SDP Memo: 027 The Infeasibility of High Quality Ionospheric Calibration of SKA1-LOW: response to comments

Document number……………………………………………………………………………….SDP Memo 27
Document Type…………………………………………………………………………………………MEMO
Revision……………………………………………………………………………………………………..1A
Author…………………………………………………………………………………………… T. Cornwell
Release Date………………………………………………………………………………………..08/09/2016
Document Classification…………………………………………………………………………Unrestricted

Lead Author	Designation	Affiliation
Tim Cornwell	Author	TCC
Signature & Date:	Signature: Email: realtimcornwell@gmail.com	

SDP Memo Note

The SDP memos are designed to allow the quick recording of investigations and research done by members of the SDP. They are also designed to raise questions about parts of the SDP design or SDP process. The contents of a memo may be the opinion of the author, not the whole of the SDP.

Executive summary

In SDP Memo 26, we presented a theoretical framework for understanding and predicting the scientific performance of low frequency arrays such as LOFAR and SKA1-LOW. We argued for a specific approach to calibration in which a smooth model for the ionosphere is determined. This allows calculation of the imaging dynamic range limit due to inevitably limited linear resolution of the ionospheric phase screen. The approach relied upon a two phase approach in which pierce point phases were solved, and then a phase screen fit using Zernike polynomials. Using this approach, we concluded that calibration of SKA1-LOW with the accuracy required for EOR imaging not possible within the life time of the telescope.

Following publication of Memo 26, we received a number of comments on ways to improve the performance, and also notice of analyses giving quite different results. We respond to these constructive comments with improvements to our modelling. We find that the conclusion of Memo 26 is still supported by the modelling: SKA1-LOW cannot be calibrated well enough that EOR imaging can be concluded in a reasonable time period.

To aid this second round of modelling, we present a new, improved approach to estimation of the phase screen that is linear, and therefore has more easily understood properties. It is apparent that the measurement equations for the visibility are non-linear in the screen phase. However, the pierce point approach does work well enough to reduce the typical phase error to much less than a radian and consequently the equations can be linearized. Each visibility then provides a (complicated and entangled) constraint on the phase screen. This allows the use of standard linear algebra techniques to solve for a compact representation of the phase screen. The linearity also means that much weaker sources can be used in the phase screen estimation – an important property since calibrating from bright sources only leads to bias (see e.g. [RD21]).

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Acronym	**Definition**
LOFAR	Low Frequency Array
LOW	Low frequency component of SKA1
SDP	Science Data Processing

1. Purpose of the document

The purpose of this of this memo is to investigate the calibration of the ionosphere above SKA1-LOW and the impact on imaging. This follows on from previous work [RD1].

2. Scope of the document

This memo has estimation of the imaging performance of the EOR as the main scientific goal. Thus the problem addressed is the imaging of a smooth extended signal on a few degree scale seen through a turbulent and evolving ionosphere and bright foreground sources.

We introduce a number of refinements of the model used in Memo 26. We present a new partially linearized approach for calibration, and analysis the performance of this approach.

3. References

[RD1] Cornwell, T.J., (2016), SKA SDP Memo 26

[RD2] http://www.timcornwellconsulting.com/Memo26/

[RD3] J. D. Bregman, “System Design and Wide-field Imaging Aspects of Synthesis Arrays with Phased Array Stations,” *Ph.D. thesis*, p. 123, Dec. 2012.

[RD4] S. Bhatnagar, T. J. Cornwell, K. Golap, and J. M. Uson, “Correcting direction-dependent gains in the deconvolution of radio interferometric images,” *Astron. & Astrophys.*, vol. 487, pp. 419–429, 2008.

[RD5] Noll R.J. “Zernike polynomials and atmospheric turbulence” J.Opt.Soc.Am. vol 66 No. 3 pp207 1976.

[RD6] H. K. Vedantham and L. V. E. Koopmans, “Scintillation noise power spectrum and its impact on high redshift 21-cm observations,” *arXiv.org*, vol. astro-ph.IM, no. 3. pp. 3099–3117, 01-Dec-2015.

[RD7] R. J. van Weeren, W. L. Williams, M. J. Hardcastle, T. W. Shimwell, D. A. Rafferty, J. Sabater, G. Heald, S. S. Sridhar, T. J. Dijkema, G. Brunetti, M. Brüggen, F. Andrade-Santos, G. A. Ogrean, H. J. A. Röttgering, W. A. Dawson, W. R. Forman, F. de Gasperin, C. Jones, G. K. Miley, L. Rudnick, C. L. Sarazin, A. Bonafede, P. N. Best, L. Birzan, R. Cassano, K. T. Chyzy, J. H. Croston, T. Ensslin, C. Ferrari, M. Hoeft, C. Horellou, M. J. Jarvis, R. P. Kraft, M. Mevius, H. T. Intema, S. S. Murray, E. Orru, R. Pizzo, A. Simionescu, A. Stroe, S. van der Tol, and G. J. White, “LOFAR Facet Calibration,” ApJS, vol. 223, no. 1, p. 2, Mar. 2016.

[RD8] L Koopmans, et al., “The Cosmic Dawn and Epoch of Reionisation with SKA.”, Proceedings of Advancing Astrophysics with the Square Kilometre Array (AASKA14) 9 -13 June, n/a 2015 p. 1.

[RD9] Mevius, M. et al., “Probing Ionospheric Structures using the LOFAR radio telescope,” vol. astro-ph.IM. 15-Jun-2016.

[RD10] S. Bhatnagar, T. J. Cornwell, K. Golap, and J. M. Uson, “Correcting direction-dependent gains in the deconvolution of radio interferometric images,” *Astron. & Astrophys.*, vol. 487, pp. 419–429, 2008.

[RD11] Noll R.J. “Zernike polynomials and atmospheric turbulence” J.Opt.Soc.Am. vol 66 No. 3 pp207 1976.

[RD12] A. Glindemann, S. Hippler, T. Berkefeld, and W. Hackenberg, “Adaptive Optics on Large Telescopes,” *Experimental Astronomy*, vol. 10, no. 1, pp. 5–47, Apr. 2000.

[RD13] S J Wijnholds, J D Bregman, and A van Ardenne, “Calibratability and its impact on configuration design for the LOFAR and SKA phased array radio telescopes”, Radio Science, 2011 vol. 46 (5) pp. RS0F07-n/a

[RD14] R J van Weeren et al., The Astrophysical Journal Supplement Series 2016 vol. 223 (1) p. 2

[RD15] Intema, H T and Van der Tol, S and Cotton, W D, “Ionospheric calibration of low frequency radio interferometric observations using the peeling scheme-I. Method description and first results”, A&A, 501, 1185–1205 (2009)

[RD16] Intema, H. T. (2014). SPAM: Source Peeling and Atmospheric Modeling. Astrophysics Source Code Library

[RD17] Dillon, J.~S. and Parsons, A.R. (2016), “Redundant Array Configurations for 21 cm Cosmology", ArXiv e-prints, 1602.06259

[RD18] Martin, Poppy L and Bray, Justin D and Scaife, Anna M M, “Limits on the validity of the thin-layer model of the ionosphere for radio interferometric calibration”, arXiv.org, 1604.03810v1

[RD19] Loi, Shyeh Tjing et al., 2016, “Density duct formation in the wake of a travelling ionospheric disturbance: Murchison Widefield Array observations”, Journal of Geophysical Research: Space Physics, 211, 1569-1586

[RD20] Sanaz Kazemi, Sarod Yatawatta, and Saleem Zaroubi (2013), “Clustered Calibration: An Improvement to Radio Interferometric Direction Dependent Self-Calibration”, Monthly Notices of the Royal Astronomical Society, Volume 430, Issue 2, p.1457-1472

[RD21] Ajinkya H Patil, Sarod Yatawatta, Saleem Zaroubi, Léon V E Koopmans, A G de Bruyn, Benedetta Ciardi, Ilian T Iliev, Maaijke Mevius, Vishambhar N Pandey, and Bharat K Gehlot “Systematic biases in low frequency radio interferometric data due to calibration: the LOFAR EoR case,” vol. astro-ph.IM. 24-May-2016.

[RD22] Glenn D. Boreman and Christopher Dainty, "Zernike expansions for non-Kolmogorov turbulence," J. Opt. Soc. Am. A 13, 517-522 (1996)

[RD23] Trott, C. (private communication)

[RD24] Wijnholds, S. (private communication)

5. Reflections on Memo 26

We have improved the model presented in Memo 26 in the following ways:

- We have used Bregman's source counts [RD3] in place of the Condon model
- We have corrected errors in the noise calculation, using the tabulations in Baseline Design version 2.
- In Memo 26, we used a diffractive scale of 14km [RD9]. Values of between 3 and 30km occur in the data of [RD9]. 90% of observations have the diffractive scale > 5km but only O(20%) have diffractive scale > 14km. Therefore, we will take a smaller number, 7km, as a representative value. "Lucky imaging" by using only observations with large diffractive scale will lead to inordinately long elapsed observing times. A more careful assessment of the impact of variable diffractive scale would be warranted in the future.

Readers offered a number of critiques of Memo 26 (see [RD2]).

- The bandwidth assumed was arguably too small (0.1MHz). Using more bandwidth brings more sources into contention. LOFAR uses between 0.187MHz [RD9] and 1 MHz [RD7]. For example, 10MHz raises the number of sources to 119 and would be expected to bring substantial improvements in the maximum value of J recoverable. However, the pierce point estimation is only stable if the number of sources is much less than the number of antennas. Thus the criticism is correct but it requires an approach that does not rely on the fitting of pierce point phases.
- I was asked if peeling is included in Memo 26. It is. All sources brighter than 10 times the minimum pierce point candidate are removed before pierce point analysis, assuming that the peeling is accurate. I have increased this to 100 times the minimum.
- The observing efficiency of 30% was widely thought to be too optimistic. I have now adopted 10%. This represents the fraction of the year that a field to be imaged is seen in the night (to avoid the sun) and above 45 degrees' elevation (the limit on elevation angle from the L1 requirements), and the ionosphere has good behaviour.
- Other analyses of the recovery of just a TID phase screen (i.e. a sinusoid) showed that that problem was well-conditioned [RD23][RD24]. That is entirely plausible because of the small number of parameters required compared to those needed to represent a turbulent phase screen. Such analysis tells us little about the stability of estimating an entire spectrum of phase errors.

Our conclusions after considering these and other suggestions and comments was that:

- We should represent the screen directly rather than via pierce points,
- We should use fewer parameters for the array core, or alternatively look for an approach that assesses how many parameters are warranted,
- We should search in λ^2 space for the optimum TEC and then average over frequency (which is what I call tracking in the memo).

6. Degrees of freedom

To emphasize the difficulty of imaging through the ionospheric screen, we can use the analysis of Memo 26 to estimate the number of Zernike terms required. The Noll formula for the residual phase error is:

Equation 1

$$\sigma_J^2 \sim 0.2944\, J^{-\frac{\sqrt{3}}{2}} \left(\frac{B}{r_0}\right)^{\frac{5}{3}} \qquad \text{rad}^2$$

This can be solved for the number of Zernike's required as a function of observing time and dynamic range (see Figure 1). For example, reaching 50 dB in 6 months observing (5 years elapsed at 10% efficiency) requires Zernike's up to 18,000. The azimuthal variable m is about 240 so the fringe around the edge of the ionospheric disk goes through 240 cycles in π 55km so the resolution is about 0.6km – close to the Fresnel radius.

However, for such large order polynomials, the power is concentrated close to the edge of the unit disk, which in Memo 26 we took to be the first zero point, and so the analysis becomes less and less sensitive to high order Zernike's. Using substantial bandwidth, say 10% or more, would provide measurements with the primary beam null in different locations.

Another possibility is to move away from the Zernike polynomial basis functions and adopt basis functions that have less extreme behaviour.

The number of 18,000 Zernike coefficients is a measure of the information needed from each 10s integration in order to calibrate the ionospheric screen well-enough that 50dB dynamic range can be reached in 6 months observing (about 5 years elapsed time). The number of pierce points can be no smaller than this number and if we were to have 512 optimally located stations, the number of sources would obey:

$$N_{sources} \gg 36$$

If the stations are not optimally distributed, then the required number of sources could be substantially higher. We have also argued that the pierce point analysis becomes ill-conditioned and thus unstable unless the number of sources is much less than the number of stations. So the constraint is:

$$N_{stations} \gg N_{sources} \gg 36$$

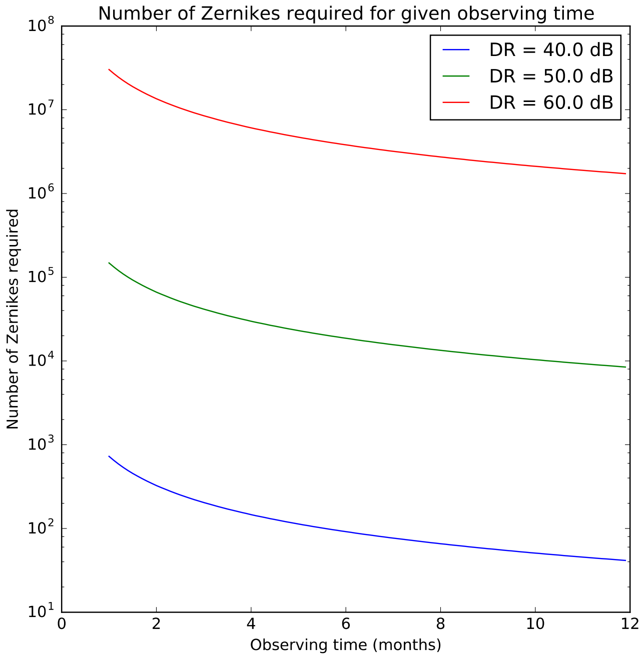


Figure 1 Number of Zernike terms needed to achieve a given dynamic range for a given observing time. The elapsed time will be about ten times longer. Thus to achieve 50dB in 6 months requires about 18,000 Zernike polynomials. At 100MHz, Fresnel diffraction starts to dominate for J>20,000.

However, all of this analysis is for a Kolmogorov spectrum whereas recently published observations of the ionosphere with LOFAR [RD9] indicate that the exponent of the phase error with distance is typically higher than the 5/3 expected from a Kolmogorov spectrum, being closer to 1.8. This represents less power in the higher order modes, leading to the possibility that the phase error may converge for more moderate numbers of Zernike polynomials. The formula for the residual phase variance is then a function only of the parameter n instead of the Noll parameter J:

Equation 2

$$\sigma_{n,m}^2 \sim \frac{(n+1)}{\pi} \frac{\sin\left(\pi\left(\frac{\beta-2}{\beta}\right)\right)\Gamma\left(\frac{2n+2-\beta}{2}\right)\Gamma\left(\frac{\beta+4}{2}\right)\Gamma\left(\frac{\beta}{2}\right)}{\Gamma\left(\frac{2n+4+\beta}{2}\right)} \left(\frac{B}{r_0}\right)^{\beta-2} \qquad \text{rad}^2$$

For consistency, we will continue to use the original Noll formula, but check it against this formula. We will find that the differences are negligible.

7. Visibility fitting

In [RD1] we described a two phase approach in which first the values for the pierce point phases were solved using a non-linear LSQ approach, and second a linear approach fit converted these to estimates of the screen. As the number of pierce points approaches the number of stations, the first step becomes unstable and the entire approach fails in estimating the screen in sufficient accuracy. However, the approach is still useful in obtaining approximate phase coherence across the field as long as the number of pierce points is much less than the number of stations.

Consider a point source of known flux S and location σ, seen through an ionospheric phase screen at height h. The screen phase obeys the following equation:

Equation 3

$$\theta(x) = T(x)\lambda^2$$

where T is the Total Electron Content along the line of sight, and λ is the observing wavelength. The visibility contribution from the point source is:

Equation 4

$$V(x_1, x_2) = Se^{j\frac{2\pi}{\lambda}(x_1 - x_2).\sigma + \lambda^2\left(T(x_1 + h\sigma) - T(x_2 + h\sigma)\right)}$$

Following this logic, we will assume that the approach described in [RD1] is used to estimate the gross values of the TEC screen so that the phase error is always much less than one radian. We will use the same symbol T to represent the residual error in the TEC. In this circumstance, the equation can be linearized as:

Equation 5

$$V(x_i, x_j) = 2\pi j \sum_S Se^{j\frac{2\pi}{\lambda}(x_i - x_j).\sigma}\lambda^2\left(T(x_i + h\sigma) - T(x_j + h\sigma)\right)$$

We cannot neglect the effects of Fresnel diffraction (see e.g. [RD6]). This term introduces an additional phasor:

Equation 6

$$e^{j\frac{\pi}{\lambda h}\left(x_i^2 - x_j^2\right)}$$

So the visibility equation becomes:

Equation 7

$$V(x_i, x_j) = 2\pi j \sum_S Se^{j\left(\frac{2\pi}{\lambda}(x_i - x_j).\sigma + \frac{\pi}{\lambda h}\left(x_i^2 - x_j^2\right)\right)}\lambda^2\left(T(x_i + h\sigma) - T(x_j + h\sigma)\right)$$

We use a least squares approach to estimate the TEC screen from the measured visibilities and the known sources. This single complex-valued equation leads to two real-valued constraint equations on the coefficients of the Zernike polynomial expansion of the TEC screen. We weight these by square root of the inverse variance so the output singular value spectrum can be interpreted as signal-to-noise ratio.

Equation 8

$$T(x_i) = \sum_{J} a_j \; Z(j, x_i)$$

Equation 9

$$V(x_i, x_j) = 2\pi j \sum_{S} S e^{j\left(\frac{2\pi}{\lambda}(x_i - x_j).\sigma + \frac{\pi}{\lambda h}\left(x_i^2 - x_j^2\right)\right)} \lambda^2 \left(\sum_{J} a_j \; Z(J, x_i + h\sigma) - \sum_{J} a_j \; Z(J, x_j + h\sigma) \right)$$

$$V(x_i, x_j) = 2\pi j \sum_{S} S e^{j\left(\frac{2\pi}{\lambda}(x_i - x_j).\sigma + \frac{\pi}{\lambda h}\left(x_i^2 - x_j^2\right)\right)} \lambda^2 \left(\sum_{J} a_J \; Z(J, x_i + h\sigma) - \sum_{J} a_J \; Z(J, x_j + h\sigma) \right)$$

$$V(x_i, x_j) = 2\pi j \sum_{S} S e^{j\left(\frac{2\pi}{\lambda}(x_i - x_j).\sigma + \frac{\pi}{\lambda h}\left(x_i^2 - x_j^2\right)\right)} \lambda^2 \left(\sum_{J} a_J \; Z(J, x_i + h\sigma) - \sum_{J} a_J \; Z(J, x_j + h\sigma) \right)$$

Equation 10

$$C1 = +2\pi \sum_{S} \left(\frac{S}{\sigma_V}\right) \sin\left(\frac{2\pi}{\lambda}\left(x_i - x_j\right).\sigma + \frac{\pi}{\lambda h}\left(x_i^2 - x_j^2\right)\right) \lambda^2 \, Z_J(x_i + h\sigma)\,;\ \forall i, j$$

$$C2 = +2\pi \sum_{S} \left(\frac{S}{\sigma_V}\right) \cos\left(\frac{2\pi}{\lambda}\left(x_i - x_j\right).\sigma + \frac{\pi}{\lambda h}\left(x_i^2 - x_j^2\right)\right) \lambda^2 \, Z_J(x_i + h\sigma)\,;\ \forall i, j$$

The term $Z\left(x_j + h\sigma\right)$ is the normalised Zernike polynomial evaluated at the pierce point for the source as seen from station j. Using Zernike polynomials has some drawbacks. First, the computational load is high, though this can be reduced by optimization and parallelization (see Appendix A). Second, the Zernike polynomials are most often used for low order aberrations. In this case, we will have need of Noll numbers of up to 18,000! However, the same equations work for other expansions, although we will not necessarily have a closed form for the residual phase error. In that case, Monte Carlo techniques can be used to find the residual phase error.

Each visibility sample provides one of these linear equations. So the total number of constraints is:

Equation 11

$$N_{constraints} = N_{stations}(N_{stations} - 1)$$

The number of degrees of freedom is given by the number of Zernike's (at the highest 20,000), and the number of constraints is twice the number of visibilities (typically 500,000). Hence this linearized approximation is very over-constrained.

In addition, by using a number of frequency channels, we can use complementary information to obtain the TEC. So the total number of constraints is:

Equation 12

$$N_{constraints} = N_{channels}N_{stations}(N_{stations} - 1)$$

Equation 13

$$N_{dof} = J_{max}$$

The total number of sources used in the linear part of the process is constrained by computational limitations. (see Appendix A for more detail of the optimisations performed). As the number of sources goes up, the number of Zernike's to be evaluated rises in proportion and also the highest Zernike required rises.

By comparison, faceting uses the visibility as the primary constraints but then derives a constraint for each station and pierce point.

Equation 14

$$N_{constraints} = N_{channels}N_{stations}N_{sources}$$

This neglects the inevitable coupling between different constraints. To control the coupling, we have argued that the number of pierce points. must be less than the number of stations, typically 512. Wavelength diversity can be used to improve the number of constraints. So we have that:

Equation 15

$$N_{constraints} \ll N_{channels}N_{stations}(N_{stations} - 1)$$

Equation 16

$$N_{dof} = J_{max}$$

Ignoring for a moment the possibility of using wavelength diversity, we can see that the real advantage of direct fitting to the visibilities is that more than $N_{stations}$ sources can be used. In Figure 2 and Figure 3, we show the number of sources available for each approach, as calculated from the visibility noise level, ionospheric coherence time, and source counts. The threshold for PPC is a source having piercing point SNR at least 10.0, and the threshold for Direct Fitting to Visibilities (DFV) is a source having image SNR (in the ionospheric coherence time) not less than 10.0. Figure 4 shows the same behaviour for two frequencies, 50MHz and 100MHz.

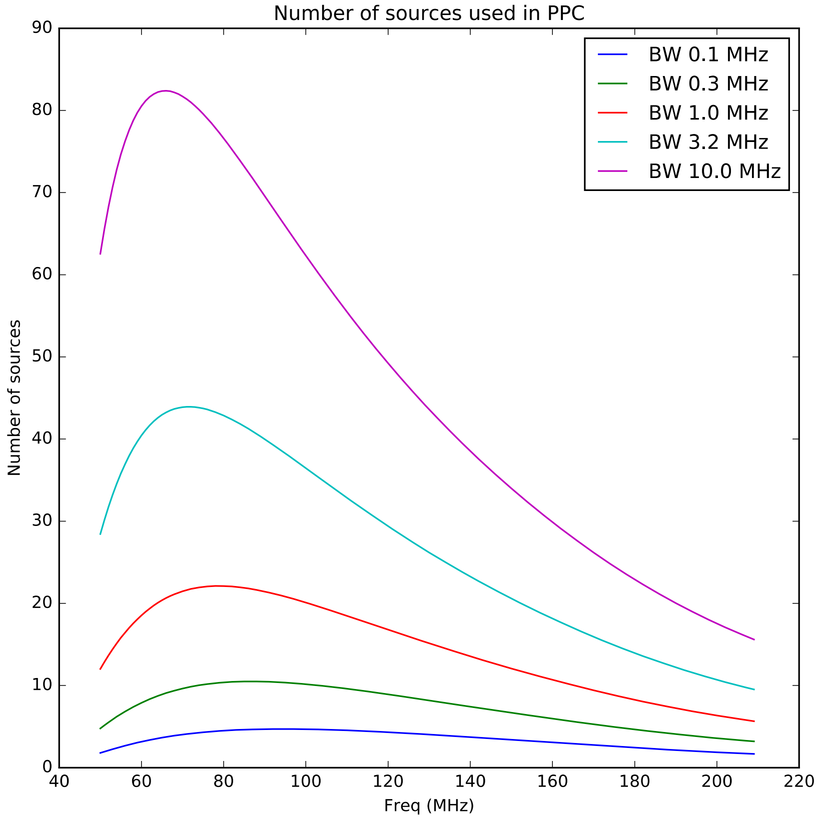


Figure 2 Number of sources available for Piercing Point Calibration (PPC) as function of observing frequency and bandwidth

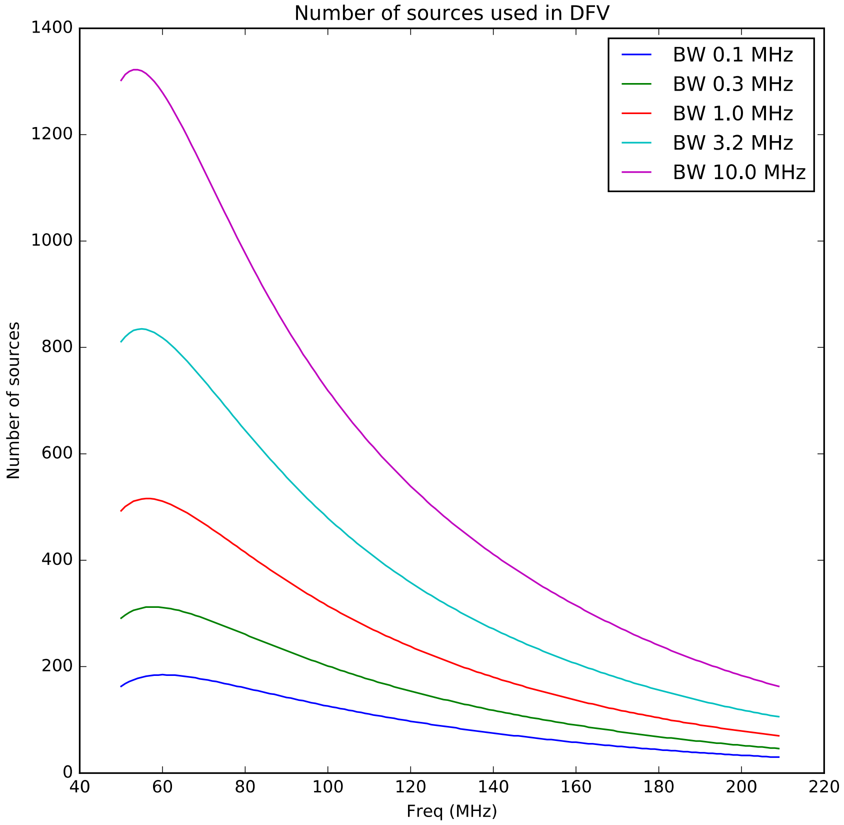


Figure 3 Number of sources available for Direct Fitting Visibilities (DFV) as function of observing frequency and bandwidth

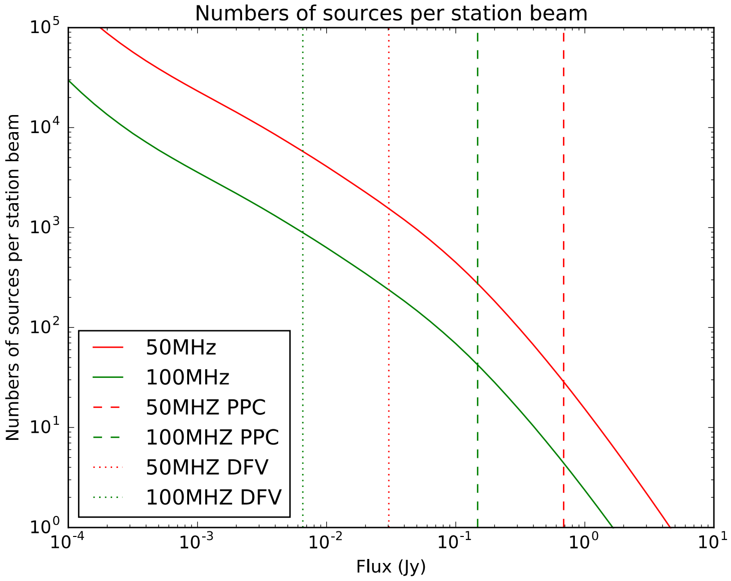


Figure 4 Number of sources per station beam and noise level for 50 and 100 MHz, with bandwidth 1 MHz . This replaces the corresponding figure in M26 where the bandwidth used was 100Khz. In addition, the source counts are those derived by Bregman [RD3]. The thresholds for Pierce Point Calibration (PPC) and Direct Fitting Visibilities (DFV) are both shown.

Table 1 Logical steps in determining the dynamic range limitation due to incomplete ionosphere phase screen estimation, solving directly for the phase screen from the visibilities.

Step	Yields	Comment
• Use pierce point analysis to make the array coherent in 10s	Linearized measurement equation in the phase screen representation	Only requires limited number of source and so should be stable
• Model phase screen by orthonormal polynomials	Orthogonal basis for phase screen	Could use a similar basis, such as discrete cosine. If so, the residual phase error might have to be computed by Monte Carlo.
• Calculate coupling matrix Z of visibility sample points to sources through phase screen	Matrix connecting orthonormal polynomial estimate to each visibility point	Typically 512^2 constraints and a maximum of about 15,000 Zernike's.
• Calculate eigenvalue of coupling matrix Z^TZ.	Ordered list of sensitivity to each eigenvector	Shows utility of array and sources in determining phase screen
• Find cut-off point J in eigenvalues	Upper limit to highest eigenvalue that can be recovered	Shows limit to recovery of fine scales
• Find integrated ionospheric phase power not modelled	Residual phase power on non-modelled finer scales	Uses Noll formula or Monte Carlo
• Calculate dynamic range limit in solution interval	Instantaneous dynamic range limit due to non-modelled finer scales	
• Assume $\sqrt{t}$ scaling to obtain necessary integration time	Integration time required to average out residual phase screen errors	Assumes that the phase screen errors integrate down coherently

Table 2 Performance using Direct Fitting Visibilities (DFV), as well as revised sensitivity , diffractive scale, and source counts. Average of ten trials (selecting different realizations of the sources). Use of the modified Noll formula (Equation 2) introduces changes at the level of difference between realisations. The maximum value of J remains well below that required to achieve 50dB in 6 months sky time.

Config	**Bandwidth** MHz	**Nsources**	**peak**	**J**	**stdphase**	**DR**	**Tsky**
LOWBD2	0.1	13	401.0	275	0.06	43.9	19.1
LOWBD2	0.3	28	249.5	299	0.05	44.0	17.7
LOWBD2	1.0	46	430.0	330	0.05	44.2	16.3
LOWBD2	3.0	70	232.2	342	0.05	44.3	15.8
LOWBD2	10.0	122	321.7	354	0.05	44.3	15.3
LOWBD2-RASTERHALO	0.1	14	346.0	567	0.04	45.2	10.2
LOWBD2-RASTERHALO	0.3	22	178.4	586	0.04	45.3	9.90
LOWBD2-RASTERHALO	1.0	44	173.1	627	0.04	45.4	9.33
LOWBD2-RASTERHALO	3.0	54	220.0	591	0.04	45.3	9.82
LOWBD2-RASTERHALO	10.0	115	285.6	646	0.04	45.5	9.10

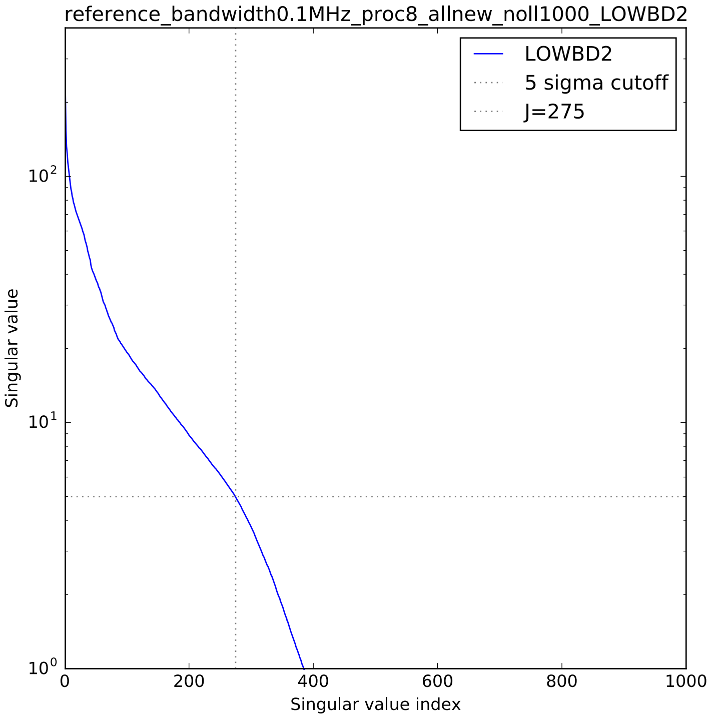


Figure 5 Singular value spectrum for 0.1MHz bandwidth. For low singular value indices, the SKA1-LOW core provides most of the sensitivity. For high singular value indices (i.e. finer scales), the halo dominates.

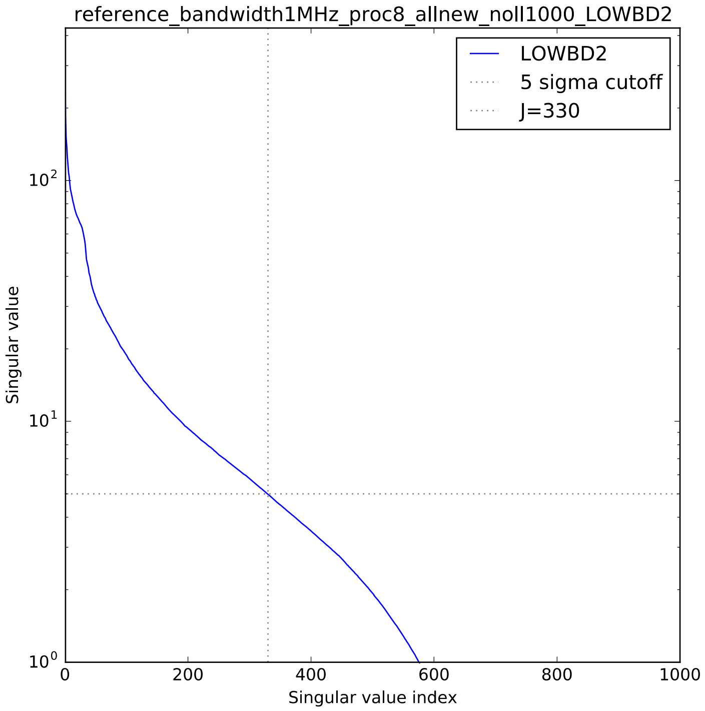


Figure 6 Singular value spectrum for 1MHz bandwidth. For low singular value indices, the SKA1-LOW core provides most of the sensitivity. For high singular value indices (i.e. finer scales), the halo dominates.

8. Conclusions

We have revisited Memo 26 in light of various comments with the goal of seeing if the conclusions still stand. We find:

- Using the Zernike analysis from Memo 26, we have shown that in order to reach the dynamic range specification of 50dB in no more than 5 years' real time, we need between 15,000 and 20,000 Zernike terms. We take this number as an estimate of the parametrization needed for the estimation of a Kolmogorov turbulence TEC screen independent of the form of parametrization.
- We have argued that the current (non-linear) approaches of cluster and facet imaging (which we collectively call Pierce Point Calibration PPC)) will not lead to stable solutions for the required number of degrees of freedom. See also [RD21] for discussion of bias and instability.
- We have proposed a two-step process for calibration of Low, composed of one non-linear step followed by a linear step. The array can be made coherent using PPC of a small number of sources, and next the full linearized measurement equations can be solved directly for the phase screen from the visibilities, a process we call Direct Fitting Visibilities (DFV). This means that weak sources can be included in the phase estimation, thus possibly reducing the bias seen by [RD21].
- Our approach allows estimating TEC instead of a phase at a single frequency. However, this will push up the time required by the number of frequency channels.
- We have performed optimisation and parallelisation of the solution of the linearized measurement equations. Performing the solution for 512 stations and 3 MHz bandwidth (74 sources) required about 400,000 thread-seconds. The bulk of the work is in calculating measurement equations connecting source-screen-station.
- Keeping up with real time would require O(40,000) cores. Hence, we conclude that while we have demonstrated viability of our calibration approach, a feasible numerical solution remains to be discovered. With the use of pre-caching certain recurring calculations, the calculation may be well suited to evaluation on a GPU.
- For the LOFAR case described by [RD21], the number of sources is 74, the number of Zernike's measurable is 97, and the on-sky time required is 47 years.
- These conclusions have been generated using the Kolmogorov spectrum. We have checked results using the observed exponent (1.8 instead of 5/3) of the structure (see [RD7]) and find changes less than the changes between realisations.
- Even having improved the model from Memo 26, with visibility based fitting, and ignoring a number of complicating factors, calibration of SKA1-LOW for EOR imaging at the specified dynamic range is still infeasible.
- For narrow bandwidths (and thus small numbers of sources), use of a heavily optimised configuration in which all outer stations are placed in a raster does cut down the required sky time by a factor or two but the total elapsed time is close to twice the lifetime of the telescope (50 years). For wider bandwidths, there is no such effect.

9. Suggestions for future work

In the previous section, we concluded that, even in the absence of certain complicating factors, the array calibration for EOR is infeasible. In Memo 26, we gave a list of the complicating factors that were not considered. We can update this list for the linear approach described here.

~~1. The relationship between phase and TEC~~[1]
2. Incomplete knowledge of the sky and antenna performance
~~3. Non-linear coupling and bias in the solution for phases of pierce points~~[2]
4. Non-linear coupling and bias in the peeling of bright sources
5. Off-zenith primary beam effects
6. The effect of sources outside the main lobe of the primary beam
7. Other non-ideal behavior of the telescope
8. The strong variability seen in ionospheric behavior (see e.g. [RD1] and [RD9])
9. Vertical structure in the ionosphere (see e.g. [RD19])
~~10. Algorithms to estimate and apply the phase screens, such as clustered calibration~~ [3]
11. Noise arising from the intra-station calibration
~~12. Fresnel effects~~[4]

The stroke-out effects have been addressed in this memo. None of the remaining effects are likely to bring any improvement in estimation of the phase screen.

The weakest point in the chain of argument presented in Memo 26 is between the phase error and the dynamic range (equation 7). Simulations to test this assumption would be highly worthwhile and could serve to undermine the conclusions of this memo and memo 26.

[1] Our approach allows for TEC estimation but we have not yet investigated sampling in frequency.

[2] We assume that the combination of pierce point fitting plus linear visibility fitting is unbiased. This strictly remains to be demonstrated.

[3] The paper by [RD21] covers the bias in sufficient detail.

[4] We have inserted the Fresnel phase terms into the relevant source-screen-station phase.

Appendix A: Optimisation

The calculations described in this memo can take days for the largest case of 512 stations, 100 sources, and 15,000 Zernike's. Hence optimisation is vital.

1. Optimisation of Python

We used the python profile tool along with the snakeviz visualisation (see Figure 7). We also found line_profiler to be very helpful (see Figure 8).

Profiling showed that the bulk of the time was going to calculation of the Zernike polynomials. The simplest way to optimise the calculation of the Zernike's is to cache intermediate results. Using functools.lru_cache is very simple:

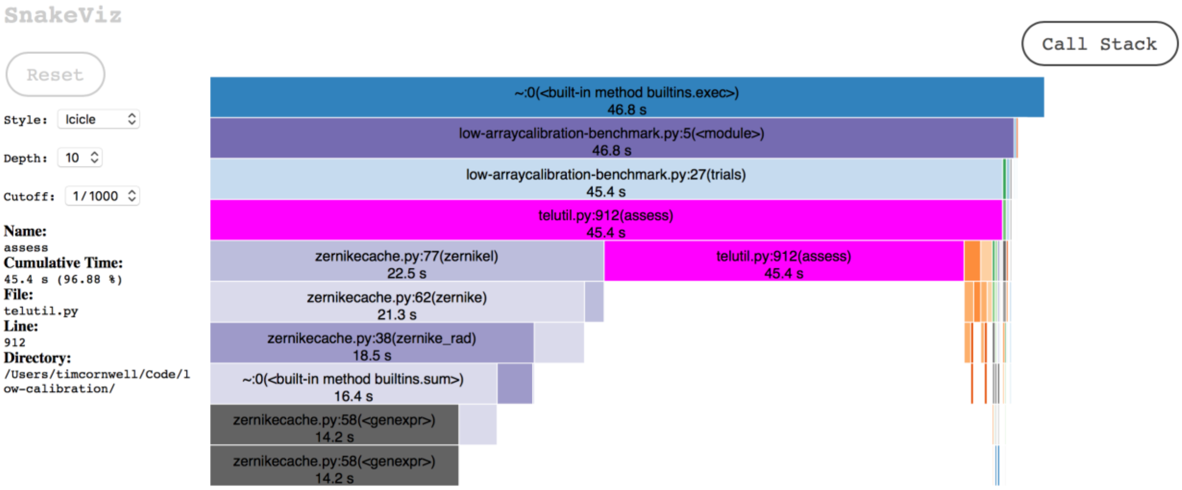
```
from scipy.misc import factorial as facbackend
from functools import lru_cache

@lru_cache(maxsize=None)
def fac(n):
    return facbackend(n)
```

This results in excellent cache hits:

```
fac CacheInfo(hits=29657, misses=55, maxsize=None, currsize=55)
pre_fac CacheInfo(hits=1080640512, misses=7428, maxsize=None, currsize=7428)
inv_sum_pre_fac CacheInfo(hits=112895229, misses=771, maxsize=None, currsize=771)
```

Use of the numpy.power function in place of the python exponentiation operator ** brings about an order of magnitude speed improvement in that operation.


SnakeViz
Call Stack
Reset
~:0(<built-in method builtins.exec>)
46.8 s
low-arraycalibration-benchmark.py:5(<module>)
46.8 s
low-arraycalibration-benchmark.py:27(trials)
45.4 s
telutil.py:912(assess)
45.4 s
zernikecache.py:77(zernikel)
22.5 s
telutil.py:912(assess)
45.4 s
zernikecache.py:62(zernike)
21.3 s
zernikecache.py:38(zernike_rad)
18.5 s
~:0(<built-in method builtins.sum>)
16.4 s
zernikecache.py:58(<genexpr>)
14.2 s
zernikecache.py:58(<genexpr>)
14.2 s

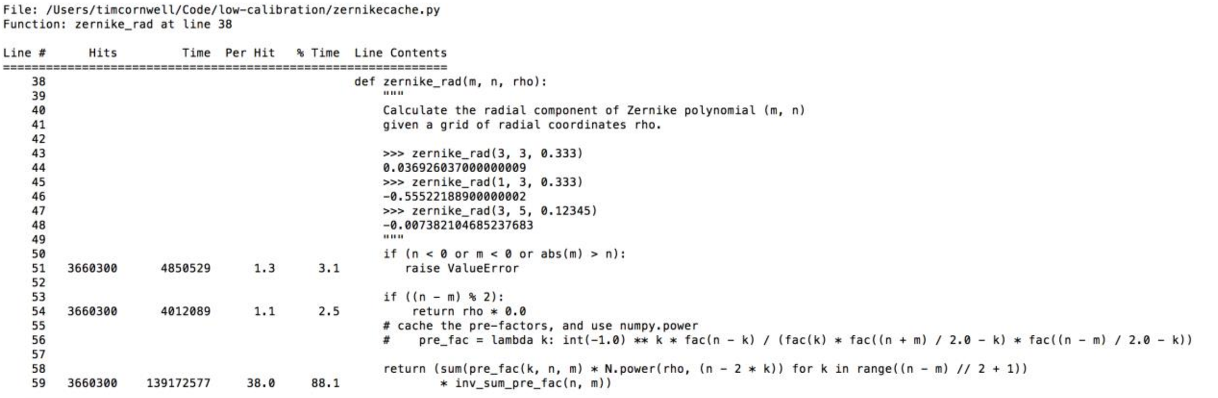
Style: Icicle

Depth: 10

Cutoff: 1/1000

Name:
assess
Cumulative Time:
45.4 s (96.88 %)
File:
telutil.py
Line:
912
Directory:
/Users/timcornwell/Code/low-calibration/

Figure 7 snakeviz output for optimised code



```
File: /Users/timcornwell/Code/low-calibration/zernikecache.py
Function: zernike_rad at line 38

Line #      Hits         Time  Per Hit   % Time  Line Contents
==============================================================
    38                                           def zernike_rad(m, n, rho):
    39                                               """
    40                                               Calculate the radial component of Zernike polynomial (m, n)
    41                                               given a grid of radial coordinates rho.
    42
    43                                               >>> zernike_rad(3, 3, 0.333)
    44                                               0.036926037000000009
    45                                               >>> zernike_rad(1, 3, 0.333)
    46                                               -0.55522188900000002
    47                                               >>> zernike_rad(3, 5, 0.12345)
    48                                               -0.007382104685237683
    49                                               """
    50                                               if (n < 0 or m < 0 or abs(m) > n):
    51   3660300      4850529      1.3      3.1          raise ValueError
    52
    53                                               if ((n - m) % 2):
    54   3660300      4012089      1.1      2.5          return rho * 0.0
    55                                               # cache the pre-factors, and use numpy.power
    56                                               #    pre_fac = lambda k: int(-1.0) ** k * fac(n - k) / (fac(k) * fac((n + m) / 2.0 - k) * fac((n - m) / 2.0 - k))
    57
    58                                               return (sum(pre_fac(k, n, m) * N.power(rho, (n - 2 * k)) for k in range((n - m) // 2 + 1))
    59   3660300    139172577     38.0     88.1              * inv_sum_pre_fac(n, m))
```

Figure 8 line_profiler output after optimisation

There is potentially a very large dot product to form the normal equations from the design matrix. We split this up by one station in the interferometers and accumulate the covariance matrix. This also avoids calculating station-to-station outer product summing over all constraints at once.

2. Parallelisation:

It is clear that these calculations of the measurement equations could be distributed quite easily and would bring roughly linear scaling. We adopted pymp since it had a suitable interface for straightforward data parallelisation. Thus the inner loops which were:

```
Covar_A = numpy.zeros([nnoll, nnoll])
for station in range(nstations):
    print('Calculating station %d to all other stations' % (station))
    A = numpy.zeros([nnoll, nstations, 2])
    for noll in range(nnoll):
        for source in range(sources.nsources):
            fx = hiono * l[source] + x
            fy = hiono * m[source] + y
            r = numpy.sqrt(fx ** 2 + fy ** 2)
            phi = numpy.arctan2(fy, fx)
            # This is a time sink: zernikecache is optimised because of this
            z = zernikel(noll, r / rmax, phi)
            z[r > rmax] = 0.0
            # All the previous operations are per station. In the code below we
            # make the transition to baselines
            vphasor = weight * phasor[source, :] * z * numpy.conj(phasor[source, station])
            vphasor[P[:,station] < rmin] = 0.0
            # In this approach, we only get access to the summed effects of all sources
            # as seen through the screen and correlated with another station.
            A[noll, :, 0] += numpy.real(vphasor)
            A[noll, :, 1] += numpy.imag(vphasor)
    # Reshape so that the dot product can work.
    A = numpy.reshape(A, [nnoll, nstations * 2])
    Covar_A += numpy.dot(A, A.T)
```

become:

```
print('Using pymp with %d processes' % nproc)
Covar_A = numpy.zeros([nnoll, nnoll])
for station in range(nstations):
    print('Calculating station %d to all other stations' % (station))
    # The A array has to be shared and locked across all threads
    A = pymp.shared.array([nnoll, nstations, 2])
    A[:,:,:]=0.0
```

```
with pymp.Parallel(nproc) as p:
    # The calculations are done in discrete ranges of the noll parameter
    for noll in p.xrange(nnoll):
        for source in range(sources.nsources):
            # This next part is done on nstations-long vectors
            fx = hiono * l[source] + x
            fy = hiono * m[source] + y
            r = numpy.sqrt(fx ** 2 + fy ** 2)
            phi = numpy.arctan2(fy, fx)
            # This is a time sink: zernikecache must be optimised
            z = zernikel(noll, r / rmax, phi)
            z[r > rmax] = 0.0
            # All the previous operations are per station. In the code below we
            # make the transition to baselines
            vphasor = weight * phasor[source, :] * z * numpy.conj(phasor[source, station])
            vphasor[P[:,station] < rmin] = 0.0
            # In this approach, we only get access to the summed effects of all sources
            # as seen through the screen and correlated with another station.
            if p.lock:
                A[noll, :, 0] += numpy.real(vphasor)
                A[noll, :, 1] += numpy.imag(vphasor)
# Reshape so that the dot product can work.
A = numpy.reshape(A, [nnoll, nstations * 2])
Covar_A += numpy.dot(A, A.T)
```

The A matrix must be shared across all processes and locked for write. With the use of p.xrange, the distribution is dynamically adjusted. Scaling is quite linear for an eight core/2 threads per core CPU.

In practice, this algorithm is probably too slow for purpose. Accumulating the constraint equations for one 10s observation takes about 2 million thread-seconds on an eight core Intel® Core™ i7-6900K running at 3GHz.