HERA Memo #133
Received 24 Oct 2024
Posted 19 Nov 2024

HERA Team Memo: Improvements to Flagging of Low-Level RFI

Joshua S. Dillon and Tyler A. Cox for the HERA Analysis Team

October 24, 2024

1 Executive Summary

We present here a small update to RFI flagging in H6C IDR 2.3 (Dillon et al., 2024)—small enough that we are not declaring a "new IDR" and are simply overwriting or "hot swapping" some files and notebooks. The key update is that we found a better statistical approach to "Round 2" flagging, which uses only cross-correlations after an initial set of RFI flags and calibration, which made it easier to pick out narrow-band, low-level outliers that are really only apparent after averaging together a whole day and examining a single spectrum.

2 Background

A second round of RFI flagging after `smooth_cal` was introduced in H6C IDR 2.2 (Dillon and Murray, 2023). It was not changed in H6C IDR 2.3 (Dillon et al., 2024). The idea was to take all calibrated visibilities, redundantly average them in unique baseline groups, delay filter them with a linear fit to DPSS, and then incoherently average the absolute values of the residuals. Using an estimator of what this would have done to pure noise, we turned this statistic into a z-score by subtracting the expectation value and then dividing by the expected standard deviation. The resulting spectra showed clear positive-going outliers—signs of RFI (see Figure 1).

However, this yielded a statistic that was somewhat biased in a couple of ways. First, the distribution was a bit off from 0 (see Figure 2)—a fact we addressed by shifting the ee and nn distributions by a single per-polarization number for the whole night. Second, as we see clearly in Figure 1 and in the whole-night version in Figure 3, data points next to flags show strong negative-going features. These are a sign of missing variance after the delay filter.

In retrospect, this makes sense. Because the DPSS fit is not constrained in the flagged channels, it can over-fit the noise in channels that neighbor them. If there are any positive-going outliers amidst the over-fit noise, they may be difficult to pick out in any cut on z-score. Nonetheless, as Figure 3 shows and as we argued in Dillon and Murray, 2023, the new technique highlights significant deviations from noise that are narrow band and/or clearly associated with other previously flagged RFI—good evidence that we are in fact finding additional low level RFI contaminating our data.

To identify that RFI, a series of flagging steps were performed:

1. For each polarization, flag any 5σ outliers.
2. For each polarization, flag any 4σ outliers that are next to prior flags. This "watershed" algorithm is run until convergence is reached.
3. Iteratively flag whole integrations or whole channels whose unflagged average z-score is greater than 1 (note this that is not the same as a 1σ outlier). This is only done for either channels or times in a given loop, based on whichever average has the largest outlier. After each flagging step, flag any channels flagged more than 25% of the time (compared to the least-flagged channel) and any integrations flagged for more than 10% channels (compared to the least-flagged integration). Continue until convergence.

The last step, which is a bit convoluted, was meant to address some of the clear outliers that remained but were only visible after averaging together a full night's z-scores. The result is shown in Figure 4.

3 Flattening Out the z-Scores

The impact of flags on the z-score spectrum makes low-level RFI hard to identify, especially near other flags (which is where we might expect it, a priori). We'd like to flatten out this metric, and it turns out that the impact of the flagging pattern is a predicable consequence of the linear filtering operator. So we take each redundant baseline group's averaged visibility, V_{i-j}, and divide by the prediction of the noise on that baseline, σ_{i-j}, which comes from the autocorrelations and the N_{samples}. Defining

$$s_{i-j} = \frac{V_{i-j}}{\sigma_{i-j}}, \tag{1}$$

we calculate the least-squares DPSS fit as

$$\mathbf{s}_{i-j}^{\text{model}} = \mathbf{A}(\mathbf{A}^\dagger \mathbf{N}^{-1}\mathbf{A})^{-1}\mathbf{A}^\dagger \mathbf{N}^{-1}\mathbf{s}_{i-j} \tag{2}$$

where $\mathbf{A}$ is set of DPSS vectors appropriate for the particular band, delay limit, and eigenvalue cutoff (Ewall-Wice et al., 2021). (In practice, we filter independently above and below FM). Because we are filtering SNRs, $\mathbf{N}^{-1}$ is diagonal in frequency space with 1 if the channel is unflagged and 0 if it is flagged. Our delay-filtered SNR is thus:

$$\mathbf{f}_{i-j} \equiv \mathbf{s}_{i-j} - \mathbf{s}_{i-j}^{\text{model}} \tag{3}$$

In practice, we set f_{i-j} in flagged channels to `np.nan`.

It turns out that the reason we were over-fitting points near flagged channels is because they have high leverage—they disproportionately affect the fit and decrease the residual noise variance. This can be computed and accounted for. We simply have to multiply element of $\mathbf{f}_{i-j}$, such that:

$$f'_{i-j} = f_{i-j}\frac{\sqrt{\pi(1-h_{i-j})}}{2} \tag{4}$$

where the h_{i-j} is defined as

$$h_{i-j} \equiv \text{diag}\left[\mathbf{A}(\mathbf{A}^\dagger \mathbf{N}^{-1}\mathbf{A})^{-1}\mathbf{A}^\dagger \mathbf{N}^{-1}\right]. \tag{5}$$

We then take the average of the magnitude of the filtered SNRs over baselines, which has an expectation value of 1. Our estimator of the z-score is thus:

$$z = \left(\frac{1}{N_{\text{ubl}}}\left[\sum_{\text{baselines}i-j}^{N_{\text{ubl}}} \left|f'_{i-j}\right|\right] - 1\right)\sqrt{\frac{\pi N_{\text{ubl}}}{4-\pi}} \tag{6}$$

which follows from the fact that f'_{i-j} is complex and $\left|f'_{i-j}\right|$ is half-normally distributed. We have verified that the statistics work with out pure-noise simulations with real flagging patterns.

The upshot of all this is that the spectrum is much flatter (see Figure 5) and much closer to Gaussian-distributed (see Figure 6). The result is even clearer in the final z-score waterfall (see Figure 7). It should be noted that part of the improvements here are due to other parameterization changes, which we address next.

4 Parameter Tweaks and Additional Flagging

Flatting the z-score spectrum is an opportunity to revisit the parameters and flagging algorithm. In the calculation of per-file z-scores, there are two parameters that we explored. The first is the delay at which to filter. As we saw in Figure 3, certain LSTs show broad-ish features around 125 MHz and at a time of [JD].35 that repeat night to night. We believe these are mutual coupling systematics exacerbated by the galaxy in the far-sidelobes. Delaying filtering at 750 ns seems to remove this feature (see Figure 7) much better than the previously used 500 ns delay, which was chosen in part because even larger filtering delays made the negative-going outliers worse.

The other parameter is the cutoff on how redundant a baseline needs to be in order to be included in the incoherent average. There is a trade-off that we still do not fully understand, where using more baselines (i.e. admitting baselines with lower redundancy) leads to a more biased distribution of z-scores (compare the peak in Figure 2 to Figure 6). This quantity is expressed in as a fractional redundancy relative to the redundantly-averaged autocorrelation of the same polarization. Previously this `MIN_SAMP_FRAC` was set to 0.05. After some experimenting with how this parameter drives the interplay between the bias and the sensitivity with which we can detect RFI, we raised it to 0.15.

With all the z-scores computed, we now move on to the `full_day_rfi_round_2` notebook, where we have revisited the algorithm for finding and flagging RFI by adding 3 new steps and modifying two existing ones slightly:

1. **NEW:** For each polarization, iteratively flag whole integrations or whole channels with an unflagged *median* z-score greater than 1. This is done almost identically as was previously done with the unflagged mean z-score and includes cuts on total flagging fractions of channels and integrations.

2. For each polarization, flag any 4σ outliers (down from 5σ).

3. For each polarization, perform iterative "watershed" flagging on 2σ outliers that neighbor flags (down from 4σ).

4. For each polarization, perform the same whole integration/channel flagging, this time on those with and unflagged *mean* z-score greater than 1 (this is exactly the same as before).

5. **NEW:** Turn the time-averaged z-score spectrum into a proper SNR by accounting for the number of unflagged times that go into each channel in the average. Now perform basically process as we described in section 3. Filter at 250 ns above and below FM, correct for the leverage factor, and then flag any 4σ outliers. This is done iteratively—we only flag outliers that are within a factor of 1.5 of the largest outlier, working our way down until none are above 4σ. The full algorithm lives `full_day_rfi_round_2` notebook (for now).

6. **NEW:** Perform one more round of 2σ watershed flagging.

While the above leads to significantly more flagging that what we were previously doing (26.986% of the waterfall for 2459861 is now flagged, compared to 22.441% after the previous algorithm), it really looks like the time-averaged z-scores are free from any compact systematics in frequency space. Our hope is that the new simultaneous inpainting algorithm (Dillon et al., 2024) will alleviate any problems associated with the considerable increase in flagging density.

References

Dillon, J. S. and S. Murray (2023). "HERA Memo #125: H6C Internal Data Release 2.2". reionization.org/memos.

Dillon, J. S., S. Murray, T. A. Cox, et al. (2024). "HERA Memo #131: H6C Internal Data Release 2.3". reionization.org/memos.

Ewall-Wice, Aaron, Nicholas Kern, Joshua S. Dillon, et al. (Jan. 2021). "DAYENU: a simple filter of smooth foregrounds for intensity mapping power spectra". In: MNRAS 500.4, pp. 5195–5213. DOI: 10.1093/mnras/staa3293. arXiv: 2004.11397 [astro-ph.CO].

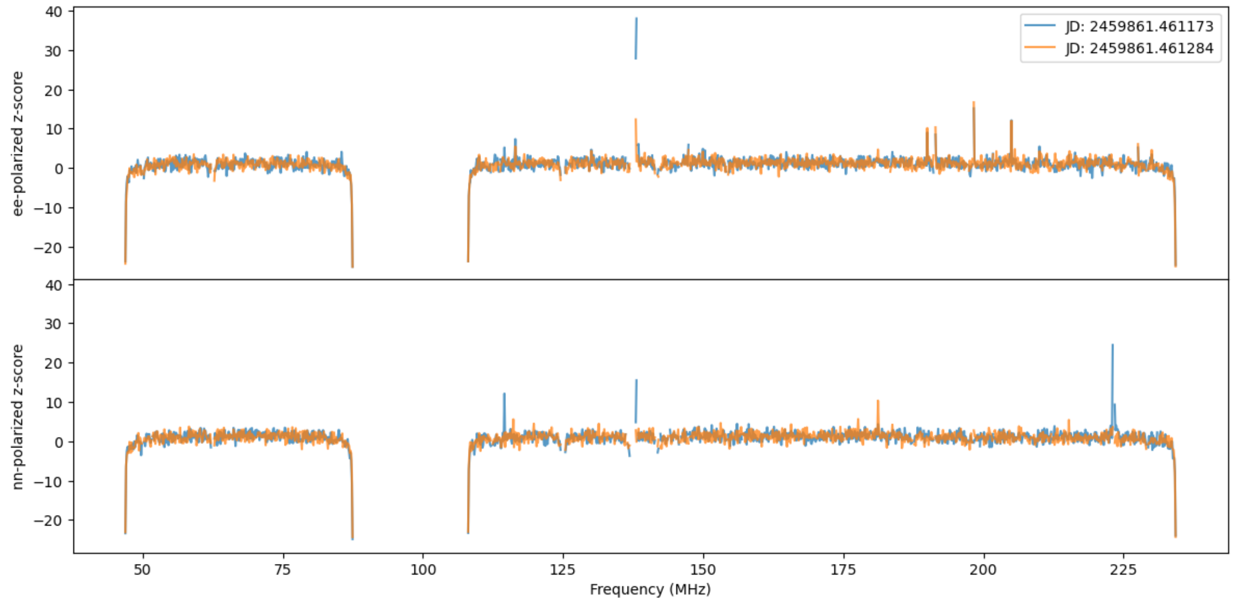

Figure 1: z-scores in a single 2-integration file after redundant-averaging, delay filtering at 500 ns, incoherently averaging, subtracting the expectation value, and dividing by the expected standard deviation. Clear positive-going outliers are previously unflagged RFI. From here.

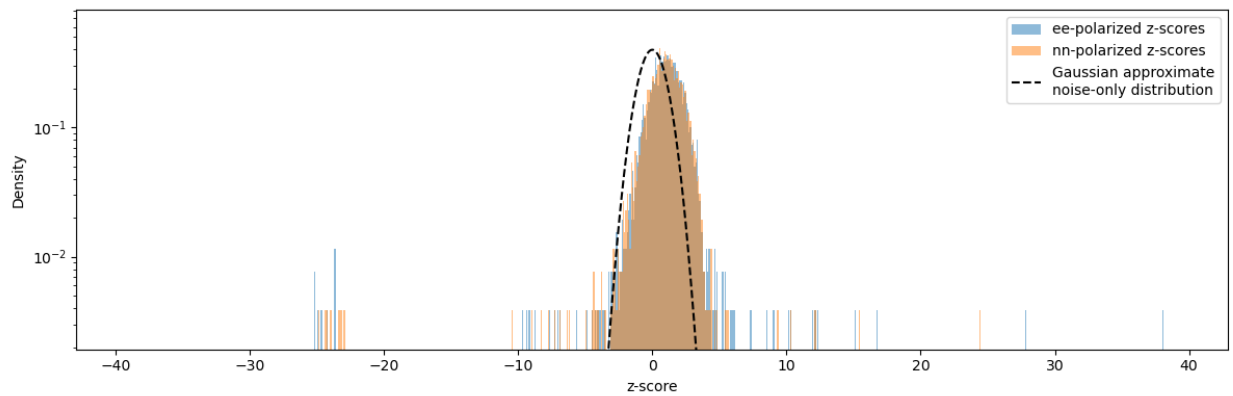

Figure 2: The distribution of z-scores from Figure 1 relative to a Gaussian distribution. From here.

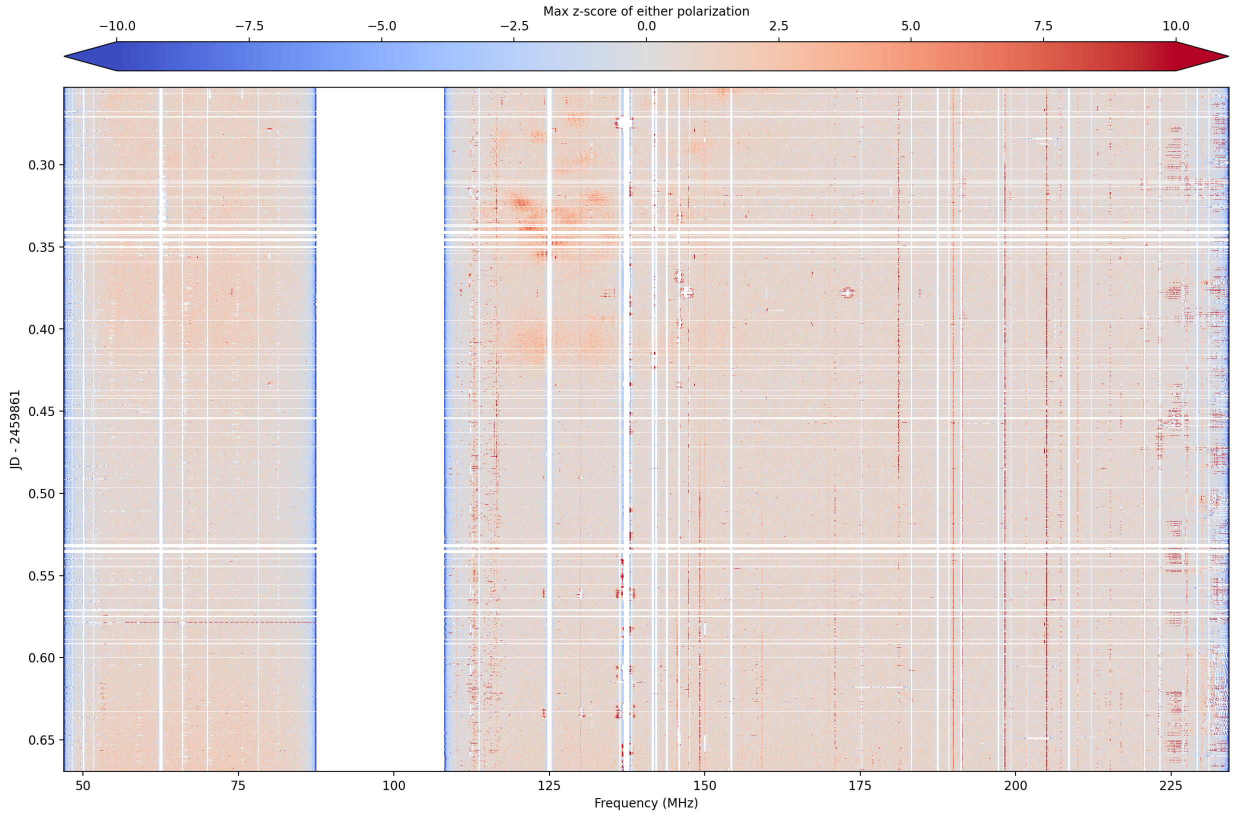

Figure 3: A full day waterfall of z-scores as in (and including) Figure 1, but where the larger of the two polarizations' values is shown. From here.

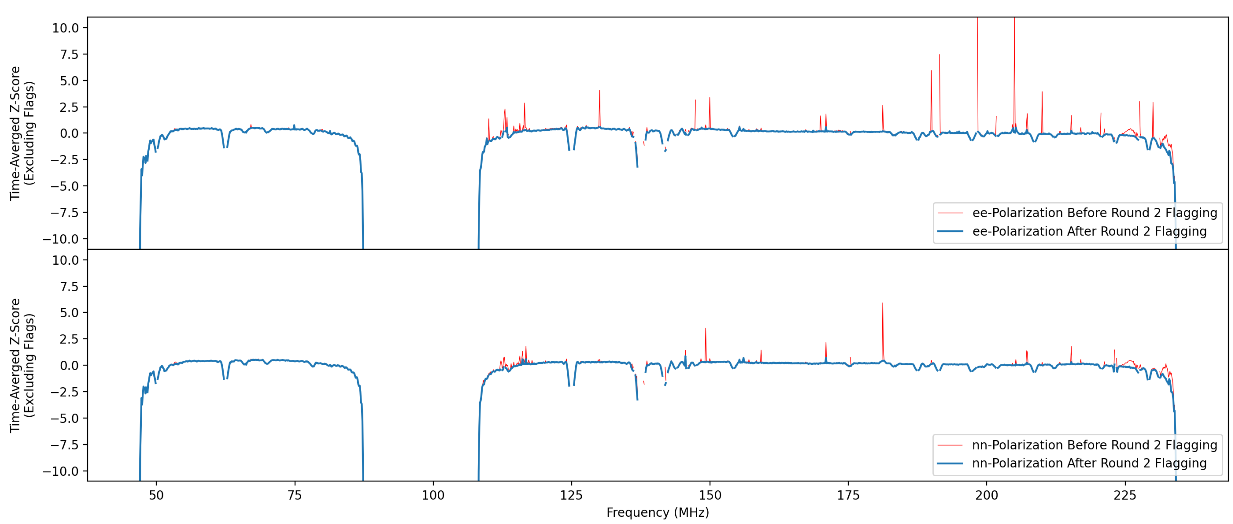

Figure 4: Averaged z-score spectra over a whole night. Note the clear impact of the flags in creating downward-going excursions. Also note both the low-level RFI that was caught (red) and some of the points in blue that were not caught, despite being clearly visible here and slightly visible in the waterfall in Figure 3. From here.

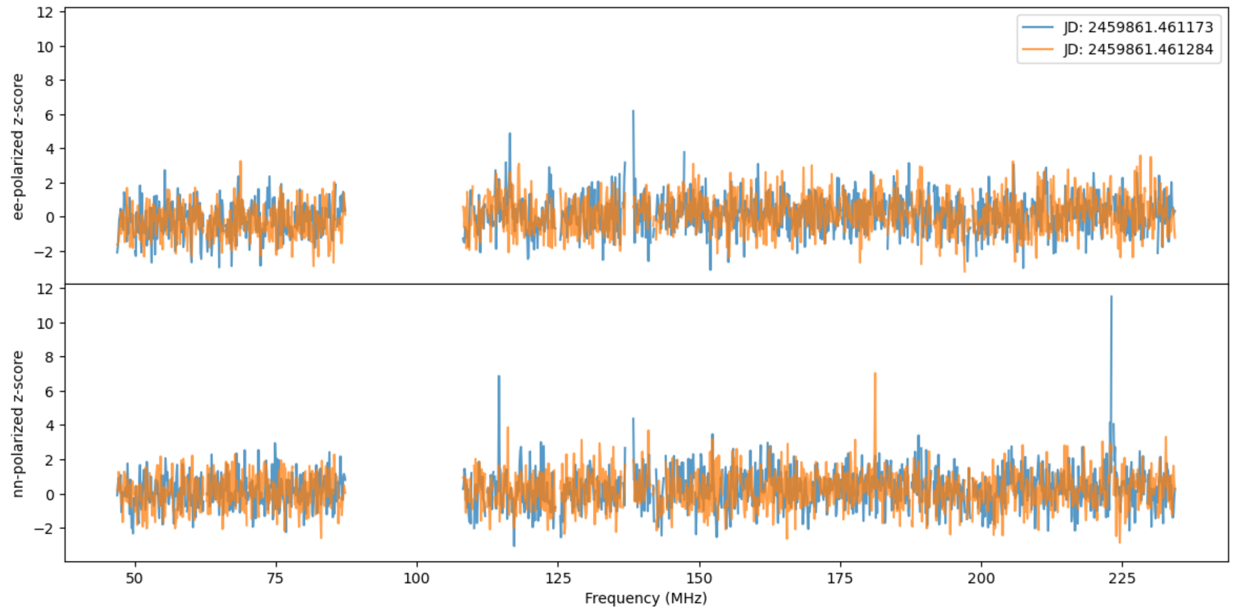

Figure 5: The new spectrum of z-scores in a single file. Compare to Figure 1, though note that the prior flags are not 100% identical and that the y-axis limits have changed. From here.

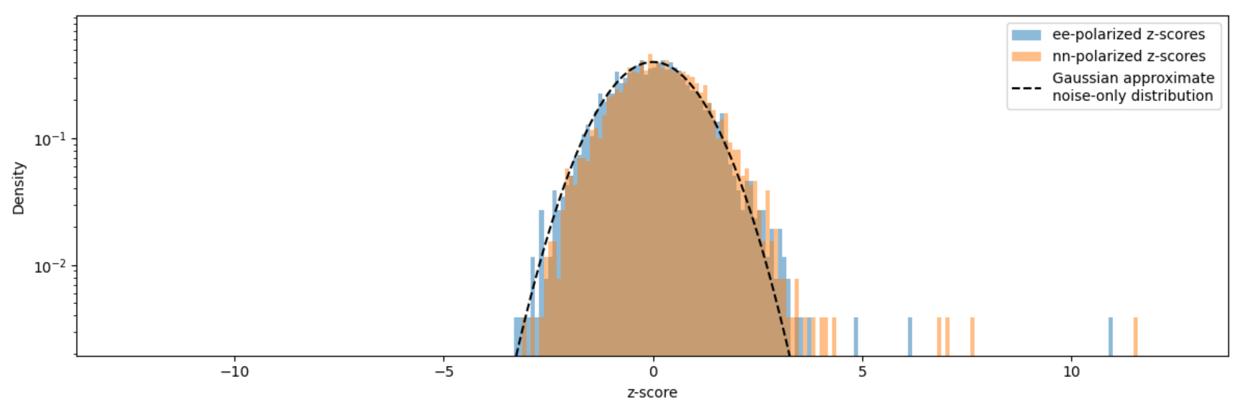

Figure 6: The new distribution of z-scores. Compare to Figure 2, though note the change in the x-axis limits. From here.

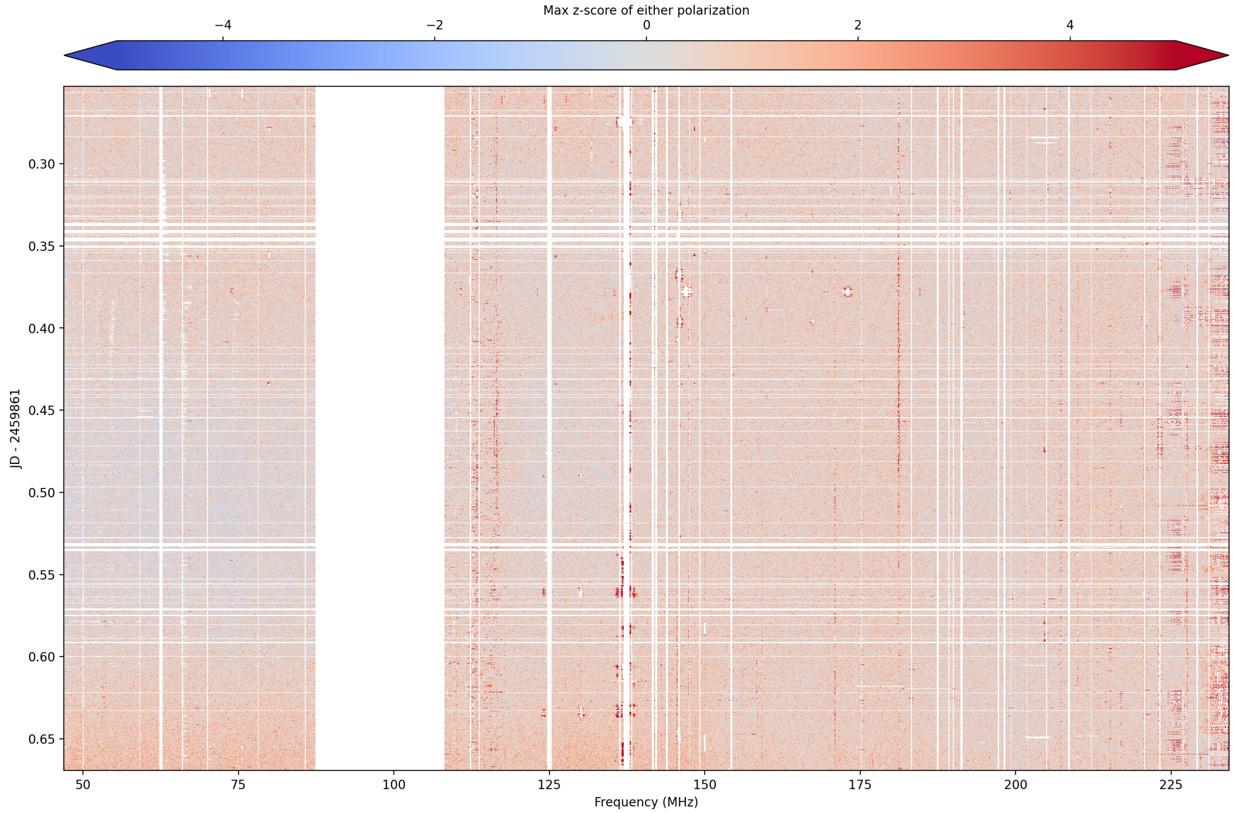

Figure 7: The z-score waterfall. Compare to Figure 3, though note the change in the color scale. From here.

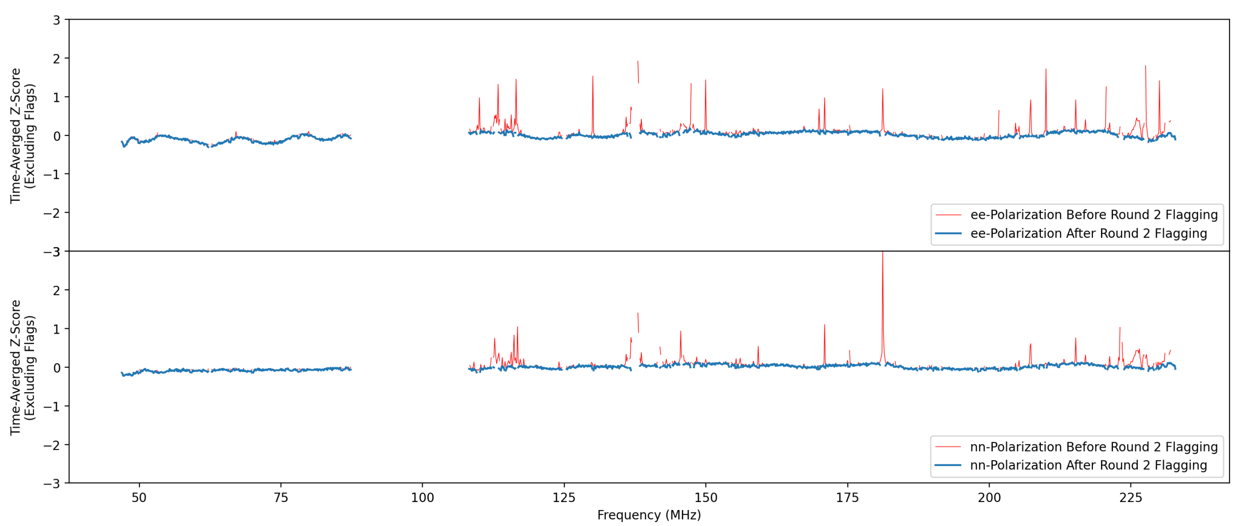

Figure 8: The time-averaged z-score, before and after round 2 flagging. Compare to Figure 4, though note the change in the y-axis limits. From here.