Publications of the Astronomical Society of Australia (PASA)
doi: 10.1017/pas.2024.xxx.

w-stacking w-projection hybrid algorithm for wide-field interferometric imaging: implementation details and improvements

L. Pratley[1]*, M. Johnston-Hollitt[2] and J. D. McEwen[1]
[1]Mullard Space Science Laboratory (MSSL), University College London (UCL), Holmbury St Mary, Surrey RH5 6NT, UK
[2]International Centre for Radio Astronomy Research (ICRAR)- Curtin University, 1 Turner Ave, Bentley, 6102, WA, Australia

Abstract

We present a detailed discussion of the implementation strategies for a recently developed w-stacking w-projection hybrid algorithm used to reconstruct wide-field interferometric images. In particular, we discuss the methodology used to deploy the algorithm efficiently on a supercomputer via use of a Message Passing Interface (MPI) k-means clustering technique to achieve efficient construction and application of non co-planar effects. Additionally, we show that the use of conjugate symmetry can increase the w-stacking efficiency, decreasing the time required to construction and apply w-projection kernels for large data sets. We then demonstrate this implementation by imaging an interferometric observation of Fornax A from the Murchison Widefield Array (MWA). We perform exact non-coplanar wide-field correction for 126.6 million visibilities using 50 nodes of a computing cluster. The w-projection kernel construction takes only 15 minutes prior to reconstruction, demonstrating that the implementation is both fast and efficient.

Keywords: techniques: image processing – techniques: interferometric – methods: data analysis

1 INTRODUCTION

The advent of wide-field interferometers such as the Murchison Widefield Array (MWA; Tingay et al., 2013; **?**), Long Wavelength Array (LWA; Ellingson et al., 2009) and the Low Frequency Array (LOFAR; van Haarlem et al., 2013) has created a number of imaging challenges. These challenges include the large number of measurements in each observation, the instrumental effects that are measurement dependent, and the large image sizes due to high resolution and wide-field of view. Additionally, these telescopes have a variety of science goals, including high priority science such as probing Galactic and extra-galactic magnetic fields (especially in low mass galaxy clusters; Johnston-Hollitt et al. 2015), and detecting the redshifted 21cm spectral line of the Epoch of Reionoization (Koopmans et al., 2015). Furthermore, the wide-field of view provides the advantage of observing many objects in a single pointing, reducing the observation time needed to survey the radio sky. If the imaging challenges are overcome, it will herald an era of unprecedented sensitivity and resolution for the low frequency sky, over extremely wide-field of views.

Non-coplanar baselines, (u, v, w), in the presence of wide-fields of view produce measurement dependent effects, i.e. a directional dependent effect (DDE) that is different for each measurement. Each w value provides a complex exponential, known as a chirp, that needs to be modelled in the image domain and applied during image reconstruction. This has been through the use of two algorithms, the w-stacking algorithm, where average w corrections are applied in the image domain to groups of measurements, and the w-projection algorithm, where average w-corrections are applied when degridding in the (u, v, w) domain. The w-stacking algorithm(Humphreys & Cornwell, 2011) has the trade off that a Fast Fourier Transform (FFT) needs to be applied for each w group. The w-projection algorithm (Cornwell, 2008) has the trade off that kernel construction can be expensive and the sup-

[*] Luke.Pratley@gmail.com

port size is large for large w values. Both algorithms have been limited to correcting individual groups of measurements for large data sets (Cornwell, 2008; Offringa et al., 2014).

Two recent developments have allowed individual correction for each data set. The first is the use of adaptive quadrature and radial symmetry to calculate w-projection kernels orders of magnitude faster than the full 2d calculation (Pratley et al., 2019d, hereafter Paper I). The second is the developments in distributed image reconstruction from state of the art convex optimization algorithms, which provide a natural framework for the Message Passing Interface (MPI) distribution of FFTs and degridding for radio interferometric imaging (Pratley et al., 2019a). Recently, an MPI hybrid w-stacking w-projection algorithm demonstrating these developments was applied on a super computing cluster, where 17.5 million measurements were individually corrected over a 25 by 25 degree field of view from an MWA observation (Paper I). Such individual correction has not been previously possible.

After reviewing the w-stacking w-projection algorithm, we provide the algorithmic details of how to distribute the measurements through a k-means clustering algorithm to improve computational performance, the use of conjugate symmetry to reduce the range of w values, and show the application of these algorithms to a larger data set to demonstrate the improvement. We end with a discussion of future strategies for kernel calculation and adapting the algorithm to model other DDEs.

The paper is laid out as follows. Section 2 introduces the wide-field interferometric measurement equation. Section 3 describes the distributed k-means clustering algorithm used to create the w-stacks and the reconstruction algorithm used to generate a sky model of the observed data. Section 4 times and compares the w-stacking w-projection algorithm before and after using conjugate symmetry, as a function of image size, w-range, and number of visibilities. Section 5 demonstrates the application of the algorithm for this implementation on an observation of Fornax A. Section 6 proposes possible improvements in kernel calculation for large data sets, and discusses how other directional dependent effects can be included into the algorithm. The work is concluded in Section 7.

2 WIDE-FIELD IMAGING MEASUREMENT EQUATION

The non-coplanar wide-field interferometric measurement equation is

$$y(u, v, w') = \int x(l, m) a(l, m) \frac{\mathrm{e}^{-2\pi i w'(\sqrt{1-l^2-m^2}-1)}}{\sqrt{1 - l^2 - m^2}} \times \mathrm{e}^{-2\pi i (lu + mv)}\, \mathrm{d}l \mathrm{d}m\,, \tag{1}$$

where (u, v, w') are the baseline coordinates and (l, m, n) are directional cosines restricted to the unit sphere. In this work, we define $w' = w + \bar{w}$, where $\bar{w}$ is the average value of w-terms, and w is the effective w-component (with zero mean), x is the sky brightness and a includes direction dependent effects such as the primary beam. The measurement equation is a mathematical model of the measurement process, i.e. signal acquisition, that allows one to calculate model measurements y when provided with a sky model x.

A number of methods can be used to solve for x given samples y, such as CLEAN (Högbom, 1974), Maximum Entropy (Ables, 1974; Cornwell & Evans, 1985), and Sparse Regularization algorithms (McEwen & Wiaux, 2011; Onose et al., 2016; Pratley et al., 2018; Dabbech et al., 2018; Pratley et al., 2019d,a). Ultimately, all interferometric measurement equations are derived from the van Cittert-Zernike theorem (Zernike, 1938) and the measurement equation can be extended to include general direction dependent effects and polarization, and to solve for x natively on the sphere (McEwen & Scaife, 2008; Smirnov, 2011; Price & Smirnov, 2015).

To make use of the FFT, the measurement equation is traditionally calculated and approximated using degridding (Fessler & Sutton, 2003; Thompson et al., 2008). The measurement equation can be represented by the following linear operations

$$\boldsymbol{y} = \mathsf{WGCFZS}\boldsymbol{x}\,. \tag{2}$$

S represents a gridding correction and correction of baseline independent effects such as $\bar{w}$, Z represents zero padding of the image, F is an FFT, G represents a sparse circular convolution matrix that interpolates measurements off the grid and the combined GC includes baseline dependent effects such as variations in the primary beam and w-component in the interpolation, and W are weights applied to the measurements. This linear operator is typically called a measurement operator $\boldsymbol{\Phi} = \mathsf{WGCFZS}$ with $\boldsymbol{\Phi} \in \mathbb{C}^{M\times N}$. Furthermore, $\boldsymbol{x}_i = x(\boldsymbol{l}_i)$ and $\boldsymbol{y}_k = y(\boldsymbol{u}_k)$ are discrete vectors in $\mathbb{R}^{N\times 1}$ and $\mathbb{C}^{M\times 1}$ in this setting. The measurement operator has an adjoint operator $\boldsymbol{\Phi}^\dagger$. The dirty map can be calculated by $\boldsymbol{\Phi}^\dagger\boldsymbol{y}$, and the residual map by $\boldsymbol{\Phi}^\dagger\boldsymbol{y}-\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}\boldsymbol{x}$.

3 DISTRIBUTED WIDE-FIELD IMAGING

In this section, we briefly describe the algorithmic details for the distributed w-projection w-stacking hybrid algorithm.

We use the interferometric image reconstruction software package PURIFY[1] (version 3.0.1, Pratley et al. 2019b) developed in C++ (Carrillo et al., 2014; Pratley et al., 2018, 2019a), where the authors have implemented an MPI distributed measurement operator. The authors have also developed MPI distributed wavelet transforms, along with MPI variations of the alternating direction method of multipliers (ADMM) algorithm in the software package SOPT[2] (version 3.0.1, Pratley et al. 2019c).

This is not the first time sparse image reconstruction has been used for wide-fields of view. In particular, the w-term is known to spread information across visibilities, increasing the effective bandwidth in what is known as the spread spectrum effect (Wiaux et al., 2009; McEwen & Wiaux, 2011; Wolz et al., 2013; Dabbech et al., 2017), increasing the possible resolution of the reconstructed sky model. But these previous works have been restricted to proof-of-concept studies. One of the advantages of sparse image reconstruction algorithms, such as ADMM, is that they can allow direct reconstruction of an accurate sky model, unlike CLEAN based algorithms that produce a restored image (Pratley et al., 2018).

3.1 w-projection w-stacking measurement operator

In the MPI w-stacking w-projection algorithm the measurement operator corrects for the average w-value in each w-stack, then applies an extra correction to each visibility with the w-projection. Each w-stack $\boldsymbol{y}_k$ has the measurement operator of

$$\boldsymbol{\Phi}_k = \mathsf{W}_k\mathsf{GC}_k\mathsf{FZ}\tilde{\mathsf{S}}_k\,, \tag{3}$$

the gridding correction, $\tilde{\mathsf{S}}_k$, has been modified to correct for the w-stack dependent effects, such as the average $\bar{w}_k$ or the primary beam

$$\left[\tilde{\mathsf{S}}_k\right]_{ii} = \frac{a_k(l_i, m_i)\mathrm{e}^{-2\pi i\bar{w}_k(\sqrt{1-l_i^2-m_i^2}-1)}}{g(\sqrt{l_i^2+m_i^2})\sqrt{1-l_i^2-m_i^2}}\,. \tag{4}$$

We choose no primary beam effects within the stack $a_k(l_i, m_i)$. $g(\sqrt{l_i^2+m_i^2})$ is the window for the anti-aliasing filter. This gridding correction shifts the relative w value in the stack. This can reduce the effective w value in the stack, especially when the stack is close to the mean $\bar{w}_k$, i.e. the value of $w_i - \bar{w}_k$ is small for all i in stack k. This reduces the size of the support needed in the w-projection gridding kernel for each stack,

$$[\mathsf{GC}_k]_{ij} = [GC]\Big(\sqrt{(u_i/\Delta u - q_{u,j})^2 + (v_i/\Delta u - q_{v,j})^2}\,, w_i - \bar{w}_k, \Delta u\Big)\,. \tag{5}$$

$(q_{u,j}, q_{v,j})$ represents the nearest grid points, and we use adaptive quadrature to calculate

$$[GC]\Big(\sqrt{u_{\mathrm{pix}}^2+v_{\mathrm{pix}}^2}, w, \Delta u\Big) = \frac{2\pi}{\Delta u^2}\int_0^{\alpha/2} g(r) \times\mathrm{e}^{-2\pi i w(\sqrt{1-r^2/\Delta u^2}-1)} J_0\left(2\pi r\sqrt{u_{\mathrm{pix}}^2+v_{\mathrm{pix}}^2}\right) r\mathrm{d}r\,, \tag{6}$$

where $g(r)$ is the radial anti-aliasing filter, Δu is the resolution of the Fourier grid of the field of view zero padded by the oversampling ratio $\alpha = 2$, and $(u_{\mathrm{pix}}, v_{\mathrm{pix}})$ are the pixel coordinates on the Fourier grid. More details can be found in Paper I.

For each stack $\boldsymbol{y}_k \in \mathbb{C}^{M_k}$ we have the measurement equation $\boldsymbol{y}_k = \boldsymbol{\Phi}_k\boldsymbol{x}$. It is clear that each stack has an independent measurement equation. However, the full measurement operator is related to the stacks in the adjoint operators such that

$$\boldsymbol{x}_{\mathrm{dirty}} = \begin{bmatrix}\boldsymbol{\Phi}_1^\dagger, & \ldots, & \boldsymbol{\Phi}_{k_{\max}}^\dagger\end{bmatrix}\begin{bmatrix}\boldsymbol{y}_1\\ \vdots\\ \boldsymbol{y}_{k_{\max}}\end{bmatrix} = \boldsymbol{\Phi}^\dagger\boldsymbol{y}\,. \tag{7}$$

We use MPI all reduce to sum over the dirty maps generated from each node. The full operator $\boldsymbol{\Phi}$ is normalized using the power method.

3.2 Clustering w-stacks

It is ideal to minimize the kernel sizes across all stacks, minimizing the memory and computation costs of the kernel. We develop an MPI k-means clustering algorithm which greatly improves performance by reducing the values of $|w_i - \bar{w}_k|^2$ across the w-stacks. Each MPI node finds the w-stack to which a visibility belongs, updating the cluster centers across all MPI nodes with each iteration. This is then followed by an all-to-all MPI operation to distribute the visibilities to their w-stacks. There already exist parallel and distributed

[1] `https://github.com/astro-informatics/purify`

[2] `https://github.com/astro-informatics/sopt`

k-means clustering algorithms for big data (Stoffel & Belkoniene, 1999; Aggarwal & Reddy, 2013). The k-means w-clustering algorithm is presented in Algorithm 1. This algorithm is necessary to reduce computation and operating memory when applying the w-projection kernels by reducing the support size of each kernel.

3.3 Conjugate symmetry

Prior to w-stacking with the k-means algorithm, conjugate symmetry may be used to restrict the w-values onto the positive w-domain. The origin of the w-effect stems from the 3d Fourier transform of a spherical shell and a horizon window, with the w component probing the Fourier coefficient of the signal along the line of sight. The sky, the horizon window, the spherical shell, and the primary beam can all be interpreted as a real valued signal. This provides a conjugate symmetry between $-|w|$ and $+|w|$, i.e.

$$y^*(u, v, -|w|) = y(-u, -v, |w|)\,. \tag{8}$$

Properties of noise remain unchanged under conjugate symmetry, meaning that measurements can be restricted to positive w, i.e. $w \in \mathbb{R}_+$. Other modelled instrumental effects may need to be conjugated, which is only important when they are complex valued signals. In particular, polarized signals, e.g. Stokes Q, U, and V, are independent real valued signals. Thus, linear polarization has a slightly different relation

$$y_P^*(u, v, -|w|) = y_Q(-u, -v, |w|) - i y_U(-u, -v, |w|)\,, \tag{9}$$

suggesting the reflection should be done to the Stokes Q and U visibliities before combination into linear polarization, and then combined with $-i$ rather than $+i$. This combination is important for accurate polarimetirc image reconstruction (Pratley & Johnston-Hollitt, 2016).

3.4 Distributed ADMM

As in Paper I, we use the alternating direction method of multipliers (ADMM) algorithm implemented in PURIFY (Pratley et al., 2018, 2019a) to solve the optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^N} \left\|\boldsymbol{\Psi}^\dagger \boldsymbol{x}\right\|_{\ell_1} \quad \text{subject to} \quad \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_{\ell_2} \le \epsilon\,, \tag{10}$$

where $\boldsymbol{\Psi}$ is a wavelet transform, the term $\left\|\boldsymbol{\Psi}^\dagger \boldsymbol{x}\right\|_{\ell_1}$ is a penalty on the number of non-zero wavelet coefficients, while $\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_{\ell_2} \le \epsilon$ is the condition that the measurements fit within a Gaussian error bound ϵ. MPI is used to distribute the wavelet transform and enforce fidelity constraints, in conjunction with w-stacking.

PURIFY (version 3.0.1, Pratley et al. 2019b) has been updated to implement the w-stacking w-projection measurement operator with MPI, k-means clustering, and conjugate symmetry to efficiently reduce the effective w-value within a compute cluster. We find that the use of conjugate symmetry allows the k-means algorithm to increase the density of the w-stack locations. This in turn reduces the effective w values that are required to be corrected for by the w-projection kernels, and greatly decreases the computational burden of the w-projection algorithm in the kernel construction.

4 EFFICIENCY OF W-STACKING WITH CONJUGATE SYMMETRY

In this section we compare the efficiency of the w-projection w-stacking algorithm before and after applying the conjugate operation to the visibilities. By restricting w to be greater than zero, we increase the density of the w-stacks, decreasing the distance $|w - \bar{w}_k|$ of each visibility from the center of a given w-stack, k. When this distance is negligible, the correction required by the w-projection algorithm is negligible. The off-set error in a particular w-stack can be further corrected by the w-projection algorithm for the following expression

$$\mathrm{e}^{-2\pi i(w - \bar{w}_k)(\sqrt{1 - l^2 - m^2} - 1)}\,. \tag{11}$$

By performing a Taylor series expansion of the exponential $\mathrm{e}^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$, we find the real part grows as $|2\pi(w - \bar{w}_k)(\sqrt{1 - l^2 - m^2} - 1)|^2/2$ and the imaginary part leads as $|2\pi(w - \bar{w}_k)(\sqrt{1 - l^2 - m^2} - 1)|$. For simplicity, we know that $(w - \bar{w}_k)$ is bounded by the w spread of the visibilities in a stack Δw_k. We can relate the spread in l and m with the radius Δr for the field of view, and find the relation $\Delta r = 1/(2\Delta u)$. Using this analysis, we want $|2\pi\Delta w_k(1 - \sqrt{1 - (\Delta r)^2})|$ to be roughly less than the small offset error η in the zero padded field of view for every visibility in the stack, k. When the w-stacks are evenly spread (which is expected to be less efficient than using k-means), we find $\Delta w_k = (w_{\max} - w_{\min})/n_{\mathrm{d}}$, where n_{d} is the number of w-stacks. This provides a bound between the number of w-stacks and offset error over a given field of view

$$n_{\mathrm{d}} \ge \frac{|2\pi(w_{\max} - w_{\min})(\sqrt{1 - (\Delta r)^2} - 1)|}{\eta}\,. \tag{12}$$

When we apply conjugate symmetry we find that the difference $(w_{\max} - w_{\min})$ is reduced to

Algorithm 1 k-means w-stacking:
The k-means algorithm sorts the visibilities into clusters (w-stacks) by minimizing the average w deviation, $(\bar{w} - w)^2$, within each cluster.We use bold variables to denote an array, subscript to denote the array element and superscript to denote the iteration. The algorithm returns two arrays: $\boldsymbol{n}$ is the array of indices that labels the w-stack for each visibility; $\bar{\boldsymbol{w}}$ is the average w value within each w-stack. The algorithm requires a starting w-stack distribution $\bar{\boldsymbol{w}}^{(0)}$, which we choose to be evenly distributed between the minimum and maximum w-values. The algorithm should iterate until $\bar{\boldsymbol{w}}^{(t)}$ has converged, which we choose to be a relative difference of 10^{-3}. Note p is the index of visibility, q is the index for w-stacks, and c is the place holder for the minimum deviation for the visibility at index p. The AllSumAll(x) operation is an MPI reduction of a summation followed by broadcasting the result to all compute nodes.

```
given $\bar{\boldsymbol{w}}^{(0)}, \boldsymbol{n}^{(0)}, w_{\mathrm{total}}, n_{\mathrm{total}}, \boldsymbol{w}_{\mathrm{sum}}, \boldsymbol{w}_{\mathrm{count}}$
repeat for $t = 1, \ldots$
    $\boldsymbol{w}_{\mathrm{sum}} = \boldsymbol{0}$
    $\boldsymbol{w}_{\mathrm{count}} = \boldsymbol{0}$
    repeat for $p = 1, \ldots$
        $m = 2(w_{\max} - w_{\min})^2$
        repeat for $q = 1, \ldots$
            $c = (\bar{\boldsymbol{w}}_q^{(t)} - \boldsymbol{w}_p)^2$
            if $c < m$ then
                $m = c$
                $\boldsymbol{n}_p^{(t+1)} = q$
            end if
        until $q > n_{\mathrm{total}}$
        $\boldsymbol{w}_{\mathrm{sum}_{\boldsymbol{n}_p^{(t+1)}}} = \boldsymbol{w}_{\mathrm{sum}_{\boldsymbol{n}_p^{(t+1)}}} + \boldsymbol{w}_p$
        $\boldsymbol{w}_{\mathrm{count}_{\boldsymbol{n}_p^{(t+1)}}} = \boldsymbol{w}_{\mathrm{count}_{\boldsymbol{n}_p^{(t+1)}}} + 1$
    until $p > w_{\mathrm{total}}$
    repeat for $q = 1, \ldots$
        $\bar{\boldsymbol{w}}_q^{(t+1)} = 0$
        if AllSumAll($\boldsymbol{w}_{\mathrm{count}_q}$) $> 0$ then
            $\bar{\boldsymbol{w}}_q^{(t+1)} = $ AllSumAll($\boldsymbol{w}_{\mathrm{sum}_q}$)/AllSumAll($\boldsymbol{w}_{\mathrm{count}_q}$)
        end if
    until $q > n_{\mathrm{total}}$
until convergence
```

$\max(|w|) - \min(|w|)$. This reduces the number of w-stacks n_{d} required to reach a level of accuracy over the image, suggesting the efficiency increase. For example, after applying conjugate symmetry to a uniform w coverage with $w_{\max} = -w_{\min}$, only half the number of stacks are needed for the same level of accuracy. In practice the k-means algorithm can also reduce the number of stacks required, when w-coverages are clustered rather than uniformly spread, which is typically the case.

We also estimate that the 2-dimensional support size within a w-stack will be bounded by the maximum of J^2 and $(2\Delta w_k/\Delta u)^2$, and it is clear that more efficient placements of w-stacks reduces memory and computation needed with the w-projection kernel. For uniform coverage, we expect that the number of 2d kernel coefficients is bounded by $(2(w_{\max} - w_{\min})/(n_{\mathrm{d}}\Delta u))^2$. This bound on support is further reduced to $(2(\max(|w|) - \min(|w|))/(n_{\mathrm{d}}\Delta u))^2$ when conjugate symmetry is applied.

Lastly, when $\eta \geq |2\pi(w - \bar{w}_k)(\sqrt{1 - (\Delta r)^2} - 1)|$ for a chosen tolerance and given visibility, we suggest that there is little advantage in using the w-projection kernel. There may be small gains in kernel construction time by assuming $w = \bar{w}_k$ to avoid calculating the w-projection kernel through adaptive quadrature when the Hankel transform of $g(r)$ has a closed form. From the work of Pratley et al. (2019d), a safe choice to bound the error is $\eta = 0.01$ but we expect this to be very conservative for most science cases. In the limit where the stacking density is high enough, this method then reduces to the standard w-stacking algorithm.

4.1 Comparison

In this section we show the increase in efficiency of the construction and application of the measurement operator using the w-projection and w-stacking algorithm before and after applying conjugate symmetry to the visibilities.

To compare the efficiency, we undertake a series of timing experiments using a range of images sizes, w-stacks, and number of visibilities.

To perform the reconstruction we used the Grace computing cluster at University College London. Each node of Grace contains two 8 core Intel Xeon E5-2630v3 processors (16 cores total) and 64 Gigabytes of RAM.[3]

Each data point was generated using 25 compute nodes and 25 w-stacks (*i.e.* one w-stack per node). The coverage was generated randomly using a Gaussian sampling density in u, v, and w. We choose the standard deviation of w to be 100 wavelengths, making the full range to be approximately ± 300 wavelengths. The field of view was kept fixed to 25 by 25 degrees while we vary the range of w, number of pixels (N), and number of visibilities (M). We repeated each timing measurement thrice and then record the average time for each experimental configuration.

We used an oversampling ratio of $\alpha = 2$, with a 2d kernel support size $J^2 = 16$ at $w = 0$, with the support size scaling as $(2(w_i - \bar{w}_k)/\Delta u)^2$. The standard deviation of the w sample density is determined by the value σ_w, which we chose values of $50, 100, 150$ wavelengths, with maximum w values of $\pm 3\sigma_w$. Figure 1 shows the time required to construct the measurement operator $\mathbf{\Phi}$ and apply $\mathbf{\Phi}^\dagger\mathbf{\Phi}$ as a function of image size, for $N = 256^2, 512^2, 1024^2, 2048^2$, and 4096^2 pixels, and $M = 10^6, 10^7$, and 10^8 visibilities. All w-projection kernels are stored across the cluster to be ready for application.

For constructing $\mathbf{\Phi}$ we find that kernel construction time is independent of image size, which is clear when the kernel construction dominates over the cost of planning the FFT. This is demonstrates the advantage of using adaptive quadrature during kernel construction for high resolution images, where the computation scales with the field of view and w only, leaving it completely independent of number of pixels in the image.

For $\sigma_w = 50$ we find that there is little improvement by applying the conjugate, which is easily explained by suggesting that 25 w-stacks is enough to efficiently cover the range of w values over $[-150, 150]$ and $[0, 150]$ wavelengths.

For the larger w ranges of $\sigma_w = 100$ and $\sigma_w = 150$ we find that applying conjugation to the visibilities increases the efficiency of the w-stacking density. This reduces the w-projection kernel size, improving the construction speed of the w-projection kernels considerably. Kernel construction is approximately 5 times faster after applying conjugation. The reduced w-kernel size also reduces the time required to perform degridding and gridding operations during image reconstruction. However, as mentioned in the previous section, these performance gains are only seen if there are many visibilities with $w < 0$.

For $\sigma_w = 150$ with $M = 10^8$, we found that not applying conjugation resulted in large kernel construction times of greater than 140 minutes, and we did not have the compute resources to measure this as a function of N. However, applying conjugation significantly reduced construction times to 30 minutes.

In Figure 2, we fixed the image size to be small $N = 256^2$ and measure construction and application times for $M = 10^6, 10^7, and 10^8$. We find that there is linear scaling in construction time as a function of M. The application times also increase with M, but it is not clear that it is linear.

We also find that in the time to apply the measurement operator, the FFT scales with image width $\sqrt{N}$, and the contribution from the application of the gridding and degridding kernels that grows with M. This is expected from the two contributions $\mathcal{O}(M)$ and $\mathcal{O}(\alpha^2 N \log \alpha^2 N)$ for the interpolation and FFT respectively. However, we expect that the application time is limited by the node with the most measurements. Also the varying kernel support sizes make it difficult to expect a clear relation for application time against the number of measurements.

4.2 Current Implementation Limitations

While we have shown performance improvements with this work, there are still limitations with the current implementation. We note that some of these limitations can be overcome. First, we pre-compute and store all of the kernels for use during image reconstruction. While this is fine for short snapshot observations, it requires a large amount of working memory, and we expect that on-the-fly calculation methods proposed later in this work may prove useful (see Section 6). Second, this implementation is bottlenecked in working memory and CPU resources by the node with the w-stack that contains the largest number of gridding kernel coefficients. An alternative method for distributing the gridding kernel coefficients and balancing computational load across the MPI nodes is described in Pratley & McEwen (2019).

5 APPLICATION TO MWA OBSERVATION OF FORNAX A

We use PURIFY (version 3.0.1, Pratley et al. 2019b) to perform wide-field image reconstruction of an

[3] More details can be found at `https://wiki.rc.ucl.ac.uk/wiki/RC_Systems#Grace_technical_specs`

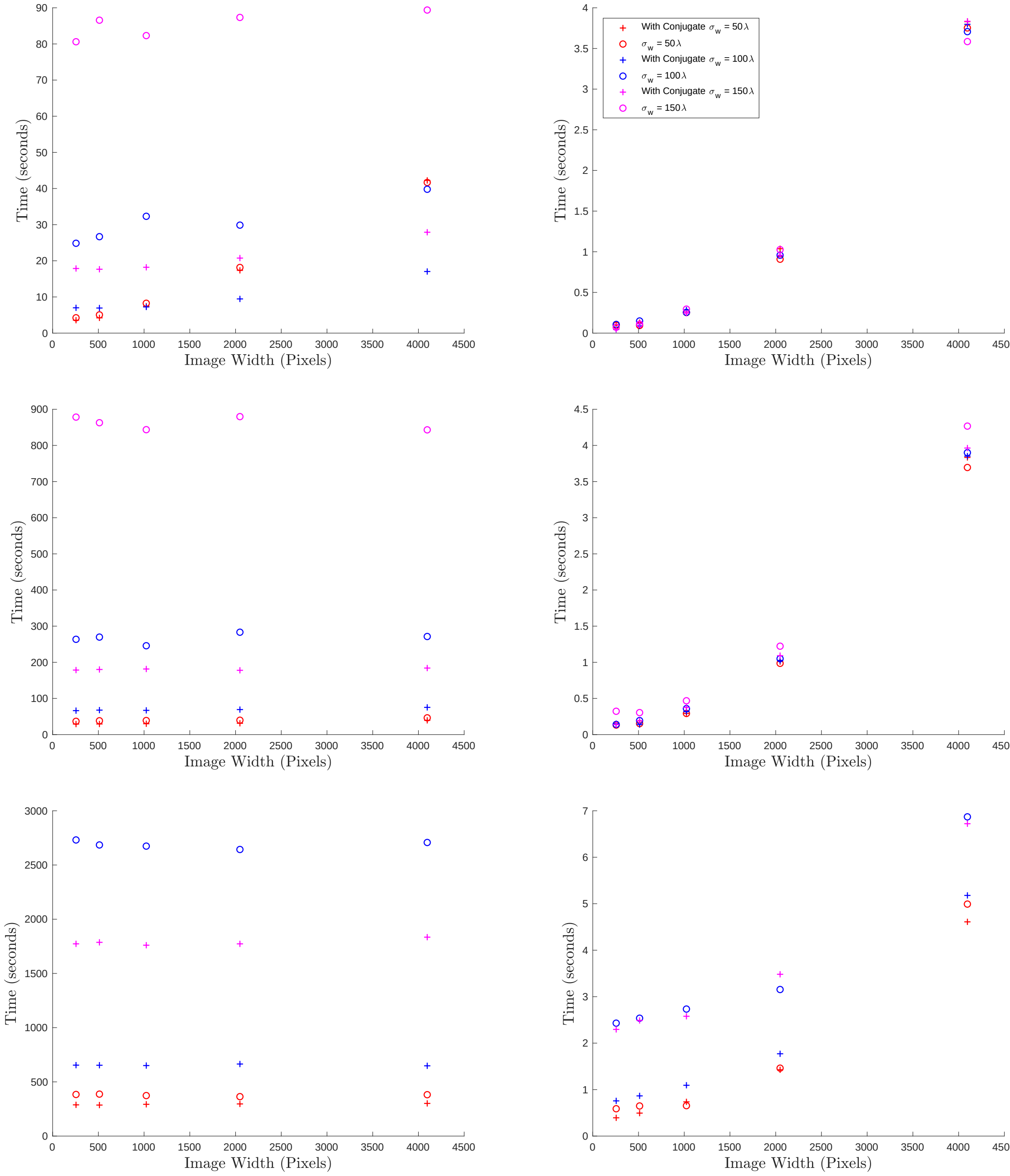

Figure 1. The left column shows plots of measurement operator ($\mathbf{\Phi}$) construction times and the right column shows plots of $\mathbf{\Phi}^\dagger\mathbf{\Phi}$ application times, as a function of image size N. The top, middle, and bottom rows show the times when using $M = 10^6, 10^7, 10^8$ visibilities. The standard deviation of the w sample density is determined by the value σ_w, which we chose values of $50, 100, 150$ wavelengths. We show results for before and after applying conjugation to the visibilities before construction, where we find improvements in performance for large w ranges due to an increase in w-stacking, as described in Section 4. The kernel construction time is independent of image size due to the use of adaptive quadrature, this is clear for large M in the middle and bottom rows. For $\sigma_w = 150$ wavelengths with $M = 10^8$, we found that not applying conjugation resulted in large kernel construction times of greater than 140 minutes (not shown). We found construction time reduces to 30 minutes after applying conjugation as shown in the figure.

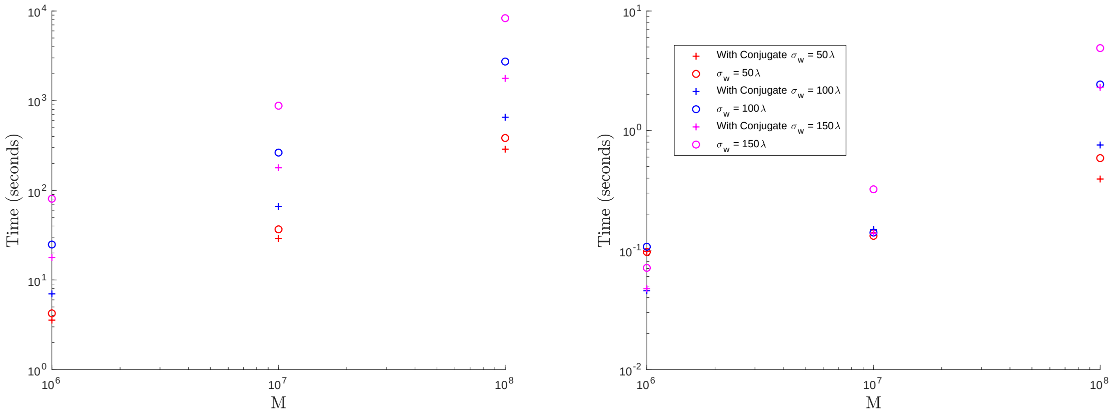

Figure 2. The left figure shows plots of measurement operator ($\mathbf{\Phi}$) construction times and the right figure shows plots of $\mathbf{\Phi}^\dagger\mathbf{\Phi}$ application times, as a function of the number of visibilities M for $N = 256^2$. The standard deviation of the w sample density is determined by the value σ_w, which we chose values of $50, 100, 150$ wavelengths. We find that the kernel construction time increases linearly with M. We find that the applying the conjugate consistently reduces the time required to calculate the w-projection kernels and can reduce the time for application.

observation of Fornax A taken with the MWA. The observation has a pointing centre of 03h 22m 41.7s -37d 12m 30s, and the integration time is 112 seconds. Fornax A was observed using XX and YY polarizations, with the visibilities transformed into Stokes I. The bandwidth was 30.72 MHz with a central frequency of 184.955 MHz and using 768 channels, which is a standard observational mode for the MWA (Prabu et al., 2015; Ord et al., 2015). The data reduction, including flagging and calibration, is as per McKinley et al. (2015).

To perform the reconstruction we use 50 nodes of the Grace computing cluster at University College London. Each node of Grace contains two 8 core Intel Xeon E5-2630v3 processors (16 cores total) and 64 Gigabytes of RAM.[4]

The reconstructed image is of 2048 by 2048 pixels, with a pixel width of 45 arc-seconds and a field of view of 25 by 25 degrees. The w values range between 0 and approximately 600 wavelengths for the total of 126.6 million visibilites, after conjugating the visibilities for negative w values, i.e. a range of 1200 wavelengths originally.

Sorting the visibilities into 50 w-stacks (one per MPI node) took a total time of under 5 seconds using the MPI distributed k-means algorithm described in Algorithm 1. If the average relative difference of each w-stack centre $\bar{\boldsymbol{w}}_i$ between k-means iterations is less than 10^{-3} we consider the algorithm has converged. We do not expect the w-projection algorithm performance to improve beyond this level of accuracy in clustering as a function of the number of iterations. In this case, the algorithm converged in 6 iterations.

It took a total of 15 minutes to construct a w-projection kernel for all visibilities, using quadrature accuracy of 10^{-6} in relative and absolute error, as described in Paper I. The w-projection kernel construction time in Paper I was 40 minutes for 50 w-stacks (over 25 compute nodes), with the same field of view and same image size, over the same range of w values, but for only 17.5 million visibilities. We find that the use of conjugate symmetry before the k-means clustering algorithm allows for more efficient computation of the w-projection kernels due to more efficient w-stacking because of the reduced range of w-values, allowing for 2.6 times faster kernel construction for approximately 7 times as many measurements (126.6 million visibilities), i.e. an overall saving of approximately 18 times.

Reconstruction time took 12 hours, with a total of 2475 iterations, with the FFT and wavelet operations contributing to much of this time due to the large image size. Note that we elected to run the reconstruction for a much longer time than needed to produce an acceptable image. We erred on the side of a higher number of iterations than strictly necessary in order to get a very high quality reconstruction.

The reconstructed image can be seen in Figure 3, which also shows the residual and dirty maps. The bright, extended source Fornax A is visible at the field centre, with the rest of the field consisting mostly of point sources. The residual map shows that the reconstruction models many of the sources in the field of view, however, the point spread function from bright sources outside the region imaged are still present in the residuals. De-

[4] More details can be found at `https://wiki.rc.ucl.ac.uk/wiki/RC_Systems#Grace_technical_specs`

spite outside sources disrupting the reconstruction, the root mean squared (RMS) value of the residual map is 15 mJy/beam, and the dynamic range of the reconstruction (as calculated in Pratley et al., 2018) is 844,000. The dynamic range is calculated by

$$\mathrm{DR} = \frac{\sqrt{N}\|\boldsymbol{\Phi}\|^2}{\|\boldsymbol{\Phi}^\dagger\left(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\right)\|_{\ell_2}} \max\{\boldsymbol{x}_k\}\,, \qquad (13)$$

i.e. the ratio of the peak of the recovered image to the root-mean-square (RMS) of the residuals for a normalized measurement operator. We note that the squared operator norm $\|\boldsymbol{\Phi}\|^2$ is the largest eigenvalue of $\boldsymbol{\Phi}^\dagger\boldsymbol{\Phi}$.

Figure 4 shows a zoom in of Figure 3, with the colour scale adjusted to show the reconstruction of Fornax A in greater detail. From the scaled residuals it is clear that this reconstruction accurately models the extended structure of Fornax A.

6 IMPROVEMENTS FOR THE FUTURE

We discuss two classes of possible improvements: kernel interpolations and correction for non-standard direction dependent effects.

6.1 Kernel interpolation

While we have shown that the use of k-means clustering and complex conjugation can aid in kernel construction, w-projection kernels can still be expensive in construction time due to the large number of coefficients in $\mathsf{\mathbf{GC}}$. This construction overhead can be further reduced using interpolation methods, such as bilinear interpolation between 1d w-planes, or parametric fitting. This may allow for on the fly calculation of kernels during imaging. We discuss how a radially symmetric kernel could affect such methods in the future.

6.1.1 w-planes: bilinear interpolation

The radially symmetric kernel allows fast and accurate calculation, while reducing the dimensions of the kernel. This allows for fast and accurate pre-sampling of the w-projection kernel directly in the uvw-domain, in some cases to a sufficient pre-sampling density that the error from linear interpolation is negligible compared to the aliasing error. While the mathematical basis for bilinear interpolation is discussed in detail in Paper I, here we present the implementation considerations.

First we make it clear that a non-radially symmetric kernel would mean pre-sampling in $(u_{\rm pix}, v_{\rm pix}, w)$, which is a computational challenge. For $N_u \times N_v$, samples in (u, v), we would have N_w w-projection planes. This requires in total $N_u N_v N_w$ samples. The total memory required in pre-samples is $16 \times 10^{-6} \times N_u N_v N_w$[Megabytes].

With radial symmetry, we show in Paper I that the w-projection kernel can be computed as a function of $(\sqrt{u_{\rm pix}^2 + v_{\rm pix}^2}, w)$. For N_{uv} radial samples in $\sqrt{u_{\rm pix}^2 + v_{\rm pix}^2}$, and N_w samples in w, we have only $N_{uv}N_w$ samples. This can be thought of as pre-computing 1d w-planes, rather than 2d w-planes. Additionally, each sample only requires a 1d integral by quadrature, reducing the pre-sampling time.

The 1d nature of the problem suggests better scaling of pre-sampling computation time and memory, allowing extremely accurate w-projection kernels. The total memory required in pre-samples is $16 \times 10^{-6} \times N_{uv}N_w$[Megabytes].

It is also worth noting that pre-sampling is only required for positive (u, v, w), since the complex conjugate can be used to estimate $(u, v, -w)$ and radial symmetry can be used for negative u and v. This leads to additional memory savings in pre-sampling.

Pre-sampling can be optimized for accuracy and storage by using an adaptive sampling density. The pre-samples could be stored permanently in cases where kernel construction is performed repetitively.

Bilinear interpolation is computationally cheap, and could make accurate on-the-fly construction of w-projection kernels possible, which could be needed for large data such as for the Square Kilometre Array (SKA) (Hollitt et al., 2017). In the case where storing the gridding kernels consumes more memory than the pre-sampled kernel, on-the-fly construction can be built into the $\mathsf{\mathbf{GC}}$ operator, where bilinear interpolation is used on application. However, memory layout of the pre-samples would be important, since the sample look-up time could reduce the speed of the calculation considerably.

6.1.2 Function fitting

Another powerful solution to improve kernel construction costs can be found from the well-known prolate spheroidal wave function (PSWF) gridding kernels, which do not have a closed form expression.

PSWFs can be defined multiple ways, such as having optimal localization of energy in both image and harmonic space, making them difficult to compute. They can be calculated directly through Sinc interpolation after solving a discrete eigenvalue problem, but this can be computationally expensive, or they can be calculated using a series expansion. However, this has not stopped radio astronomers using the PSWFs for decades, ever since the work of Schwab (1978, 1980) described a custom made PSWF that

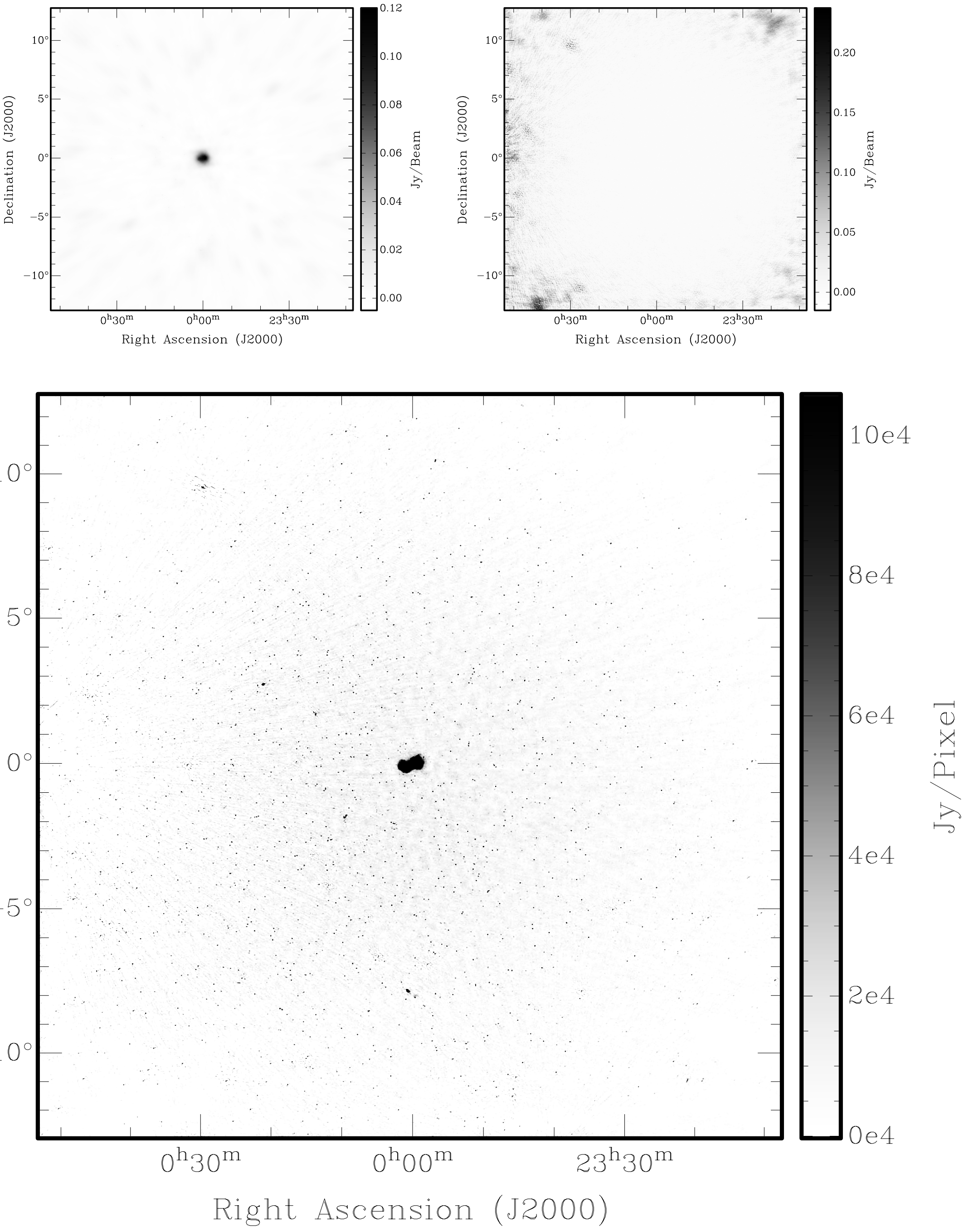

Figure 3. The dirty map (Top Left), residuals (Top Right), and sky model reconstruction (Bottom) of the 112 second MWA Fornax A observation centered at 184.955 MHz, using 126.6 million visibilities and an image size of 2048^2 (each pixel is 45 arcseconds and the field of view is approximately 25 by 25 degrees). This image was reconstructed using the MPI distributed w-stacking-w-projection hybrid algorithm, exploiting conjugate symmetry and the k-means clustering algorithm for distribution of w-stacks presented herein, and using the radial symmetric w-projection kernels, in conjunction with the ADMM algorithm. The dynamic range of the reconstruction is 844,000. The RMS of the residuals is approximately 15 mJy/beam over the entire field of view. The residuals are larger at the edges of the image due to side lobes of sources outside the field of view.

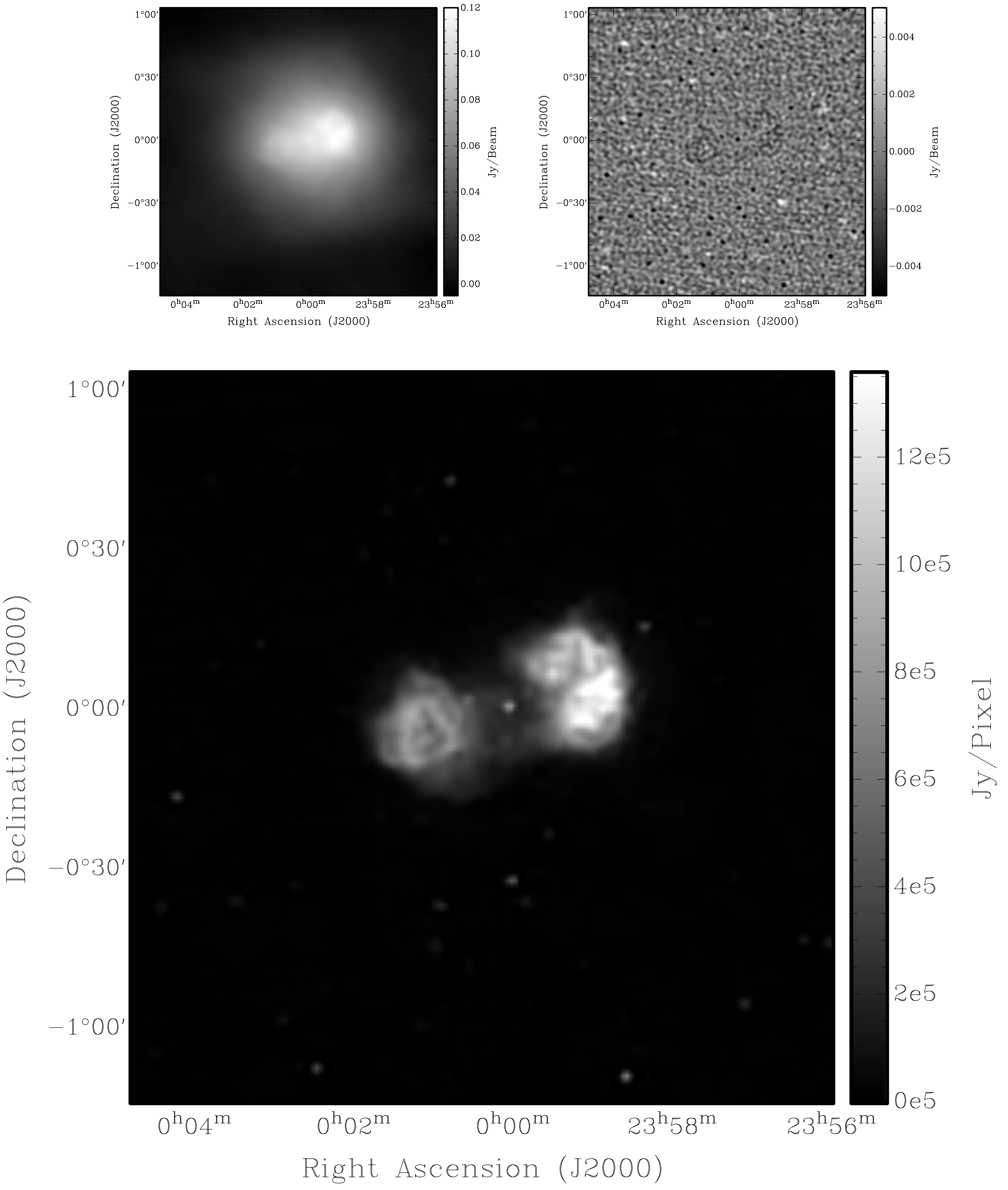

Figure 4. Same as Figure 3 zoomed view centered on Fornax A, showing the recovered structure of the double lobed radio galaxy. The residuals have been scaled to show the details. The residuals over the zoomed region have an RMS of 1.2 mJy/beam.

has been used in CASA (McMullin et al., 2007), AIPS (Greisen, 2003), MIRIAD (Sault et al., 1995), and PURIFY (Carrillo et al., 2014). In Schwab (1978, 1980), a rational approximation is used to provide a stable and accurate fit to the PSWF, which has stood the test of time.

A similar approach can be used to provide an accurate fit to w-projection kernels. Put simply, it is possible to fit a radially symmetric kernel as a function of three parameters $\left(\sqrt{u_{\rm pix}^2 + v_{\rm pix}^2}, w, \Delta u\right)$, i.e. polynomial fitting. This has various advantages over the pre-sampling method, such as reduced storage, no pre-sampling time, and reduced look up time (which could be critical for on-the-fly application). However, stability and reliability of the fit is not guaranteed and would require further investigation.

6.2 Additional direction dependent effects

The 1d radially symmetric kernel framework can be used in conjunction with general 2d kernels that model DDEs. It is clear that the 1d w-projection kernel derivation can be extended to other analytic radially symmetric baseline dependent effects, i.e. a function of r or $\sqrt{u^2 + v^2}$ only. But this does not stop the inclusion of more general baseline dependent effects, such as the spectral and polarimetric primary beams and time dependent ionospheric models. Generating these models will require computation that may or may not be worse than the non-coplanar baseline effects, which are telescope dependent. Non-coplanar baseline effects are a special case, where the effects need to be modeled on each baseline and can be modeled in stacks of visibilities. However, in many cases DDE models are station dependent, suggesting the computation is not as extreme as the non-coplanar case. Additionally, these effects may apply to groups of visibilities in time, frequency, and polarization, reducing the number of effects that need to be modeled.

In the worst case scenario, each baseline will have different DDEs, which can be included by further convolutions (since convolution is commutative)

$$\begin{aligned}&[GC](\sqrt{u_{\rm pix}^2 + v_{\rm pix}^2}, w) \to \\ &\quad D_{ij}(u, v, w) \star [GC](\sqrt{u_{\rm pix}^2 + v_{\rm pix}^2}, w)\,,\end{aligned} \tag{14}$$

where $D_{ij}(u, v, w)$ is a model of the DDEs in the uvw-domain between two stations ij. Typically if $D(u, v, w)$ is band limited, the additional convolution can be performed with a discrete convolution, since $[GC](\sqrt{u_{\rm pix}^2 + v_{\rm pix}^2}, w, \Delta u)$ is also smooth. The discrete convolution has computational complexity $\mathcal{O}(J_{GC}^2 J_D^2)$, where J is the width of each kernel. If D is separable in (u, v), then this can be reduced greatly to $\mathcal{O}(J_{GC}^2 J_D)$.

The computation of $D(u, v, w)$ may require modeling in the image domain with an FFT for each baseline or it may be known analytically in (u, v, w). In the case where $D_{ij}(u, v, w) = D_j(u, v, w) \star D_i^\star(u, v, w)$ is separable into station dependent effects, it greatly reduces the modeling computation from $N_{\rm Ant}(N_{\rm Ant} - 1)/2 \to N_{\rm Ant}$ kernel constructions.

The w-stacking distribution structure can be applied to model other effects, such as time dependent primary beam and ionospheric models. Distributing the visibilities into (time) t, (frequency) ν, and (polarization) p DDE-stacks could alleviate some of the challenges of $D \star GC$ construction; this applies whenever a DDE can naturally be applied to a group of baselines. For a given DDE-stack, we can apply the stack's DDE model directly in the image domain. This can be efficiently done using recent developments in the work of van der Tol et al. (2018).

7 CONCLUSION

We have discussed details of the w-stacking w-projection algorithm implementation, including details of the k-means clustering, introduction of conjugate symmetry to improve the computational efficiency of the current algorithm, and possible extensions to the current algorithms and code base to further improve efficiency and accuracy of the reconstructions.

We measured the time to pre-compute and apply an implementation of the MPI w-stacking w-projection algorithm. We found that the use of conjugate symmetry greatly improves the w-stacking efficiency, which reduces the cost in w-projection kernel construction and application. It is also clear that using adaptive quadrature allows kernel construction that is independent of image size, making it efficient for large high resolution images.

We use the MPI distributed ADMM implementation in PURIFY to reconstruct an MWA observation of Fornax A, recovering accurate sky models of the complex source Fornax A and of point sources over the entire 25 by 25 degree field of view. We find that we can construct w-projection kernels for 7 times the number of measurements, 2.6 times faster than the time taken in Paper I (an overall saving of approximately 18 times), using the same image size, field of view, and range of w values.

We conclude the work with proposals to modify the implementation of the 1d radial w projection kernels for large data sets, such as the use of kernel interpolation and the inclusion of non radially symmetric directional dependent effects. Accurate correction of wide-field and instrumental effects is critical in the era of next generation radio interferometers and are vital to achieving science goals ranging from the detection of the Epoch of Reionisation to accurately reconstructing cosmic magnetic fields.

ACKNOWLEDGEMENTS

We thank Dr Benjamin McKinley for providing the calibrated MWA data of Fornax A. This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC, grants EP/M011089/1). The authors acknowledge the use of the UCL Grace High Performance Computing Facility (Grace@UCL), and associated support services, in the completion of this work.

Facilities: MWA

REFERENCES

Ables J. G., 1974, A&AS, 15, 383

Aggarwal C., Reddy C., 2013, Data Clustering: Algorithms and Applications. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, Taylor & Francis

Carrillo R. E., McEwen J. D., Wiaux Y., 2014, MNRAS, 439, 3591

Cornwell T. J., 2008, IEEE Journal of Selected Topics in Signal Processing, 2, 793

Cornwell T. J., Evans K. F., 1985, A&A, 143, 77

Dabbech A., Wolz L., Pratley L., McEwen J. D., Wiaux Y., 2017, MNRAS, 471, 4300

Dabbech A., Onose A., Abdulaziz A., Perley R. A., Smirnov O. M., Wiaux Y., 2018, MNRAS, 476, 2853

Ellingson S. W., Clarke T. E., Cohen A., Craig J., Kassim N. E., Pihlstrom Y., Rickard L. J., Taylor G. B., 2009, IEEE Proceedings, 97, 1421

Fessler J. A., Sutton B. P., 2003, IEEE Transactions on Signal Processing, 51, 560

Greisen E. W., 2003, in Heck A., ed., Astrophysics and Space Science Library Vol. 285, Information Handling in Astronomy - Historical Vistas. p. 109, doi:10.1007/0-306-48080-8_7

Högbom J. A., 1974, A&AS, 15, 417

Hollitt C., Johnston-Hollitt M., Dehghan S., Frean M., Butler-Yeoman T., 2017, in Lorente N. P. F., Shortridge K., Wayth R., eds, Astronomical Society of the Pacific Conference Series Vol. 512, Astronomical Data Analysis Software and Systems XXV. p. 367

Humphreys B., Cornwell T. J., 2011, SKA Memo, 132

Johnston-Hollitt M., et al., 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), p. 92

Koopmans L., et al., 2015, Advancing Astrophysics with the Square Kilometre Array (AASKA14), p. 1

McEwen J. D., Scaife A. M. M., 2008, MNRAS, 389, 1163

McEwen J. D., Wiaux Y., 2011, MNRAS, 413, 1318

McKinley B., et al., 2015, MNRAS, 446, 3478

McMullin J. P., Waters B., Schiebel D., Young W., Golap K., 2007, in Shaw R. A., Hill F., Bell D. J., eds, Astronomical Society of the Pacific Conference Series Vol. 376, Astronomical Data Analysis Software and Systems XVI. p. 127

Offringa A. R., et al., 2014, MNRAS, 444, 606

Onose A., Carrillo R. E., Repetti A., McEwen J. D., Thiran J.-P., Pesquet J.-C., Wiaux Y., 2016, MNRAS, 462, 4314

Ord S. M., et al., 2015, PASA, 32, e006

Prabu T., et al., 2015, Experimental Astronomy, 39, 73

Pratley L., Johnston-Hollitt M., 2016, MNRAS, 462, 3483

Pratley L., McEwen J. D., 2019, arXiv e-prints, p. arXiv:1903.07621

Pratley L., McEwen J. D., d'Avezac M., Carrillo R. E., Onose A., Wiaux Y., 2018, MNRAS, 473, 1038

Pratley L., McEwen J. D., d'Avezac M., Cai X., Pérez-Suárez D., Christidi I., Guichard R., 2019a, Astronomy and Computing, submitted, arXiv:1903.04502

Pratley L., McEwen J. D., d'Avezac M., Carrillo R., Christidi I., Guichard R., Pérez-Suárez D., Wiaux Y., 2019b, PURIFY, doi:10.5281/zenodo.2587838, `https://doi.org/10.5281/zenodo.2587838`

Pratley L., McEwen J. D., d'Avezac M., Carrillo R., Christidi I., Guichard R., Pérez-Suárez D., Wiaux Y., 2019c, SOPT, doi:10.5281/zenodo.2584256, `https://doi.org/10.5281/zenodo.2584256`

Pratley L., Johnston-Hollitt M., McEwen J. D., 2019d, ApJ, 874, 174

Price D. C., Smirnov O. M., 2015, MNRAS, 449, 107

Sault R. J., Teuben P. J., Wright M. C. H., 1995, in Shaw R. A., Payne H. E., Hayes J. J. E., eds, Astronomical Society of the Pacific Conference Series Vol. 77, Astronomical Data Analysis Software and Systems IV. p. 433

Schwab F. R., 1978, VLA SCIENTIFIC MEMO-

RANDUM 129, Suppression of Aliasing by Convolutional Gridding Schemes. National Radio Astronomy Observatory, Charlottesville, Virginia

Schwab F. R., 1980, VLA SCIENTIFIC MEMORANDUM 132, Optimal Gridding. National Radio Astronomy Observatory, Charlottesville, Virginia

Smirnov O. M., 2011, A&A, 531, A159

Stoffel K., Belkoniene A., 1999, in Amestoy P., Berger P., Daydé M., Ruiz D., Duff I., Frayssé V., Giraud L., eds, Euro-Par'99 Parallel Processing. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1451–1454

Thompson A. R., Moran J., Swenson G., 2008, Interferometry and Synthesis in Radio Astronomy. Wiley

Tingay S. J., et al., 2013, PASA, 30, 7

Wiaux Y., Puy G., Boursier Y., Vandergheynst P., 2009, MNRAS, 400, 1029

Wolz L., McEwen J. D., Abdalla F. B., Carrillo R. E., Wiaux Y., 2013, MNRAS, 436, 1993

Zernike F., 1938, Physica, 5, 785

van Haarlem M. P., et al., 2013, A&A, 556, A2

van der Tol S., Veenboer B., Offringa A. R., 2018, A&A, 616, A27