EVLA Memo 238
Efficacy of Flagging for SRDP Images

Toni Norton, Drew Medlin, Amy Kimball

May 2025

1 Introduction

The Science Ready Data Products (SRDP) project was initially created to provide users with science quality data and images through the efforts of the Pipeline Operations team, given that the observation followed specific setup guidelines, which included being compatible with the pipeline. VLA Data Analysts would perform quality assessment of SRDP-compliant VLA observations that have been calibrated by the pipeline and create additional flags if necessary. The type of flagging included, but was not limited to: RFI, spectral windows of too high/low gain amplitude, gain phase jumps, time ranges of decorrelation, high weights, DTS hardware issues, and more. A question was posed of whether the flagging being performed made any significant difference to the quality of the image produced. This question kick-started an investigation into the SRDP process, wherein images created from flagged datasets were compared to images created with no additional flagging, excepting automatic pipeline or online flags.

For the purpose of this memo, a "flagged" image or observation refers to the original images generated and archived by the Pipeline Operations team that include additional flags added by Data Analysts. "Unflagged" refers to the same source from the same observation, calibrated and imaged with no additional flagging.

The RMS and the dynamic range of the images were the heuristics chosen to decide the effectiveness of flagging. RMS is the measure of noise in an image, the lower the better. Dynamic range is the measure of the noise in relation to the peak brightness of the source/image. In this case, a higher dynamic range is preferable, as it implies the source is much brighter than the noise of the image. The maximum pixel value–the brightest pixel in the image–was also recorded for both flagged and unflagged images.

The following data analysts participated in the formation and completion of this investigation: Drew Medlin, Aaron Lawson, Toni Norton, Audrey Zinn, James Khor, Tierra Candelaria, and Efrain Arzaga.

2 Data

217 SRDP-compliant observations with flags applied by Data Analysts were collected. Many of these had multiple images, either due to different sources observed in the same execution block, or the same source observed in different bands. Data from all four configurations at the VLA were evaluated, and the breakdown of images from each configuration was as follows: 156 images from A configuration, 85 from B configuration, 65 from C configuration, 18 from D-to-C configuration change, and 38 from D configuration. There were also 3 images from BnA-to-A configuration. Broken down by band, there were 186 in C band, 87 in X band, 31 in Ku band, 27 in K band, 19 in Ka band, and 10 in Q band. This yielded a total of 365 images used for comparison.

The data were drawn from April of 2023 to February of 2024. Observations before October 2023 (namely, A and B configuration data) were calibrated using CASA version 6.4.1, while observations after (C and D configurations) were calibrated with CASA 6.5.4. Along with this CASA version switch, the self-calibration imaging pipeline was released for Operations use, so observations completed with the self-cal pipeline may have self-cal images. Self-calibration refers to the process of using the target image as a model to further calibrate the complex gains of the target source, and can only be completed when an image has a sufficient Signal-to-Noise (S/N) ratio. Regular-calibration (reg-cal) is the standard imaging that uses the bandpass and complex gain calibrators for calibration, and is completed for all images produced by the SRDP project.

Self-calibration images were used when available, over their reg-cal counterparts, due in part to the extra details provided in the self-cal stages of the weblog. For observations without, or before the introduction of the new pipeline, reg-cal images were utilized instead. Self-cal images were only compared to other self-cal images, and reg-cal compared to reg-cal.

The original flagged images were downloaded from the archive and the original imaging weblogs were accessed through the Workspaces cache directory on the NM lustre system. Unflagged datasets were obtained using the CASA Integrated Pipeline (CIPL) workflow to calibrate the observation from the beginning with none of the analyst-added flags. Once the calibration was completed, each was run through the imaging pipeline via a pipescript to generate the unflagged images.

3 Methods

To find the significance of flagging, the RMS and dynamic range were compared between the flagged and unflagged images. Three different RMS values were used, two per image. The first was the RMS of the image obtained from the CARTA Statistics widget when viewing the full image. This was used for every single image. The second and third were used depending on whether self-calibration was performed on both the flagged and unflagged images. If self-calibration was present, the RMS value was pulled from stage 7 of the

imaging weblog, the Final RMS. If there was no self-calibration, the RMS was taken from stage 6 of the imaging weblog, the non-pbcor image RMS. The RMS values provided by the weblog only had 2 (stage 6) or 3 (stage 7) significant figures, which was often not sensitive enough to show a difference between the flagged and unflagged RMS.

The percent delta was calculated between the flagged and unflagged images using the following equation:

$$\frac{flagged\ RMS - unflagged\ RMS}{unflagged\ RMS} * 100$$

The dynamic range likewise was calculated two different ways depending on the presence of self-calibration. The self-cal stage of the imaging weblog supplies a value for the dynamic range of the image. This value was used when available for both flagged and unflagged images. The percent delta was calculated for the dynamic range in this way:

$$\frac{flagged\ DR - unflagged\ DR}{unflagged\ DR} * 100$$

If the dynamic range was not provided by the weblog, it was instead calculated using the RMS and maximum pixel value, each taken from the CARTA Statistics widget when viewing the full image. The percent delta was then calculated using these equations:

$$flagged\ DR = \frac{flagged\ Max\ pixel}{flagged\ RMS}$$

$$unflagged\ DR = \frac{unflagged\ Max\ pixel}{unflagged\ RMS}$$

$$\frac{flagged\ DR - unflagged\ DR}{unflagged\ DR} * 100$$

At the beginning of the investigation, only observations using the self-cal pipeline were being considered. Even with the self-cal pipeline, not every source has enough S/N to enable self-calibration; in such cases, the pipeline does not perform that step and does not include a precalculated dynamic range. While RMS was still recorded, the DR for these particular images was skipped over. Once pre-self-cal pipeline observations were added to the investigation, the method for finding DR evolved, and reg-cal images were no longer skipped over. This resulted in a slight discrepancy between the number of RMS percent differences and the number of DR percent differences, seen later in the memo.

4 Results

When comparing the RMS values, a positive percentage means the unflagged image had a lower RMS. In general, lower RMS is desirable in images as it

implies less noise. Of 365 images, 76.16% had a positive percent difference. Thus, flagging data—even when a scan, spectral window, etc. is deemed "bad"—usually impacted the RMS negatively, adding more noise to the image. However, the average percent difference of RMS for all 365 images was only 0.79%. The better conclusion from this result is that flagging has very little, almost negligible impact on the final image product.

A similar idea was reflected in the results regarding dynamic range. Of the 332 images where a dynamic range difference was recorded, 74.4% had a negative percent difference. A negative percentage in this case means that the unflagged image had a higher dynamic range. Because a higher dynamic range is the more desirable outcome, the added flags seemed to do more harm than good to the overall image. However, similar the RMS comparison, the average difference was only -1.34%. Once again, these percentages were so small that it really made no significant difference.

The maximum pixel values were also recorded, but the difference in brightness was negligible between flagged and unflagged images. Additionally, there was no correlation between whether flagging, or a higher or lower RMS/DR, resulted in a higher or lower peak brightness.

When separated by band or configuration, the results remain consistent. The full breakdown of the results summary is below:

Overall			
	RMS of Full Image	RMS from Weblog	Dynamic Range
Total Images	365	365	332
Median % Δ	0.55%	0.00%	-0.63%
Average % Δ	0.79%	0.72%	-1.34%
% Better than Flagged	76.16%	31.78%	74.10%
% Worse than Flagged	23.84%	9.32%	25.90%
Count over 5%	34	30	10
Count under -5%	21	23	32

Per Band				
		RMS of Full Image	RMS from Weblog	Dynamic Range
Average	C	1.69%	1.39%	−1.56%
	X	−0.33%	−0.45%	−0.71%
	U	−1.66%	0.63%	−2.97%
	K	1.15%	0.86%	−0.60%
	A	0.55%	0.16%	0.43%
	Q	1.08%	−0.20%	−2.91%
% Better than Flagged	C	79.03%	43.55%	73.66%
	X	67.82%	20.69%	66.67%
	U	87.10%	25.81%	67.74%
	K	74.07%	18.52%	44.44%
	A	68.42%	15.79%	47.37%
	Q	80.00%	10.00%	50.00%
% Worse than Flagged	C	20.43%	7.53%	22.04%
	X	32.18%	17.24%	24.14%
	U	12.90%	0.00%	16.13%
	K	25.93%	7.41%	37.04%
	A	31.58%	10.53%	31.58%
	Q	20.00%	10.00%	30.00%

Per Configuration				
		RMS of Full Image	RMS from Weblog	Dynamic Range
Average	A	1.76%	1.38%	−1.72%
	B	−0.22%	−0.36%	−1.48%
	C	−0.64%	0.88%	−0.82%
	D>C	0.38%	0.26%	−2.32%
	D	1.80%	0.44%	0.11%
% Better than Flagged	A	83.97%	49.36%	73.72%
	B	61.18%	31.76%	72.94%
	C	75.38%	4.62%	60.00%
	D>C	66.67%	11.11%	33.33%
	D	31.58%	2.63%	28.95%
% Worse than Flagged	A	15.38%	2.56%	24.36%
	B	38.82%	31.76%	24.71%
	C	26.15%	0.00%	16.92%
	D>C	33.33%	0.00%	22.22%
	D	5.26%	0.00%	2.63%

*"% Better than Flagged" is the percentage of unflagged images that had a lower RMS (positive Δ) or a higher dynamic range (negative Δ) than their flagged counterparts. Percentages of Better and Worse may not add to 100% because the RMS difference from the weblogs was often 0.

Images that had a difference of more than 5% for RMS, or less than −5% for dynamic range, were deemed significantly worsened by flags. In Ku band, almost 13% of the 31 images had a significant difference. C band had the

second most impacted images, where between 9 to 11% of the 186 images had a difference over 5% and/or less than −5%. Similarly, B configuration had the highest percentage (nearly 13%) of significant RMS difference while C config had the highest percentage (13.5%) of significant DR differences.

Conversely, images were deemed to be significantly improved by flags when the RMS difference was less than −5% or when the DR difference was greater than 5%. X band saw the greatest improvement by flags, with respect to RMS, with almost 14% of the 87 images showing a significant difference. Across both categories, 10.5% of 27 Ka images showed significance. B config was the only configuration with more than one or two images improved with flagging, and it only showed in RMS difference. 20% of the 85 B config images had less than −5% difference in RMS. The majority of these images (12 of the 17) came from the same observation, a mosaic where one spectral window with RFI was flagged. These also accounted for all 12 of the significantly improved X band images.

It should be noted that a significant RMS difference did not necessarily beget a significant DR difference. For example, of the 34 total images where the RMS difference was greater than 5%, only 10 of them also had a DR difference less than −5%. A fourth of the significantly worsened Ku images were significant in both categories; less than a sixth of the significantly worsened B config images were significant in both.

5 Discussion

While these results show that flagging has little to no impact on the quality of the image, there were a few instances that jumped out in opposition.

Of the 217 observations, two had high weights (in the millions) on a scan that prevented tclean from being able to clean the image. Once flagged, these images were able to be cleaned. Further investigation into high weights is being completed to determine the threshold of impact, but this project revealed the importance of flagging weights that are exceptionally high.

Another example of helpful Data Analyst intervention was when the pipeline used a poor choice for the reference antenna. When this happens, an entire baseband may be flagged on every antenna, resulting in a loss of up to half the data. Normal operation procedures call for rerunning the pipeline to ignore the antenna as a reference, whether the observation falls into the SRDP category or not, so this practice should continue.

Aside from other extreme or rare cases, which are and should continue to be dealt with individually, flagging small errors or deviations in data is shown to be unnecessary. Most often it has no effect on the data at all, though in some cases it can worsen or improve the images. From the Overall table, there were 34 images where flagging significantly worsened the data (more than 5%), while only 21 images where flagging significantly improved it. It is almost a net-zero difference such that only a few will need Data Analyst intervention.