CAMBRIDGE
UNIVERSITY PRESS

Research Article

Verifying the Australian MWA EoR pipeline II: Fundamental limits of the `AusEoRPipe` and the impact of instrumental effects

Jack Laurence Bramble Line[1,2] , Cathryn Trott[1,2] , Nichole Barry[1,2], Dev Null[1,2] and Christopher H. Jordan[1,2]

[1]International Centre for Radio Astronomy Research, Curtin University, Perth, WA, Australia and [2]ARC Centre of Excellence for All Sky Astrophysics in 3 Dimensions (ASTRO-3D), Perth, Australia

Abstract

Detection of the weak cosmological signal from high-redshift hydrogen demands careful data analysis and an understanding of the full instrument signal chain. Here, we use the `WODEN` simulation pipeline to produce realistic data from the Murchison Widefield Array (MWA) Epoch of Reionisation experiment and test the effects of different instrumental systematics through the `AusEoRPipe` analysis pipeline. The simulations include a realistic full sky model, direction-independent calibration, and both random and systematic instrumental effects. Results are compared to matched real observations. We find that, (i) with a sky-based calibration and power spectrum approach we have need to subtract more than 90% of all unresolved point source flux (10 mJy apparent) to recover 21-cm signal in the absence of instrumental effects; (ii) when including diffuse emission in simulations, some k-modes cannot be accessed, leading to a need for some diffuse emission removal; (iii) the single greatest cause of leakage is an incomplete sky model; and (iv) other sources of errors, such as cable reflections, flagged channels, and gain errors, impart comparable systematic power to one another and less power than the incomplete sky model.

Keywords: Astronomy data analysis; Reionisation; GPU computing

1. Introduction

Exploration of the first billion years of the history of the Universe promises to shed light on the growth and evolution of the first generations of stars and galaxies and the transformation of the cosmos from a neutral to a predominantly ionised intergalactic medium (IGM). One of the primary observational avenues for this period is the hyperfine transition from primordial neutral hydrogen gas that fills the IGM, which can be observed at low radio frequencies (50–200 MHz) with radio telescopes. The hydrogen brightness temperature encodes the radiation and thermal properties of the IGM over time and space; at early times, fluctuations are dominated by heating of the gas from the first stars and galaxies, while at later times, the lack of signal from ionised regions of hydrogen gas dominate the spatial fluctuations (Furlanetto, Peng Oh, & Briggs 2006).

The cosmological 21-cm signal is obscured by considerably brighter foreground emission from active galactic nuclei (AGN) and star forming galaxies, as well as our own Galaxy. Crucially, these are synchrotron and free-free emitters, with continuum spectra, allowing a spectral distinction to be made between them and the spectrally structured 21-cm line emission. This fundamental difference forms the basis for discriminating foreground contaminating power from the signal of interest for all experiments attempting this measurement (Koopmans et al. 2015; HERA Collaboration et al. 2023; Beardsley et al. 2016; Trott et al.

2020; Barry et al. 2019; Patil et al. 2017; Mertens et al. 2020). The brightest and closest of these foreground sources can be resolved and measured individually, and these form the basis for a sky-based approach to calibrating low-frequency radio data. The calibration model and our understanding of the low-frequency radio sky are critical for obtaining clean and accurate data. In addition to careful calibration, there are other systematic effects that prevent a complete and pristine dataset. These include spectral channels flagged due to radio frequency interference (Wilensky et al. 2023), time steps flagged due to radio frequency interference (RFI) or instrument issues, ionosphere refraction of the signal, and an incomplete sky model, whereby some of the sky flux is missing from the calibration model (Barry et al. 2019). These act to make both calibration data and science data inaccurate. Understanding the impact of these effects is crucial for (i) obtaining the cleanest set of data, (ii) prioritising effort to address particular systematics, and (iii) confidence in the robustness of the reported 21-cm signal power. In Line et al. (2024), the end-to-end data processing pipeline was tested to ensure there is no signal loss from our techniques. In this paper, realistic simulations are used to test the impact of different random and systematic errors and provide guidance on their importance for the next generation of analysis tools.

In Paper I of this series (Line et al. 2024), the `AusEoRPipe` was tested against signal loss with a model 21-cm signal, ensuring that the methodology does not bias the signal power. In this paper, we focus attention on the impact of calibration, and instrumental systematics on the ability to detect the 21-cm signal. We focus on the MWA EoR high-band frequency range (167–198 MHz), as this is covered by the 21-cm sky model (detailed in Section 3.1). We focus on the MWA phase I layout. Due to computational

constraints, we limit ourselves to the EoR0 field (centred at RA,Dec = $0^h, -30°$). We only consider Stokes I sky models as this not only reduces computational costs, but the calibration catalogue used by the `AusEoRPipe` is already Stokes I only. We further constrain ourselves to zenith beam-pointings; in doing so, we have the computational resources to simulate multiple observations closely spaced in LST. This allows us to experiment with averaging calibration solutions, as the MWA instrument should be somewhat stable over the space of half an hour (see Section 6 for averaging results). Even with these constraints, we are able to perform simulations that test the fundamental limits of 21-cm recovery with the `AusEoRPipe` in the presence of foregrounds, as well as the impact of instrumental effects on the `AusEoRPipe` (Section 5).

The paper is structured as follows. In Section 2, we overview the `AusEoRPipe` and strategy for simulating visibilities. In Section 3, we detail the point, diffuse, and 21-cm sky models used for simulating visibilities. In Section 4, we investigate the effects of calibration and subtraction on simulated data containing no instrumental effects; we add instrumental effects in Section 5. In Section 6, we explore averaging calibration solutions over time. Finally, we discuss and conclude our work in Sections 7 and 8.

2. Pipeline and simulation strategy

2.1 `AusEoRPipe` overview

For a full overview of the `AusEoRPipe`, including various software packages, see Line et al. (2024). Here we include the most pertinent details.

MWA observations are undertaken in 2-min snapshots, sampled at 2 s and 40 kHz resolution. Direction-independent calibration is applied at native resolution, before the data are averaged to 8 s and 80 kHz for further analysis. During the 2-min snapshot, the phase centre is fixed in celestial coordinates, and the beam pointing centre is fixed, meaning the sky drifts a small amount through the primary beam. The calibration is performed on a per-channel basis. This choice for the vanilla version of the `hyperdrive` software was made in response to the spectral polynomial fitting that was performed by the earlier RTS software (Mitchell et al. 2008), which imparted systematic power into the power spectrum (PS), and could not handle cable reflections. In this work, we maintain the per-channel calibration strategy and present the results within that framework.

Direction-independent calibration uses a sky model constructed from unresolved (point) and extended sources that are above the horizon at the observation time. The sky model is constructed from a combination of GLEAM (Wayth et al. 2015; Hurley-Walker et al. 2017) and LoBES (Lynch et al. 2021) for point and multi-component sources, and additional custom models for A-Team radio galaxies and supernova remnants (Cook, Trott, & Line 2022; Line et al. 2020). Within the EoR fields, this catalogue is 90% complete to 32 mJy (Lynch et al. 2021). Notably, the calibration sky model does not include diffuse emission. This omission is handled by only calibrating with baselines longer than 30λ, where the diffuse power is sub-dominant. Nonetheless, this is a shortcoming of the calibration sky model. The same sky model can then be used for foreground subtraction, whereby model visibilities formed from the catalogue are directly subtracted from the calibrated visibilities. This is a good path for removing foreground power, but assumes that there are no direction-dependent (DD) effects such that the model deviates from the measurements. DD calibration, such as peeling (Mitchell et al. 2008), is not implemented for this work.

2.2 Simulation strategy

We make one further concession to computational and development time and only focus on instrumental effects that are independent of direction upon the sky. These kinds of effects, such as visibility flagging and internal cable reflections, can be added to visibilities post simulation. In this way, a base set of simulations can be used to test a number of instrumental effects in isolation, for little extra compute. Sky directional effects, such as the ionosphere and directional RFI, must be simulated during the calculations of the visibilities themselves. This is outside the scope of this paper and left for future work.

2.3 Data selection

In all `AusEoRPipe` testing, we aim to reproduce as many real effects as possible. As such, we simulate with observational settings matching real MWA data for comparison. We select a set of 15 EoR0 zenith pointings from 2015, spanning observation GPS times 1125938488 to 1125940192 (UTC 2015-09-10 16:41:11 - UTC 2015-09-10 17:09:35). Note that GPS start time is used for the observation identifier within the MWA. These observations are known to produce good calibration solutions and have little ionospheric activity (Jordan et al. 2017), allowing for a reasonable comparison to simulation. We only simulate a single pointing (zenith) to reduce computational load, particularly with the diffuse and 21-cm emission. In Trott et al. (2020), the behaviour of individual versus combined pointings in the PS did not show significant differences, justifying use of a single pointing for this work. We simulate all data at a 2 s, 40 kHz to match the resolution of real data, ensuring any effects of averaging to 8 s, 80 kHz when applying calibration solutions or creating power spectra are captured. All power spectra shown come from the north-south aligned dipoles, because these are shown to consistently produce cleaner power spectra in previous MWA publications with this pipeline (Trott et al. 2016). A weighted FFT is used throughout for spectral analysis.

3. Sky models

Broadly there are three regimes of sky signal we are interested in. The first is the 21-cm signal itself. The second we describe as discrete (compact) sources. We define any singular astrophysical object as a discrete source and so include extended A-team sources here. These are sources that contribute sky power on smaller angular scales and can be used as calibrator sources. The third we call the diffuse, which broadly covers all large angular-scale emission, predominantly coming from synchrotron emission from the Milky Way. Both the discrete and diffuse affect calibration and signal recovery in different ways, so including both is paramount to pipeline testing. We can take advantage of the additive nature of visibilities, allowing us to simulate a base set of 21-cm visibilities, and add on the two foreground models in stages to test these differences in isolation.

The three sky models used in the work are detailed in the following subsections. All are visualised as seen from the MWA when the EoR0 field is transiting the meridian in Fig. 1.
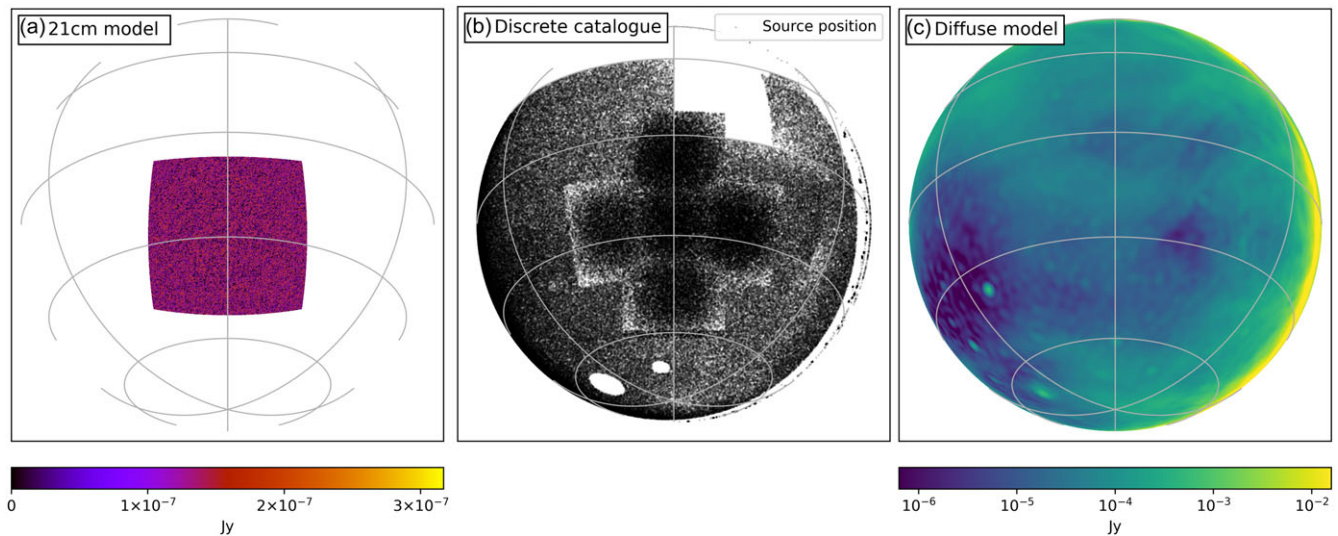
Figure 1. All-sky orthographic projections of the sky models, centred at RA, Dec $= 0°, -27°$ (-27° is zenith for the MWA), where: (a) shows a slice of the 21-cm sky model at 167 MHz; (b) shows the positions of all sources in the discrete sky model; (c) shows the diffuse sky model at 200 MHz

3.1 21-cm model

We use the 21-cm model described in Paper I (Line et al. 2024), which was designed to match EoR0 high band. It covers a $\sim 50 \times 50$ squared degree field, at a spectral resolution of 80kHz, and angular resolution of $\sim$27 arcsec. The model is a TAN FITS projection with each pixel represented as a point source in the sky model, with the $\sim$27 arcsec resolution over-sampling the $\sim$2 arcmin MWA resolution. The visibilities are analytically predicted from a point source via assuming a point source is a dirac-delta function and applying that to the measurement equation. For the exact mechanics of the WODEN simulator, please see Line (2022). For details of how the 21-cm model was projected and interpolated onto the sky, along with the ability of the AusEoRPipe to recover the expected PS, please see Line et al. (2024).

3.2 Discrete foreground model

For this model we use the current calibration catalogue as used by the AusEoRPipe. This hybrid catalogue uses GLEAM as a base (Hurley-Walker et al. 2017), with LoBES (Lynch et al. 2021) covering the EoR0 field, and the work of Procopio et al. (2017) covering the EoR1 field. LoBES is used to supplant the EoR0 field as it was created with the MWA Phase II array layout, offering a resolution of $\sim$80 arcsec, compared to the $\sim$2 arcmin resolution of GLEAM, which was created with the MWA phase I layout. As a result, LoBES is 90% complete at 32 mJy whereas GLEAM is 90% complete at 170 mJy. This difference is visualised in the denser central fields in Fig. 1b.

A number of extended sources are modelled including Fornax A (Line 2022), and Galactic supernova remnants (Cook et al. 2022). All source spectral energy distributions are modelled by either a power-law or curved power-law model, making all discrete sources spectrally smooth. Most sources are represented as point sources in the sky model, with a number Gaussian components used for sources with more structure on the sky. The most complicated sources are represented by Shapelets, including Fornax A. Again, all visibilities are analytically calculated from the point source, Gaussian, or Shapelet parameters. The visibility calculations for a Gaussian and Shapelet component are again detailed in (Line 2022). Fig. 1b shows the sky coverage of this model; the missing area in the top right is due to the underlying GLEAM catalogue coverage. In total, the discrete sky model has 338 797 sources.

3.3 Diffuse foreground model

The diffuse sky model is based on an m-mode analysis map as described in Kriele et al. (2022), made using the Engineering Development Array 2 (EDA2 Wayth et al. 2022). The EDA2 is a low angular resolution telescope based at the same site as the MWA, and using the same MWA dipoles, making the sky coverage perfectly matched for our purposes. The original Stokes I map was imaged at 159 MHz into a HEALPix (Górski et al. 2005) $N_{side} = 64$ projection (a HEALPixel resolution of $\sim$0.9°). Kriele et al. (2022) also produced an accompanying spectral index map. To match the resolution of the MWA, we take the original Stokes I and spectral index maps and upgrade to $N_{side} = 2\,048$ ($\sim$1.7 arcmin) using the healpy (Zonca et al. 2019) ud_grade function. We then smooth the map using the healpy.smoothing function with a 0.9° kernel, to remove the edges of the original HEALPixels. It should be noted that this diffuse map includes all sky emission, meaning all discrete sources as described in 3.2 are present. The consequence being any simulation containing both sky models will result in some double counting of flux. However, for the purposes of this paper, both models combined need not perfectly match the sky; as long as they produce comparable power to that seen in real data, they can be used in pipeline validation.

4. Pristine simulation

We first test the AusEoRPipe in the presence of foregrounds, without systematic instrumental errors. We simulate a single EoR0 zenith observation 1125939344, which has an LST$= 359.9°$, setting the EoR0 field close to directly at zenith. We simulate all three sky models and check the ability of the AusEoRPipe in recovering
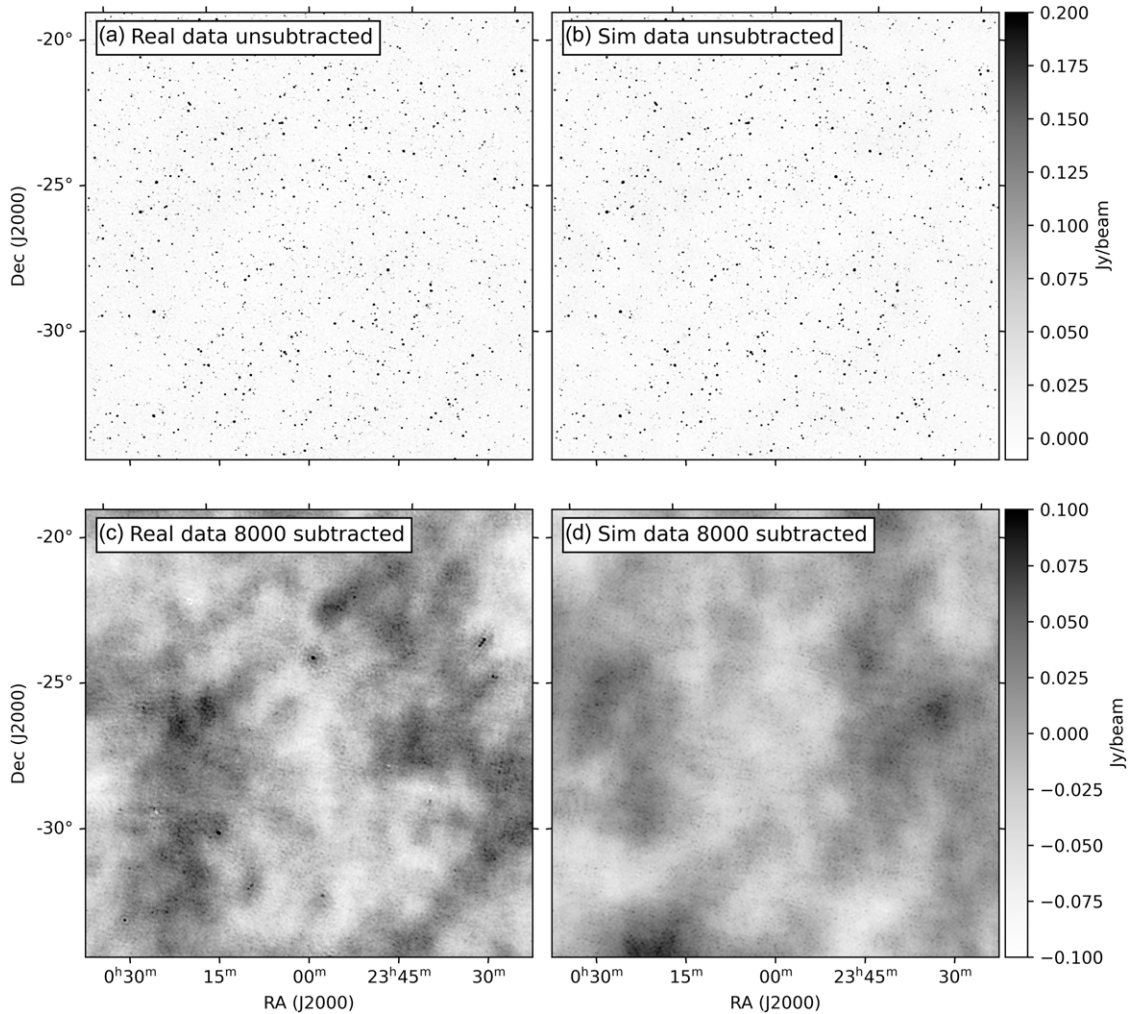
Figure 2. Comparison of a real zenith pointed 2 minsnapshot to simulated data. Both real and simulated data were averaged to 8 s, 80 kHz, and the full bandwidth imaged via `WSClean`. The top row were both imaged with Briggs 0 weighting, where (a) shows the real data after calibration and (b) shows the simulated data with no calibration. The bottom row were both imaged with natural weighting, where (c) shows the real data after subtracting 8 000 sources and (d) shows the simulated data with the same 8 000 sources subtracted.

the 21-cm with and without calibration involved. After combining all three sky models, we image the visibilities using `WSClean` (Offringa & Smirnov 2017; Offringa, Mertens, & Koopmans 2019) and compare to real data in Fig. 2. To compare the diffuse model, we subtract the same 8 000 discrete sources from the real and simulated data. This should leave the similar amounts of discrete power in both the real and simulated data, down to the completion level of the LoBES catalogue. This qualitative comparison shows excellent agreement with the discrete model to real data. The diffuse model does not align perfectly, but shows similar power on the sky, as well as angular distribution of that power.

4.1 Direct subtraction

In this subSection we test how much discrete sky model flux must be subtracted before we can recover the 21-cm signal. We start by creating a two minute simulation containing the full discrete and 21-cm sky models as a visibility test bed. With the EoR0 field centre

at zenith, there are 227 585 discrete sources above the horizon, all of which are present in the test bed visibilities. We then generate three sub sky models to subtract from the test bed, by cutting the sky model at different flux thresholds (10^{-1}, 10^{-2}, 10^{-3} Jy). These flux thresholds are in reference to the apparent flux, where we have weighted the discrete sky model by the primary beam to calculate the apparent flux of all sources at the central frequency of 182 MHz. See Table 1 for the resultant numbers after these cuts were applied.

We simulate these three sub sky models through `WODEN` and then subtract them from the test bed directly in visibility space. Resultant 1D PS are shown in Fig. 3. Before averaging from a 2D to a 1D PS, the foreground wedge is avoided by excluding any modes where $k_\parallel < 0.08\,h\mathrm{Mpc}^{-1}$, $k_\perp > 0.06\,h\mathrm{Mpc}^{-1}$, and performing a horizon line cut (discarding any modes in the wedge as marked by the solid black line as shown in Fig. 10. Fig. 3 shows that the apparent flux cut at 10^{-3} Jy is necessary to reliably recover the 21-cm signal.

Note this is a flux cut on this particular discrete sky model; this is not saying that leaving all sources below 10^{-3} Jy in real data will

Table 1. Cuts placed on discrete sky model and resultant number of sources and their contribution to the apparent flux. We include a column detailing the total number of sources above the horizon without a flux cut for reference.

Apparent flux cut (Jy)	Num. sources	Fraction of total apparent flux (%)	Fraction of total num. sources (%)
10^{-1}	3 027	55.3	1.3
10^{-2}	26 359	88.6	11.6
10^{-3}	79 766	98.5	35.0
No cut	227 585	100.0	100.0
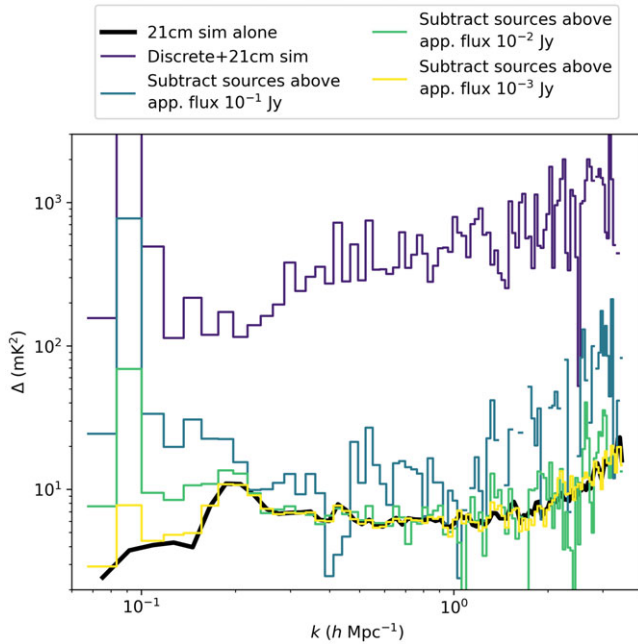
Figure 3. 1D PS depicting data from a single zenith EoR0 observation where only the discrete foregrounds were included are shown. All PS were obtained after performing a wedge cut.

Figure 4. 1D PS depicting data from a single zenith EoR0 observation where both discrete and diffuse foregrounds were included.

allow a detection. This result is limited by the completeness of this discrete sky model, and these simulations do not include confusion noise. It should also be noted that as we cut by apparent flux, at the lower the flux cut thresholds, we are adding more sources outside the main lobe of the MWA primary beam. These sources that are further from field centre contribute to higher $k_\parallel$ values due to projection effects. So although they contribute less overall power than sources in the main primary beam lobe, they contribute power closer to the so-called 2D PS 'window', an area of the PS expected to have the least contamination from foreground sources.

We repeated this experiment after adding the diffuse sky model to the test bed. We call this combination of 21-cm, discrete, and diffuse sky models as the full sky model. We find that subtracting the three sub discrete sky models makes little difference to the results and instead subtract the entire discrete sky model. Results are shown in Fig. 4. As mentioned in Section 3.3, the diffuse sky model contains the entire visible radio sky and so may contain duplicate flux at higher k-modes (small angular scales). Residual power left at high k-modes after subtraction therefore may be false power. However, the large residual power seen at lower k-modes comes exclusively from the diffuse model. These results
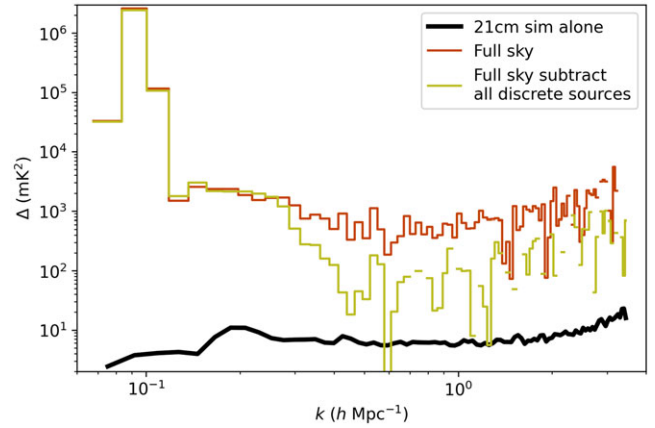
show for this single observation, the 21-cm signal is unrecoverable at low k-modes without some removal of power from the diffuse foregrounds.

4.2 Calibration and subtraction

We repeat the subtraction experiment in Section 4.1 but now include calibration. We use `hyperdrive di-calibrate` to calibrate the full sky model visibilities, using a calibration catalogue of the apparent brightest 8 000 sources in the discrete sky model. We then use `hyperdrive solutions-apply` to apply those calibration solutions to the full sky model visibilities. We then subtract the uncalibrated discrete sky model visibilities from the calibrated visibilities.[a] To summarise, we calibrate using an incomplete discrete sky model and then subtract the entire discrete sky model from the visibilities, thereby including a systematic error in the calibration model. We also repeat this procedure using a combination of just the discrete and 21-cm sky models. If we had perfect calibration, we would expect this to result in a perfect recovery of the 21-cm signal. Results are shown in Fig. 5. These results show that calibration induces up to three orders of magnitude of power into the window, when using a single observation. This comes from percent-level amplitude fluctuations in the calibration solutions, that vary quickly as a function of frequency. This couples power at low k-modes from the spectrally smooth foregrounds, into higher k-modes. This leakage completely masks the 21-cm signal. We note here that applying the calibration solutions to the 21-cm simulation alone does not bias the recovery of the signal. Only a small fraction of power in the sky is perturbed, and so the 21-cm alone is greater than the leakage caused by calibration solutions from itself.

5. Instrumental effects

In this Section we detail the direction independent random and systematic errors and how they are simulated. We use a test bed of 30 min of EoR zenith simulations (15 contiguous 2 min snapshots), containing both the diffuse and discrete sky models. We apply each error in isolation to this

---

[a]This last step is equivalent to using `hyperdrive vis-subtract` using the entire calibration catalogue. We re-use existing `WODEN` simulations to save on compute.
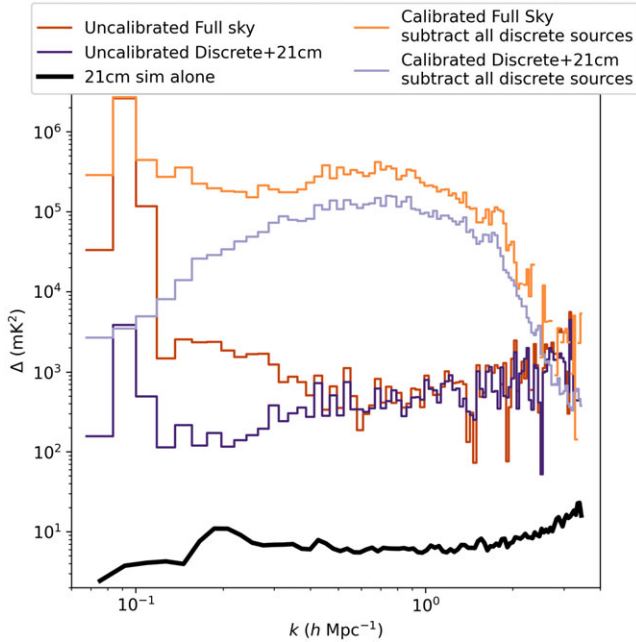
Figure 5. 1D PS depicting data from a single zenith EoR0 observation, showing the effects of calibration on leakage into the window. Note that the dark purple and brown lines showing uncalibrated simulations are colour-coded to lines showing the same simulations in Figs. 3 and 4, for easy comparison.

test bed and investigate the ability of `hyperdrive` to calibrate them in Section 5.5. All instrumental effects were added using `add_instrumental_effects_woden.py`[b]

5.1 Edge and centre channel flagging

The legacy MWA correlator used a polyphase filterbank (Tingay et al. 2013) to frequency channelise data. The entire bandwidth was split across 24 'coarse bands', each of 1.28 MHz bandwidth, with a frequency-dependent bandpass imparted by the filterbank. These coarse bands suffered from low SNR at the edges, and aliasing causes the central channel to also degrade. For this reason, calibration is more effective when flagging these channels. Given we are testing 40 kHz resolution data, the first systematic instrumental effect is to simply flag the first two, central, and final two channels in each coarse band.[c] These regular spectral flags cause harmonic modes, linking foreground power from low to high $k_\parallel$. The current version of the CHIPS PS processing software can optionally include a non-uniform FFT calculator to handle these missing channels. Besides flagging for bandpass effects, real data are also flagged for malfunctioning receiving elements and RFI. We leave investigating these flagging effects for future work.

5.2 Tile based gain error

Each receiving element (often called a 'tile' for MWA,) in a dual polarisation interferometer can have a gain (g_x, g_y) and leakage (D_x, D_y) term for each polarisation. Each visibility is a correlation

---

[c]This is easily achieved with `hyperdrive` by passing the optional argument `-fine-chan-flags-per-coarse-chan 0 1 16 30 31`.

---

of two tiles. `WODEN` implements any tile gains and leakages by multiplying the visibility between tiles 1 and 2 through the following operation:

$$
\begin{bmatrix} V'_{12\,XX} & V'_{12\,XY} \\ V'_{12\,YX} & V'_{12\,YY} \end{bmatrix} =
$$

$$
\begin{bmatrix} g_{x1} & g_{x1}D_{x1} \\ g_{y1}D_{y1} & g_{y1} \end{bmatrix} \begin{bmatrix} V_{12\,XX} & V_{12\,XY} \\ V_{12\,YX} & V_{12\,YY} \end{bmatrix} \begin{bmatrix} g^*_{x2} & g_{x2}D^*_{x2} \\ g_{y2}D^*_{y2} & g^*_{y2} \end{bmatrix}^T \quad (1)
$$

where V is a visibility, $*$ means the complex conjugate, T is the transpose, and V' is the visibility after the tile gains and leakages have been applied. The leakage terms are calculated via Equation A4.5 from Thompson, Moran, & Swenson (2017):

$$
D_x = \Psi - i\chi \quad (2)
$$

$$
D_y = -\Psi + i\chi, \quad (3)
$$

where Ψ, χ are alignment errors of the dipoles. This equation is really designed for single antennas, but in the MWA case, one could assume all dipoles in a tile are aligned perfectly to the mesh, and the mesh is slightly offset. This would mean the alignment errors for all dipoles are the same.

In this work, we apply a different random gain error to each tile and polarisation. We draw the amplitudes of g_x, g_y from a uniform distribution between 0.7 and 1.3, $U(0.7, 1.3)$. We add a phase slope to g_x, g_y as a function of frequency with a maximum phase offset drawn from $U(-60°, +60°)$. We draw Ψ from $U(0, 0.02°)$ and χ from $U(0, 0.05°)$. The random gain errors are kept constant across all 15 observations and are therefore a systematic error. The manifestation of the gain errors can be seen in Fig. 11.

5.3 Cable reflections

Mismatched impedance at coaxial cable ends can cause internal reflections that setup standing waves, adding frequency-dependent ripples to the visibilities. We follow the formalism from Beardsley et al. (2016) to define the cable reflection gain seen by tile i for polarisation pol as

$$
R_{\text{pol},i}(\nu) = R_{0,i} \exp\left(-2\pi i\nu\tau_i\right), \quad (4)
$$

where $R_{0,i}$ is the complex reflection coefficient, and τ_i is the time delay caused by the cable length connected to tile i. The time delay is given by

$$
\tau_i = \frac{2l_i}{0.81c}, \quad (5)
$$

where l_i is the length of the cable connected to tile i, and c is the speed of light. The factor 0.81 comes from the velocity factor of the cable, which we again take from Beardsley et al. (2016).

The amplitudes of these cable reflections have been measured to average between 0.02 and 0.1 using Fast Holographic Deconvolution - FHD (Barry, private communication). We therefore draw the amplitude of $R_{0,i}$ from $U(0.02, 0.1)$ and add a random phase offset of $U(0.0, 180°)$ for each tile.

5.4 Noise

Thermal noise on cross-correlations from an interferometer are estimated from both internal receiver temperature (T_{rec}) and the
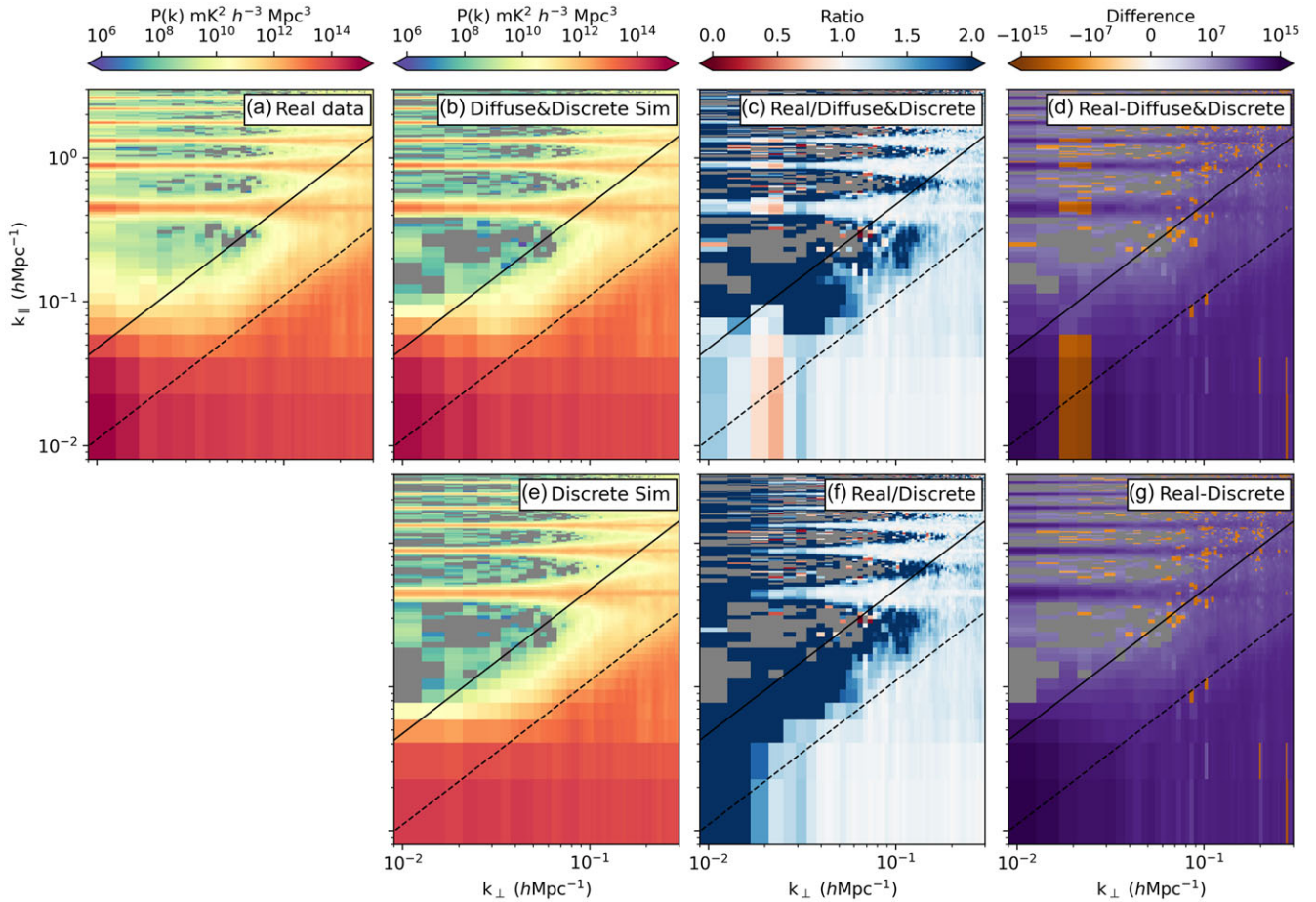
Figure 6. Comparison of an integration over 15 real two-minute zenith observations to simulated data. All real and simulated data have been calibrated using 10 000 sources, with the real data calibrated using 8 000 sources. In the ratios, blue means more power in the real data, red means less power. In the differences, purple means more power in the real data, and orange less. Negative pixels in *a), b), e)* are shown in grey. These pixels have also been masked in the ratio and differences.

sky temperature (T_{sky}), via Equation 6.50 in Thompson et al. (2017):

$$\sigma_{cross} = \frac{\sqrt{2} k_b (T_{sky} + T_{rec})}{A_{eff} \sqrt{\Delta \nu \Delta t}}, \qquad (6)$$

where A_{eff} is the effective area of the tile, $\Delta \nu$ is the channel width, and Δt is the integration time. σ_{cross} describes the standard deviation of a zero mean Gaussian noise distribution. Throughout this paper, we set $T_{sky} = 228$ K, $T_{rec} = 50$ K, $A_{eff} = 20.35$ m^2. A different realisation of the noise is added to each simulated observation, inducing a random error.

5.5 Applying instrumental effects

We take each effect detailed in Section 5 and apply them to the 15 zenith observations in isolation, and all in combination. We calibrate all simulations using the apparent 10 000 brightest sources in the sky for the given LST of that simulation. We integrate all 15 observations containing all instrumental effects into 2D PS and compare to real data in Fig. 6. We compare simulations containing both the discrete and diffuse sky models, as well as only the discrete sky model. Qualitatively it can be seen from the top row that the added instrumental errors match the real data, given

the matching power in coarse band harmonics, and leakage at $k_{\parallel} \sim 0.8$. However, there is clearly less leakage into the window from the simulations, indicating further unmodelled systematics. As expected, Fig. 6 shows missing power at the large spatial scales and along the horizon when only simulating the discrete sky model. hyperdrive calibrates cable reflections well due to its per-channel calibration strategy.

The manifestations of each instrumental effect in isolation are shown in 1D PS in Fig. 7. In general, all instrumental effects add some systematic power, but the application of calibration exacerbates this. Source subtraction removes overall power in foreground-dominated modes, but does not help to correct instrumental errors. In most cases, random errors add overall power, whereas systematic errors (e.g. flags and cable reflections) impart structured power.

5.6 Incomplete sky model

In Section 4.1, completeness of the sky model used for subtraction revealed that discrete sources above an apparent flux density of 10^{-3} Jy needed to be removed to detect the 21-cm signal. In addition, Section 4.2, which used a set calibration sky model, showed that calibration alone can introduce significant power. Here we
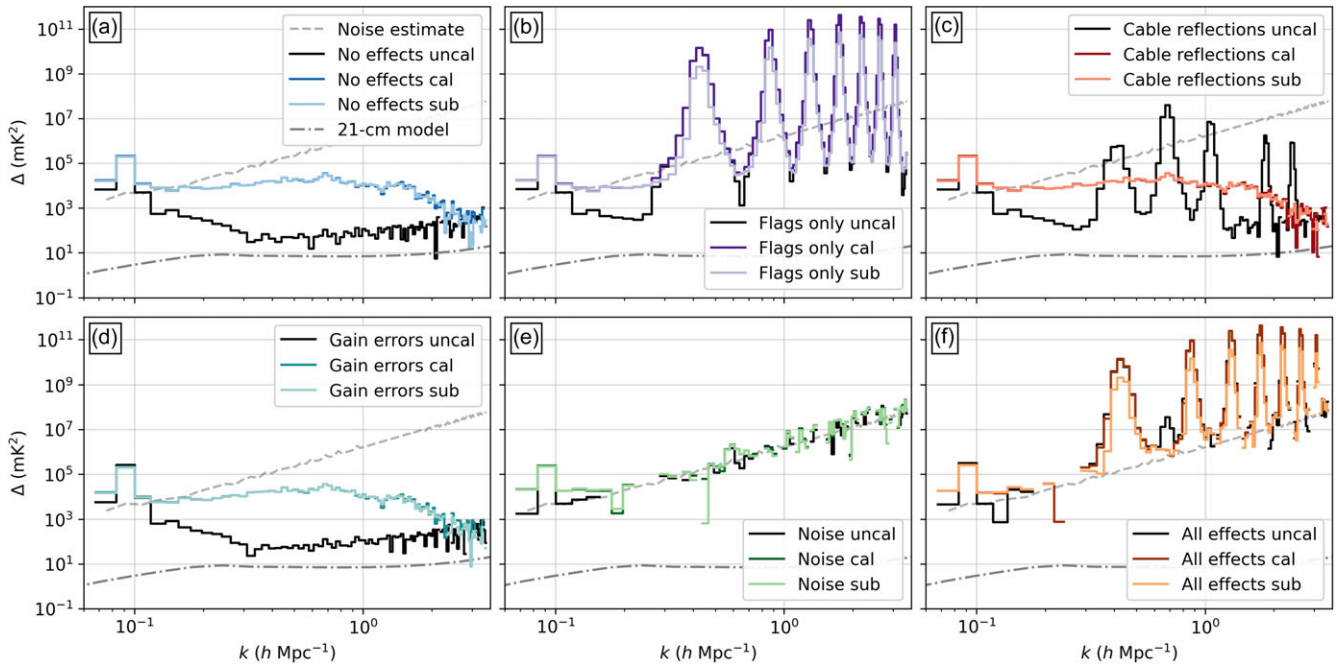
Figure 7. Comparison of different instrumental effects and their manifestation in the window. These 1D PS were made from 15 zenith observations, the 2D PS of which are shown in Fig. 6. Wedge cuts have been applied before averaging into 1D. In general, all instrumental effects add some systematic power, but the application of calibration exacerbates this. Source subtraction removes overall power in foreground-dominated modes, but does not help to correct instrumental errors.
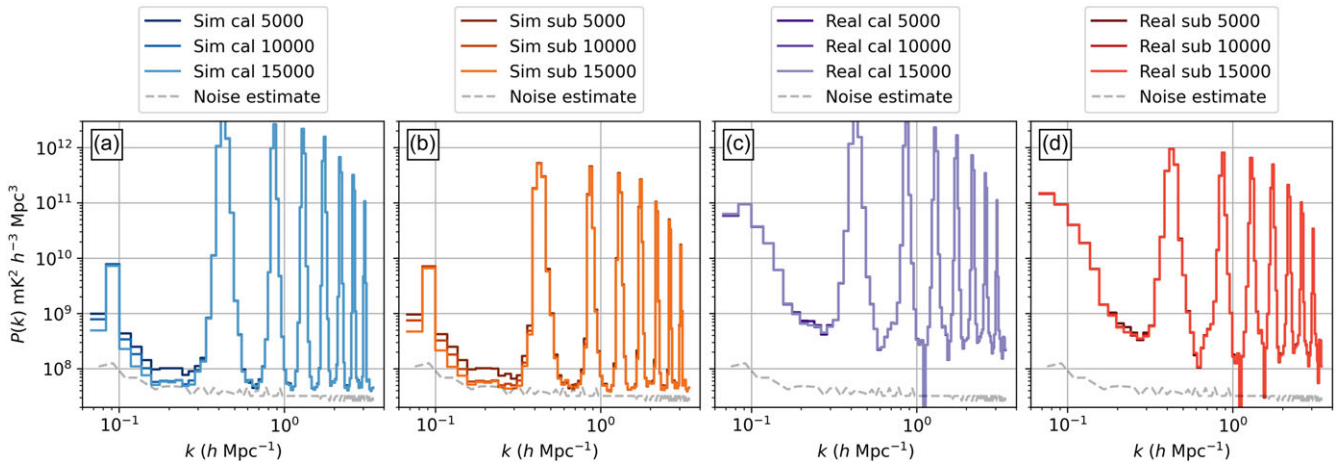
Figure 8. Comparison of calibrating and subtracting with various numbers of sources, with simulated data on left, real data on the right, when instrumental errors are included. Power spectra are made from integrating 15 EoR0 zenith observations. The simulations are noiseless as the effects of changing the calibration are below the noise threshold. Calibration was run on simulations including noise, so noise effects are carried into calibration solutions and then applied to a noiseless simulation.

test whether, in the presence of instrumental effects, does increasing the number of calibrator sources improve leakage In this test, all dipoles are assumed to be functional (`hyperdrive` only has to calculate one primary beam, and so it can do 15 000 calibration sources in about 15 min. With real data, we flag dead dipoles and have to calculate numerous primary beam patterns, which is slower. So in these tests on real data, I have assumed all dipoles are alive. This might introduce a different calibration systematic.)

A full discrete + diffuse sky model is simulated, with 5 000, 10 000, or 15 000 sources used for calibration and subtraction.

A matched set of 15 zenith EoR0 observations are then calibrated/subtracted with the same parameter sets. Fig. 8 displays the results as 1D power spectra. The simulations are noiseless as the effects of changing the calibration are below the noise threshold. Calibration was run on simulations including noise, so noise effects are carried into calibration solutions and then applied to a noiseless simulation.

As observed previously, the simulated data outperform the real data, with significantly more systematic power in the latter. In the simulated datasets, inclusion of more calibration (and subtraction)
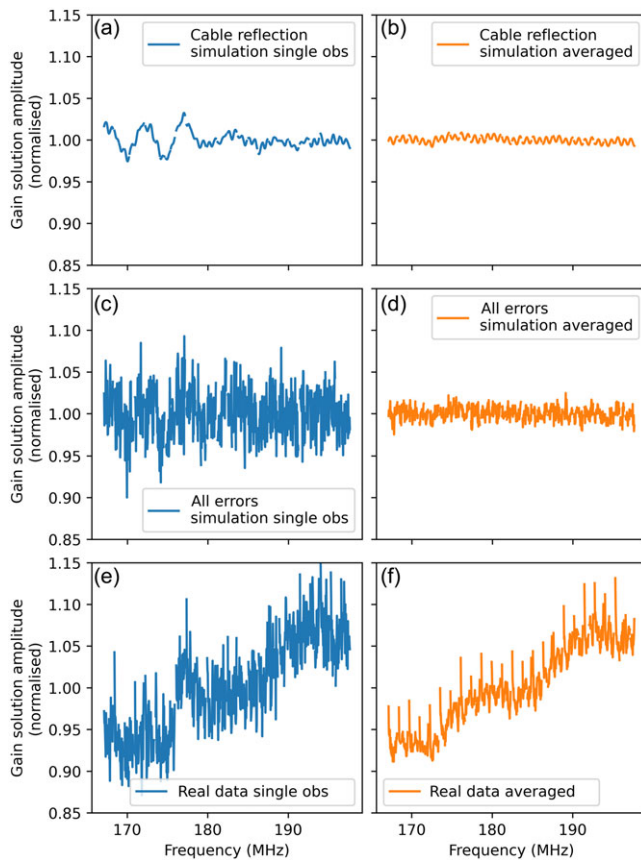
Figure 9. Calibration amplitudes showing the effects of averaging calibration solutions over 15 observations. Calibrations for a single tile are shown; other tiles show similar behaviours. Left column shows a single observation, right the average over 15 observations. Top row shows the simulation only containing cable reflection errors; middle row shows the simulation containing all instrumental errors; bottom shows real data.

sources, does reduce the leakage into the EoR window. Similar results are observed for the real data, but the relative amplitude of improvement is reduced compared with the simulations.

6. Calibration averaging

The results above included data simulated with thermal noise. In a 2-min snapshot, with calibration undertaken on 2-s cadence and 40 kHz spectral resolution, the visibilities contain ~30 Jy of noise per channel, limiting the calibration precision. Ideally, one would want to average calibration solutions over time to reduce the noise, but this can only be achieved if the solutions are stable over time. The MWA has been shown to be a stable system (Jordan et al. *submitted*), but changes in the beamformer settings to re-point the telescope can interrupt this. These individual pointings contain 15 observations (30 minutes long). It is reasonable to assume that solutions may be averaged over an individual pointing.

Fig. 9 shows the effects of averaging for the calibration solution amplitudes of a single tile. The top row shows the only real frequency-dependent amplitude effect, which are cable reflections. The calibration errors coming from an incomplete sky model are obvious in Fig. 9a, which seem to average out in Fig. 9b to leave cable reflection ripples. The noisy calibration solutions obviously get less noisy with averaging.

Although it is reasonable to average over a pointing, in reality there are complications like RFI, satellite passes, equipment failure etc that mean we may be averaging data with underlying signal differences, which could lead to bias. Fig. 10 shows what happens when you average. Blue in the ratio (and purple in the difference means) less power in the window when you average the solutions, a.k.a less leakage. Of interest however is at low $k_{\parallel}, k_{\perp}$ in the real data (where the diffuse power resides), the averaged solutions result in less power. This can be seen by the blue in the ratio and purple in the difference in the bottom left pixels in Fig. 10. It is hard to disentangle what is causing this reduction in power at large scales in the real data. Both simulations and real data show improvement in the EoR window with the averaging, which suggests that this is an avenue worth pursuing. Care must be taken in understanding what exactly is causing the reduction in both leakage and large-scale power for the real data.

7. Discussion

The work presented here is focussed on the MWA Australian pipeline, but many of the lessons are generally applicable to interferometers attempting experiments requiring high dynamic range. The interplay of different effects can exacerbate the impact of systematic errors, while some approaches are more well-suited to removing some systematics compared to others. For example, channel-based calibration handles cable reflections well, but is prone to imprinting spectral structure from calibration solutions on data, if smoothing is not used. A hybrid approach will be required. A per-channel calibration strategy has the benefits of allowing large degrees of freedom for spectral structure, but at the cost of higher thermal noise. With an expectation that the pure instrumental gain response should have minimal spectral structure, one may take an intermediate approach where some spectral regularisation of parametric fitting is undertaken. Li et al. (2019) used redundant calibration, along with a tile-based model for the structure imparted by cable reflections, whereas Yatawatta et al. (2009) and Yatawatta (2010) used regularisation to obtain smooth gain solutions across frequency. Some of these avenues have been explored with Hyperdrive, but none are currently used in production mode.

WODEN has been carefully designed to produce realistic MWA datasets, with many of the real instrumental effects observed in our data. The inclusion of diffuse emission in the simulations, currently an unused sky component in the MWA hyperdrive calibration and subtraction pipeline, has demonstrated the significance of its effect on our ability to detect the 21-cm signal. Despite the careful treatment and modelling of the full signal chain, there is clear evidence that there are unmodelled systematics that remain in the data (see, for example, Fig. 6). The key conclusions that can be drawn from this work are as follows:

- With a sky-based calibration, uv-gridding, and PS approach, we have need to subtract more than 90% of all discrete flux to recover 21-cm signal in absence of instrumental effects;
- When including diffuse emission in simulations, we are never able to access some k-modes (from 30 min of data), leading to a need for some diffuse emission removal;
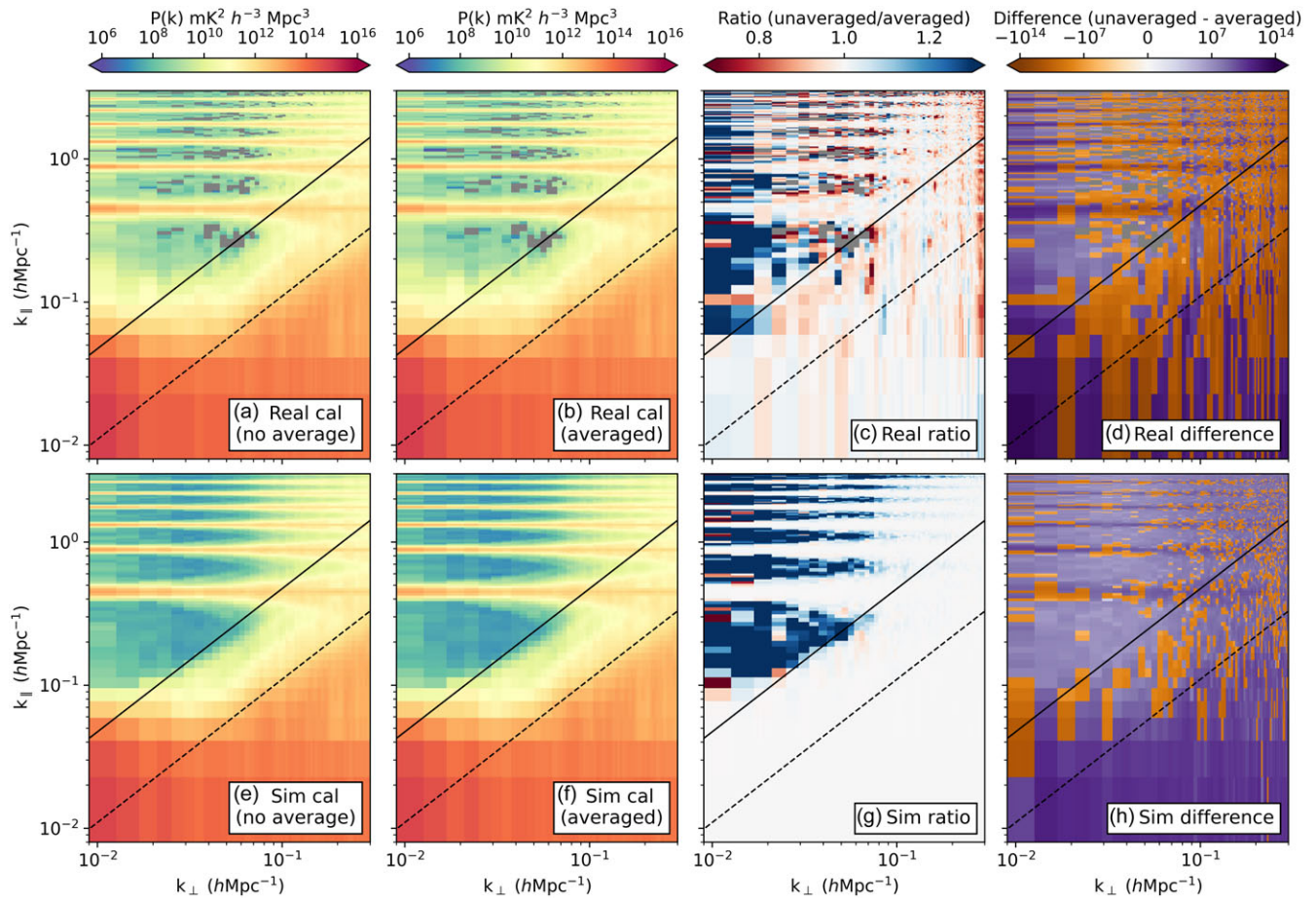- The single greatest cause of leakage is an incomplete sky model;

Figure 10. Effects of averaging calibration solutions over 15 observations for real data (top row) and simulated data (bottom row). The data in the simulation contain both discrete and diffuse sky models, but do not contain noise, in an attempt to better reveal any systematic bias involved in averaging. The calibration solutions applied to the simulation were derived from a simulation that did contain noise however, so the effects of noise are captured in the calibration solutions.

- `hyperdrive` easily treats simple tile gain errors and cable reflections.

The third point is worth discussion; without a higher-resolution and more sensitive southern hemisphere sky model, the MWA *sky-based* calibration is currently limited to the mJy-level source detections available from LoBES and GLEAM-X. Early SKA arrays will have good sensitivity, but only array releases close to the full array will have the angular resolution and sensitivity increase needed to improve our discrete source sky model.

Based on the analysis presented here, there are likely three major updates needed in the `AusEoRPipe` for a detection:

- A gain smoothness enforcement in the calibration algorithm, similar to the approach of LOFAR (Yatawatta et al. 2009). This should intrinsically cause less leakage, but needs to be applied in a way that does not cause bias or signal loss;
- Fitting of cable reflections, similar to the FHD approach (Beardsley et al. 2016). This might allow averaging

and/or fitting of calibration solutions to further reduce leakage;
- Treatment of the diffuse foregrounds. A foreground fitting approach, for example, like GPR (Mertens et al. 2020), and in a way that is robust against 21-cm signal loss.

There are many options for future work to extend this analysis, including:

- Increasing the field pointings that are simulated to match observations (i.e. away from zenith-pointed beams);
- Include frequency-dependent tile amplitude gain effects for more realistic instrumental effects;
- Varying the short baseline cutoff to optimise for diffuse calibration;
- Including full polarisation simulations and calibration;
- Including the refractive effects of the ionosphere;
- Including missing dipoles (different primary beams per tile) - this analysis is currently being undertaken.

8. Conclusion

We used WODEN, a full-sky simulator designed to produce realistic simulations of MWA EoR data, using discrete, 21-cm, and diffuse sky sources. We added known direction-independent MWA instrumental effects to this simulated data. The simulations were used to test the impact of different effects on MWA data, and in particular, on the PS of brightness temperature fluctuations of primordial hydrogen. We find that use of an incomplete sky model for data calibration has the largest impact on the ability to detect the EoR signal. Other effects such as channel and timestamp flagging, and cable reflections have less of an impact and do not prevent 21-cm science. We also find that sources down to less than 10 mJy need to be removed for success, and that failing to treat diffuse emission removes signal accessibility for some modes.

During this work we made extension use of the kvis (Gooch 1996) and DS9 (Joye & Mandel 2003) FITS file image viewers.

References

Barry, N., et al. 2019, ApJ, 884, 1. doi: 10.3847/1538-4357/ab40a8. arXiv: 1909.00561 [astro-ph.IM].

Beardsley, A. P., et al. 2016, ApJ, 833, 102. doi: 10.3847/1538-4357/833/1/102. arXiv: 1608.06281 [astro-ph.IM].

Cook, J. H., Trott, C. M., & Line, J. L. B. 2022, MNRAS, 514, 790. doi: http://dx.doi.org/10.1093/mnras/stac1330 10.1093/mnras/stac1330. arXiv: 2205.04644 [astro-ph.IM].

Furlanetto, S. R., Peng Oh, S., & Briggs, F. H. 2006, PhR, 433, 181. ISSN: 0370-1573. doi:https://doi.org/10.1016/j.physrep.2006.08.002. https://www.sciencedirect.com/science/article/pii/S0370157306002730.

Gooch, R. 1996, Astronomical Data Analysis Software and Systems V, Vol. 101, Astronomical Society of the Pacific Conference Series, ed. G. H. Jacoby, & J. Barnes, 80.

Górski, K. M., Hivon, E., Banday, A. J., Wandelt, B. D., Hansen, F. K., Reinecke, M., & Bartelmann, M. 2005, ApJ, 622, 759. doi: 10.1086/427976. arXiv: astro-ph/0409513 [astro-ph].

HERA Collaboration, et al. 2023, ApJ, 945, 124. doi: 10.3847/1538-4357/acaf50.

Hurley-Walker, N., et al. 2017, MNRAS, 464, 1146. ISSN: 0035-8711. doi: 10.1093/mnras/stw2337. http://mnras.oxfordjournals.org/lookup/doi/10.1093/mnras/stw2337.

Jordan, C. H., et al. 2017, MNRAS, 471, 3974. ISSN: 0035-8711. doi: 10.1093/mnras/stx1797. http://academic.oup.com/mnras/article/471/4/3974/3979467/Characterization-of-the-ionosphere-above-the.

Joye, W. A., & Mandel, E. 2003, Astronomical Data Analysis Software and Systems VII, Vol. 295, Astronomical Society of the Pacific Conference Series, ed. H. E. Payne, R. I. Jedrzejewski, & R. N. Hook, 489.

Koopmans, L., et al. 2015, Advancing Astrophysics with the Square Kilometre Array (aaska14), 1. doi: 10.22323/1.215.0001. arXiv: 1505.07568 [astro-ph.CO].

Kriele, M. A., Wayth, R. B., Bentum, M. J., Juswardy, B., & Trott, C. M. 2022, PASA, 39,e017. doi: 10.1017/pasa.2022.2. arXiv: 2201.04281 [astro-ph.IM].

Li, W., et al. 2019, ApJ, 887, 141. ISSN: 1538-4357. doi: 10.3847/1538-4357/ab55e4. arXiv: 1911.10216. https://iopscience.iop.org/article/10.3847/1538-4357/ab55e4.

Line, J. L. B., et al. 2020, PASA, 37, e027. doi: 10.1017/pasa.2020.18. arXiv: 2005.09316 [astro-ph.IM].

Line, J. L. B., Trott, C. M., Cook, J. H., Greig, B., Barry, N., & Jordan, C. H. 2024, PASA, 1. doi: 10.1017/pasa.2024.31.

Line, J. 2022, JOSS, 7, 3676. doi: 10.21105/joss.03676.

Lynch, C. R., et al. 2021, PASA, 38, e057. doi: 10.1017/pasa.2021.50. arXiv: 2110.08400 [astro-ph.CO].

Mertens, F. G., et al. 2020, MNRAS, 493, 1662. doi: 10.1093/mnras/staa327. arXiv: 2002.07196 [astro-ph.CO].

Mitchell, D. A., Greenhill, L. J., Wayth, R. B., Sault, R. J., Lonsdale, C. J., Cappallo, R. J., Morales, M. F., & Ord, S. M. 2008, ISTSP, 2, 707. ISSN: 1932-4553. doi: 10.1109/JSTSP.2008.2005327. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4703504.

Offringa, A. R., Mertens, F., & Koopmans, L. V. E. 2019, MNRAS, 484, 2866. doi: 10.1093/mnras/stz175. arXiv: 1901.04752 [astro-ph.IM].

Offringa, A. R., & Smirnov, O. 2017, MNRAS, 471, 301. ISSN: 13652966. doi: 10.1093/mnras/stx1547.

Patil, A. H., et al. 2017, ApJ, 838, 65. ISSN: 1538-4357. doi: 10.3847/1538-4357/aa63e7. http://arxiv.org/abs/1702.08679%20http://stacks.iop.org/0004-637X/838/i=1/a=65?key=crossref.e87ceb9a46c4eee2c879aae7024ce160.

Procopio, P., et al. 2017, PASA, 34, e033.ISSN: 1323-3580. doi: 10.1017/pasa.2017.26. https://www.cambridge.org/core/product/identifier/S1323358017000261/type/journal_article.

Thompson, A. R., Moran, J. M., Jr. & Swenson, G. W. 2017, Interferometry and Synthesis in Radio Astronomy (3rd edn.). doi: 10.1007/978-3-319-44431-4.

Tingay, S. J. J., et al. 2013, PASA, 30, e007. ISSN: 1323-3580. doi: 10.1017/pasa.2012.007. http://arxiv.org/abs/1206.6945%20http://www.journals.cambridge.org/abstract_S1323358012000070.

Trott, C. M., et al. 2020, MNRAS, 493, 4711. doi: 10.1093/mnras/staa414. arXiv: 2002.02575 [astro-ph.CO].

Trott, C. M., et al. 2016, ApJ, 752, 34. doi: 10.3847/0004-637X/818/2/139. http://arxiv.org/abs/1601.02073%20http://dx.doi.org/10.3847/0004-637X/818/2/139.

Wayth, R. B., et al. 2015, PASA, 32, e025. ISSN: 1323-3580. doi: 10.1017/pasa.2015.26. http://arxiv.org/abs/1505.0604%20http://www.journals.cambridge.org/abstract_S1323358015000260.

Wayth, R., et al. 2022, JATIS, 8, 011010. doi: 10.1117/1.JATIS.8.1.011010. arXiv: 2112.00908 [astro-ph.IM].

Wilensky, M. J., et al. 2023, ApJ, 957, 78. doi: 10.3847/1538-4357/acff bd. arXiv: 2310.03851 [astro-ph.CO].

Yatawatta, S. 2010, 2010 IEEE Sensor Array and Multichannel Signal ProcessingWorkshop, SAM 2010, 2, 69. doi: 10.1109/SAM.2010.5606769.

Yatawatta, S., Zaroubi, S., de Bruyn, G., Koopmans, L., & Noordam, J. 2009, 2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop (IEEE), 150. doi: 10.1109/DSP.2009.4785912.

Zonca, A., Singer, L., Lenz, D., Reinecke, M., Rosset, C., Hivon, E., & Gorski, K. 2019, JOSS, 4, 1298. doi: 10.21105/joss.01298. https://doi.org/10.21105/joss.01298.

Appendix A. Calibration solutions

Figure A1. Calibration gain amplitudes and phases from a simulated two minute snapshot. These demonstrate the constant gain and flat phase slopes added to the simulation. The underlying simulation was of both the diffuse and discrete sky models and only contained gain errors with no other instrumental effects. Calibration was performed using 10 000 sources through `hyperdrive`. Any spectral structure in the amplitudes comes from the incomplete sky model rather than injected gains.