldsymbol{\psi}_N\right)^2 + \frac{\lambda}{2}||\mathbf{w}||^2, \tag{D1}$$

where $\mathbf{w}$ are the weights and second term is a standard l_2 regularization. However, as the inverse function is highly non-linear, we firstly transform $\boldsymbol{\psi}_N$ into another (possibly infinite) feature space ϕ, and learn the transformation $\widetilde{\mathbf{w}}$ back to y. The minimization is then:

$$\sum_i \left(y_i - \widetilde{\mathbf{w}}^T \phi\left(\boldsymbol{\psi}_N\right)\right)^2 + \frac{\lambda}{2}||\widetilde{\mathbf{w}}||^2 . \tag{D2}$$

One can prove that the feature space ϕ does not have to be accessed and is only implicitly defined by the kernel $\widetilde{\kappa}\left((\boldsymbol{\psi}_N)_i, (\boldsymbol{\psi}_N)_j\right) = \phi\left((\boldsymbol{\psi}_N)_i\right) \cdot \phi\left((\boldsymbol{\psi}_N)_j\right)$.

In Fig. D3, we show the results of such procedure. We can see that the learned mapping is indeed non-linear and the initial distribution is well preserved. However, the inverse transform is not exact and differences between distributions of initial and recovered samples can be clearly seen (Fig. D4).

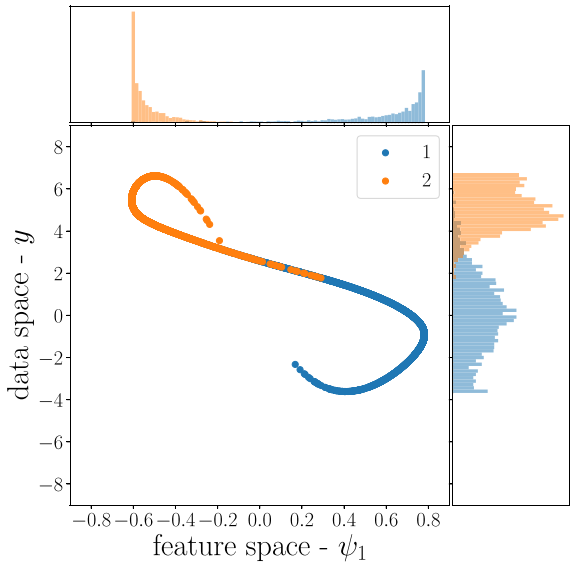

Figure D3. Mapping from the feature space $\boldsymbol{\psi}_N = (\psi_1, \psi_2, ...)$ back to the data space, y. Kernel ridge regression feature space ϕ is never accessed and only defined by the kernel $\tilde{\kappa}$. Samples from the two underlying modes are separated in colour for better visualization.

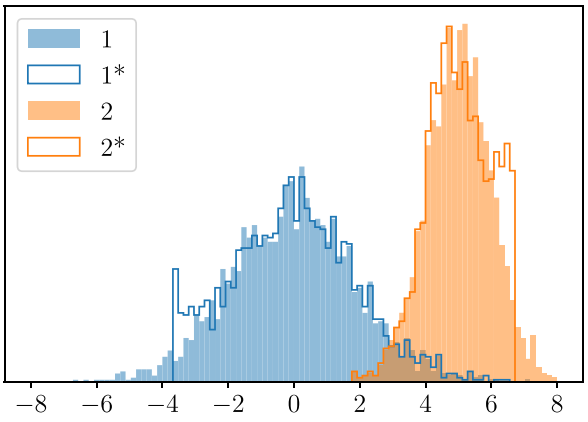

Figure D4. Histogram of initial samples (marked with 1 and 2) and samples recovered after inverse transformation from ψ_N to y (marked with $1*$ and $2*$).

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