Dancing with the stars

Albert, J.G.

Chapter 4

On the Feasibility of Probabilistic Ionospheric Screens for LoTSS

J.G. Albert
R.J. Van Weeren
H.T. Intema
AND
H.J.A. Röttgering

On the feasibility of probabilistic ionospheric screens for LoTSS

J. G. ALBERT, R. J. VAN WEEREN, H. T. INTEMA, AND
H. J. A. RÖTTGERING

In preparation for submission

Producing direction dependent calibrated radio images is vital for low-frequency wide-field radio surveys such as the LOFAR Two-Metre Sky Survey (LoTSS). The current facet-based direction dependent calibration of LoTSS is limited by two main factors: the sparsity of suitable in-field calibrators, and ill-conditioning when many in-field calibrators are used. A screen-based method of direction dependent ionospheric calibration was recently proposed in Albert et al. [2020b] which was found to surpass the state-of-the-art direction dependent calibration on a single randomly selected LoTSS data set, using fewer calibrators than used by the LoTSS calibration. The proposed method potentially poses a logical next LoTSS survey improvement if it can be shown to be feasible and robust on a larger sample set. In this paper we apply the method to the deep Lockman Hole data set (12 observations and 100 hours in total) which spans varying times-of-day, season, and varieties of ionosphere. We present a neural-network based outlier detection method that significantly improves over the original outlier detection method. We quantity the improvement to image artefacts around inter-calibrator sources by computing equivalent integration-time gain. We find that for observations with low and high ionospheric activity that the method provides, respectively, an average equivalent integration-time gain near bright sources of 1.3, and 2.0. In two of the observations the improvements are less significant. We suggest that during these two observations small-scale ionospheric structure limits the screen-based method. In no cases is the resulting image quality worse than LoTSS. We apply the same method to calibrate and image the full data set and find that the same improvements scale to deep observations. Our method produces a zoo of ionospheric doubly differential total electron content screens supporting that the ionosphere has many different distinct behaviours. We discover an unmodelled systematic in the Jones scalars, which results in residual artefacts that limit the effectiveness of method. Using a combination of real and simulated Jones scalar data we propose and implement a preliminary correction for it. The proposed correction is found to partially correct for the systematic, and indicates that the general approach is correct and can be improved to completely account for the systematic. Using these results we determine that the method is capable of extending the LoTSS calibration and imaging pipeline, and is also promising for < 100 MHz data.

Acknowledgements. J. G. A. and H. T. I. acknowledge funding by NWO under 'Nationale Roadmap Grootschalige Onderzoeksfaciliteiten', as this research is part of the NL SKA roadmap project. J. G. A. and H. J. A. R. acknowledge support from the ERC Advanced Investigator programme NewClusters 321271. R. J. vW. acknowledges support of the VIDI research programme with project number 639.042.729, which is financed by the Netherlands Organisation for Scientific Research (NWO). This research has made use of the University of Hertfordshire high-performance computing facility and the LOFAR-UK compute facility, located at the University of Hertfordshire and supported by STFC [ST/P000096/1]. J. G. A. thanks Aleksandrina Skvortsova for spending many hours clicking through outlier data.

4.1 Introduction

One of the primary objectives of the Low-frequency array [LOFAR; van Haarlem et al., 2013] is to be a wide-field low-frequency surveying instrument of the entire northern hemisphere with unprecedented high sensitivity and resolution. There are numerous challenges involved in realising this goal, which requires efficiently storing and processing petabytes of data. First among these, any good wide-field low-frequency survey requires good direction dependent (DD) calibration and imaging [Cohen and Röttgering, 2009]. Great progress has been made to meet the DD calibration and imaging challenge. The calibration program killMS [Tasse, 2014b, Smirnov and Tasse, 2015] applies a clever sparsification of the optimisation problem to overcome computational limitations of DD calibration, and uses an extended Kalman filter to regularise the Jones matrices which improves the conditioning of the problem. The imaging program DDFacet [Tasse et al., 2018] is a wide-field wide-band non-coplanar deconvolution algorithm with spatially varying point-spread-function, and allows externally defined Jones matrices to be applied. Together these two programs supply the tools that make up the state-of-the-art DD calibration and imaging solution used by the LOFAR Two-Metre Sky-Survey [LoTSS; Shimwell et al., 2019].

The LoTSS is planned in a tiered program of successive image quality, sky-coverage, and depth. The first data release [DR1; Shimwell et al., 2019] provided the first tier of the survey, presenting images of 5% of the northern sky, with an excellent median sensitivity of $S_{144\mathrm{MHz}} = 71\ \mu\mathrm{Jy\,beam}^{-1}$ and point-source completeness of 90% at integrated flux density of 0.45 Jy. The coming second data release (DR2) will present improved DD calibrated images of a much large fraction of the northern sky, and should improve the median sensitivity as well as improve the detection of diffuse emission. Future data releases will go deeper, and use long baselines to achieve resolutions of $0.3''$. Since we have internal access to the unreleased LoTSS-DR2 archive, we'll make reference to DR2.

Despite enabling one of the most ambitious radio surveys ever conducted, the state-of-the-art DD calibration and imaging method is fundamentally limited by 1) the sparsity of available calibrators in the field of view, and 2) the ill-conditioning of the system. When the angular separation between calibrators is too great, the ionospheric distortions are directionally under-sampled, leading to non-isoplanaticity [Fellgett and Linfoot, 1955], and when there are too many degrees of freedom this leads to 'self-cal bias'[1] [Grobler et al., 2014]. These two limitations are not mutually exclusive; even when the field of view is densely populated with bright calibrators, selecting too many potentially leads over-parametrisation and flux absorption. Indeed, the LoTSS-DR2 choice of 45 calibration directions was chosen to balance these two aspects.

The scattering effects of the ionosphere can be seen as a diffuse halo around sources, which are not necessarily visible above the noise [Vedantham and Koopmans, 2015] and can significantly impact studies sensitive to the power-spectrum of faint diffuse emission [e.g. Harrison et al., 2016, Patil et al., 2017, Vernstrom et al., 2017, van Weeren et al., 2019]. Some part of this scattering can be corrected via DD calibration, however there is always a remaining component which adds noise to the image, which depends on the spatio-temporal power-spectrum of ionospheric FED and the availability of bright enough calibrators. For LOFAR this noise-like component is expected to be of a level comparable to thermal noise [Vedantham and Koopmans, 2016]. The DD correction performed by killMS is a facet-based

[1] The term 'self-cal bias' was coined by Ger de Bruyn.

algorithm that assumes each facet is an isoplanatic patch. When this is true, then all sources in a facet can be calibrated with a single number per antenna, to good approximation, removing most ionospheric effects and leaving only a thermal noise-like halo residual. However, since the ionosphere is highly dynamic often the isoplanatic assumption is violated and a facet fails to be well calibrated.

Recently, Albert et al. [2020b] proposed a probabilistic DD calibration and imaging approach for ionospheric calibration of LOFAR high-band antennae (HBA; 115–189MHz) radio interferometric data that alleviates the issue of the sparsity of calibrators. The method is based on inferring doubly differential total electron content [DDTEC; described in Albert et al., 2020b] using a probabilistic non-diffractive tomographic technique [Albert et al., 2020a]. The method was shown to significantly improve DD dispersive phase errors between calibrators on a single randomly selected observation taken from the LoTSS-DR2 archive. In particular, the method was shown to reduce the root-mean-squared residuals by 32% within $90''$ of inter-calibrator, bright (peak flux > 100 mJy beam^{-1}) sources in comparison with the archival LoTSS-DR2 image. For comparison, the reduction in DD scattering artefacts, around these inter-calibrator sources, is equivalent to observing for approximately twice as long. If this result can be expected when the method is applied to most observations, then it prompts the possibility of reprocessing the entire LoTSS-DR2 archive (13 000 hours of observing time) with these inter-calibrator ionospheric artefacts suppressed. However, before proposing such a computationally expensive reprocessing the feasibility of the method on multiple nights needs to be accessed, as well as an understanding and amelioration of the systematic biases of the method.

Understanding the systematic biases of the method is vital for releasing a product that the scientific community can trust. Such systematic biases include flux absorption due to an incomplete sky model and over-parametrisation, dispersive phase and amplitude errors due to an incomplete beam model, and incorrect DDTEC inference due to model misspecification of the Jones scalars. Since this method introduces no extra degrees of freedom and maintains the same calibrator sparsity as LoTSS, we expect the same source completeness properties as the current LoTSS-DR2 processing, and we do not investigate it here.

In this paper we examine 1) the method effectiveness in a larger sample of observations as well as on a multi-epoch (100 hours) imaging application, 2) the robustness to ionospheric conditions, and 3) systematic biases due to model misspecification. In Section 4.2 we describe the data and calibration procedure. In Section 4.3 we assess the method individually on 12 data sets with identical points taken from the LoTSS archive. We also combine all data sets and produce a deep image, which we compare to LoTSS. We then we quantify the robustness to differences in ionospheric conditions. Finally, in Section 4.4 we investigate a systematic due to model misspecification using a combination of real and simulated Jones scalar data, and then present a correction for this systematic.

4.2 Method

In this paper, we are primarily focused on assessing our method for robustness to ionospheric conditions. To that end, we limit the variation due to calibrator distribution, sky model completeness, field declination, and proximity to bright sources. The optimal experimental setup is therefore to compare the method on the same field over multiple observations. We

select the deep Lockman Hole (162°, 58°) data set (12 observations and 100 hours in total) for this purpose, whose observation dates and times are shown in **Table** 4.1 along with total daily Sun-spot count. The selected data set can be divided into two sample sets with six samples in each category: high Sun-spot count at dawn (Summer of 2018), and low Sun-spot count at dusk (Spring of 2015).

Sun-spot counts are known to correlate strongly with ionospheric free electron density (FED) due to extreme ultra violet ionising radiation from the Sun [Kiepenheuer, 1946], and therefore are an approximate measure of the expected ionospheric conditions. As can be seen from the Sun-spot counts, the observations from Summer of 2018 are expected to have a lower FED compared to the observations from Spring of 2015. We note that high Sun-spot count, and thus FED, does not necessarily imply temporal or spatial characteristics of the ionosphere, as these qualities are largely influenced by the bulk motion properties of the Earth's atmosphere. In fact, the local time-of-day has the largest impact on the temporal and spatial properties of the ionosphere, for example it has been suggested that scintillation is more pronounced near sun-rise due to increased FED variation [e.g. Spoelstra, 1983]. Given the different season and time of day of the data sets, they probe two vastly different ionosphere varieties.

Table 4.1: Date and time of 12 Lockman Hole 8 hour observations selected for method analysis

Obs. ID	Start time (UTC)	Total daily Sun-spots
667218	2018-09-13 07:05:31	0 ± 0
667204	2018-09-12 07:06:28	14 ± 1.1
664480	2018-08-19 08:39:05	14 ± 0.9
664320	2018-08-15 08:49:19	13 ± 0.9
659948	2018-07-12 11:08:29	0 ± 0
659554	2018-07-10 11:11:19	0 ± 0
342938	2015-05-08 14:50:43	150 ± 10.2
340794	2015-04-25 17:08:19	69 ± 6.2
299961	2015-03-24 17:47:39	108 ± 5.9
294287	2015-03-21 19:11:19	26 ± 2
281008	2015-03-14 18:26:58	55 ± 4.7
274099	2015-03-08 20:11:19	28 ± 1.9

Note: Sun-spot uncertainties are standard deviations of measurements from solar observatories around the world.

For the purpose of discussion in this current work we provide a summary of the steps of the method from Albert et al. [2020b]:

1. *Subtract and solve step.* Subtract a good model of the sky from the visibilities, except for a set of bright calibrators. Solve against the isolated calibrators.

2. *Smooth and slow-resolve.* Smooth the Jones scalars, and resolve on a long time scale to simultaneously improve the conditioning and solve the holes problem. For description of the hole problem see Shimwell et al. [2019]. The phases are smoothed in frequency by fitting a phase model, and amplitudes in time using a median filter with 15 min. filter window.

3. *Measure DDTEC.* Infer the DDTEC of the Jones scalars with a variational hidden Markov model. Perform outlier detection and flagging on the inferred DDTEC.

4. *Infer DDTEC screen.* Apply model marginalised Gaussian process regression on the measured DDTEC, with a physically motivated DDTEC covariance function, to infer DDTEC for a screen of directions covering the field of view.

5. *Image.* Image the original visibilities with a concatenation of the smoothed and screen solutions.

With the exception of isolating the calibrators before solving, the first two steps can be seen as equivalent to the LoTSS DD calibration, and steps 3 and 4 can be seen as an extension of the LoTSS DD calibration pipeline. A good quality of an extension to an existing pipeline is that it does not worsen the data product with respect to the preceding part of the pipeline. The calibrator selection criteria in step 1 of the original method was peak flux > 0.3 Jy beam^{-1} separated by at least $6'$. Depending on the field, this selection could result in less than 20 to more than 50 in-field calibrators, which significantly impacts the inter-observational DDTEC screen consistency. To that end, in this paper we choose to standardise step 1 by choosing the same number of calibrators as used by LoTSS – the brightest 45 directions separated by at least $6'$. Our calibrator selection is not not exactly the same as in LoTSS. In LoTSS, the direction of calibrators is not necessarily centred on a source, whereas our calibrator directions must precisely encircle a bright source. This can be seen in **Figure** 4.6a, where the LoTSS calibrators do not centre on any particular source. This is because LoTSS calibrates against all the sources within a facet, and we calibrate against isolated bright sources. However, by choosing the same number of calibrators, and ensuring they are sufficiently separated, the average distance between our calibrators is the same as in LoTSS.

The Lockman Hole data sets have been previously jointly calibrated and imaged as part of LoTSS-DR2. Therefore, we have a deep sky model that is a factor of approximately 3.5 deeper than the sky model for an individual eight-hour observation. We use this deep sky model to perform calibrator isolation in step 1. In DDFacet, deconvolution is performed using hybrid matching pursuit with a genetic algorithm that locally optimises deconvolution around each clean component [Tasse et al., 2018]. The sky model resulting from this process may contain scattered negative components, especially around bright sources, which are called sky model artefacts. Failure to include these sky model components in the subtraction mask can cause undefined distortion of the Jones scalars solved for. We manually ensure that the subtraction mask in step 1 encompasses all sky model components of the calibrator source. In particular, radio galaxy 3C244.1 is the brightest source in the field, and has scattered sky model components that would fall outside the default subtraction mask, a circle with radius $120''$ around each calibrator. We used a $240''$ radius mask for this calibrator.

4.2.1 An improved outlier detection

As stated in Albert et al. [2020b], outlier detection and flagging of the DDTEC inferred in step 3 is extremely important for the performance of the method, since the Gaussian process regression (step 4) depends on accurate estimates of the DDTEC uncertainty. That is, the method is robust to outlier DDTEC values so long as the corresponding uncertainties are large enough to reflect this. In the original method outlier detection was done with a heuristic

method using radial basis functions that detects outliers based on spatial similarity, however this method resulted in too many false negatives when applied to the Lockman Hole data sets. To this end, we manually identified 4531 outliers and 74460 non-outliers then designed and trained a neural network to perform classification.

The input to the neural network is the time series of DDTEC mean and log-uncertainty of the K nearest neighbours of an optical pathway under consideration, as shown in **Figure** 4.1. We choose $K = 15$ (1/3 calibrator directions in the field of view), i.e. a DDTEC value is classified as an outlier based on the nearest 15 calibrators in the antenna's field of view. This choice of input reflects how humans visually determine what is an outlier. In particular, in order for a DDTEC value to be classified as an outlier by a human we require 1) that it visually 'looks out of place' in the context of the DDTEC screen, or 2) the temporal evolution of that optical path 'looks out of place' when compared with neighbouring calibrators. These two criteria correspond to being spatially and temporally 'out of place'. It is important that we consider these two criteria, because sometimes when there are many outliers in an antenna's field of view the spatial information alone cannot discriminate between outlier and non-outlier.

The neural network architecture is an ensemble of eight 10-layer residual temporal à trous convolutional neural networks [CNN; e.g. He et al., 2015]. Each layer has a kernel of shape *(kernel, filters)* corresponding respectively to how wide the convolution window is, and the number of features, similar to channels in an image. We choose *kernel=3*, and a *filter=48*. The non-linearities of each CNN is the function $f(x) = \max(0, x)$. After we apply the non-linearity we down sample the data according to a rule known as a pooling operation. Each layer is a residual layer, i.e. the input to the layer is added to the output after applying the non-linearity and pooling operation. Each layer is an à trous convolution [Yu and Koltun, 2016] which has inspiration from wavelet decomposition. In à trous convolution the kernel window is expanded by a parameter called the dilation rate. Roughly speaking this controls how sensitive the layer is to diffuse signal. Each member of the ensemble has a different combination of pooling operation and dilation rate which gives each member a unique structure that is sensitive to different types and scales of structure in the data. The set of eight combinations of pooling and dilation rate is $\{\text{MaxPool},\text{AvgPool}\} \times \{1,2,3,4\}$. The idea of using an ensemble of classifiers is that it averages away the bias of a single classifier, which is the motivation behind many ensemble-based methods such as random forests.

A final layer projects the input to *filter=1*, which is interpreted as the logit of a binary classification problem. To facilitate learning we add a constant to the output of each CNN equal to the log prior probability of the training set, which is -3.6. This means that when the CNN outputs zero the logits will correspond to the prior logits and the neural network only needs to learn deviations from the prior. The loss function of each CNN is weighted binary cross-entropy. We weight the positive classifications by the ratio of non-outliers to outliers, which is approximately 36, so that the average gradient magnitudes of all outlier examples is equal to that of the non-outlier examples. This ensures that we do not preferentially learn to classify non-outliers.

We then jointly train the whole ensemble with Adam stochastic gradient descent [Kingma and Ba, 2014] using a mini-batch size of 32, time-segments of 50 minutes, and learning rate of 10^{-4}. We mask out the optical paths without ground truth labels. We train for 20 epochs which takes approximately 6 hours on a 32 core CPU. Following training we set the classification threshold of each member of the ensemble to the optimal value using the receiver operating characteristic (ROC) curve The ROC curve is a plot of the FPR against FNR as a function of

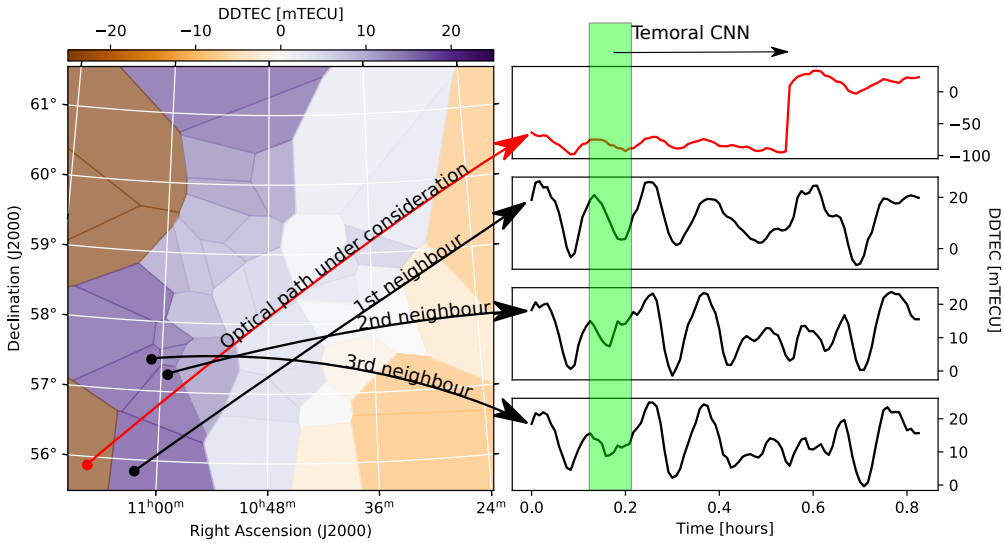


Figure 4.1: Pictorial description of the input to the neural network. For each optical pathway we take the K nearest neighbours in the field of view, measured by great circle separation, and stack the time series of measured DDTEC and log-uncertainty. An ensemble of temporal CNNs are applied to the input and the outputs are trained for logistic regression per time step. The ensemble median classification is used to identify outliers.

Table 4.2: Comparison of outlier detection methods.

Obs. ID	Neural network method				Heuristic method			
	FNR	FPR	Est. FN	Est. FP	FNR	FPR	Est. FN	Est. FP
667218	1.5%	11.2%	728	274 474	45.2%	5.0%	21 999	122 109
667204	0.4%	4.4%	125	113 679	54.8%	6.1%	19 387	155 669
664480	0.1%	10.4%	49	275 810	58.7%	11.4%	22 768	300 110
664320	1.4%	18.3%	1 500	461 773	49.1%	8.2%	51 741	206 502
659948	1.5%	6.0%	1 247	150 594	45.8%	5.1%	38 239	128 304
659554	0.6%	7.5%	452	195 082	39.3%	7.2%	28 220	187 883
342938	2.2%	11.2%	1 056	320 300	42.3%	5.7%	20 422	161 647
340794	5.9%	5.4%	6 148	152 023	43.7%	4.2%	45 270	118 631
299961	2.1%	2.8%	806	76 993	43.6%	3.2%	17 068	88 911
294287	12.4%	6.9%	25 327	185 568	43.7%	6.9%	89 487	186 491
281008	1.6%	4.2%	532	119 309	42.4%	3.7%	13 705	107 378
274099	4.7%	9.9%	17 329	250 393	48.8%	5.9%	181 955	150 008
Total	2.6%	7.0%	55 298	2 575 997	48.0%	5.5%	550 261	1 913 642

Note: On average there are 2 749 800 optical paths per observation.
Note: FNR is false-negative rate, FPR is false-positive rate, FN is false negatives, and FP is false positives.

the threshold. We choose the threshold for each member of the ensemble that minimises FNR plus FPR. At deployment time we take the median classification from all members of the ensemble as the classification.

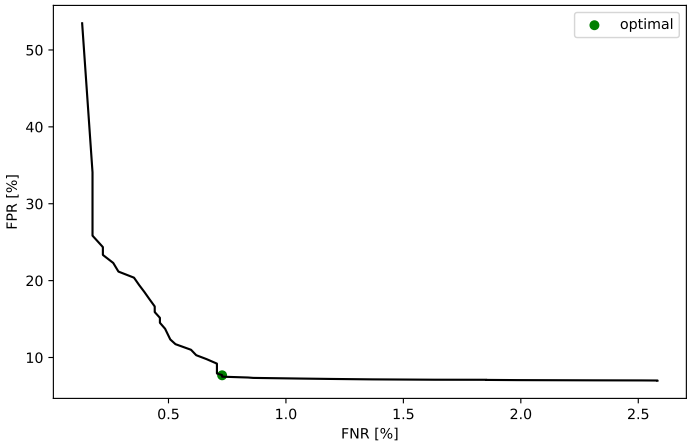


Figure 4.2: Receiver operating characteristic curve as a function of the threshold for the number of neural-network identified outliers in an antenna's field of view. We flag an antenna's entire field of view if the number of neural-network identified outliers is above a threshold. We value low FNR without significantly raising FPR. The optimal threshold of 30 directions gives a FNR of 0.7% and FPR of 7.7%.

A comparison of performance of the heuristic method to the neural network is shown in **Table** 4.2. The average false-negative rate (FNR) of the heuristic and neural network methods is 48% and 2.6% respectively. False negatives (FN), being outliers that are missed, have a much larger impact on performance than false positives (FP), which are accidentally flagged non-outliers. Therefore, the neural network method provides an order of magnitude improvement.

Using the FNR, FPR and the number of neural network classified outliers we can estimate the number of FN and FP in the neural network classifications. Despite the significant improvement of the outlier detection, we estimate that approximately 55 298 outliers were missed. We observe that most of these FN occur when there are a large number of outliers detected in the same field of view. This suggests that when the field of view has too many outliers it is difficult for the neural network to decide which data points are not outliers.

We use this intuition to motivate a secondary rejection step. We reject an antenna's entire field of view if the number of neural-network identified outliers in the field of view is above a certain threshold. When an antenna's entire field of view is flagged we flag all corresponding baselines in the visibilities involving the antenna under consideration. This rejection mechanism requires no extra computation since it uses the results of the neural network classifications. **Figure** 4.2 shows the receiver ROC curve for this rejection mechanism. The ROC curve plots the FPR against FNR as a function of the threshold. We identify a

threshold that meets our requirements of low FNR without raising the FPR significantly. From the ROC curce we choose an optimal threshold of 30 directions. This results in a FNR of 0.7%, which is a factor of almost four in reduction, and a FPR of 7.7%, which is unchanged.

4.3 Image improvements and ionospheric robustness

The method significantly improved the DD errors in 10/12 observations, while 2/12 showed only moderate improvements. **Figure** 4.3 shows typical examples of how inter-calibrator DD errors are corrected by the method. In observations 274099 and 297287 many inter-calibrator DD effects still remained after applying the method, despite verifying that that screens looked sensible. Note that the resulting image quality in these observations is still at least as good, or better, than the LoTSS-DR2, i.e. DD errors did not get worse.

There are several possible explanations why these two observations did not show significant improvements. One is that the calibrator layout, with an average inter-facet separation of $38'$, may under-sample small scale ionospheric structure during these observations. This separation is equivalent to ionospheric scales of 2 km to 5 km, which occasionally occur in the ionosphere [Yeh and Swenson, 1959, Mevius et al., 2016]. Another possibility is that non-stationary structure due to travelling ionosphere disturbances (TIDs) [e.g. van der Tol, 2009] may have been prevalent in these observations. Our method currently does not support non-stationary disturbances, though extensions are possible to account for such behaviour. Alternatively, the thin-screen approximation described in Albert et al. [2020b] may not be valid in these observations. Koopmans [2010] shows that wide-field low frequency arrays should take into account the 3D ionosphere. In our application we take a full 3D non-diffractive tomographic model in the limit of a thin-ionosphere for computational reasons, although there are possible optimisations that might make 3D modelling feasible. Finally, we observe in **Table** 4.2 that these two observations have a much higher number of estimated FN, suggesting that the Jones scalar data may have been of sub-standard quality resulting more missed outliers.

In **Figure** 4.3 we observe that the spoke-like patterns, which originate from uncorrected ionospheric scattering, are corrected by the method. We observe that around all corrected sources, especially the right-most one, there are radially asymmetric arc-like patterns. These arc-like patterns appear in varying degrees in all observations. We analyse and propose a correction for these arc-like patterns in Section 4.4.

To quantify improvement to DD errors we measure the root-mean-squared residuals (RMSR) around inter-calibrator bright sources before and after applying the screen. To measure RMSR, we select the brightest 66 non-calibrator compact sources across the field of view and mask an annulus with inner radius set to avoid the source flux and outer radius set to $45''$. If there are nearby compact or diffuse structures these are excluded from the mask. The RMSR is then computed as the square-root of the mean of the square of the flux density in the mask. We denote RMSR in the screen-corrected image as σ_{screen}, and $\sigma_{\text{no-screen}}$ as the RMSR in the image with same calibration procedure as LoTSS-DR2. We refer to $\sigma_{\text{background}}$ as the background noise.

From the RMSR we compute the equivalent integration-time gain (EITG), as the relative extra amount of integration time that would be required to achieve the RMSR of the screen, σ_{screen}, when using the LoTSS-DR2 calibration method. Assuming that scattering effects occur

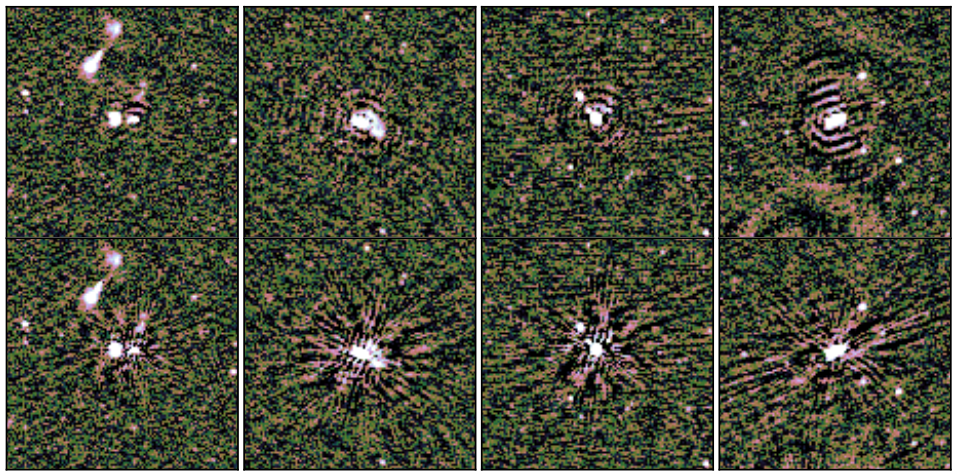


Figure 4.3: Visual example of how the method corrects inter-calibrator DD errors. These sources are taken from observation 342928. The lower panels are images with 45-direction calibration using the same method as LoTSS-DR2, and the upper panels are images with DDTEC screens applied. The mean background noise is 70 μJy beam^{-1}.

uniformly random throughout an observation then the EITG is given by $(\sigma_{\text{no-screen}}/\sigma_{\text{screen}})^2$.

Figure 4.4 shows a histogram of the EITG for the two sample sets as well and also plots the EITG of the individual observations against the total daily Sun-spot counts. We observe that the EITG in both sample sets are peaked above one and have long tails extending to higher values. There is an apparent difference in the histograms of the two sample sets. The high Sun-spot (Spring 2015) sample set has a mean EITG of 2.0 while the low Sun-spot (Summer 2018) sample set has a mean EITG of 1.3, and the tail of the prior one at larger EITG is heavier. Given that the Spring 2015 is closer to Solar maximum, has high Sun-spot counts, and is at dusk, the ionospheric FED is likely much larger than in the Summer 2018 observations, therefore scattering is likely stronger. Indeed, the scattering artefacts in the Summer 2018 images are far weaker. These results suggests that the method corrections are more significant during times of high ionospheric FED and less significant when the ionosphere FED is lower, which are equivalent to stronger and weaker scattering respectively. This is also compatible with the result of Albert et al. [2020b].

This result is partially explained by the fact that observations with less scattering have fewer DD errors to correct, therefore the maximal achievable EITG is lower. This necessarily constrains the EITG in the Summer 2018 sample set to lower numbers than the EITG in the Spring 2015 sample set, regardless of the model accuracy. Thus, we cannot rule out the possibility that the model is not accurately modelling the underlying DDTEC in the Summer 2018 sample set. For example, there is likely more turbulence near dawn [e.g. Spoelstra, 1983] which produces small scale structure that the average calibrator spacing of 38′ does not resolve. Since the scattering effects are very weak in the Summer 2018 observations,

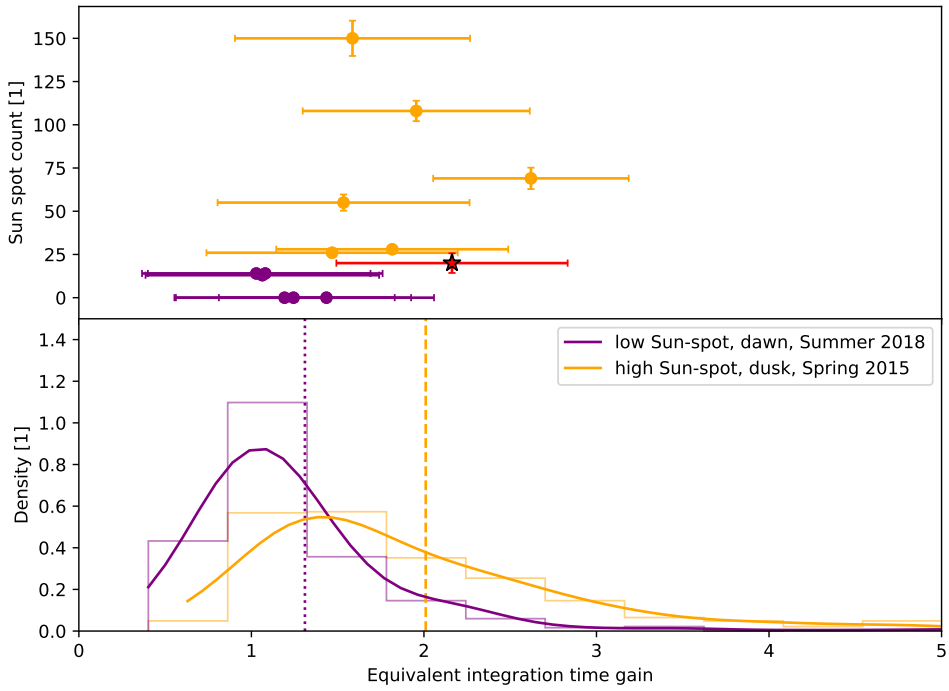


Figure 4.4: Upper panel: Total daily Sun-spot count individual observations plotted with respect to the EITG. The red star is the data point from Albert et al. [2020b]. Lower panel: Histograms and kernel density estimation of ETIG aggregated per sample set. The dashed line indicates the mean of 2.0 for the histogram for the Spring 2015 sample set. The dotted line indicates the mean of 1.3 for the histogram for the Summer 2018 sample set.

sometimes with DDTEC values only a few times the uncertainty, it is not possible to investigate this possibility.

We can effectively characterise the model efficacy independent of the ionospheric conditions by considering the EITG as a function of the strength of ionospheric scattering. As a proxy for ionospheric scattering strength we use the ratio $\sigma_{\text{no-screen}}/\sigma_{\text{background}}$. **Figure** 4.5 plots EITG against this proxy of ionospheric strength. We observe that when the $\sigma_{\text{no-screen}}$ is close to the background noise the EITG is approximately unity, indicating that there is no improvement from the model. Conversely, when the $\sigma_{\text{no-screen}}$ is an order of magnitude larger than the background noise the EITG is in the vicinity of three. There is a large scatter in the relation, however a power-law fit indicates that the EITG $\propto \sqrt{\sigma_{\text{no-screen}}/\sigma_{\text{background}}}$. The ideal model that completely removes all scattering effects would produce a power-law index of two. The deviation from this ideal relation is likely explained by at least two things. Firstly, there is always a low-level halo of speckles around sources, proportional to the brightness of the source, due to ionospheric scattering on sub-solution interval timescales. Secondly, as mentioned already, in Section 4.4 we investigate a systematic which contributes to the residuals around sources.

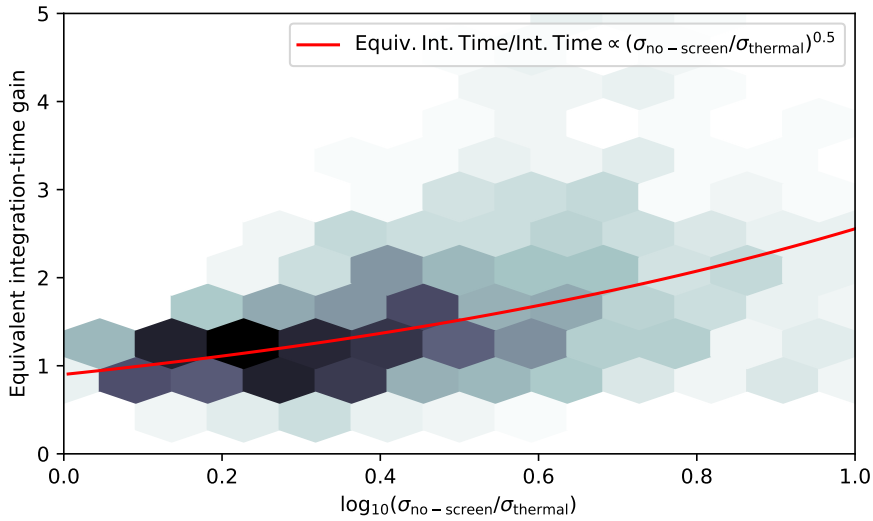


Figure 4.5: 2D Histogram of the EITG and ionospheric scattering strength proxy, $\sigma_{\text{no-screen}}/\sigma_{\text{background}}$, aggregated over all observations. The red line shows a power-law fit.

The scatter also slightly goes below one for $\sigma_{\text{no-screen}} \lesssim 3\sigma_{\text{background}}$. This can be explained by the enhanced effect of small errors or systematics at low noise levels. Specifically, suppose the flux of the pixels in the masked regions where RMSR is measured is composed of the background flux and the artefact flux, with root-variances $\sigma_{\text{background}}$ and $\sigma_{\text{artefact}} = \alpha\sigma_{\text{background}}$ respectively. The artefact flux can be due scattering as well as modelling systematics. Then, since the background and artefacts are independent we have that the EITG is,

$$\text{EITG} = \frac{1+\alpha^2_{\text{no-screen}}}{1+\alpha^2_{\text{screen}}}, \tag{4.1}$$

where $\alpha_{\text{no-screen}}$ and α_{screen} correspond to before and after applying the screen model, respectively. Note that α_{screen} contains systematics imposed by modelling, which $\alpha_{\text{no-screen}}$ is free of. When $\alpha_{\text{no-screen}}$ and α_{screen} are small then a modelling error can lead to EITG less than one. For example, suppose $\sigma_{\text{no-screen}} = 3\sigma_{\text{background}}$, then a modelling error of $2\sigma_{\text{background}}$ implies EITG less than one.

We jointly imaged all observations producing a deep image. **Figure** 4.6 shows the comparison between our deep image and the LoTSS-DR2 archival deep image. Since the calibrator layout is different between the LoTSS-DR2 image and our image, there are two comparisons we make.

The first is a comparison of scattered flux near calibrators. Since LoTSS calibrates against all flux in a facet, the solution does not perfectly correct any particular source. Instead, the solution partially corrects all sources in the facet, with boundaries indicated by cyan lines in **Figure** 4.6a. Compare this with our calibrator sources, inside red circles, which are all well corrected. This reflects the fact that we isolate the calibrators before solving.

One of the benefits of not isolating calibrators is that all solutions in a facet will be partially improved, and this is a good strategy when a screen-based method is not available. If we were to image with only the isolated calibrator solution then we would see very clean calibrators but worse inter-calibrator scattering than in the LoTSS image. Therefore, when no screen-based method is available calibrating against a facet is appropriate. However, calibrating against a whole facet can fail entirely when the isoplanatic patch size becomes much smaller than the facet.

The second comparison we make is that of the scattered flux around inter-calibrator sources. **Figure** 4.6b shows some examples of the improvements made by our model. We observe that the spoke-like artefacts are mostly gone, indicating that the effects of the ionosphere have been mostly removed. The majority of the remaining artefacts are from the modelling systematic mentioned earlier, which are also visible in the individual observations, in **Figure** 4.3. These artefacts are typically lower level than the scattering related artefacts, moreover we expect to be able to characterise and remove this modelling systematic.

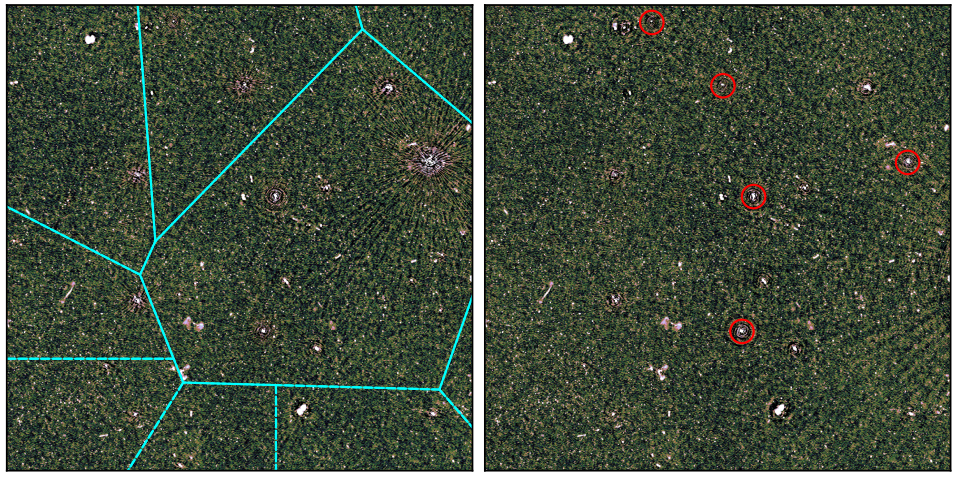
4.4 A non-ionospheric systematic and correction

In **Figures** 4.3 and 4.6b we observe asymmetric arc-like artefacts around inter-calibrator sources after applying our screen-based calibration. These artefacts do not appear uniformly across the field of view, and are less severe in some observations. For example, they were not noticeable in Albert et al. [2020b]. The asymmetric pattern suggests that this is a phase-effect, and the arc-like pattern suggests that it is varies on the scale of hours. If it varied on shorter timescales then the patterns would be more radial as ionospheric scattering artefacts are. Therefore, this suggests that the phase component of the Jones scalar model is missing a slowly varying component.

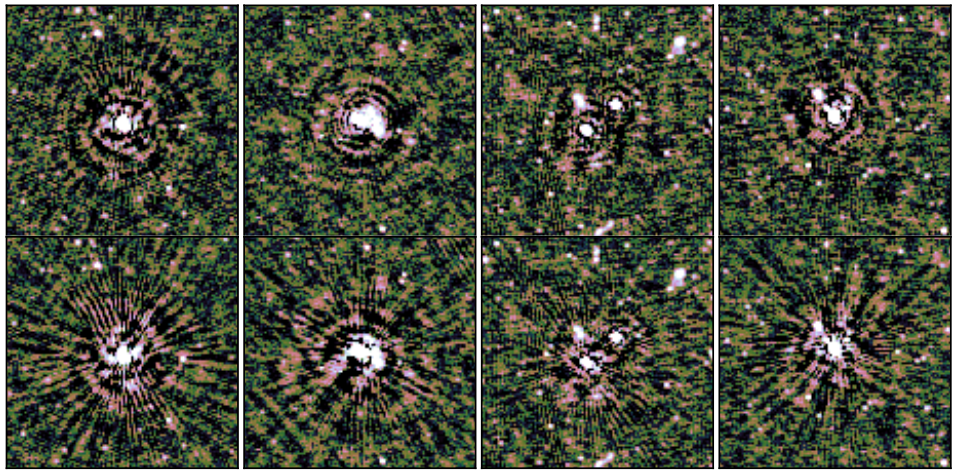
We observe that both in LoTSS-DR2 archival images as well as our images without screens applied, e.g. the LoTSS-DR2 deep image in **Figure** 4.6 and the lower panels in **Figure** 4.3, that there are no arc-like artefacts. This implies that the Jones scalar smoothing performed in LoTSS calibration and step 2 of our method can account for this systematic. In LoTSS calibration the phase model is a two-parameter linear model, a TEC-like $\propto \nu^{-1}$ term and a constant term. In step 2 of our method the phase model is a three-parameter linear model, a TEC-like $\propto \nu^{-1}$ term, a constant term, and a clock-like $\propto \nu$ term. Both LoTSS and step 2 of our method assume slowly changing amplitudes. Note, we introduced a clock-like term upon noticing residual direction independent components in the Jones scalars [see Albert et al., 2020b]. This is fundamentally why we perform directional referencing before inferring DDTEC.

The performance of the method depends on the Jones scalars being modelled correctly, since a DDTEC screen inferred from biased DDTEC will lead to biased screens. The model assumes that the only DD contributions to the Jones scalars are slowly varying amplitudes due to beam errors, and phases $\propto \nu^{-1}$ due to non-diffractive weak scattering in the ionosphere. The temporal aspect of the Jones scalars is modelled with a hidden Markov model using variational inference assuming a linear phase model containing only DDTEC [Albert et al., 2020b].

Global optimisation is guaranteed via a clever basin-hopping routine. This is based on the analytic form of the variational expectation in [Albert et al., 2020b]. The variational



(a) Central region of deep image. Left: LoTSS-DR2 deep archival. Right: our deep image.



(b) Cut-outs of inter-calibrator sources. Bottom row: LoTSS-DR2 deep archival. Top row: our deep image.

Figure 4.6: Comparison between the LoTSS-DR2 archival deep Lockman Hole image and the same data calibrated with our screen-based method. Our deep image does not apply the systematic correction proposed in Section 4.4. Cyan lines and red circles correspond to the LoTSS calibration regions and to our calibrators, respectively. The mean background noise in both the LoTSS-DR2 archival deep image and our deep image is 27 μJy beam^{-1}.

expectation has many local maxima due to phase wrapping. Since we have the analytic form we know precisely how far apart the local minima are, therefore we can simply hop between them to find the global maximum (which is unique).

We search for a slowly varying unmodelled phase systematic, by plotting the residual between the posterior mean phase and the phase of the data. The upper panel of **Figure** 4.7 shows the residual phase over time of a single antenna and direction following Jones scalar modelling. We observe a low-order (in the sense of polynomials), symmetric in frequency, slowly varying underlying structure in the residuals. The phase residuals in this example can be quite large, reaching approximately 0.4 radians.

We perform linear regression for each time-slice and plot the first regression coefficient (the slope) in the lower panel. The regression coefficient has units of radians/MHz and can be converted to an effective phase residual by multiplying by half of the bandwidth. The effective phase residual scale is indicated on the right axis, and reflects the scale of the residuals in the upper panel. From the effective phase residual we can see that this residual changes slowly over the observation, on the scale of hours. This suggests that it is related to the arc-like artefacts.

The regression coefficient is insensitive to phase residuals originating from noise and therefore is a good measure of the systematic. **Table** 4.3 summarises the effective phase residuals for each observation in terms of percentiles of the effective phase residuals. We observe that the distribution of effective phase residuals is heavily tailed. While half of the effective phase residuals are below approximately 0.07 radians, 20% of effective phase residuals are above approximately 0.15 radians and 10% are above approximately 0.20 radians.

Table 4.3: Effective phase residuals per observation.

Obs. ID	Effective phase residual (rad)		
	50-%ile	80-%ile	90-%ile
667218	0.06	0.13	0.20
667204	0.06	0.13	0.19
664480	0.06	0.13	0.19
664320	0.07	0.15	0.23
659948	0.06	0.14	0.21
659554	0.06	0.13	0.20
342938	0.07	0.14	0.19
340794	0.07	0.16	0.23
299961	0.06	0.12	0.18
294287	0.06	0.15	0.23
281008	0.05	0.11	0.16
274099	0.07	0.16	0.25

The phase residuals must originate from one or more DD systematics, since we have directionally referenced the Jones scalars. This is why we infer doubly differential total electron content instead of differential total electron content. **Figure** 4.8 shows an example of the effective phase residual plotted over a field of view. We observe that the effective phase residuals are clearly correlated over direction.

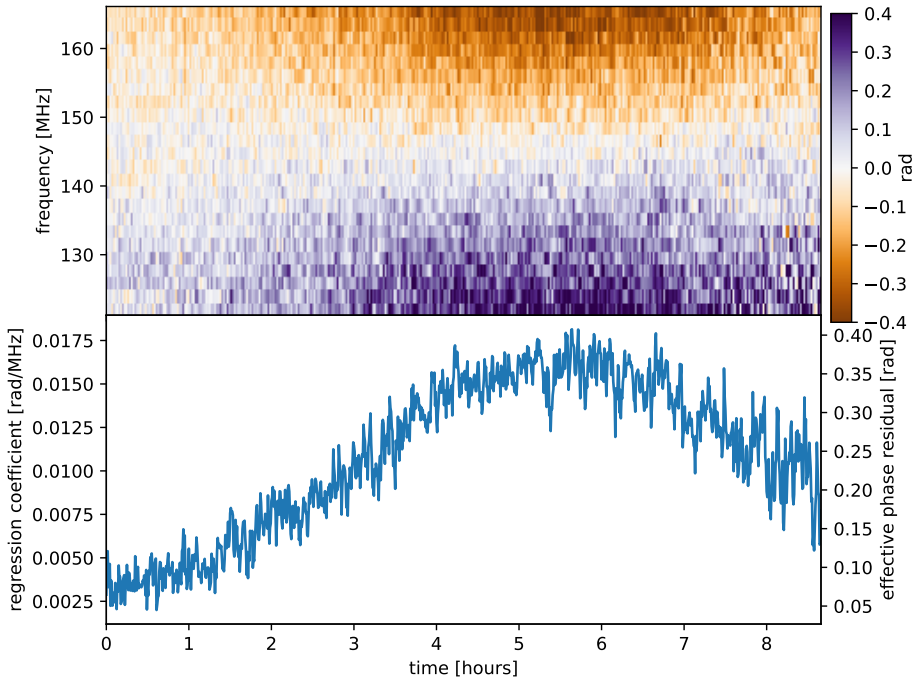


Figure 4.7: Example of unmodelled systematics in the phase residuals. The upper panel shows the phase residuals for a single antenna and direction. The lower panel shows the first regression coefficient (slope) of the phase residuals and the corresponding effective phase residual on the right axis.

Given that there is a correlation over direction and that the effect changes on the time scale of hours there are only a few known candidates for this systematic. The most logical cause is that the beam model is inaccurate. The beam of a LOFAR station is quite complicated since the tiles of the array exhibit electromagnetic coupling, which are not perfectly modelled. Since the beam shape is a function of phase tracking centre, an inaccurate beam model would introduce an effect that changes slowly over the course of an observation.

We propose that the phase residuals can be heuristically modelled by the addition of a constant-in-frequency term in the Jones phase model which changes slowly in time. This proposal is based on the fact that both LoTSS calibration and step 2 of our method produce images without arc-like patterns and both include a constant-in-frequency term in the phase model. Albert et al. [2020b] show how to perform variational Bayesian inference with a hidden Markov model [HMM; Rabiner and Juang, 1986] with any arbitrary linear phase model. In step 3 of the method we use a single linear term, i.e. DDTEC. Here we consider the effects of introducing a constant-in-frequency term to the linear phase model. In particular, we consider the bias-variance trade-off of introducing an additional degree of freedom.

We simulate 100 independent Jones scalar sequences using a two-component linear phase model with DDTEC plus a constant-in-frequency term. We use the HMM to simulate

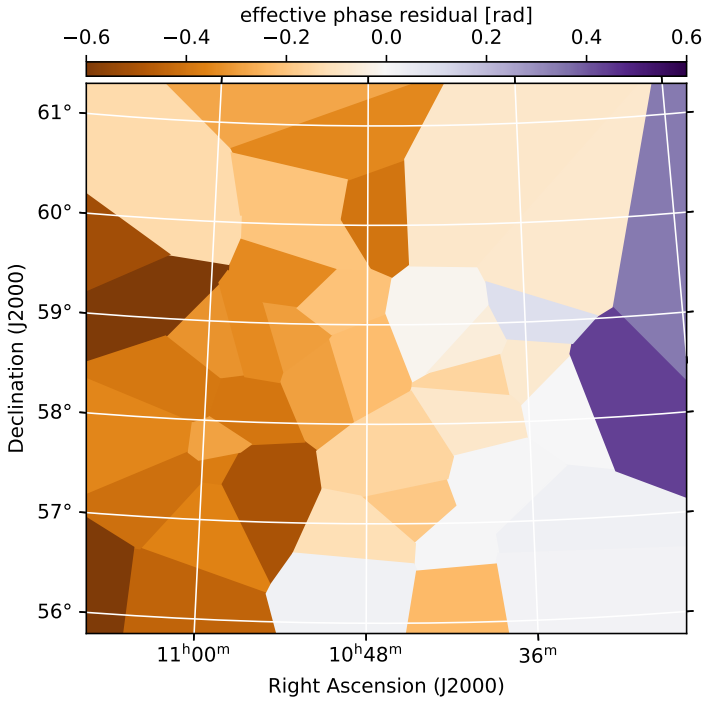


Figure 4.8: Example of the effective phase residual from observation 659948 exhibiting correlation among directions.

the Jones scalars, drawing from the empirical HMM parameter distribution for DDTEC, i.e. the variance of the Gaussian steps and observational uncertainty. We choose a Gaussian step variance of $(0.04\ \text{rad})^2$ for the constant-in-frequency term, which was selected to produce effective phase residuals that resemble the lower panel of **Figure** 4.7.

Given these simulated Jones scalars we perform HMM variational inference with two different linear phase models: 1) a DDTEC-only linear phase model, and 2) a DDTEC plus constant-in-frequency linear phase model. The first model is exactly the same one used in step 3, and the second model is the exact same one that the Jones scalars are simulated with. We then compute the residual between the posterior mean DDTEC and the ground truth.

Figure 4.9 plots the DDTEC residual for both linear phase models as a function of the ground truth constant-in-frequency term. We also compute the effective phase residual for the simulated data and find that there is a scaling relation of 0.157 times the ground truth constant-in-frequency term. The effective phase residual scale is shown on the top axis. We observe that the single component linear phase model (DDTEC-only) results in a residual that depends linearly on the ground truth constant-in-frequency term, with a slope of approximately $-16.7\ \text{mTECU rad}^{-1}$. The variance of the scatter is very small, $(0.75\ \text{mTECU})^2$. We observe that the two component linear phase model results in residuals with no significant dependence on the ground truth constant-in-frequency term. The variance of the scatter is an order of magnitude larger that the single component model, $(15\ \text{mTECU})^2$.

The first model results in high-bias low-variance residuals, while the second model results in low-bias high-variance residuals. The high-bias of the first model follows because the missing component in the phase model shifts the global optimum away from the ground truth, whereas the high-variance of the second model follows because we have introduced twice as many degrees of freedom given the same amount of information in the data. The high-bias of the first model implies that for antennas with phase residuals such as the one in **Figure** 4.8, that the inferred DDTEC is offset by as much as 60 mTECU, and this offset changes very slowly over the course of an observation. We suspect that this gives rise to the asymmetric arc-like patterns. Furthermore, while the effective phase residuals might seem deceptively small, $\ll 1$ rad, the effect on the inferred DDTEC is quite large.

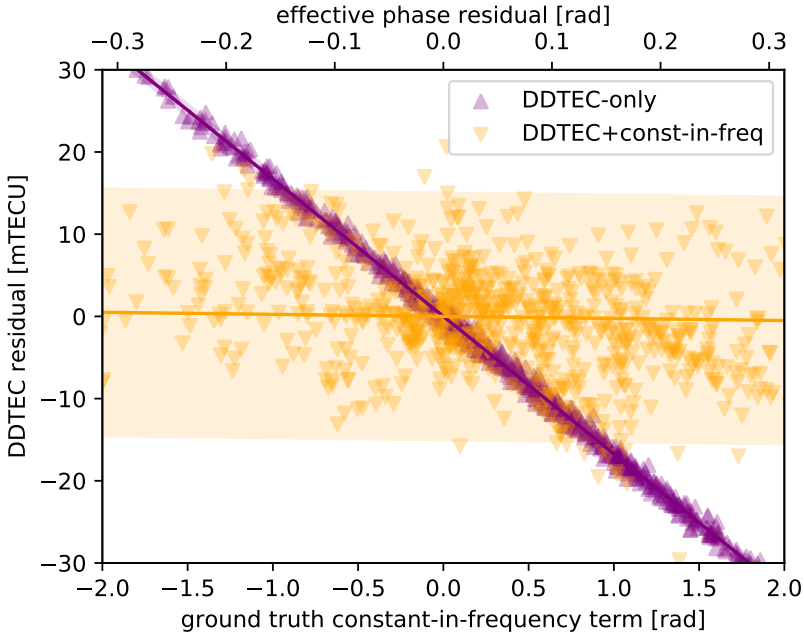


Figure 4.9: Deviations of the inferred DDTEC from the ground truth for two linear phase models. The single component model (purple) is the DDTEC-only model, and the two component model (orange) is the DDTEC and constant-in-frequency model. The shaded regions are 95% confidence regions.

The high-variance of the second model implies that we cannot simply add the constant-in-frequency term to the HMM in step 3, since the resulting DDTEC would be useless for screen-based modelling. We propose an alternative method of debiasing based on the effective phase residual and a strong prior on the smoothness. We use the scaling relation between the effective phase residual and the ground truth constant-in-frequency term to get a noisy estimate of the constant-in-frequency term. We then apply a median filter over a time window of 2.5 hours with reflecting boundaries to produce a slowly varying smooth estimate of the constant-in-frequency term. We then subtract this estimate of the constant-in-frequency term and resolve again with the DDTEC-only model. This method requires running

the HMM inference twice, once to get the effective phase residuals, and a second time once the phases have been debiased. Since this systematic is DD, we also interpolate the constant-in-frequency term to the screen directions. We apply nearest-neighbour interpolation as a first-order approximation.

Figure 4.10 shows a comparison of inter-calibrator sources significant arc-like artefact before and after debiasing. We observe the arc-like artefacts are suppressed, however a low-level diffuse artefact appeared. This confirms that the arc-like artefacts were indeed due to these phase residuals. The fact that the arc-like pattern disappears, despite the resulting diffuse artefacts suggests that our general approach to debiasing is correct, however there is still work to be done to perfect the method. For example, good outlier flagging proved critical for the DDTEC before inferring the screen, therefore perhaps outlier detection and flagging with these constant-in-frequency terms using spatial information would improve the results. Also, the timescale of 2.5 hours may be too long. Given that the resulting diffuse artefact is large scale, they likely stem from the treatment on short baselines, i.e. central antennae.

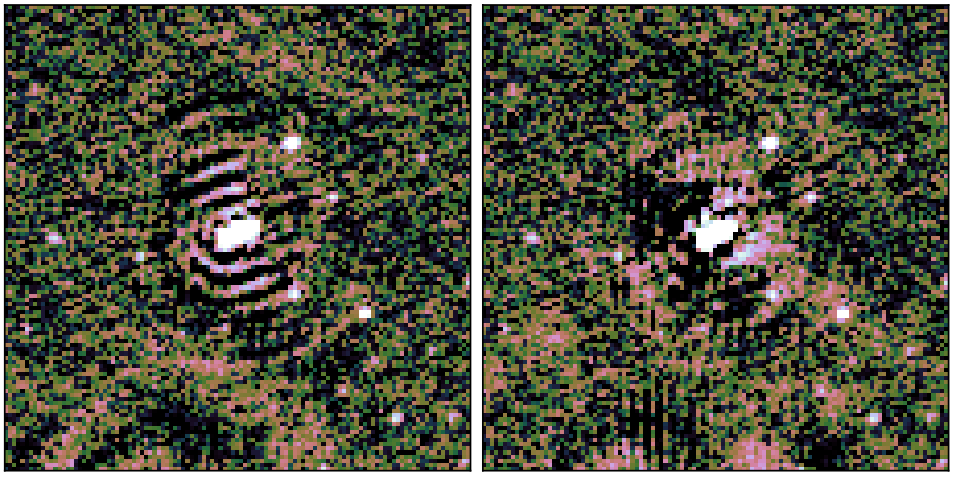
4.5 Information in DDTEC screens

An interesting product of this method is the rich view of the ionosphere's behaviour. **Figure** 4.11 shows a random selection of DDTEC screens across the Lockman Hole data set, organised according to similarity of features. The posterior uncertainty in these DDTEC screens is $\lesssim 1$ mTECU. The screens were randomly selected from antennae at least 1 km from the reference antenna. Because they are randomly sampled they provide a cross section of the common types of ionosphere varieties. We see that features ranges from simple gradients across the field of view, to wave-like patterns, to higher-order aperiodic features. This supports the fact that there are many distinct behaviours in the ionosphere [e.g. Mevius et al., 2016, Jordan et al., 2017]. This is one of the most detailed views of the ionosphere ever made in terms of angular (sub-arcminute) and temporal (sub-minute) resolution, as well as precision of the DDTEC measurements (sub-mTECU). This zoo of DDTEC screens also allows us to explore the limits of our screen-based method. In particular, it enables us to explore in the future how sparse the calibrators can be selected such that we can still recover these screens.

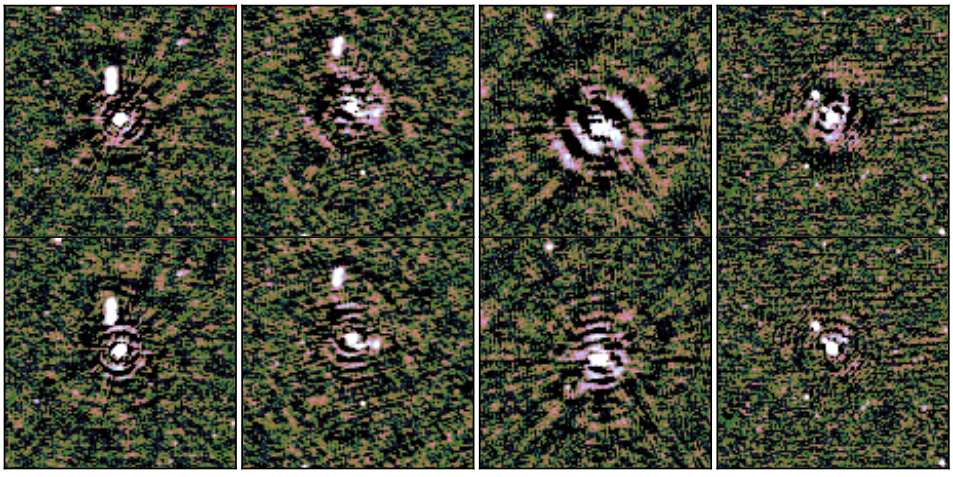
One of the tantalising future applications of this method is DD calibration of LOFAR-LBA. While it is beyond the scope of this paper to give an in-depth prediction for how feasible this will be, we can already draw conclusions from the variety of structures seen in **Figure** 4.11. Specifically, we note that roughly 20% of the DDTEC screens have wave-like patterns and another 20% have rough small-scale structure. Such prevalence of small-scale structure suggests that LOFAR-LBA, with typically less than ten bright calibrators with good quality Jones scalars per field of view[2], may struggle to perform DD calibration. This suggests DD calibration of LOFAR-LBA may benefit from the planned LOFAR expansion which would allow dual HBA-LBA observations, where the DDTEC screens from HBA can be applied to LBA.

As observed in Albert et al. [2020a], the full tomographic application is able to perform inference on sub-calibrator spacing. This suggests that a valuable future direction would be to make the full tomographic method computationally feasible for application to LOFAR-LBA.

[2] Private communications with W. Williams.



(a) Left: before applying debiasing. Right: After applying debiasing. Source comes from observation 342938.



(b) Cut-outs of inter-calibrator sources. Bottom row: before applying debiasing. Top row: After applying debiasing. Sources come from observations (left to right): 281008, 340794, 299961, 342938.

Figure 4.10: Comparison of inter-calibrator sources with significant arc-like patterns before and after applying our proposed debiasing correction.

On top of this, we would also like to understand what further improvements to the screen model would be beneficial at better handling some of the more difficult ionospheres, e.g. non-stationary FED covariances, or curvature of Earth.

Since the ionospheric model has undergone several approximations [see Albert et al., 2020b] for computational efficiency, we no longer perform tomography. Therefore, it is not possible to learn about most of the ionosphere's physical parameters. In principle, it is still possible to constrain the ratio of irregularity correlation scale to height of the ionosphere. However, since we have performed a FED hypothesis marginalisation it is not straight forward to interpret these parameters. Furthermore, as shown in Section 4.4, the modelling systematic can result in DDTEC that differ from the ground truth as much as 60 mTECU. It would therefore be premature to interpret the DDTEC screens for ionospheric science.

4.6 Conclusion

We have tested the probabilistic screen-based method of Albert et al. [2020b] on 12 observations (100 hours) of Lockman Hole LOFAR HBA data, which covers a wide range of ionospheric conditions, and compared the results to the calibration procedure of LoTSS-DR2. We find that 10/12 images showed significant improvement beyond the method of LoTSS-DR2, while 2/10 observations showed as-good or only slightly better improvements. We suggest that these moderate improvements are due to unresolved small scale structure in the ionosphere, or non-stationary structure. This implies that the method robust to most ionospheric conditions.

We have quantified the improvements from screen-based modelling in terms of equivalent integration-time gain around bright sources, which is the extra relative observation time that would be needed using the LoTSS-DR2 calibration method to reduce the scattering artefacts to the level that the screen achieves. We find that when the free electron density is high, due to increased Sun activity, measured in terms of Sun-spots, the equivalent integration-time gain is on average two. That is, the method achieves scattering effects equivalent to an observation twice as long. We also jointly imaged all observations, and compared this the to deep LoTSS-DR2 image. We find that similar improvements also scale to deep images, though modelling systematics build up and increase the noise around bright sources.

We discovered a modelling systematic that results in asymmetric arc-like artefacts around inter-calibrator sources. The origin of the phase component is not yet known, though we suggest it could be related to beam model inaccuracies. We propose that it can be accounted for as an additional DD constant-in-frequency phase component in the Jones phase model. Using a combination of real and simulated data we show that including a constant-in-frequency term in the HMM results in high-variance DDTEC inference unsuitable for screen inference. Therefore, an alternative method was proposed to account for the bias. We propose a simple filtering method, based on the slowly varying nature of the systematic, to remove it from the data. We show that this simple procedure removes the arc-like patterns, but introduces low-level diffuse artefacts. This suggest the general approach to accounting for the systematic is correct, but more work is needed to perfect it.

Once this systematic is robustly accounted for, the results of this paper suggest our screen-based method is feasible for reprocessing the entire LoTSS archive. Future improvements to the screen-based method, that would improve the robustness and also make it applicable to

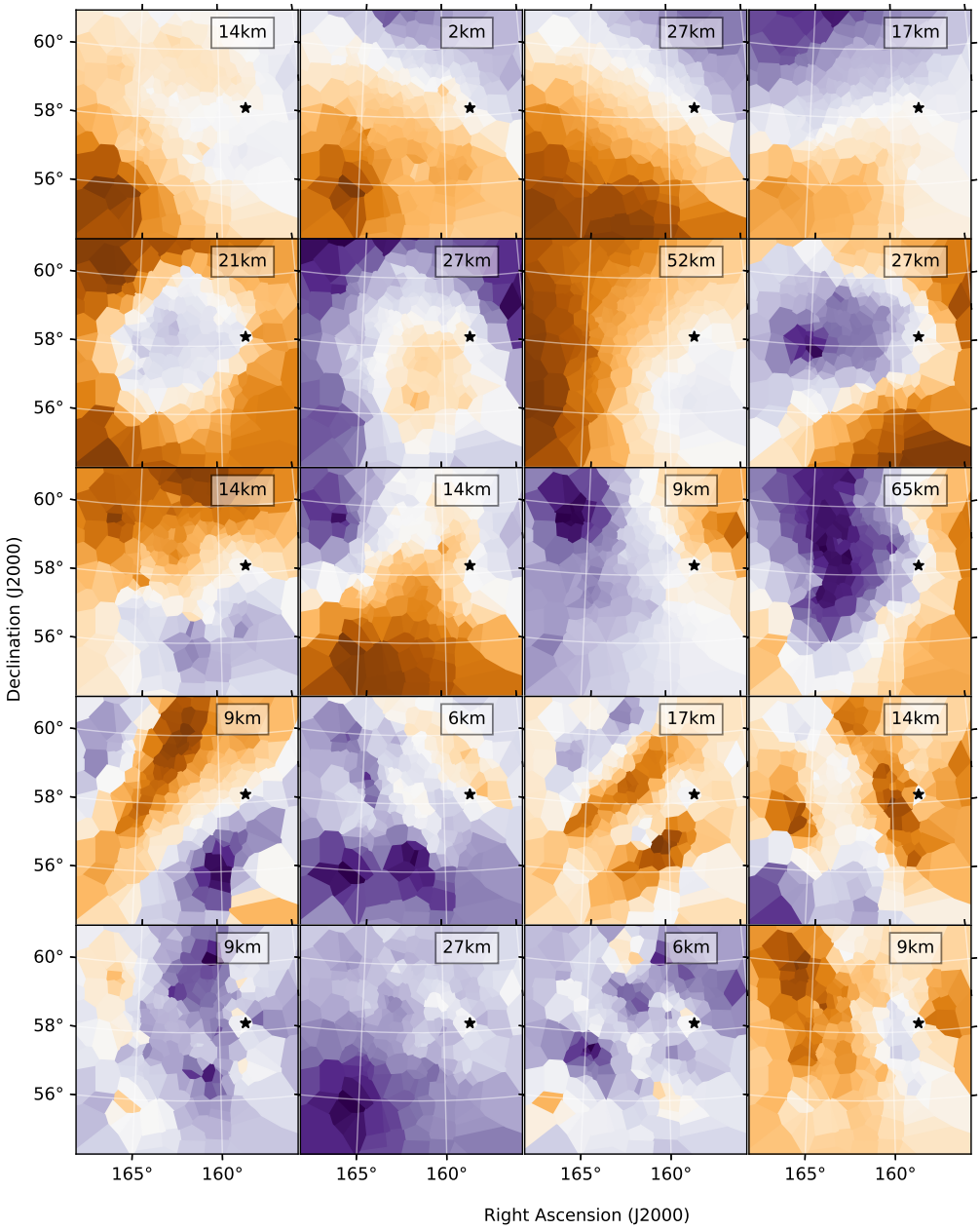


Figure 4.11: *Caption continued on next page.*

Figure 4.11: A zoo of DDTEC screens. Each panel is a DDTEC screen with normalised scale. The posterior uncertainty in these DDTEC screens is $\lesssim 1$ mTECU. The screens are randomly taken throughout the Lockman Hole data set from antennae at least 1 km from the reference antenna. They have been organised according to similar features ranging from simple gradients, to wave-like structures, to rough nearly uncorrelated structures. The black star is the reference direction. The distance from the reference antenna is shown in the box.

LOFAR-LBA, would be to relax the thin-layer approximation so that the model becomes fully (non-diffractive) tomographic as originally proposed in Albert et al. [2020a].