ALMA Memo 626
Estimates of ALMA WSU Data Properties

Amanda A. Kepley,[1] Crystal Brogan,[1] John Carpenter,[2] María Díaz Trigo,[3]
Bunyo Hatsukade,[4, 5, 6] and Jonathan Antognini[2]

[1] *National Radio Astronomy Observatory, 520 Edgemont Road, Charlottesville, VA 22903, USA*
[2] *Joint ALMA Observatory, Avenida Alonso de Córdova 3107, Vitacura, Santiago, Chile*
[3] *European Southern Observatory, Karl-Schwarzschild-Str. 2, 85748 Garching bei München, Germany*
[4] *National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan*
[5] *Graduate Institute for Advanced Studies, SOKENDAI, Osawa, Mitaka, Tokyo 181-8588, Japan*
[6] *Institute of Astronomy, Graduate School of Science, The University of Tokyo, 2-21-1 Osawa, Mitaka, Tokyo 181-0015, Japan*

(Dated: 2024 January 31)

ABSTRACT

The ALMA Wideband Sensitivity (WSU) will double and eventually quadruple the correlated bandwidth of ALMA with the goal of providing 0.1 $\mathrm{km\,s^{-1}}$ spectral resolution across the entire available bandwidth for all bands with improved sensitivity by upgrading the receivers, digital electronics, and correlator. This upgrade will result in an overall increase in the number of channels per data set, which will lead to increases in data rates, visibility data volumes, and product sizes. This memo uses the properties of data taken in ALMA Cycle 7 and 8 to estimate what the properties of WSU data might be over two future ALMA cycles allowing us to quantify the impact of the WSU. We provide estimates for two different stages: early WSU, where some bands are upgraded and the correlator can provide two times the current bandwidth, and later WSU, where all bands are upgraded and the correlator can provide four times the current bandwidth. We find that, in general, the distribution of the WSU data properties have a narrow peak with a long tail of values that extend out to several orders of magnitude beyond the time-weighted mean. The values in the long tail dominate the requirements for transfer, storage, and processing of the data. Our assumptions, most notably that of identical spectral windows across the band, mean that the WSU data properties presented here represent a conservative estimate of the final WSU data products. There are also systematic uncertainties associated with extrapolating the science use cases requested by the PIs today to those that may be requested 5-10 years in the future, especially those associated with Bands 1 and 2. We discuss several potential options for reducing the impact of the more extreme WSU projects. However, we cannot fulfill the scientific requirements of the WSU without processing projects in the long tail.

1. INTRODUCTION

The goal of the ALMA Wideband Sensitivity Upgrade (WSU) is to double and eventually quadruple the correlated bandwidth of ALMA (Carpenter et al. 2022a), based on the recommendations of the ALMA 2030 Development Roadmap (Carpenter et al. 2018). To do this, the receivers, digital electronics, and correlator of ALMA will be upgraded with the goal of providing 0.1 $\mathrm{km\,s^{-1}}$ spectral resolution across two to four times the current bandwidth. This functionality will allow observers to observe more lines at high velocity resolution in a single observation compared to today. The science case for the WSU is centered around three major science themes that require increased bandwidth and sensitivity to keep ALMA at the forefront of scientific discovery: Origins of Planets, Origins of Chemical Complexity, and Origins of Galaxies (Carpenter et al. 2018). These three themes encompass an extensive range of scientific cases that will exploit the ALMA 2030 capabilities including, but not limited to, kinematic signatures of planet formation, physical and chemical structure of disks, discovery of new complex organic molecules, chemical inventories of cold cores and protostars, chemistry and dust content around evolved stars, unbiased redshift surveys of galaxy, and imaging of black holes with VLBI (Carpenter et al. 2018). However, the full range of science done by ALMA will extend beyond these three themes since it is a general purpose PI-driven instrument.

The peak data rates for WSU are presented in Carpenter et al. (2023). The goal of this memo is to estimate the properties of the ensemble of data that will be produced by the WSU to guide the necessary data transfer, storage, computing, and hardware requirements downstream of the correlator. To do this, we extrapolate from the properties of the ALMA data taken in Cycles 7 and 8 using the method outlined in Section 2, which builds on the initial work done to estimate the WSU data rates by Carpenter et al. (2022b). The resulting estimated data properties including data rates, volume of visibility data, product sizes, and cube properties are described in Section 3.1 through 3.4. The main feature of all these results is that the distributions in general have a long tail that dominates the date rates, total data volume, and product size. This tail is not unique to the WSU; it is also seen in current BLC/ACA data. However, the tail extends to higher values for the WSU because of the overall increase in number of channels. The assumptions underlying these estimates are conservative, in particular the assumption of identical spectral windows across the band. There are also systematic uncertainties associated with extrapolating the science use cases requested by PIs today with those that might be requested 5-10 years in the future, especially for Bands 1 and 2. We discuss some potential options for reducing the impact of the projects in the long tail, while retaining the necessary scientific capabilities in Section 4. We summarize our findings in Section 5.

2. METHOD

2.1. *Sample Selection and Initial Parameter Estimation*

To build our sample, we downloaded the metadata from the public ALMA Archive interface for projects with project codes starting with 2019 or 2021, which correspond to Cycles 7 and 8 respectively, using the *astropy* package *astroquery* (Ginsburg et al. 2019). This sample includes projects with any data taken for them in the archive regardless of whether the data was public. From this sample, we excluded the projects in the following four categories: solar system, solar, very long baseline interferometry (VLBI), and total power. Solar system projects (1.5% of the total MOUSes) were excluded because the archive reports their fields of view as the area of the sky over which the science target was observed, not the area of the sky around the science target, thus greatly overestimating the imaged field of view. Solar (0.45% of the total MOUSes) and VLBI (0.5% of the total MOUSes)

Table 1. WSU Database Inputs

Variable	Unit	Description
mous	···	Member Observing Unit Set UID
proposal_id	···	Proposal Code
schedblock_name	···	Scheduling block name
cycle_info	···	Cycle the data were obtained in
array	···	12-m or 7-m array
science_keyword	···	Scientific keyword
scientific_category	···	Science category
band	···	Band
ntarget	···	Number of Targets in MOUS
target_name	···	Name of target (per source only)
s_fov	°	Field of View returned by the Archive
s_resolution	″	Angular Resolution returned by the Archive
mosaic	···	Was the source mosaicked?
L80	m	80th percentile baseline
blc_npol	···	Number of polarizations for BLC/ACA observations
blc_nspw	···	Number of spectral windows for BLC/ACA observations
blc_specwidth	kHz	Spectral width for BLC/ACA observations
blc_freq	GHz	Frequency for BLC/ACA observations
blc_velres	$\frac{\mathrm{km}}{\mathrm{s}}$	Velocity resolution for BLC/ACA observations
blc_nchan_max	···	Maximum number of channels in a spw
blc_nchan_agg	···	Total aggregate number of channels across all BLC/ACA spectral windows (does not take into account overlap between windows)
blc_bandwidth_max	GHz	Maximum spw bandwidth
blc_bandwidth_agg	GHz	Total aggregate bandwidth across all BLC/ACA spectral windows (does not take into account overlap between windows)
blc_tint	s	Integration (dump) time for BLC/ACA
blc_ntunings	···	Number of tunings for BLC/ACA observations
bp_time	s	Total time per MOUS spent on bandpass calibrations
flux_time	s	Total time per MOUS spent on flux calibration
phase_time	s	Total time per MOUS spent on phase calibration
pol_time	s	Total time spent per MOUS on polarization calibration
check_time	s	Total time spent per MOUS on the check source
target_time	s	Time on science target (per source only)
target_time_tot	s	Total time spent per MOUS observing the science target(s)
cal_time	s	Total time spent per MOUS observing the calibration targets
time_tot	s	Total time spent per MOUS observing calibration and science targets
Assumed Values		
nant_typical	···	Assumed typical number of antennas
nant_array	···	Assumed total number of antennas in the array
nant_all	···	Assumed total possible number of antennas over all arrays
Calculated Values		
pb	″	Calculated primary beam (HPBW)
imsize	pixels	Estimated image size
cell	″	Estimated cell size for imaging
nbase_typical	···	Calculated typical number of baselines
nbase_array	···	Calculated number of baselines for entire array
nbase_all	···	Calculated number of baselines for all arrays
weights_all	···	Observing time based weights (per mous only).
Information from Pipeline (all per MOUS only)		
predcubesize	Gbyte	Predicted cube size with no mitigation
mitigatedcubesize	Gbyte	Mitigated cube size
initialprodsize	Gbyte	Initial product size
mitigatedprodsize	Gbyte	Mitigated product size

projects were excluded because they have different processing requirements than we consider here. We also do not include total power data in our estimates, although an estimate of the total power data rate distribution is provided in Appendix A. The metadata included information on the field of view, angular resolution, array (12-m or 7-m), whether or not the project is a mosaic, information on the frequency, bandwidth, spectral resolution for each spectral window (spw), and number of polarization products, bandwidth of each spw, and the number of targets in the Member Observing Unit Set (MOUS).[1] It also includes information on the scientific category and science keywords for each project.

Because this meta-data is determined from the proposal and observing information rather than derived by the ALMA Pipeline, it does not include the effects of any size mitigations that may have been applied by the Pipeline during processing (Kepley et al. 2023a; Hunter et al. 2023). However, the meta-data has been transformed to meet the International Virtual Observatory Alliance (IVOA) standards, which are wavelength-based rather than frequency-based. For our calculations, we have converted the resulting values to the units relevant for our estimates. The spw information was converted from wavelengths to frequencies. The resulting instrumental properties like number of channels was reverse engineered based on the table in the ALMA Technical Handbook (Cortes et al. 2022) that gives the number of channels for a given usable bandwidth and spectral resolution. Finally, we calculated the velocity resolution for each spectral window from the spectral resolution and the observing frequency.

We also calculated the image properties including primary beam, cell size, and imsize for our sample. The primary beam was calculated via

$$pb_{7m}[''] = 33.3\,(300/\nu_{spw}[GHz]) \tag{1}$$
$$pb_{12m}[''] = 19.4\,(300/\nu_{spw}[GHz]) \tag{2}$$

(Schieven 2022). The predicted cell size was calculated assuming a circular beam and five pixels per beam:

$$cellsize[''] = resolution['']/5.0 \tag{3}$$

Most interferometeric observations will not have circular beams: the shape of the beam depends on the declination of the science target and when and how long it is observed for. However, we are interested in the approximate image properties, so the assumption of circular beam will give us a reasonable estimate of the cell size. The assumption of five pixels per beam is consistent with best imaging practices for interferometric data and is the ALMA Imaging Pipeline default. Only in extreme cases where the data cannot otherwise be imaged is the number of pixels per beam reduced to 3 or 3.25 if robust = 2 (Hunter et al. 2023). Reducing the pixels per beam can result in poorer fits to the psf, increased the chances of divergence, and, in general, poorer image reconstructions.

The predicted unmitigated image sizes were generated using a heuristic similar to that used by the ALMA Pipeline (Hunter et al. 2023) to determine the ALMA image size. For single fields, the imsize

[1] Each MOUS is formed out of one or more executions of a single scheduling block. The science targets within each MOUS share the same calibration strategy.

is given by the following

$$s_{0.2pb} = 1.1\,(1.12/1.22)\sqrt{\frac{-\ln(0.2)}{\ln(2.0)}} \tag{4}$$

$$imsize[pixels] = \frac{3600.0\,fov[degrees]\,s_{0.2pb}}{cellsize['']} \tag{5}$$

The factor $s_{0.2pb}$ scales the image size down to the 0.2 level of the primary beam. For mosaics, the calculation is

$$imsize[pixels] = \frac{3600.0\,fov[degrees] + 0.7\,pb}{cellsize['']} \tag{6}$$

The above equation was chosen to the analogous to the logic in the ALMA Pipeline, which takes the largest distance between the pointing centers in a row and adds 1.5 times the primary beam to image out to approximately 0.2 of the mosaic primary beam. Here since we have the FOV instead of the distance between pointing centers the factor is reduced to 0.7. The image sizes produced by the above equation were cross checked against image sizes calculated by the Pipeline and they were in good agreement.

The scientific category returned by the ALMA Science Archive does not map directly to the five proposal categories currently used in the time allocation review process: circumstellar disks, exoplanets and the solar system, cosmology and the high redshift universe, galaxies and galactic nuclei, ISM, star formation and astrochemistry, and stellar evolution and the Sun. We have mapped the individual MOUSes to the current proposal categories by using the science keyword information, which provides a unique correspondence.

In addition to the meta-data obtained from our ALMA Archive search, we obtained additional information on our sample from a variety of sources. The eightieth percentile baseline ($L80$) was not available via *astroquery* so we obtained it from the web interface for the ALMA Archive. This value provides information about the Array configuration for ALMA 12-m data: more extended arrays will have larger $L80$ values. From the Phase 2 Generation (P2G) group (R. Simon, private communication), we obtained information on the integration (dump) time currently used in operations and the number of tunings for each MOUS. These values were used to estimate the data rate and determine the number of spectral windows per tuning for each science target in our initial sample. Finally, we obtained the total integration time for each scan intent and source combination in the individual execution blocks that make up the MOUSes via a query of the internal archive metadata (F. Stoehr, private communication). The sources included both the science targets and calibrators. These values were used to calculate the total observing time per MOUS, the total time on each science target, and the total time spent on calibration including bandpass, flux (if different from the bandpass calibrator), phase, check source (if needed) and polarization (if needed), but not including system temperature, WVR, or pointing observations. Since these times are derived from scans, they also do not include any overhead from scan set ups, etc. We also included information that was extracted from the Pipeline weblog html for Cycle 7 (R. Indebetouw and A. Lipnicky, private communication) and Cycle 8 (I. Toledo, private communication). This information includes the predicted maximum cube size, mitigated maximum cube size, predicted initial product size, and mitigated product size. All of this information is available only per-MOUS since the ALMA Pipeline currently processes per MOUS. It is also only available for the MOUSes that were both calibrated and imaged by the

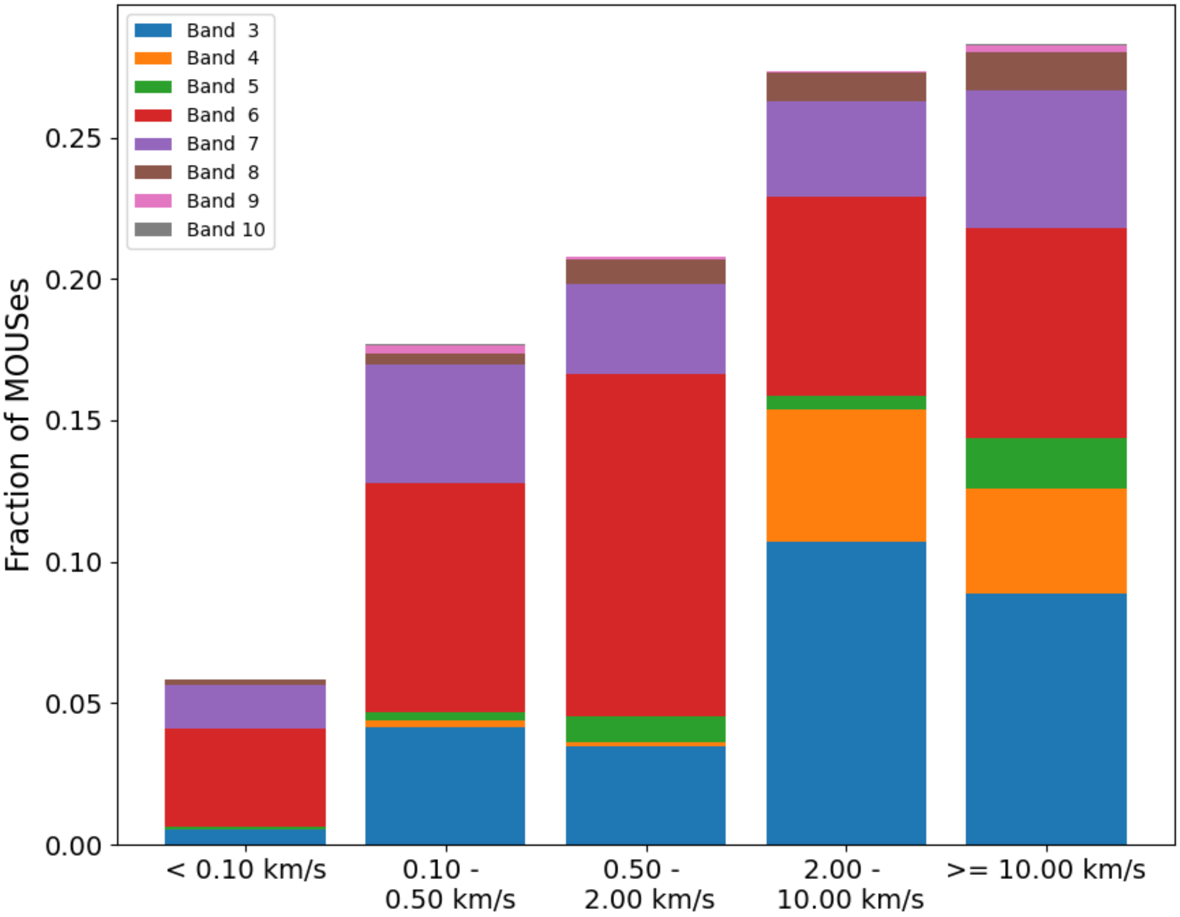


Figure 1. Distribution of the velocity resolutions from the finest spectral resolution spw per MOUS in our initial sample of ALMA cycle 7 and 8 data sets. The fraction per receiver band in each bin is shown.

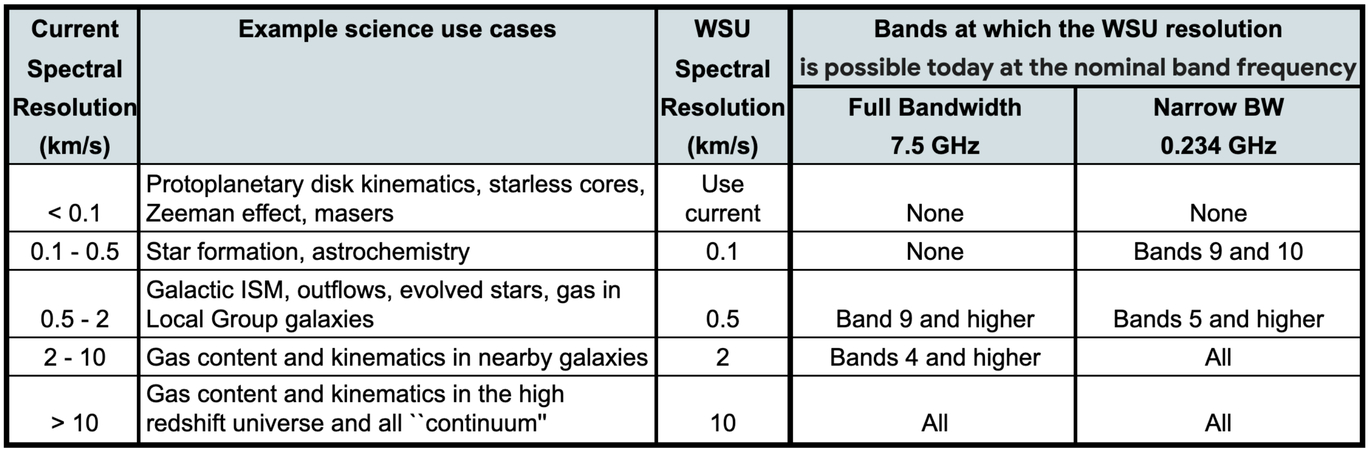
Pipeline (4718 out of 5200 total MOUSes). In general, unless otherwise noted, we plot quantities for the entire sample rather than just the Pipeline-processed subset.

We summarize the inputs to our databases in Table 1. Early versions of this database were used in Kepley et al. (2023a) and Kepley et al. (2023b).

2.2. *Estimation of WSU values*

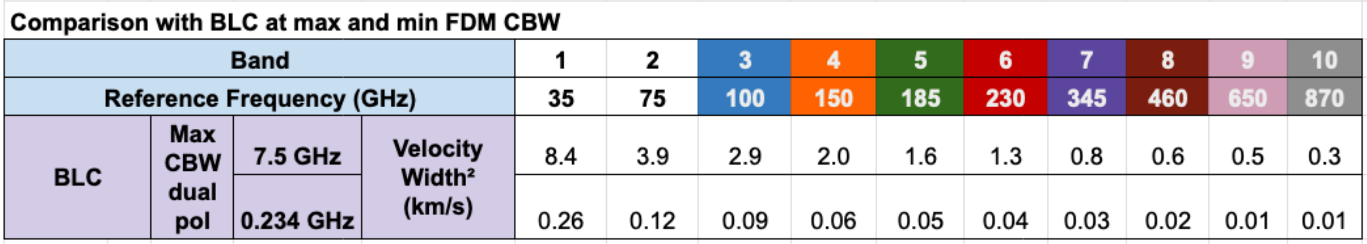
The next step is to transform the values in our database for ALMA Cycle 7 and 8 to what might be requested by PIs with the expanded capabilities afforded by the WSU: in particular access to significantly more spectral channels, so that a trade-off between required spectral resolution and bandwidth will rarely be necessary in the future. We define the WSU era as starting when the new WSU correlator for the 12m and 7m-arrays called the Advanced Technology ALMA Correlator (ATAC) is first used for science observations late this decade. Since the WSU will not add antennas to the array, we assume that the 12-m Array will have 47 antennas and the 7-m Array with have

Table 2. Velocity Bins Used to Convert Current Spectral Resolution to WSU Spectral Resolution



Current Spectral Resolution (km/s)	Example science use cases	WSU Spectral Resolution (km/s)	Bands at which the WSU resolution is possible today at the nominal band frequency	
			Full Bandwidth 7.5 GHz	Narrow BW 0.234 GHz
< 0.1	Protoplanetary disk kinematics, starless cores, Zeeman effect, masers	Use current	None	None
0.1 - 0.5	Star formation, astrochemistry	0.1	None	Bands 9 and 10
0.5 - 2	Galactic ISM, outflows, evolved stars, gas in Local Group galaxies	0.5	Band 9 and higher	Bands 5 and higher
2 - 10	Gas content and kinematics in nearby galaxies	2	Bands 4 and higher	All
> 10	Gas content and kinematics in the high redshift universe and all ``continuum"	10	All	All

Table 3. Highest spectral resolution possible today for different ALMA bands.



Comparison with BLC at max and min FDM CBW													
Band				1	2	3	4	5	6	7	8	9	10
Reference Frequency (GHz)				35	75	100	150	185	230	345	460	650	870
BLC	Max CBW dual pol	7.5 GHz	Velocity Width^2 (km/s)	8.4	3.9	2.9	2.0	1.6	1.3	0.8	0.6	0.5	0.3
		0.234 GHz		0.26	0.12	0.09	0.06	0.05	0.04	0.03	0.02	0.01	0.01

Note: CBW = Correlated Bandwidth

10 antennas.[2] Both values are typical of the number of antennas seen in operations today. We also assume that the image size and cell size will not change as a result of the WSU since the antenna sizes and configurations used by ALMA are not anticipated to change. Finally, we assume that the total number of hours available for science observing with ALMA is unlikely to change significantly in the WSU era since it is set by the length of the cycle and the amount of hardware and software maintenance required (as well as periods of poor weather). Thus we assume the current allocation for science observing time per cycle (∼4,300 hours) will remain the same for the WSU era. What will change in the WSU era are the values related to the correlator setup: total bandwidth, total number of channels, bandwidth per spectral window, spectral resolution and corresponding frequency resolution, number of channels per spectral window, integration (dump) time, and number of polarization products.

The total bandwidth, and thus the total number of channels, will evolve as the receivers are gradually updated to wider bandwidth and ATAC is upgraded from its initial deployment with 2x BW correlation capacity to 4x BW. To capture this aspect of the WSU project, we divide the WSU timeline into two stages: early and later. In the early WSU stage, some receivers (Bands 2, 6v2, and 8v2) will be upgraded to provide at least 2 times the current bandwidth (16 GHz) with a goal of 4 times the current bandwidth (32 GHz) and a new correlator, ATAC, will provide two times the bandwidth available today (16 GHz). We note that the upper end of the new wideband Band

[2] While the 12-m and 7-m arrays have a total of 50 and 12 antennas, respectively, observations with all the antennas in the array are rare because at any one time a few antennas are out of the array for maintenance.

2 receiver encompasses the current Band 3 frequency range (84-116 GHz). We thus assume that all Band 3 frequency range observations are done using the new Band 2 receiver, and thus have 16 GHz of bandwidth available in the early WSU, as this is the most conservative with respect to data rate and volume. All other receivers will retain their current bandwidth. We note that Bands 9 and 10 already provide 16 GHz of bandwidth. The science operations for this stage are estimated to begin in approximately 2029. In the later stage, ATAC will be upgraded to correlate four times the bandwidth. All receivers are assumed to have 32 GHz of bandwidth, with the exception of Band 1. That band is limited to 16 GHz because of constraints from atmospheric transmission and antenna optics.

For WSU, the options for the spectral resolution and number of channels are constrained by the properties of ATAC. This correlator provides 80 (early WSU: 2x bandwidth) to 160 (later WSU: 4x bandwidth) frequency slices. Each frequency slice has 200 MHz (usable) bandwidth with 14880 channels that can be binned from the native channel bandwidth of 13.5 kHz to coarser spectral channels. Frequency slices with the same requested level of channel binning that are frequency contiguous can be stitched together in the correlator to provide wider spectral windows. Due to the FFX architecture of ATAC, each channel is nearly independent of its neighbors and the use of a windowing function to avoid spectral aliasing is not anticipated (i.e., such as the online Hanning smoothing employed to improve the spectral quality of the 64-input Correlator, also referred to as the BLC, data which is of a FXF architecture).

As indicated above, ATAC can flexibly combine anywhere from 1 to 80 frequency slices into individual science spectral windows as long as they are frequency contiguous and have the same level of channel binning.[3] The widest potential spectral is therefore 16 GHz per pol in bandwidth. Some science use cases would benefit from such wide contiguous spectral coverage. For example, in Band 10, a 2 GHz spectral window is only 690 $\mathrm{km\,s^{-1}}$ wide, which makes it difficult to detect very broad lines from AGN outflows (Carpenter et al. 2022a). However, we anticipate that the use of such wide spws will likely be restricted to coarse resolution, as a limit to the maximum data rate per spw will likely be imposed to facilitate downstream processing (i.e., the maximum number of channels employed for a single spw cannot be greater than some number, to be determined). In this memo, we assume that all spectral windows are made up of 10 frequency slices to form windows 2.0 GHz wide, similar to the maximum spectral window bandwidth we have today (1.875GHz). Recently the ALMA project has adopted a cap of 80,000 channels per spectral window. This is broadly in agreement with our assumptions here, but would limit Band 1 spectral windows to a maximum bandwidth of 1 GHz. We discuss the implications of this decision in Section 4.2.

The number of spectral windows is determined by taking the total bandwidth at each WSU stage and dividing by the spectral window bandwidth (2.0 GHz). Although again we note that the ALMA project has not made a decision on spectral window widths and thus the final number of spectral windows will likely differ from what is presented here. In the early WSU, there will then be between 4 and 8 spectral windows. In later WSU, up to 16 spectral windows could be available depending on the final bandwidth of the receivers. We note again that the number of spectral windows depends on our assumption about the width of the individual spectral windows.

[3] The total number of frequency slices for the four times bandwidth expansion will be 160, but only 80 windows (16 GHz per pol) are possible per sideband for two sideband receivers; the data from different sidebands cannot be stitched.

Table 4. Minimum n_{avg} for 0.1 $\mathrm{km\,s^{-1}}$ velocity bin

Band	1	2	3[a]	4	5	6	7	8	9	10
minimum n_{avg}	1	2	3	5	6	8	10	15	20	32

[a] Referred to as Band 2 (high) in Carpenter et al. (2023)

We estimate the spectral resolution that might be requested by WSU PIs by using the finest spectral resolution requested per MOUS to assign the MOUS to one of the five velocity bins shown in Table 2. In general, the velocity bins have been defined so that they are each representative of a science case, although the correspondence is not always one-to-one. The distribution of our sample in each bin is shown in Figure 1. We then assume all MOUSes in a bin will use the smallest velocity resolution in an individual bin for WSU observations. For example, for projects in the 2 to 10 $\mathrm{km\,s^{-1}}$ bin, we assume that they will request 2 $\mathrm{km\,s^{-1}}$ spectral resolution for all spectral windows. Varying the spectral resolution between spectral windows introduces only minor changes to our estimates. See Section 4.1 for details.

This scheme is a minor improvement on IST Data Rate memo scenario 2 (Carpenter et al. 2022b). We have added an additional bin to account for a common trade-off made today by Band 6 observations to sacrifice spectral resolution for increased bandwidth. While this procedure may slightly overestimate the requested spectral resolution, it will likely not be a significant overestimate given the scientific limitations imposed by the current ALMA capabilities.

We emphasize that the finer velocity resolution bins in Table 2 are unavailable today over the full potential correlated bandwidth at all bands and in some cases unavailable even at the narrowest possible bandwidth for the lowest frequency bands. Table 3 shows the highest spectral resolutions possible today with BLC. Comparing those spectral resolutions to those listed in Table 2, we see that today we cannot obtain 0.1 $\mathrm{km\,s^{-1}}$ velocity resolution over the full bandwidth for any band (a key science driver for the WSU) and only for Bands 9 and 10 in the narrowest possible spectral windows. Therefore, in the absence of the need to give up bandwidth when higher spectral resolution is needed, we expect that the WSU will generally lead PIs towards using higher spectral resolution compared to today (with the exception of continuum-only projects).

We convert the estimated PI-requested WSU spectral resolution into a WSU frequency resolution. First we convert the current spectral resolution into frequency resolution ($\Delta\nu$). Then the number of WSU channels that give that resolution is

$$n_{avg} = floor(\Delta\nu/13.5\,\mathrm{kHz}) \tag{7}$$

We enforce a band-dependent minimum n_{avg} for data sets with 0.1 $\mathrm{km\,s^{-1}}$ velocity resolution using the values derived in the Data Rate Ramp Up Working Group report Table 2a (Carpenter et al. 2023). See Table 4 for a list of the values. Next we derive the total number of channels per 2.0 GHz spectral window ($\mathrm{BW_{spw}}$)

$$n_{chan,spw} = 10^6 BW_{spw}[GHz]/(13.5\,\mathrm{kHz}\;n_{avg}) \tag{8}$$

We adopt an integration time of 3.072s for the 12m-array and 9.984s for the 7m array. See Carpenter et al. (2023) for a more extended discussion on why these values were selected. In brief, the 12m value was chosen as a compromise between reducing data rates with longer integration times versus allowing better WVR corrections and self-calibration to reduce atmospheric decoherence. Today, although ~ 3 seconds is recommended for long-baselines, ~ 6 seconds is typically used for FDM data for all configurations to reduce the overall data rate. The 9.984s value for the 7m-array (which does not have WVR receivers) is similar to the value used today (10.08s).[4]

Although ATAC will always produce all four polarization products, we assume that the XY and YX polarizations will be pruned from the data for the cases where the PI does not request the relevant polarization calibration.[5] Therefore, the number of polarization products produced by WSU will be the same as today with one exception: single polarization cases. Single polarization is used today to gain a factor of two in velocity resolution at the expense of $\sqrt{2}$ lower sensitivity. Unlike the BLC, ATAC does not need to trade spectral resolution to achieve dual or full polarization. Thus, single polarization observing modes are not expected to be employed in the WSU era and therefore, single polarization are treated as dual polarization in this analysis.

Finally, we have removed from the sample cases that need full polarization at the highest spectral resolution across the entire WSU bandwidth (Band 3 with 0.1 $\mathrm{km\,s^{-1}}$ channels) since we are unaware of any science cases that would require this: there are only a few spectral lines that could be used to probe magnetic fields through the Goldreich-Kylafis or Zeeman effects. This removal only affected 8 MOUSes out of the 5200 present in our sample. In a future memo, we will investigate what the effects of non-uniform spectral windows and re-incorporate these cases into our sample.

2.3. *Band 1 and 2 estimates*

Our sample was derived from ALMA Cycles 7 and 8 and thus does not include any observations from either of the newest ALMA Bands 1 (35-50 GHz, available midway through Cycle 10) and 2 (67-116 GHz, presently in early stages of initial deployment). However, these Bands are important to include the current analysis, because as the lowest frequency ALMA bands, they will be the most challenging in terms of the number of channels required to span the full bandwidth for a given velocity resolution. For the same velocity resolution and total bandwidth, Band 1 requires approximately twice the number of channels as Band 2. However, it is notable that even after future upgrade, Band 1 cannot have more than 16 GHz total bandwidth per polarization in a single sideband (1SB) configuration due to constraints from the atmospheric transmission and antenna optics, while Band 2 is already capable of 32 GHz per polarization: IF=16 GHz per sideband per polarization in a two single band (2SB) configuration. Therefore, they will need a similar number of total channels to cover their full IF bandwidths for a given spectral resolution.

Given the large impact of these bands on the resulting data property estimates, we have made some additional assumptions to enable us to include contributions from these bands in our analysis, at least to first order. We first assume that the total time available for observations on ALMA remains fixed and thus any new receivers reduce the time spent observing with other receivers. We assume that Bands 1 and 2 will each have 5% of the total time on the 12m array and 3% of the total time on

[4] In the WSU era, high time cadence observations might use significantly shorter integration times than those adopted here. However, shorter integration times would require a trade off in the number of channels employed (through spectral binning to coarser spectral resolution) to reduce the overall data rate.

[5] Polarization calibration imposes a significant observational overhead and thus is not feasible for cases where the science does not require it.

the 7m array based on an analysis of band usage trends from Cycle 4 through 10. These values are consistent with the usage patterns of Bands without low-level CO transitions (Bands 4, 5, 8, 9, and 10). Furthermore, we assume that this time can be taken from Bands 3, 6, and 7, which are the most popular current receiver bands, in equal proportion. We randomly select MOUSes from these bands from our sample and discard them from the sample to make room for Band 1 and 2 MOUSes. The second assumption is that Bands 1 and 2 will have a similar distribution of science use cases to the current Band 3 proposal statistics. To generate replacement Band 1 and 2 MOUSes, we randomly selected MOUSes from the Band 3 projects remaining in the sample. The total number of MOUSes selected was set to match the assumed percentages of Band 1 and 2 MOUSes for the 12m and 7m array given above. For the selected Band 3 MOUSes, we set their frequencies to fiducial values for Band 1 or 2 (39 or 75 GHz) and their total bandwidth to the appropriate value for the early and later WSU stages. We then recalculated the field of view, the resolution, the primary beam, the cell size, frequency resolution, number of channels, and number of spectral windows. As above, we remove any projects that would request full polarization at the highest spectral resolutions (0.1 $\mathrm{km\,s^{-1}}$) in Bands 1 and 2 on the basis that they are not realistic science use cases. We generated 50 realizations of our sample with Bands 1 and 2 to characterize the distributions that result from the inclusion of these Bands.

2.4. *Derived WSU Data Properties*

Using the estimated WSU data properties described in Section 2.2, we calculate the following data estimated properties:

- the data rates,
- the visibility rates,
- the visibility data volumes,
- the estimated cube sizes
- the estimated product sizes

We note that our assumptions on the size of spectral windows has no effect on either the data rate and visibility data volume calculations presented in this memo. Of the above quantities, the estimated cube sizes depend most strongly on our assumption of 2 GHz bandwidth for the spectral windows. The estimated product size also does depend on the total number of spectral windows, but very weakly (see below).

The instantaneous or peak data rate is given by the following

$$datarate\,[GB/s] = (2.0 n_{byte}\, n_{apc}\, n_{base} + 4 n_{ant}) * n_{chan}\, n_{pol} / t_{int}[s]/10^9 \tag{9}$$

where n_{byte} is 2 for cross-correlations (16-bit), n_{apc} is the number of WVR streams and is assumed to be 1, n_{ant} is the number of antennas, n_{base} is the number of baselines and is equal to $n_{ant} * (n_{ant} - 1)/2$, n_{chan} is the number of channels over all spectral windows ($n_{chan} = n_{chan,spw}\, n_{spw}$), t_{int} is the integration time.

The volume of visibility data per MOUS is then

$$datavol_i[GB] = datarate_i * time_{obs,i} \tag{10}$$

and the total data volume is

$$datavol[PB] = \sum_{i=0}^{i=n_{mous}} datavol_i[GB]/10^6 \tag{11}$$

This estimate only includes the flux, bandpass, phase, and check source calibrators as well as the science targets. It does not include the pointing, focus, and Tsys observations. We separately calculate the visibility data volume for the calibrators that are present and the science targets as well as the combined visibility data volume. We note that the above assumes that all the data for the calibrators and science targets use the same visibility integration time and channelization, which may not be the case for the WSU (see discussion in Section 4).

We assume that, as today, we image each science target and spectral window separately, which results in one cube per science target/spw combination.[6] The science target cube sizes are then

$$cubesize[GB] = 4\, imsize^2 n_{chan,spw}/10^9 \tag{12}$$

where $n_{chan,spw}$ is the number of channels per spectral window. The continuum image size (also known as the mfs image) for science targets is

$$mfssize[GB] = 4\, imsize^2/10^9 \tag{13}$$

In both of the equations above, the factor of 4 corresponds to the bytes per pixel. The total product size for an individual science target is the sum of the cube and mfs image size over all spectral windows

$$productsize_{src}[GB] = 2\ (cubesize + mfssize)\, n_{spw} \tag{14}$$

where n_{spw} is the number of spectral windows per WSU stage and the factor of two accounts for the fact that both the images and primary beams are delivered to the PI. We note that the product size only depends weakly on our assumptions about the number of spectral windows. The total size of the cubes, which only depends on the total aggregate number of channels, will dominate the estimate. This product size also is only for the science targets. It does not include any continuum images or cubes of the calibration sources. The product size estimated here is the actual product size that would be produced on disk following imaging. When stored in the Archive, products like the primary beam file may be compressed to save on storage space; astronomical images usually do not compress well since they are random noise to first order. Today the overall effect of this compression is to reduce the factor of 2 in Equation 14 to a factor of 1.3 (F. Stoehr, private communication).

All the above quantities are calculated per science target. We have also aggregated the quantities to produce a per-MOUS version of the database. In this database, the individual science target properties are removed (science target name and time per science target). The total product size per MOUS is calculated by summing this value over all science targets in the MOUS.

2.5. *Summary of Final Database Properties*

[6] Image cubes are not currently made for calibrators.

Table 5. Overview of Databases

Version	Row Quantity	Number of Rows		
		Cycle 7	Cycle 8	Total
Initial Sample	MOUS/source/spw	51,368	50,107	101,475
WSU (Per Source)	MOUS/source	11,501	10,842	22,343
WSU (Per MOUS)	MOUS	2,712	2,480	5,192

Table 6. Calculated BLC/ACA Data Properties

Variable	Unit	Description
blc_cubesize	Gbyte	Maximum unmitigated BLC cube size (per mous)
blc_cubesize_sum	Gbyte	Sum of unmitigated BLC cube sizes (per mous)
blc_productsize	Gbyte	Unmitigated BLC product size (per src for src database and per mous for mous database)
blc_datarate_typical	$\frac{\text{Gbyte}}{\text{s}}$	BLC/ACA data rate
blc_visrate_typical	$\frac{\text{Gvis}}{\text{h}}$	BLC/ACA visibility rate
blc_datavol_typical_target	Gbyte	BLC/ACA visibility data volume for a single science target (per source)
blc_datavol_typical_target_tot	Gbyte	BLC/ACA visibility data volume for science target(s) (per mous)
blc_datavol_typical_cal	Gbyte	BLC/ACA visibility data volume for calibrators (per mous)
blc_datavol_typical_total	Gbyte	BLC/ACA visibility data volume for all science targets and calibrators (per mous)
blc_nvis_typical_target	Gvis	Total number of BLC/ACA visibilities for a single science target (per source database only)
blc_nvis_typical_target_tot	Gvis	Total number of BLC/ACA visibilities for all science target(s) (per mous value)
blc_nvis_typical_cal	Gvis	Total number of BLC/ACA visibilities for calibrators (per mous)
blc_nvis_typical_total	Gvis	Total number of BLC/ACA visibilities for calibrators and all science targets (per mous)

The result of the above calculations is three databases: the initial sample (from Cycles 7 & 8), the per-science-target estimated WSU properties, and the per-MOUS estimated WSU properties. The three different databases are flat files that have different numbers of rows. For the initial sample database, each row describes a unique MOUS, science target, and spectral window combination. Thus the total number of rows is the product of the number of science targets times the number of spectral windows per MOUS summed over all MOUSes. For the per-science-target estimated WSU properties database, each row is a unique combination of MOUS and science target, since the spectral windows are all assumed to the same. The total number of rows is the number of science targets per MOUS summed over all MOUSes. For the per-MOUS estimated WSU properties database, each row contains the aggregated information per-MOUS. The total number of rows in this database corresponds to the total number of MOUSes.

We provide an overview of the databases, the per-row quantities, and number of rows in Table 5. We summarize the resulting estimated quantities for both current ALMA (Table 6) and ALMA WSU (Table 7). For the columns labeled “both”, we are taking the weighted average of all 12m and 7m data with the weights being the fraction of the observing time for each MOUS divided by the total observing time for all MOUSes.

Table 7. WSU Data Properties

Variable	Unit	Description
Assumed WSU Properties		
wsu_freq	GHz	Frequency at center of bandwidth for WSU observations
wsu_npol	···	Number of polarization products
wsu_tint	s	integration (dump) time
wsu_bandwidth_early	GHz	total bandwidth for early WSU
wsu_bandwidth_later	GHz	total bandwidth for later WSU
wsu_bandwidth_spw	GHz	bandwidth per spectral window
wsu_nspw_early	···	number of spectral windows for early WSU
wsu_nspw_later	···	number of spectral windows for later WSU
wsu_specwidth_stepped2	kHz	assumed ATAC spectral resolution
wsu_chanavg_stepped2	···	number of ATAC channels that will be averaged to provide that resolution
wsu_velres_stepped2	$\frac{\mathrm{km}}{\mathrm{s}}$	corresponding velocity resolution
Stage Independent Calculated WSU Properties		
wsu_nchan_spw_stepped2		number of channels per spectral window
mfssize	Gbyte	size of one mfs image (same for both WSU and BLC/ACA)
wsu_cubesize_stepped2	Gbyte	size of one WSU cube
Stage Dependent Calculated WSU Properties		
wsu_productsize_[early,later]_stepped2	Gbyte	product size both per source and per mous
wsu_datarate_[early,later]_stepped2_typical	$\frac{\mathrm{Gbyte}}{\mathrm{s}}$	WSU data rate
wsu_visrate_[early,later]_stepped2_typical	$\frac{\mathrm{Gvis}}{\mathrm{h}}$	WSU visibility rate
wsu_datavol_[early,later]_stepped2_typical_target	Gbyte	WSU visibility data volume for a single science target observation (per src)
wsu_datavol_[early,later]_stepped2_typical_target_tot	Gbyte	WSU visibility data volume for all science target observations (per mous)
wsu_datavol_[early,later]_stepped2_typical_cal	Gbyte	WSU visibility data volume for all calibrator observations (per mous)
wsu_datavol_[early,later]_stepped2_typical_total	Gbyte	WSU visibility data volume for calibrator and science observations (per mous)

Table 8. Overview of Data Rate Properties for BLC/ACA

		BLC/ACA		
		12m	7m	both
Data Rate	Median (MB/s)	5.73	0.25	2.29
	Time Weighted Average (MB/s)	9.22	0.32	5.37
	Maximum (MB/s)	45.84	0.89	45.84
Number of Channels	Median	2,185	4,518	3,677
	Time Weighted Average	4,626	6,128	5,276
	Maximum	76,800	24,576	76,800

Table 9. Overview of Data Volume Properties for BLC/ACA

		BLC/ACA		
		12m	7m	both
Visibility Data Volume (Total)	Median (GB)	16.93	0.93	7.29
	Time Weighted Average (GB)	121.44	12.25	74.22
	Maximum (GB)	1672.92	83.03	1672.92
	Total per cycle (TB)	86.46	2.28	88.74
Visibility Data Volume (Science)	Median (GB)	11.07	0.54	4.59
	Time Weighted Average (GB)	84.76	8.58	51.82
	Maximum (GB)	1130.05	62.39	1130.05
	Total per cycle (TB)	59.98	1.52	61.50
Product Size (Total)	Median (GB)	8.79	0.27	2.53
	Time Weighted Average (GB)	528.47	3.21	301.32
	Maximum (GB)	43905.78	85.57	43905.78
	Total per cycle (TB)	620.96	1.68	622.64

Table 10. Overview of Data Rate Properties for WSU

		Early WSU			Later WSU		
		12m	7m	both	12m	7m	both
Data Rate	Median (GB/s)	0.060	0.001	0.015	0.136	0.002	0.035
	Time Weighted Average (GB/s)	0.220	0.005	0.127	0.513	0.011	0.296
	Maximum (GB/s)	1.741	0.026	1.741	3.481	0.052	3.481
Number of Channels	Median	19,807	24,627	22,818	46,385	51,566	47,509
	Time Weighted Average	73,764	114,205	91,250	171,226	244,219	202,786
	Maximum	592,592	592,592	592,592	1,185,184	1,185,184	1,185,184

Table 11. Overview of Data Volume Properties for WSU

		Early WSU			Later WSU		
		12m	7m	both	12m	7m	both
Visibility Data Volume (Total)	Median (TB)	0.155	0.004	0.061	0.366	0.008	0.153
	Time Weighted Average (TB)	3.170	0.178	1.876	7.427	0.378	4.379
	Maximum (TB)	88.656	3.283	88.656	177.312	6.565	177.312
	Total per cycle (PB)	2.067	0.036	2.103	4.815	0.077	4.892
Visibility Data Volume (Science)	Median (TB)	0.101	0.002	0.038	0.254	0.005	0.092
	Time Weighted Average (TB)	2.367	0.128	1.399	5.439	0.268	3.203
	Maximum (TB)	73.900	2.428	73.900	147.800	4.857	147.800
	Total per cycle (PB)	1.530	0.025	1.555	3.500	0.053	3.553
Product Size (Total)	Median (TB)	0.052	0.001	0.016	0.127	0.003	0.038
	Time Weighted Average (TB)	5.376	0.058	3.076	11.525	0.119	6.592
	Maximum (TB)	563.690	0.829	563.690	1127.379	1.658	1127.379
	Total per cycle (PB)	5.891	0.031	5.922	12.643	0.064	12.707

Table 12. Ratio Per Mous Between Estimated WSU data and BLC/ACA data

	Visibilities		Products	
WSU Stage	Median	Maximum	Median	Maximum
Early	7	260	4	1900
Later	18	690	9	3800

3. RESULTS

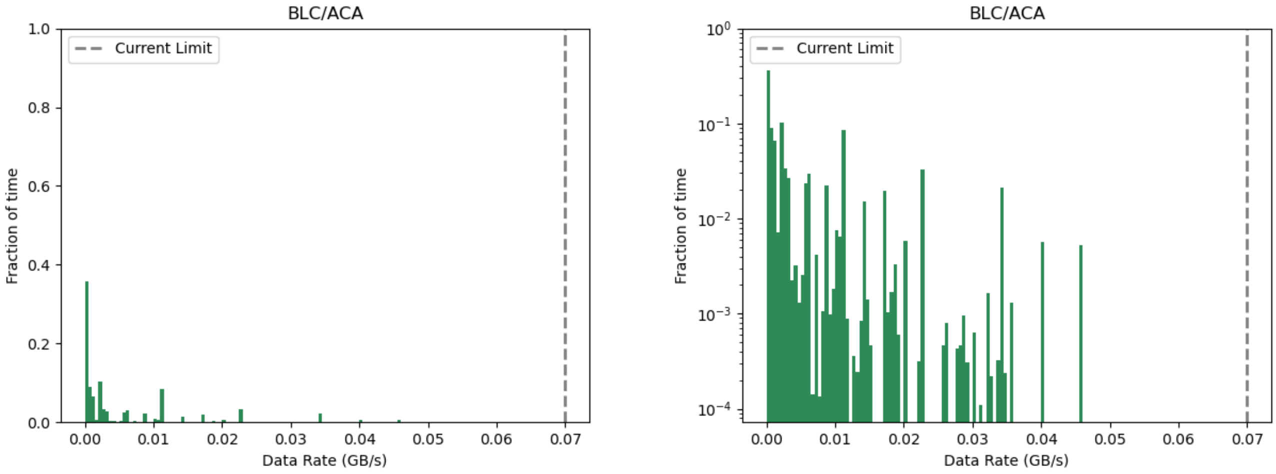
In this section, we describe the estimated properties of two cycles of WSU data including data rate, visibility data volume, and product size and compare them to present day limits on these quantities. We also provide estimates of number of spectral windows and cube sizes.

We have calculated the median, time-weighted average, and maximum for the data rate, aggregate number of channels, visibility data volumes (both total and science target only), and product size for each our realizations of the per-MOUS sample including Bands 1 and 2. The time-weighted average of these values is calculated by weighting each MOUS in the sample by the total observing time required for that MOUS. The observing time is the sum over all scans in a MOUS of the scan time for the science targets and the bandpass, phase, flux (if a separate flux calibrator is used), check source (if necessary) and polarization calibrators (if necessary). It does not include time spent on pointing, on system temperature or WVR measurements or on any overhead associated with setting up the observations. With this weighting, MOUSes with higher data rate that take relatively little observing time are down-weighted compared to MOUSes with the same data rate that take more observing time. In addition, we have also calculated the total visibility data volumes (both total and science target only) and product sizes for a single cycle by calculating the total over the two cycles in our sample and dividing by two. Since our sample includes both a long baseline and short baseline cycle, this should represent an average effective total per cycle. We then calculate the mean value of each of the above quantities over the 50 realizations of our sample. These values are given in Table 10 and 11.

For comparison with the WSU values, we present the same values derived for the current BLC/ACA in Tables 8 and 9. These statistical values were derived using the same methology as the statistical values for the WSU but the input values appropriate for the current ALMA. Table 12 gives the per MOUS ratio between the WSU estimates and the BLC/ACA input sample for the visibility data volume and products, i.e., the ratio between what the same project would produce today versus in the WSU era.

3.1. *Correlator Data Rate Estimates*

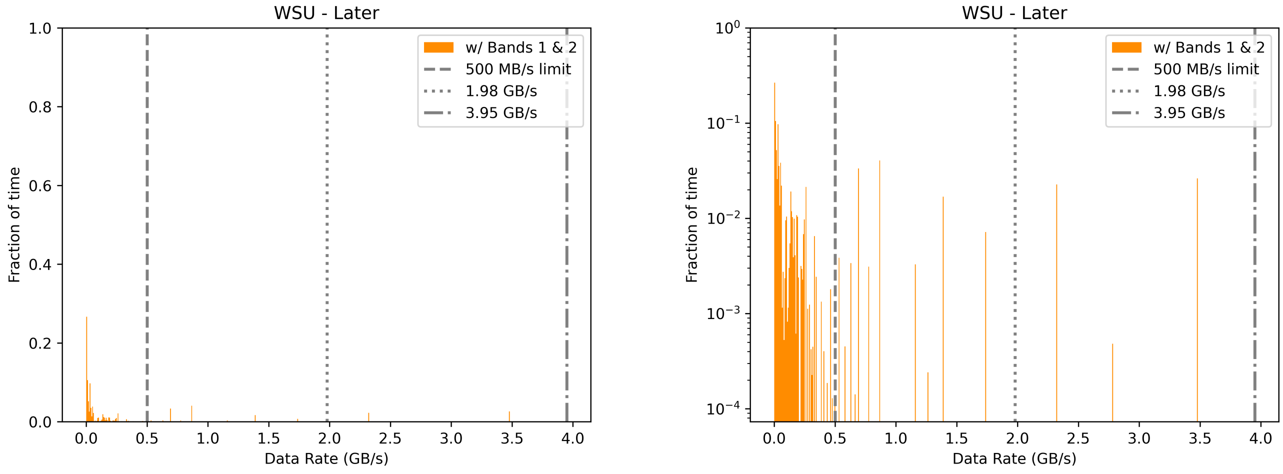
We start by examining the distribution of correlator data rates. Figure 2 shows a histogram of the current distribution of data rates from ALMA with the maximum data rate allowed (70 MB/s) shown as a dashed black line. The distribution of data rates peaks has a relatively low median of 2.3 MB/s, but has a long tail which extends out to much higher rate rates with a maximum rate of 46 MB/s, which is less than the maximum allowed data rate (70 MB/s). This discrepancy is by



a

b

Figure 2. Distribution of current BLC/ACA correlator data rates from ALMA over two ALMA cycles weighted by the fraction of observing time. Panels (a) and (b) show the same data, but have different y-axis scalings. The y-axis scaling in panel (a) is linear, while the y-axis scaling in panel (b) is logarithmic to better show the distribution of the long tail. The current data rate limit for the correlator (70 MB/s) is shown as a dotted vertical line. The current ALMA data rates do not presently approach the 70 MB/s limit by design. See associated text for explanation.



a

b

Figure 3. Distribution of estimate WSU correlator data rates over two ALMA cycles weighted by fraction of observing time for the later WSU stages. The data rates for the early WSU show a similar trend. The y-axis scaling for the left hand panel is linear, while the y-axis scaling for the right hand panel is logarithmic to better show the distribution of the long tail. Several notional data rate caps are shown with gray vertical lines.

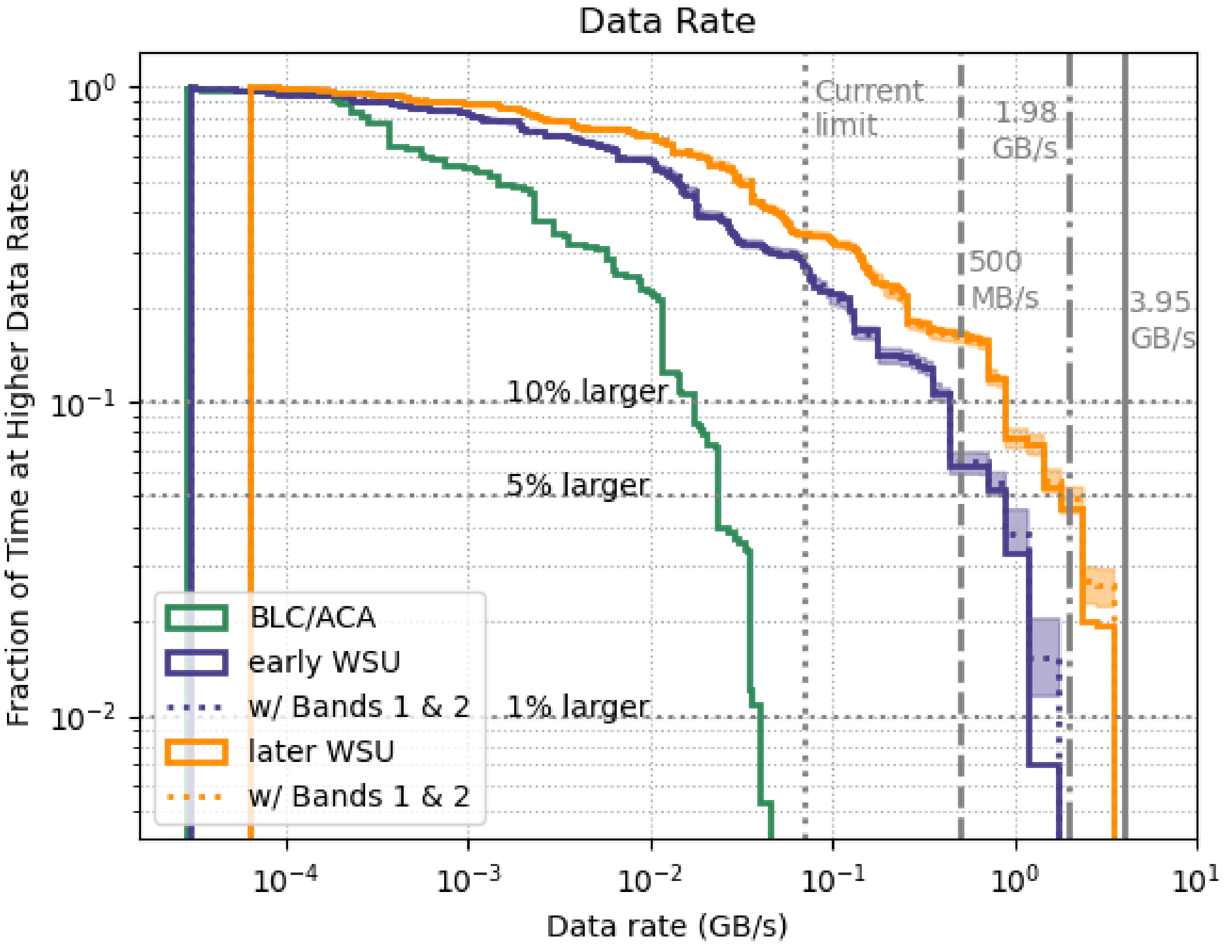


Figure 4. Complementary cumulative distribution of data rates weighted based on observing time for the current BLC/ACA and early and later WSU showing the fraction of observing time that the data rate will exceed the given value on the x-axis. The solid lines indicate the initial WSU estimates and the dotted lines indicate the median WSU estimates including Bands 1 and 2 with the shaded region indicating the upper and lower bounds of these estimates. Dotted horizontal lines indicate the thresholds for 10%, 5%, and 1% of the observing time. The current 70 MB/s data rate limit, a notional 500 MB/s limit and the 12m array peak data rates calculated in Carpenter et al. (2023, 1.98 GB/s for early WSU and 3.95 GB/s) are shown. In early WSU, 30% of the observing time will correspond to rates higher than the current data rate limit of 70 MB/s, while 6-10% of the observing time will correspond to data rates higher than 500 MB/s. These numbers will increase to 35% and 18% in the later WSU stages.

design. Phase 2 of the OT sets a 6s visibility integration time for full channelization BLC FDM modes by default, even for long baselines though the recommended value is 3s to help minimize decorrelation (e.g., tickets ICT-4139, ICT-11125, ICT-5938, and SCIREQ-2203). The reason for this that 3s integrations would overrun the current data rate limit when more than 40 antennas are used in the main array.

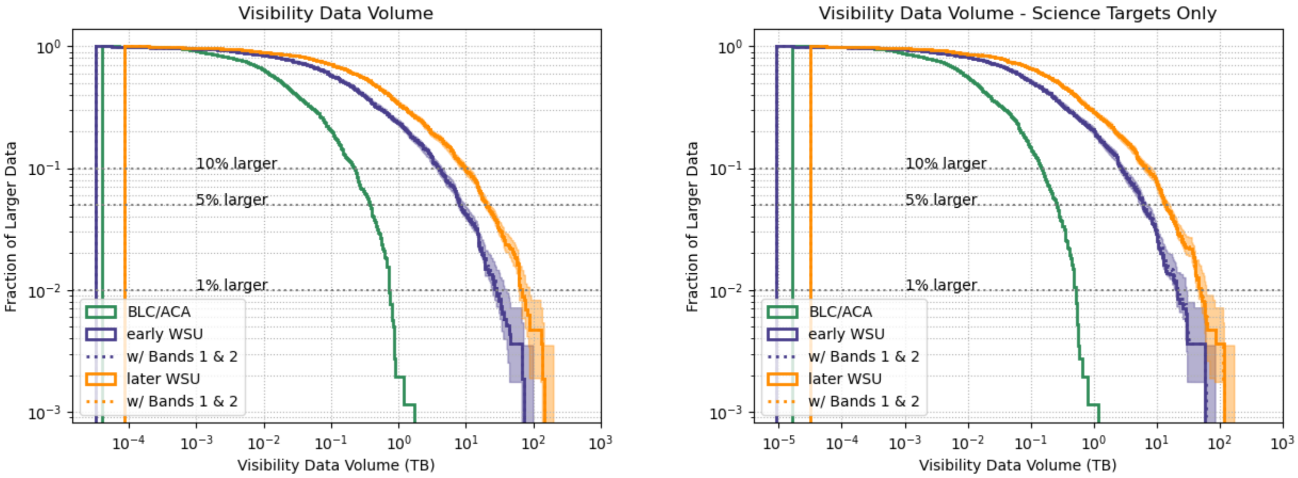
The trend of a strongly peaked distribution with a long tail extending out several orders of magnitude is also seen in the estimated later WSU correlator data rate shown in Figure 3. However, the overall magnitude of the values is much greater with a median data rate of 15-35 MB/s and a maximum data rate of 1.7 to 3.5 GB/s. These maximum data rates are slightly lower than the peak data rates for the 12m array of 1.98 GB/s for early WSU and 3.95 GB/s given in the Data Rate Ramp Up Working Group report Carpenter et al. (2023). The difference is because we have used a "typical" number of antennas for each array since we are trying to mimic the data rates the WSU would most likely produce, while Carpenter et al. (2023) uses the maximum number of antenna available in each array since that report is interested in the maximum possible data rate for the WSU.

While the projects in the long tail of the distribution will be some of the most challenging observations to process, they also represent a key part of the science case for the WSU, namely high (0.1-0.2 km/s) spectral resolutions over the full bandwidth. These observations are not possible with the current ALMA correlator (see Table 2) and are necessary for the three key science cases for the WSU. *We cannot fulfill the scientific requirements of the WSU without processing the projects in the long tail of data rates.*

Figure 4 compares the distribution of data rates weighted by observing time for the current ALMA correlators (BLC/ACA) and early and later WSU. The latter now includes estimates for ALMA Bands 1 and 2 as shaded regions. Since the high correlator data rate projects will be the most challenging, we plot the complementary cumulative distribution, i.e., the probability that an observation has a higher data rate, $P(X >= x)$, to highlight the differences in the tails of the distribution. We see that the largest difference in the tail of the distribution is between the BLC/ACA and early WSU. The later WSU stage moves the entire distribution to the right (more observing time spent at higher rates). From this plot, we see that in the early WSU stage approximately 30% of the observing time will have rates higher than the current 70 MB/s limit and 6-10% of the observing time will be at rates greater than 500 MB/s. These numbers will increase to 35% and 18% in the later WSU stage.

3.2. *Visibility Data Volume Estimates*

We use the WSU correlator data rates and the observing time from Cycles 7 and 8 to estimate the visibility data volumes for WSU, as described in Section 2.4. Recall that this observing time is the sum over the science target, bandpass, flux (if separate), phase, check (if necessary), and polarization (if necessary) calibrators and does not include the system temperature, WVR, and pointing observations. As a check on the estimated values, we compare the actual volume of visibility data in the ALMA Archive for ALMA Cycles 7 and 8 with our calculated values for the BLC/ACA. The total measured visibility data volume in the ALMA Archive for Cycles 7 and 8 is 197 TB (F. Stoehr, private communication). We estimate that the total volume of visibility data in the Archive for Cycles 7 and 8 is 181 TB, which is within 10% of the measured value. Note that we expect that our estimate of the visibility data volume would be less than the actual total because we exclude a small number of projects from our database that did not meet our criteria to be included in the sample (see Section 2.1).



(a)

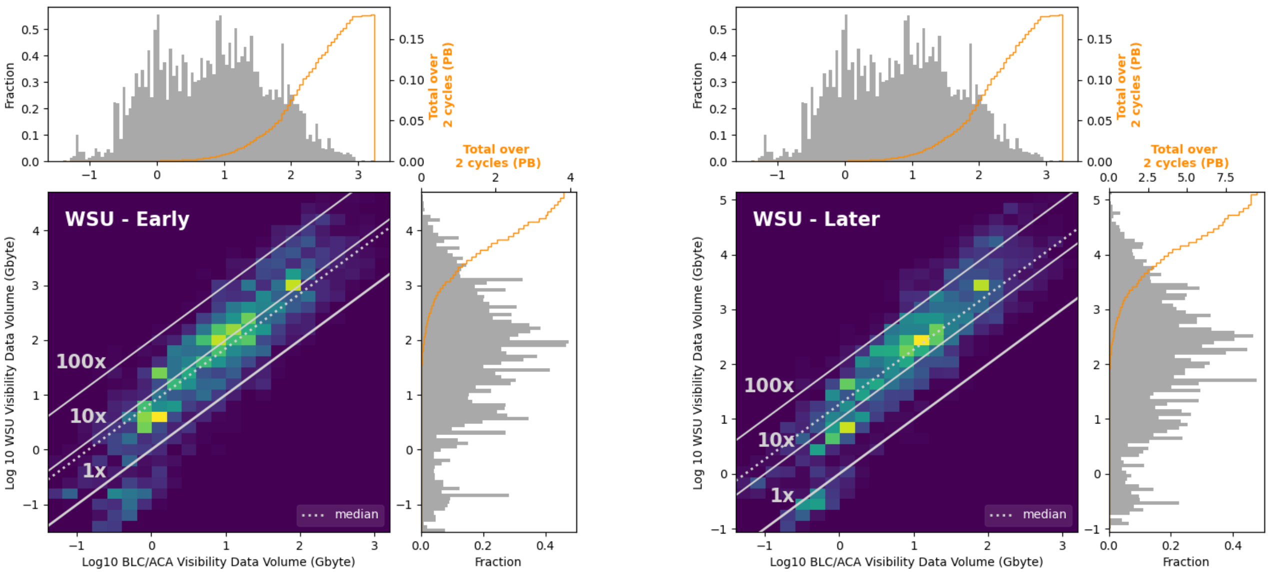
(b)

Figure 5. Complementary cumulative distribution of the visibility data volume for the BLC/ACA and the early and later stages of the WSU (including estimates for Bands 1 and 2) showing the fraction of times that the visibility data volume will be larger than the value shown on the x-axis. Panel (a) shows the total visibility data volume and panel (b) shows the visibility data volume for the science data only. The solid lines indicate the initial WSU estimates and the dotted lines indicate the median WSU estimates with Bands 1 and 2 with the shaded region indicating the upper and lower bounds of these estimates. Dotted horizontal lines indicate thresholds for 10%, 5%, and 1% of the MOUSes. For WSU, 15-35% of MOUSes will have visibility data volumes greater than that of the largest BLC/ACA project today (slightly over 1 TB).

Figure 5 compares the distribution of visibility data volumes for the BLC/ACA and the early and late stages of the WSU. The volume of visibility data from the largest WSU projects will increase by a factor or 50-100 compared to the BLC/ACA and 15-35% of MOUSes will have visibility data volumes greater than that of the largest BLC/ACA project today (slightly over 1 TB).

In addition to looking at the distributions of the data related quantities, we also compare the estimated early and later WSU visibility data volume to the current BLC/ACA values for each MOUS in Figure 6. It is important to note that these comparisons only include the WSU estimates for the entire original sample. They do *not* include the additional estimates for Bands 1 and 2 since Bands 1 and 2 were not in the original sample and thus have no direct comparison to today's data.

Depending on the stage, we find that the median increase in visibility data volume is 7 to 18 times the BLC/ACA value, but that some MOUSes have estimated WSU visibility data volumes up to 690 times the BLC/ACA value. Most of the increase in visibility data volume comes from projects in the tail of the distribution, again demonstrating the "tail wagging the dog" behavior seen in general for WSU data. As an aside, some of the WSU visibility data is smaller than the corresponding BLC/ACA visibility data. This difference is due to projects from the BLC/ACA with the coarsest velocity resolutions ($\geq 10\mathrm{km\,s^{-1}}$, i.e., TDM) that were observed in Cycle 7 and 8 with smaller integration times than we assume here: 1.008s for 7m data instead of 9.984s and 2.048s for 12m data instead of 3.072s. The difference in integration time dominates because these projects have the lowest numbers of aggregrate channels for both the BLC/ACA and WSU.



(a)

(b)

Figure 6. A two dimensional histogram showing the distribution of visibility data volume for the BLC and early (a) and later (b) stages of the WSU. The ratio of the WSU visibility data volume to the BLC/ACA visbility data volume is shown as solid diagonal lines with the ratios indicated next to the line. The median ratio is shown as a dotted line. The points below a ratio of 1 are due to lower integration values being used for the BLC/ACA than we assume here for the WSU for a subset of projects with low numbers of aggregate channels. The plot to the right of the main panel shows the distribution of the WSU visibility data volumes as a gray histogram with the cumulative distribution of data volume shown as an orange line. The top plot shows the equivalent for current BLC/ACA data. Since we are comparing present day data with future WSU data, we have included our Band 1 and 2 estimates for the WSU since there is no comparable present-day data.

The above results assume that the calibrator data will be taken at the same spectral resolution as the science target data. However, one possibility that is being considered for the WSU is to take the calibrator data at a lower spectral resolution if the spectral resolution of the science target is less than 1 $\mathrm{km\,s^{-1}}$. Doing this will reduce the overall volume of the calibrator data. Figure 7 shows the ratio of the calibrator and total data volume per MOUS. The calibrator visibility data volume is roughly 1/3 that of the total data visibility data volume, although with a large range. Reducing the spectral resolution of the calibrator data will directly reduce the number of channels and thus the data rate for the calibrators and the overall calibrator data volume. What will not change in this scenario is the science target visibility data volume, since that will be taken at the PI-requested spectral resolution.

We show the distribution of the science target only visibility data volumes in panel (b) of Figure 5. As one might expect the curves have moved to slightly lower values since we are only considering the science target data. However, the increase in the maximum size of the visibility data is still roughly

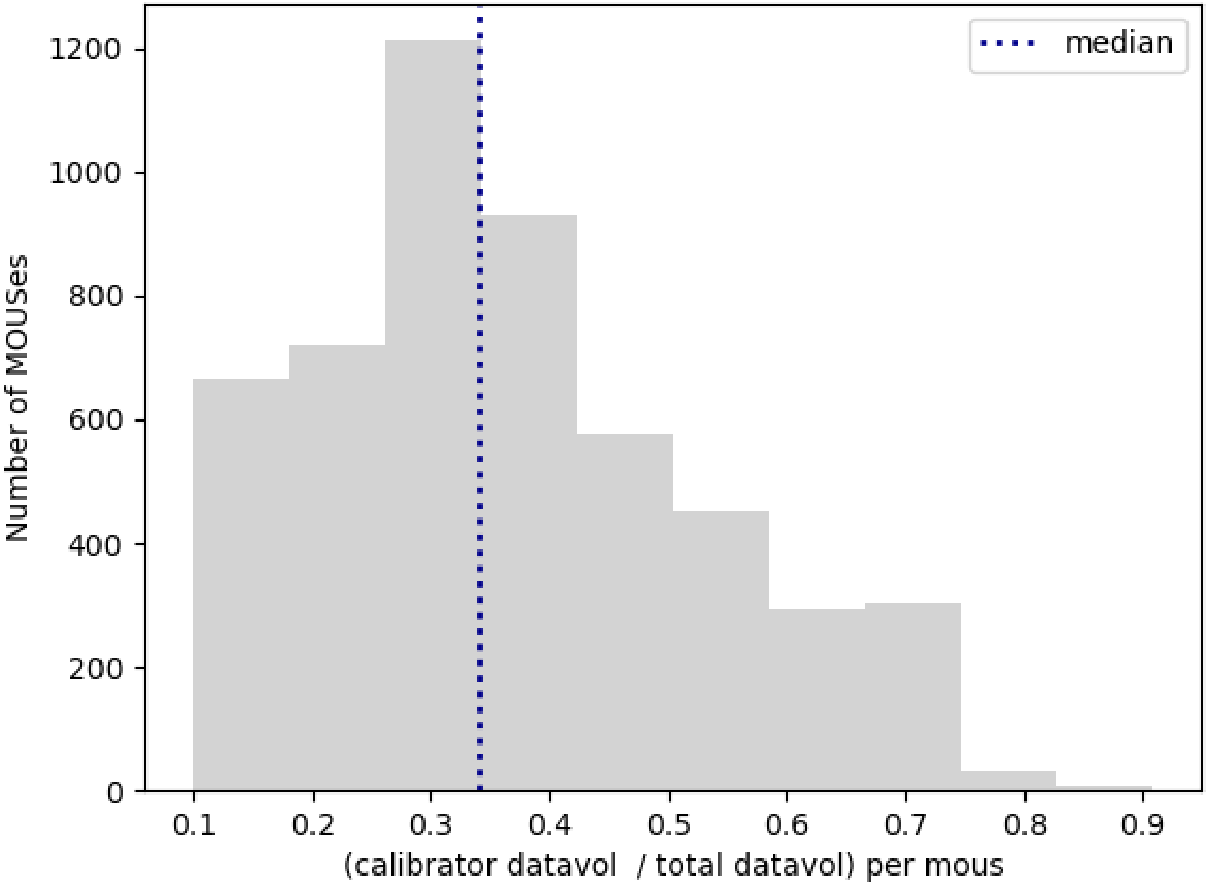


Figure 7. Ratio of the calibrator and total data volume per MOUS. The median is indicated by a dotted line. Recall, however that our science archive-based analysis does not include the pointing, WVR, or Tsys data so the actual fraction will be somewhat higher per MOUS.

two orders of magnitude. The fraction of WSU data larger than the largest current visibility data set (∼1 TB) has reduced slightly to 10-20% larger than the largest current visibility data volume.

3.3. *Product size estimates*

The raw visibility data produced by the correlator will be calibrated and imaged to produce the products used to do science. Here we define the products as continuum images and line cubes products for the science targets. We assume that a continuum image and cube is produced by each science target and spectral window (spw) combination. Furthermore, we assume that each product is imaged down to the 0.2 primary beam level, as is currently done in the ALMA Pipeline, and all channels are imaged for each cube, as required by the ALMA Archive. We note that the estimated number of spectral windows only has a minor effect on the estimated product size since the overall product size is dominated by the total size of the input cubes with the continuum images having only a small effect. The total size of the input cubes is just the sum of channels across all spectral windows multiplied by the image area, so it is independent of the estimated number of spectral windows. We

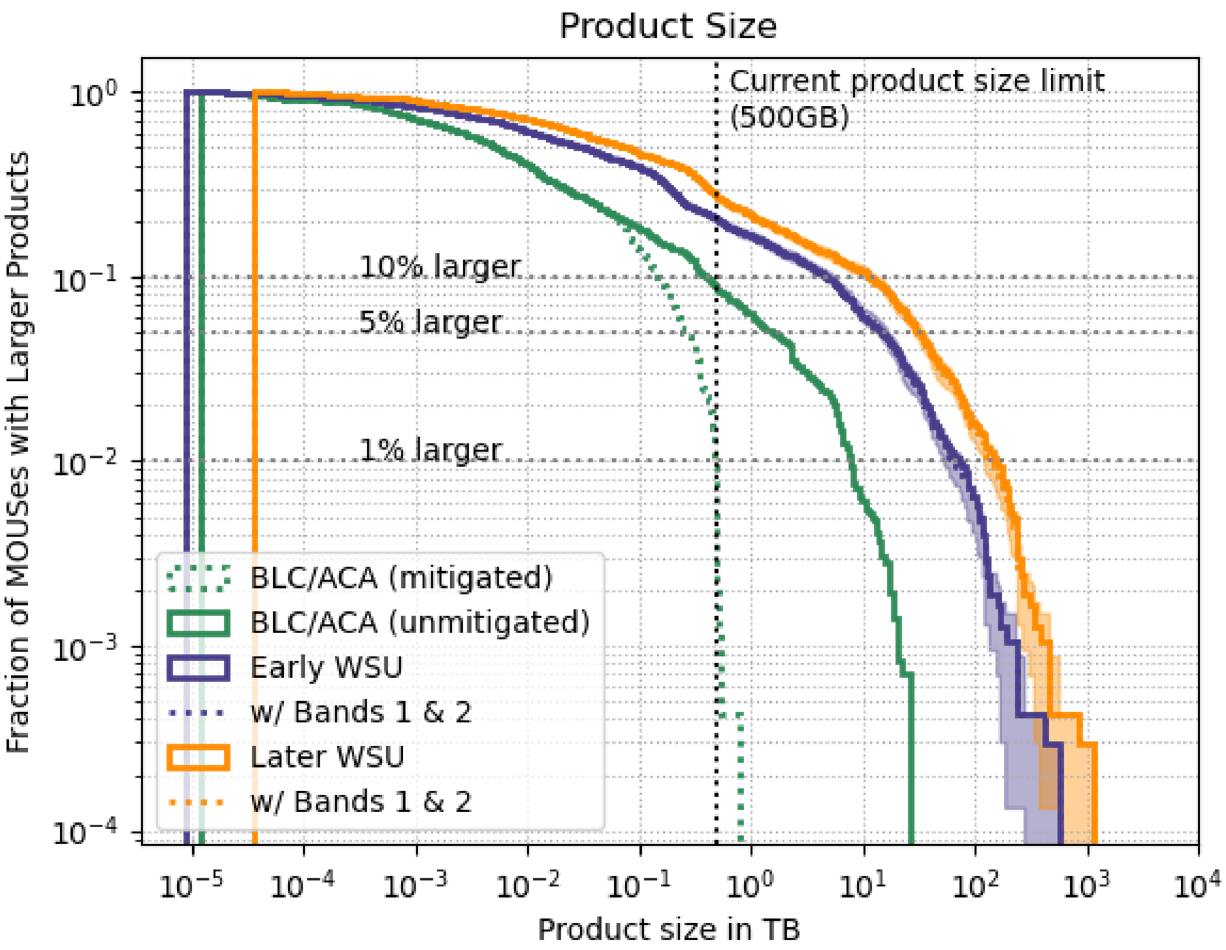
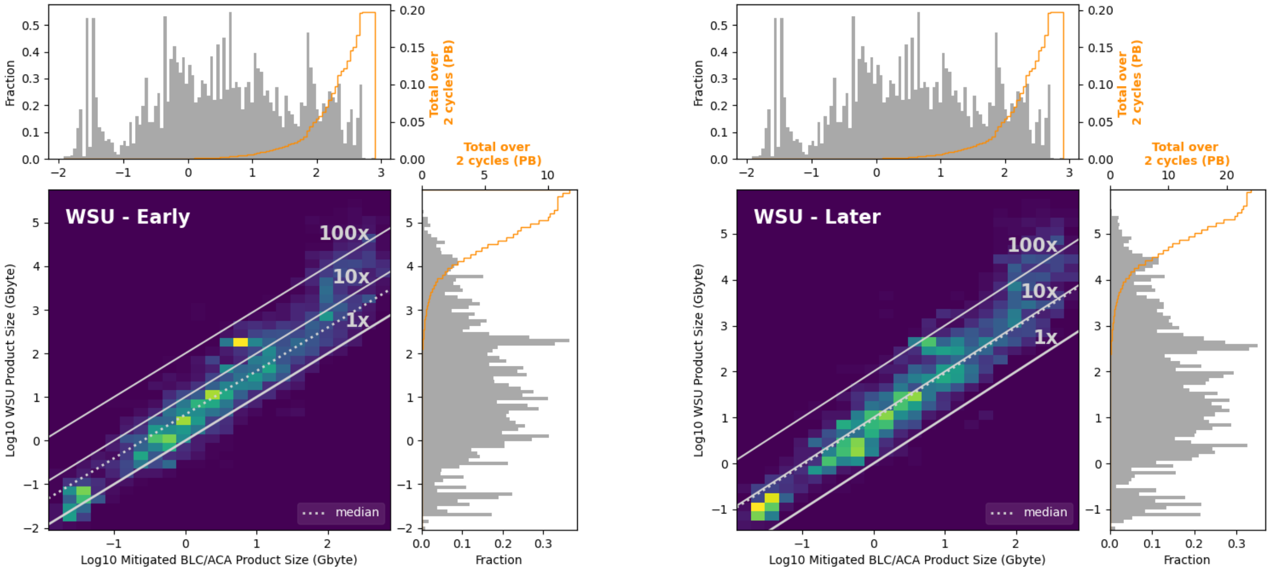


Figure 8. The complementary cumulative distribution of products sizes for the BLC/ACA (mitigated), BLC/ACA (unmitigated), and the early and later stages of the WSU. The solid lines for the WSU indicate the initial WSU estimates and the dotted lines for the WSU indicate the median WSU estimates with Bands 1 and 2 with the shaded region indicating the upper and lower bounds of these estimates. The current default maximum product size (500 GB) is indicated as a dotted vertical line. The thresholds for 1%, 5%, and 10% of the products are shown as dotted gray horizontal lines.

do not include any calibrator continuum images in our estimate of the product size, since they are minor contribution to the total product size.

Today the ALMA Pipeline mitigates, or reduces, both the quality and number of products produced for the MOUSes requiring the most demanding processing (see Hunter et al. 2023 for details on the mitigation heuristics). In Cycle 7, 19% of 12-m MOUSes were mitigated with the greatest impacts for projects in the longest baselines (Kepley et al. 2023a). Of the mitigated MOUSes, 24% were mitigated in a way that reduces the quality of the product, 22% in a way that reduces the number of science targets and spectral windows imaged, and 54% in a way that both reduces the quality of the product and reduces the number of science targets and spectral windows imaged (Kepley et al. 2023a).



(a)

(b)

Figure 9. A comparison of current mitigated ALMA product sizes from the BLC/ACA to the unmitigated estimated product sizes from early (a) and later (b) stages of WSU is shown in the main panel. The ratio between the mitigated BLC/ACA product sizes and the estimated unmitigated WSU product sizes are shown as solid gray lines with the ratio indicated next to the line. The median ratio is indicated with a dotted line. The panel to the right of the main panel shows a histogram of the distribution of unmitigated WSU data product sizes shaded in gray and the cumulative distribution of data product sizes as an orange line. The panel above the main panel shows the plot for the current mitigated BLC/ACA data. Since we are comparing present day data with future WSU data, we have included our Band 1 and 2 estimates for the WSU since there is no comparable present-day data.

To validate our product size estimates, we compare the total measured product size in the Archive today with the estimated total product size for the BLC/ACA from our database. At first glance, there appears to be a factor of 10 discrepancy between the two values with the measured total product size in the ALMA Archive for Cycles 7 and 8 of 125 TB (F. Stoehr, private communication) and an estimated total BLC/ACA product size from our database of 1270 TB. However, the difference between the actual total product sizes in the Archive is consistent with a combination of the effects of mitigation and compression of the primary beam images, with mitigation being the dominant effect. To demonstrate this, we compare the unmitigated and mitigated product size calculated by the Pipeline for all Pipeline-processed projects in Cycles 7 and 8 with the values estimated here. The total unmitigated product size calculated by the Pipeline over Cycles 7 and 8 is 1390 TB, which is consistent within a factor of 10% with our estimate based on the number of channels and estimated number of pixels in the cubes (1270 TB). The Pipeline-calculated value of the total size of all the *mitigated products* is 198 TB, confirming that the total size of the mitigated products is indeed a factor of 10 lower than the unmitigated products. However, the Pipeline-calculated value for the total *mitigated* product size is still larger than the measured total product size in the ALMA Archive for Cycles 7 and 8 (125 TB). In the Archive, however, the primary beam files are compressed which reduces the increase in product size from the factor of 2 used in the calculations in the Pipeline and this memo to a factor of 1.3 (i.e., the primary beam images add 30% to the total volume instead of doubling it). Rescaling the total product size calculated by the Pipeline by the ratio of these factors (1.3/2.0) that gives a total of 128.6 TB over Cycles 7 and 8, which is consistent with the measured total product size in the Archive, 125 TB, given above.

The distribution of product sizes for both present day ALMA and WSU is shown in Figure 8 for both current mitigated products as well as what the distribution of the unmitigated products would be. The current product size limit (500 GB) is indicated. When we compare the distribution of the mitigated BLC/ACA data to the distribution of the WSU data, we see an increase in the maximum product size by almost three orders of magnitude. However, when we compare to the distribution of the unmitigated BLC/ACA data to the WSU data, the jump in the maximum product size is only on the order of 1.5 orders of magnitude. What mitigation is effectively doing is reducing the tail of the product size distribution at the expense of reduced number and quality of data products for the most challenging cases to process. The product size distribution changes less with WSU stage. The later WSU stage increases the overall distribution slightly, but much less than the transition between the BLC/ACA and early WSU. For the early WSU, 20% of the observing time will produce products larger than the current product size limit (500 GB). This value will increase to 30% for the later stage of WSU.

We compare the ratio of the estimated WSU product sizes and the *mitigated* BLC/ACA product sizes today in Figure 9. Note that this comparison does not include the Band 1 and 2 estimates since they do not have comparable estimates from ALMA Cycles 7 and 8. The median increase in the product size is relatively modest: 4 in early WSU and 9 in later WSU. However, the maximum product size increases substantially though by 1900 (early WSU) to 3800 (later WSU). The increase in the total volume of products is again driven by the data sets in the long tail of the distribution. Note that this comparison is between the mitigated BLC/ACA products and the WSU products, which are presumed to be unmitigated. The difference would be reduced if the unmitigated BLC/ACA products were used in the comparison instead.

3.4. *Number of Spectral Windows and Cube Size Estimates*

The previous estimates of the data rate and volume of visibility data required no assumptions about the properties of the WSU spectral windows and the product size estimate was only weakly dependent on the assumed number of spectral windows. For those calculations, we have assumed that the windows were all identical and spanned the available bandwidth. Estimates of the number of spectral windows and the resulting sizes of the WSU cubes, however, do require some additional assumptions about the properties of the WSU spectral windows since each science target and spw combination will likely constitute an individual cube imaging job. As detailed in Section 2.2, we assume that the WSU spws are each 2.0 GHz wide, which similar to the current maximum spectral window bandwidth (1.875 GHz). We discuss the implications of changing this assumption in Section 4.

Figure 10 shows the distribution of spectral windows by WSU stage. We note that for the early WSU stages the maximum total number of spectral windows does not increase significantly, although the distribution does shift to the left with 70% of cases having 8 spectral windows. All cases have 16 spectral windows for later WSU. The estimates of the number of spectral windows presented in this memo are highly dependent on future decisions by the ALMA project on how PIs are allowed to

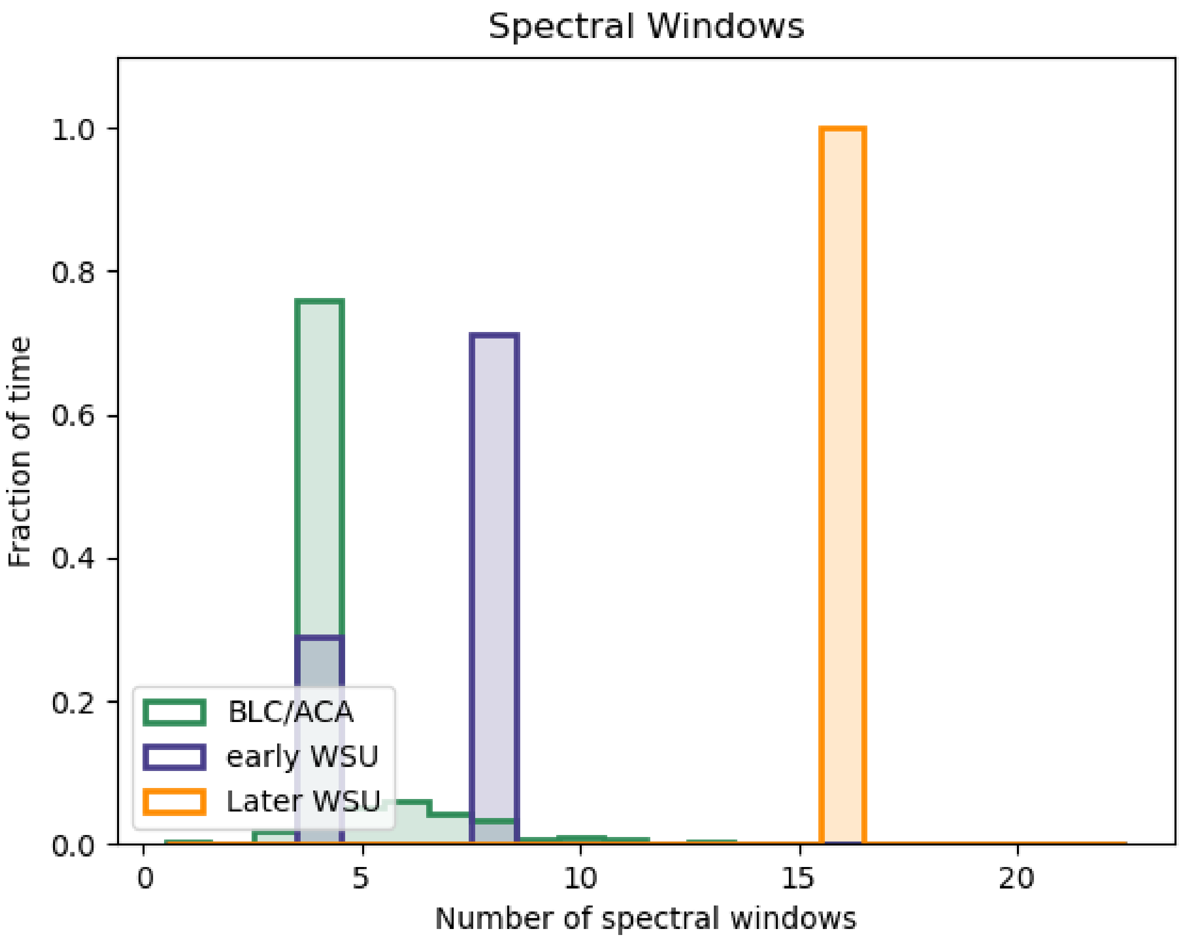


Figure 10. Distribution of spectral windows for the BLC/ACA and early and later WSU stages. The distribution of spectral windows for WSU stages should be considered a only rough guide since it dependent on future decisions by the ALMA project on how PIs are allowed to setup their spectral windows within the configurations allowed by ATAC. Please see Section 4.2 for additional discussion.

setup their spectral windows within the configurations allowed by ATAC and should be considered only rough guides to the number of spectral windows in the WSU era. See Section 4.2 for further discussion.

The complementary cumulative distributions for the maximum cube sizes per MOUS for BLC/ACA (both mitigated and unmitigated) and size of a single representative cube per MOUS for WSU are shown in Figure 11. The increase in bandwidth between early WSU and later WSU for some receivers only affects the number of spectral windows, and thus the number of cubes produced, not the size of the individual cubes. Thus the size of a *single* representative cube per MOUS does not change with the WSU stage. The product size discussed in Section 3.3 is essentially the total size of the cubes since the total size of the cubes is much greater than the total size of the continuum images. The difference between the maximum mitigated BLC/ACA cube sizes produced today and those that would be produced in the WSU era is about a factor of 50. However, the difference between

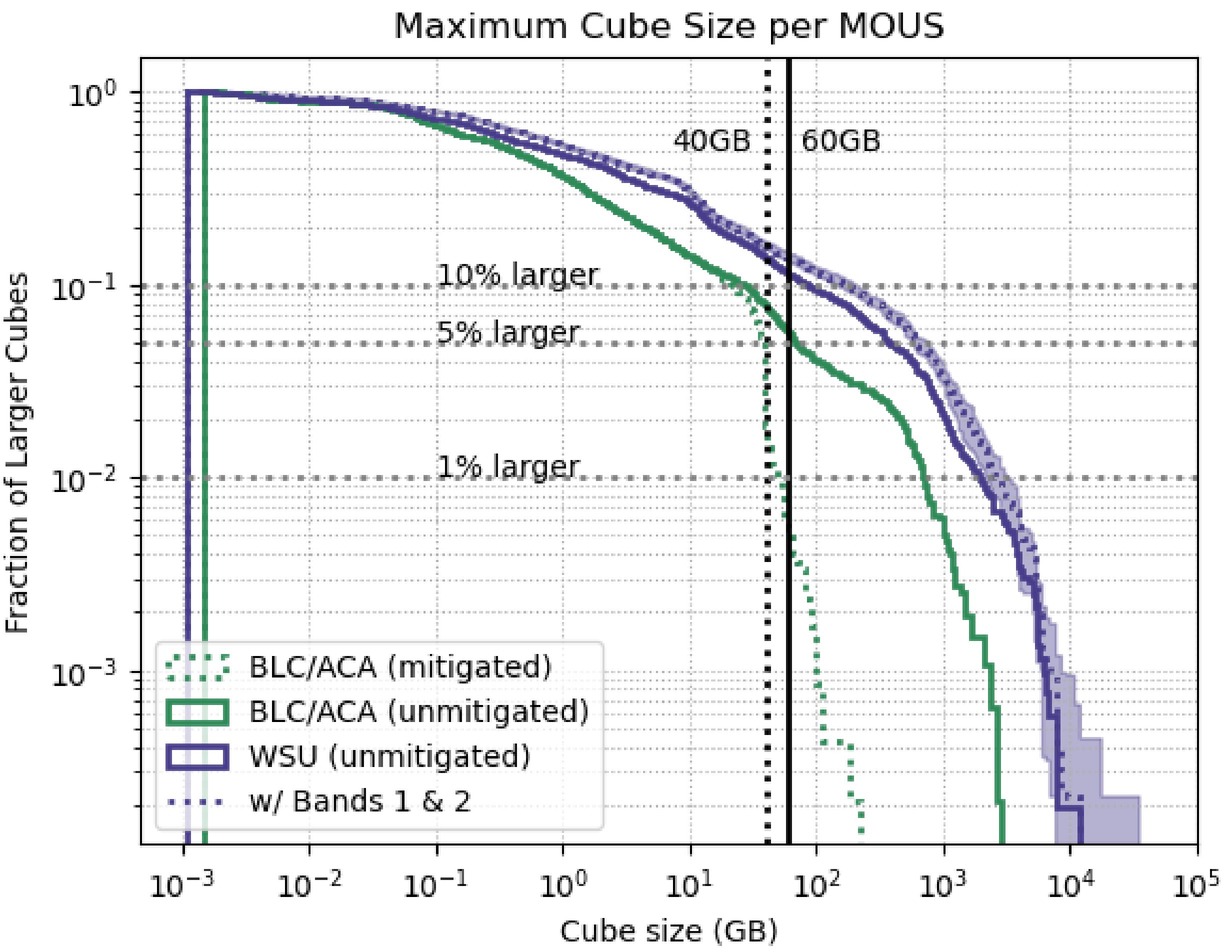


Figure 11. The complementary cumulative distribution for the size of a *single* representative cube per MOUS for the BLC/ACA (mitigated and unmitigated) and WSU (unmitigated) showing the fraction of observing time that will be spent making cubes larger than the value on the x-axis. For our estimates, we assume that the spectral windows in a MOUS are identical and have the same fixed 2.0 GHz width. The size of a single representative cube is independent of WSU stage; the difference between early and later WSU is only in the number of spectral windows and thus the number of cubes produced. The current threshold at which cubes start to be reduced in quality (40 GB) is shown as a vertical dotted line and the maximum cube size (60 GB) is shown as a solid vertical line. The cubes beyond the cube size mitigation limit were generated by manually lifting the cube size limit in operations. The individual WSU cubes will be greater than the current cube size limit of 60 GB approximately 10% of the time.

the maximum cube size is reduced to only a factor of 4 when comparing the unmitigated BLC/ACA cube size distribution to the WSU cube size distribution.

The cube size mitigation done by the ALMA Pipeline (see Hunter et al. 2023 and Kepley et al. 2023a) significantly reduces the size of the largest cubes ALMA is capable of producing. However, this size reduction is at the expense of reducing the quality of the delivered cubes. The cube size mitigations in place today reduce the imaged field of view for single fields, reduce the number of pixels per synthesized beam, and bin channels by 2 if they have not already been binned at the correlator. We emphasize that the effects of cube size mitigation are not uniform across all science cases but have a significant impact those requesting the highest spatial and spectral resolutions. In cycle 7, 100% of the MOUSes in the longest baseline configuration (C43-9) were mitigated (Kepley et al. 2023a). The individual WSU cubes will be greater than the current cube size limit of 60 GB approximately 10% of the time. The cube size limit was set to avoid a high rate of failures for processing large cubes for the typical ALMA processing allocation. See Kepley et al. (2023a) for further discussion.

A careful examination of Figure 11 reveals that today some cubes are produced that are larger than the current cube size limit. This is accomplished by manually lifting the mitigation limits in operations. However, lifting the cube size limit is not a general solution to the problem posed by the largest ALMA cubes produced today. While it is possible to increase this limit slightly, it is ultimately set by the specifications of imaging nodes (Kepley et al. 2023a). Increasing this limit significantly increases the risk of random imaging failures in operations based on the experience of two of the authors (A. Kepley and C. Brogan). In addition, currently CASA and the Pipeline do not gain from parallelization beyond ~ 17 cores. Parallelization across nodes also shows no performance improvements due to fundamental limitations in CASA (Kepley et al. 2023b).

The addition of Band 2 to our estimates does not increase maximum size of the cubes shown in Figure 11. The reason is two-fold. First, we are assuming a fixed width for our spectral windows of 2.0 GHz. Second, we have some very rare Band 3 cases in our database that require n_{avg} equal 2 to reach spectral resolutions less than 0.1 $\mathrm{km\,s^{-1}}$, which is the same n_{avg} as the case that sets the peak data rate: Band 2 with 0.1 $\mathrm{km\,s^{-1}}$ velocity resolution. Observations with the same n_{avg} and a fixed bandwidth have the same number of channels in a cube. We do have some Band 1 cases that require n_{avg} equal to 1, i.e., greater than the n_{avg} for the most challenging Band 2 case. These cases show up as an increase in the probability for very large cubes (shaded blue area).

The overall effect of including Bands 1 and 2 is instead to increase the number of cubes in the 100 GB-3 TB range rather than increase the maximum cube size. However, our estimates for the cube size produced by observations in Bands 1 and 2 may be an underestimate. These cube sizes are extrapolated from current Band 3 observations, which are restricted by the current correlator to have relatively wide channels (3 $\mathrm{km\,s^{-1}}$) for observations over the whole band. In the WSU era, the restrictions on channel size introduced by the current correlator will be lifted. Therefore, we may produce more cubes with higher spectral resolution in Bands 1 and 2 than indicated here, and thus more large cubes overall.

Quantifying the three dimensional size of the cubes for the WSU is necessary to determine what is the most efficient method of parallelization. The two dimensional distribution of the linear image size in pixels versus number of channels is shown in Figure 12 for the current BLC/ACA (unmitigated) and in Figure 13 for the early and later stages of the WSU, *excluding* the estimated Band 1 and 2 cube sizes. The total cube size increases to the upper right hand corner of the plot with the

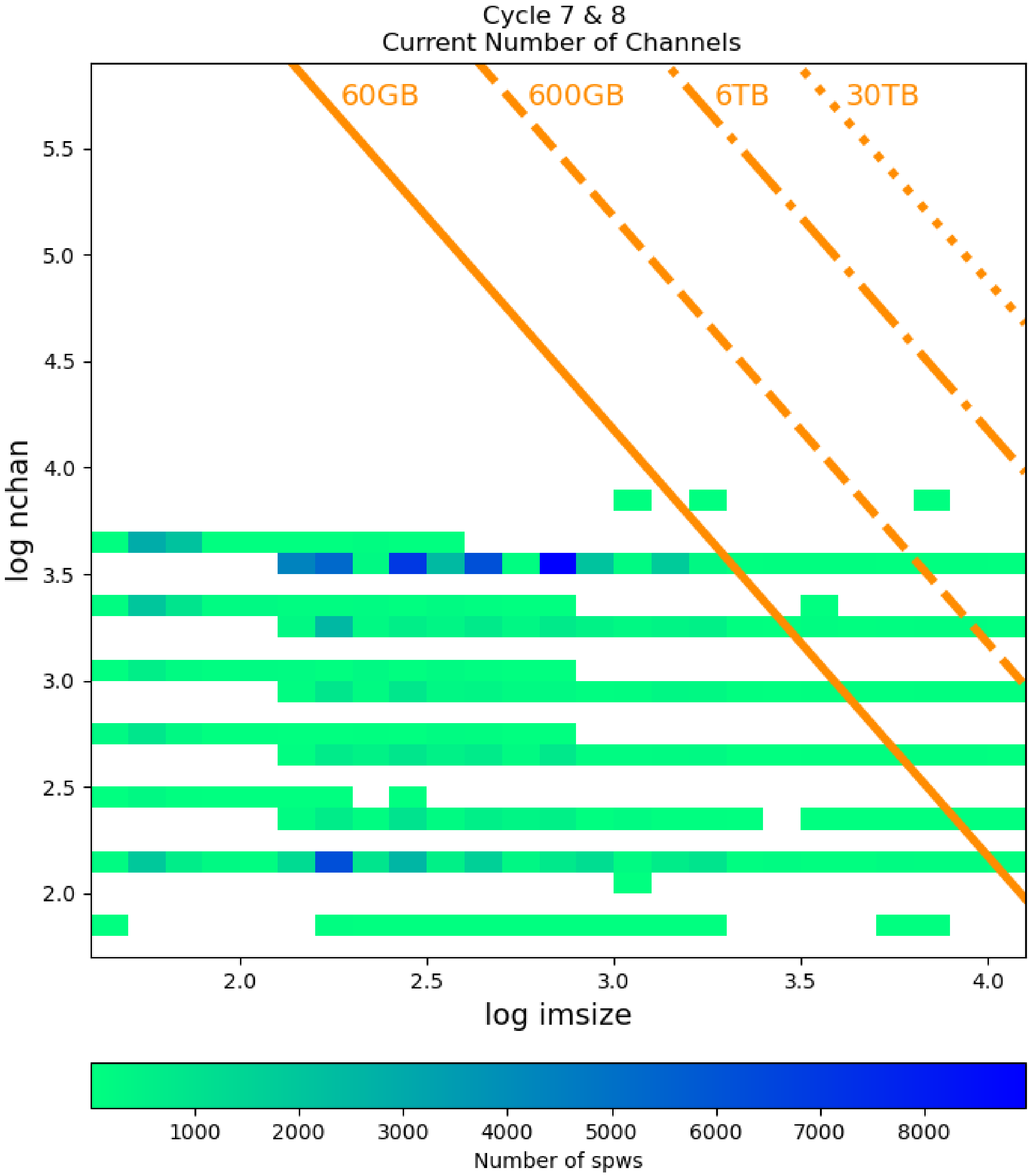
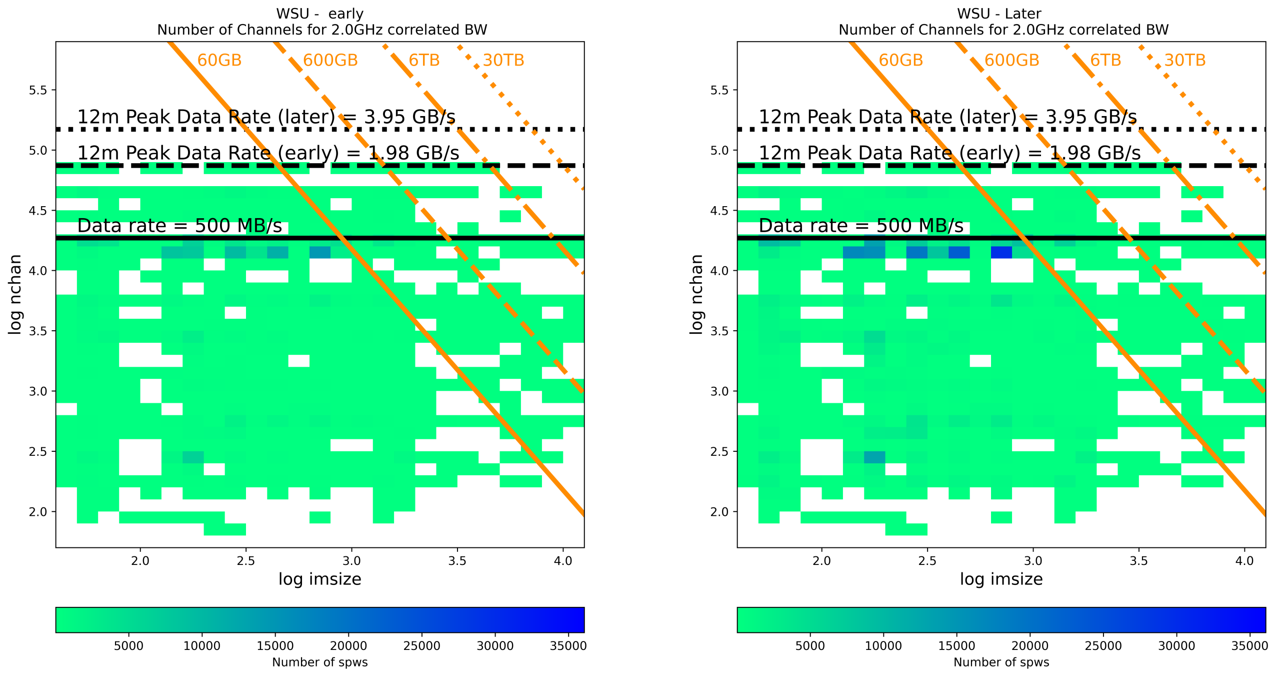


Figure 12. The two dimension distribution of spectral windows as a function of the logarithm of the number of channels and the logarithm of the one dimension image size in pixels for the unmitigated ALMA cubes produced by the current BLC/ACA. Orange lines indicate total cube sizes: 60 GB (current cube size limit), 600 GB (10 times the current cube limit), 6 TB (100 times the current cube limit), and 30 TB (500 times the current cube limit).



(a)

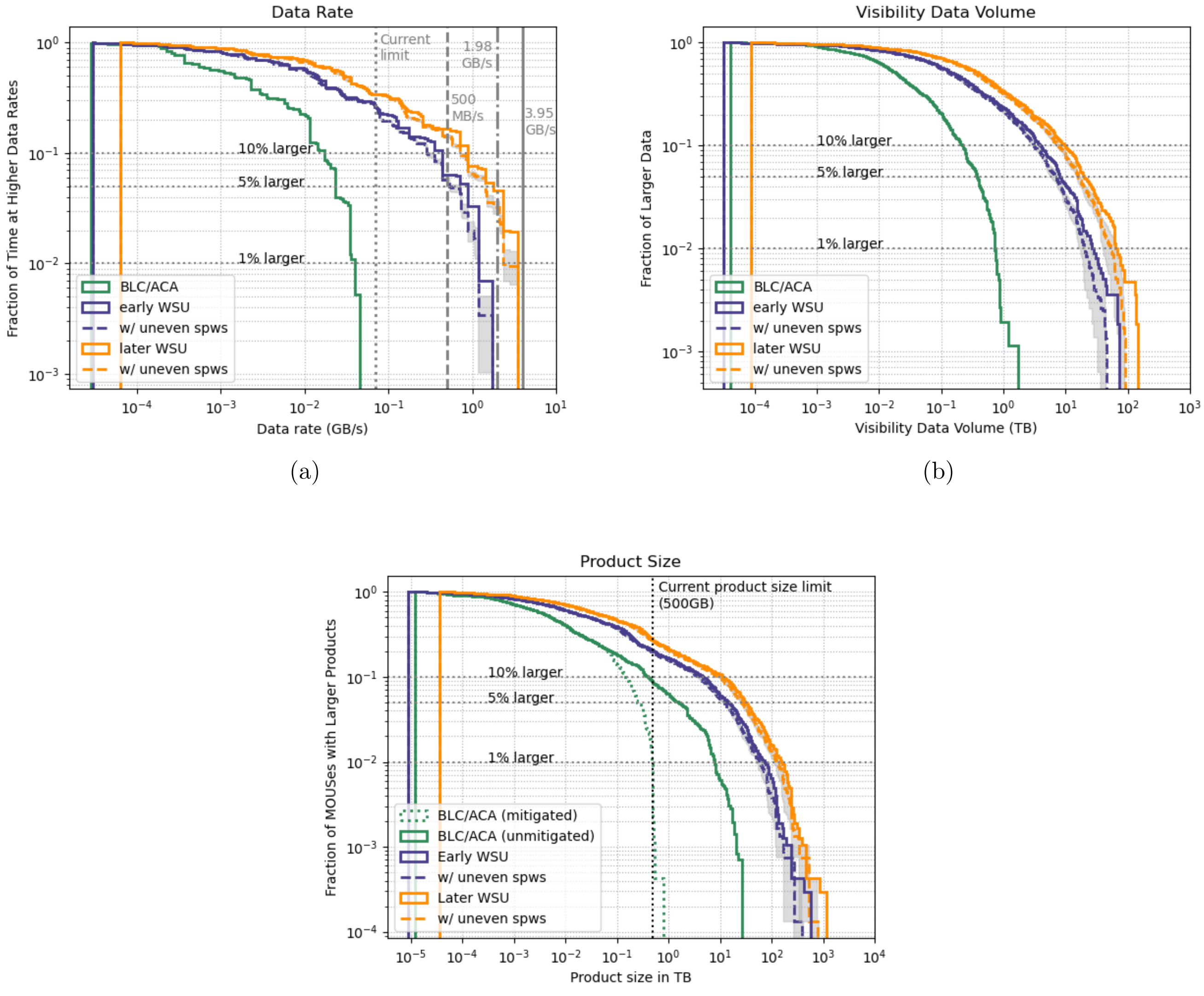
(b)

Figure 13. The two dimensional distribution of spectral windows as a function of the logarithm of number of channels and the logarithm of the one dimensional image size in pixels for early (a) and later (b) WSU. Orange lines indicate total cube sizes: 60 GB (current cube size limit), 600 GB (10 times the current cube limit), 6 TB (100 times the current cube limit), and 30 TB (500 times the current cube limit). The maximum ATAC data rate is shown as the dotted horizontal line. The data rate for Band 2 observations with 0.1 $\mathrm{km\,s^{-1}}$ channels is shown as a dashed horizontal line. A potential data rate limit of 500 MB/s is shown as a solid horizontal line. The peak 12 m array data rates from Carpenter et al. (2023) are shown as dashed (early WSU) and dotted (later WSU) horizontal lines.

current mitigation limit indicated (60 GB). Today the maximum possible unmitigated cube size is 7680 channels, with 3680 being more common, but will increase by an order of magnitude to ∼60,000 channels. While many cubes will have similar sizes to those today, the number of cubes above the current mitigation limit will increase significantly for the WSU. In addition, there may be more large cubes than previously due to the contribution of ALMA Bands 1 and 2, which will fall in the upper portion of this diagram.

4. DISCUSSION

In Section 3, we have presented the most comprehensive estimate to date of the ensemble of WSU data properties over two future ALMA cycles. We have found that the data rates, visibility data volumes, and product sizes are dominated by projects in the “long” tail of the distributions of these quantities. ALMA needs to be able to process these projects in order to fulfill the scientific goals of the WSU. In this section, we discuss some potential avenues to alleviate some of the impacts of these



(a)

(b)

(c)

Figure 14. The data rate (a), visibility data volume (b), and product size (c) complementary cumulative distributions for both the original assumption of identical spectral windows and a case where half the MOUS in the three lowest spectral resolution bins have had half their spectral windows replaced with ones at lower spectral resolution. The gray shaded areas show the range of values produced over 50 simulations.

projects, while retaining the data properties necessary to do the science, as well as some systematic uncertainties related to our estimates.

4.1. *Uneven Spectral Window Resolutions*

Our estimates assume that all spectral windows have identical properties. However, we have the freedom to set the width of the spectral windows within the parameters allowed by ATAC. The smallest possible spectral window is one frequency slice (200 MHz), while the largest possible spectral window is 80 frequency slices (16 GHz). It may be advantageous to have the spectral windows with different properties to reduce the overall data rate and align more closely with the scientific goals of the PI. For example, one might imagine a project looking at a science target where few lines are expected, but the rest is anticipated to be continuum emission. In that case, the setup most closely aligned with the PI science would be to have high spectral resolution windows over the line region(s) and lower spectral resolution windows for the continuum regions. Spectral line focused projects may also benefit by allowing PIs to place their highest spectral resolution windows where the brightest and/or narrowest lines are and lower spectral resolution windows where the lines are expected to be broader and/or faint. However, there will still likely be a subset of projects that need uniform high spectral resolution across the full bandwidth.

To estimate the effects of spectral windows with uneven spectral resolutions on the data rate, visibility data volume, and product size distributions, we have taken half the MOUSes in each of the three highest spectral resolution bins (0.5-2.0 $\mathrm{km\,s^{-1}}$, 0.1-0.5 $\mathrm{km\,s^{-1}}$, and <0.1 $\mathrm{km\,s^{-1}}$) and assigned half the spectral windows to be the current resolution and half the spectral windows to be the resolution of the top end of the bin. For example, for a MOUS in the 0.1-0.5 $\mathrm{km\,s^{-1}}$ bin, half the spectral windows would have a resolution of 0.1 $\mathrm{km\,s^{-1}}$ and the other half have a resolution of 0.5 $\mathrm{km\,s^{-1}}$. We then recalculate the data rate, total visibility data volume, and the product size. We repeated this experiment 50 times, randomly selected different MOUSes each time.

Figure 14 compares the distribution of our original estimates with those estimated above assuming that the spectral windows have uneven spectral resolution. The overall effect of replacing some of the originally high resolution spectral windows with lower spectral resolution windows is to slightly lower the number of projects in the "long tail" of the WSU distribution, but the reduction is not significant. These calculations are not affected by our assumption of a fixed spectral window bandwidth. The data rate and total visibility data volume only dependent on the aggregate number of channels across all spectral windows and the total product size depends most strongly on the aggregate number of channels across all spectral windows since it is dominated by the size of the cubes.

4.2. *Total Number of Spectral Windows*

As this memo was being completed, the ALMA project made a decision to limit the maximum total number of channels in a spectral window to 80,000 and for the ALMA subsystems to be designed to handle the maximum number spectral windows that could potentially be provided by ATAC (80 for early WSU and 160 for later WSU). We emphasize that no decisions have been made on what capabilities to offer to users. The number of channels per spectral window presented in this memo is consistent with this decision. However, it is important to note that this decision did not limit the *bandwidth* that could be covered by a single spectral window. For a given velocity resolution, the frequency width of a channel increases with increasing observing frequency. If the bandwidth per spectral window is fixed, as assumed in this memo, then at higher frequencies there will be fewer channels per spectral window and more spectral windows required to cover the WSU bandwidth. If the maximum number of channels is fixed instead, as per the recent decision by the ALMA project, then it is possible at higher frequencies to have wider bandwidth spectral windows with more channels per spectral window requiring fewer spectral windows to cover the WSU bandwidth. As an example, let us consider observations at at Band 2 (75 GHz) and Band 8 (460 GHz) with a velocity resolution of 0.1 $\mathrm{km\,s^{-1}}$. At Band 2, the maximum number of channels per spectral window sets the limit on the frequency width of the spectral windows of 2 GHz (74400 channels) and thus would require 16 spectral windows to cover the full bandwidth (32 GHz). At Band 8, however, the maximum number of channels possible sets the limit of the bandwidth of a spectral window of 16 GHz (79360), meaning that only two spectral windows would be needed to cover the full bandwidth (32 GHz). Thus the estimates presented in the memo may be biased towards having more spectral windows with fewer numbers of channels than what may be produced by the WSU in the future. However, we caution that ultimately the actual number of spectral windows for the WSU depends strongly on the future decisions made by the PIs to achieve their science goals and may differ significantly from these (and other future) estimates.

4.3. *Data Rate Caps*

The ALMA project may wish to consider programmatic caps on the data rate that are less than the maximum technically feasible data rate to reduce processing, data reduction, and Archive storage requirements. We would like to emphasize here that the effect of a data rate cap is to limit the aggregate number of channels, i.e., the number of channels summed across all spectral windows for a particular observation. While this will reduce the overall data rates, a data rate cap in and of itself does not solve the issue of the most challenging cube imaging cases (c.f. Figure 13). The reason is that data rate limits only reduce the maximum number of channels in a cube. Since the cube size depends on both the number of channels and the image size, the latter of which is governed by the configuration and number of pointings, a significant number of large cubes could still be produced. In addition, the data rate caps do not affect all science cases equally. The science impact of any data rate cap must thus be carefully considered, since a data rate cap is most likely to affect and could even make unfeasible those projects that would benefit most from the WSU, i.e., high spectral resolution observations over wide bandwidths. A cap on the number of projects requesting high data rates could be a potential option for accommodating these projects, while reducing the processing, data reduction, and Archive storage requirements.

4.4. *Spectral Resolution of Calibration Data*

As we discuss in Section 3.2, these estimates assume that both the calibration and science data have the same spectral resolution. Significant reductions in the volume of visibility data may be possible by reducing the spectral resolution of the calibration data since the calibration data represents on average a third of the overall visibility data volume, albeit with a large range (see Section 3.2). For example, the spectral information in the phase calibration observations are not used to calibrate the observations and they may be able to be taken at lower spectral resolution than required for the science observations. The wider bandwidths provided by the WSU may also offer the possibility of reducing the integration time on calibrator targets, although that strongly depends on how the spectral windows are defined. As of the writing of this memo, there are several efforts currently underway to investigate how to reduce the volume of the calibration data without compromising their quality. We will incorporate the results of these investigations in a future version of these calculations.

In addition, once the data is calibrated (both regular and self-calibration) and continuum subtracted, the science visibility data could potentially be time and frequency averaged to reduce the size of the science target visibilities and the required system performance necessary to grid the data for imaging. Any averaging, however, must be done in a way that does not reduce the science quality of the data.

4.5. *Product Size Mitigation*

In ALMA operations today, we mitigate, or reduce, the size and number of products produced for projects using heuristics that do not take into account the science goals of the project (Hunter et al. 2023). This situation negatively impacts a significant fraction of ALMA users with 20% of 12m projects in ALMA Cycle 7 receiving less optimal and/or an incomplete set of data products (Kepley et al. 2023a). However, PIs often naturally reduce the size of their images depending on their science purposes. They might choose to only image channels where there is line emission, average channels together to increase their S/N, and/or reduce the imaged field of view. These PI-based mitigations are based on the science to be performed and knowledge of the properties of the science target. It may be possible to mitigate the overall product size while retaining more useful scientific information if the science target properties can be taken into account during the mitigation process. However, this must be balanced against the value of the products for archival research, which may differ from the PIs needs. For example, one scientist's empty cube may be another scientist's very constraining limit.

One might also consider mitigating the overall size of the products produced through compression, down-sampling, and/or reduced numerical precision. However, these operations have the potential to introduce irreversible reductions in data quality. In the ALMA Scientific Specifications and Requirements (ALMA-90.00.00.00-3002-A-SPE) document, Requirement 2.13 ("Data Accuracy") states the following "Data storage and processing must not result in a significant loss in sensitivity (from, e.g., not storing enough bits per number)." These techniques should only be employed after comprehensive studies demonstrating that they are not detrimental to broad range of ALMA science.

4.6. *High Time Resolution*

The analysis presented in this memo has focused on observations that do not require high time resolution. If integration times significant shorter than 1 s are used, observations with coarse spectral resolutions and relatively few channels could still generate considerable data volumes (for visibilities and potentially for products). However, we currently estimate that the fraction of projects using such short integration times would be quite low due to the need of a high signal-to-noise per integration time and thus would not alter significantly the data volume distribution.

4.7. *Evolution of Scientific Inquiry*

Finally, the estimates presented here are based on a snapshot of the scientific requirements of the PIs of accepted ALMA proposals from 2019 through 2021. New discoveries like gravitational waves or fast radio bursts have the potential to alter the scientific landscape and change the types of observations applied for by PIs in the WSU era. In addition to new discoveries, there is also the more gradual evolution of science as on-going research opens up new questions and closes off other paths of inquiry. This evolution also has the possibility to change the types of WSU observations taken. We are biased towards accepted projects that had some observations taken for them. Projects that were rejected through the time allocation process are not included in our sample, which does not mean that they would be rejected in a future cycle. We also likely under-sample projects that are difficult to schedule due to either weather, configuration, or science target constraints. But again that does not mean that these projects would not be viable either now or in the WSU era. All of these effects have the potential to alter the distribution of scientific use cases requested by the PI in the WSU era from the estimates presented here.

5. SUMMARY

This memo presents an estimate of the ensemble of WSU data properties for two future ALMA cycles. We used all projects that had data taken for them in ALMA Cycles 7 and 8 to provide the basis of our estimates, with the exception of solar system, solar, VLBI, and total power projects. We have included estimated data rates for the latter as an Appendix (see Appendix A). Using the ATAC capabilities and a nominal receiver roll out plan, we estimate the WSU data properties for two different WSU stages: early and later. We present estimates of the data rates, visibility data volumes, product sizes, number of spectral windows, and cube sizes. We find that, in general, these quantities are dominated by the long tails of their distributions. This long tail is also present for the current BLC/ACA, although its magnitude is smaller because the maximum number of channels that can be produced by the current correlators is much less than can be produced by ATAC (15,000 vs. 1.2 million).

Our main results are as follows:

- For early WSU, 30% of the observing time is estimated have rates higher than the current 70 MB/s limit and 6-10% of the observing time is estimated to have rates greater than 500MB/s. These estimates increase to 35% and 18% for later WSU. The time-weighted average aggregate number of channels will increase from today by a factor of 17 (early WSU) to 38 (later WSU) , while the maximum number of channels will increase by factors of 8 (early WSU) to 15 (later WSU).

- We estimated that 15-35% of data sets will have visibility data volumes greater than the largest BLC/ACA project today (∼1 TB). The median increase in the visibility data volume is estimated to be 7 (early) to 18 (later) times the current size for the same data set. Some MOUSes have estimated visibility data volume up to 690 times larger than the equivalent project today.

- For early WSU, 20% of the observing time is estimated to produce projects larger than the current product size limit (500 GB) and this value is estimated to increase to 30% for later WSU. The median increase per project in product size is 4 for early WSU and 9 for later WSU, but the maximum increase per project can be 1900 times (early WSU) and 3800 times (later WSU).

- Using an assumption of 2 GHz spectral windows across the full bandwidth, we find that the maximum number of channels will increase from 3840 (7680 in rare cases) to 60,000 channels. However, this estimate depends strongly on decisions on how spectral windows will be set up by users in the future, which is still an active area of discussion.

We close with some potential avenues to reduce the overall data rates and volumes while retaining the scientific integrity of the data and some cautions about the systematic uncertainities in these estimates.

The authors would like to thank Felix Stoehr, Richard Simon, Remy Indebetouw, Andy Lipnicky, and Ignacio Toledo for providing some of the data used in this analysis. They would also like to thank the Data Processing and Data Accessibility Working Group for comments on early versions of this work. A. Kepley would like to thank Jeff Kern for suggesting that she take on this project, Ryan Loomis and Remy Indebetouw for useful conversations, and Thodori Nakos for hosting her at JAO where a portion of this work was done. This work made use of Astropy:[7] a community-developed core Python package and an ecosystem of tools and resources for astronomy (Astropy Collaboration et al. 2013; Astropy Collaboration & Price-Whelan 2018; Astropy Collaboration et al. 2022).

REFERENCES

Astropy Collaboration, & Price-Whelan, A. M. 2018, 156, 123, doi: 10.3847/1538-3881/aabc4f

Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, 558, A33, doi: 10.1051/0004-6361/201322068

Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, 935, 167, doi: 10.3847/1538-4357/ac7c74

Carpenter, J., Brogan, C., Iono, D., & Mroczkowski, T. 2022a, ALMA Memo Series, 621. https://library.nrao.edu/public/memos/alma/main/memo621.pdf

—. 2022b, ALMA IST Estimated WSU data rates memo

Carpenter, J., Brogan, C., Trigo, M. D., et al. 2023, Data Rates Ramp-Up Plan Working Group Report: Peak Data Rates, Tech. rep., ALMA-05.00.00.00-3051-1-GEN

Carpenter, J., Iono, D., Testi, L., et al. 2018, ALMA Memo Series, 612. https://library.nrao.edu/public/memos/alma/main/memo612.pdf

Cortes, P. C., Remijan, A., Hales, A., et al. 2022, ALMA Cycle 9 Technical Handbook. www.almascience.org.

Ginsburg, A., Sipőcz, B. M., Brasseur, C. E., et al. 2019, The Astronomical Journal, 157, 98, doi: 10.3847/1538-3881/aafc33

Hunter, T. R., Indebetouw, R., Brogan, C. L., et al. 2023, Publications of the Astronomical Society of the Pacific, 135, 074501, doi: 10.1088/1538-3873/ace216

Kepley, A. A., Lipnicky, A., Rao Venkata, U., & Indebetouw, R. 2023a, ALMA Memo Series, 263. https://library.nrao.edu/public/memos/alma/main/memo623.pdf

Kepley, A. A., Madsen, F., Robnett, J., & Rowe, K. S. 2023b, NAASC Memo Series, 121. https://library.nrao.edu/public/memos/naasc/NAASC_121.pdf

Schieven, G. 2022, ALMA Doc. 9.1, ver. 1. www.almascience.org.

[7] http://www.astropy.org

APPENDIX

A. MEAN DATA RATES FOR TOTAL POWER ARRAY

This section estimates mean data rates and data volumes for the Total Power (TP) Array observations. The data rate for the Total Power Array is estimated using the following formula,

$$datarate\,[GB/s] = 4n_{ant} * n_{chan} * n_{pol}/t_{int}[s]/10^9. \tag{A1}$$

The data volume per MOUS is

$$datavol_i\,[GB] = datarate_i * time_{obs,i}, \tag{A2}$$

and the total data volume is

$$datavol\,[TB] = \sum_{i=0}^{i=n_{mous}} datavol_i\,[GB]/10^3. \tag{A3}$$

The current plan is to upgrade the TP spectrometer to process the increased bandwidth and provide the requisite spectral resolution for the WSU. Although the specifications of the upgraded spectrometer (Total Power GPU spectrometer; TPGS) are not yet finalized, the same number of channels as in the ATAC calculation are adopted here, assuming that the same velocity resolution as in the 12-m and 7-m arrays will be achieved. We use the MOUSes of the approved Cycles 7 and 8 projects to estimate the data rates and the number of channels per spw in the WSU era. We assume the TP array will have four antennas ($n_{ant} = 4$), the number of polarizations of two ($n_{pol} = 2$), and the total number of hours available for observation is unlikely to change significantly. We estimate the spectral resolution that PIs might request by dividing the finest spectral resolutions into five bins. The distribution of the spectral resolution per MOUS used in the Cycle 7 and 8 observations is shown in Figure 15.

Since the Cycle 7 and 8 database does not include observations from either Bands 1 and 2, we attempt to have them in our estimation. We assume that Bands 1 and 2 will have 3% of the total observing time. Following the estimations of the 12-m and 7-m arrays, we assume that this time can be randomly taken from Bands 3, 6, and 7 in equal proportion. The second assumption is that Bands 1 and 2 will have a similar distribution of science use cases to Band 3. We select randomly from Band 3 projects and set their frequencies to the fiducial values for either Band 1 or 2 (39 or 75 GHz) and their total bandwidth to the appropriate value for the two WSU stages. We generated 1,000 realizations of our sample with Bands 1 and 2 included to characterize the distribution of these estimates.

The estimated median data rate, time-weighted average data rate, maximum data rate, and aggregate number of channels are presented in Table 13. The time-weighted average is calculated by weighting each MOUS by the observing time. The overview of data volume properties for TP Array is presented in Table 14. Figure 16 shows the distribution of data rates weighted by observing time for Cycles 7 & 8, early WSU, and later WSU.

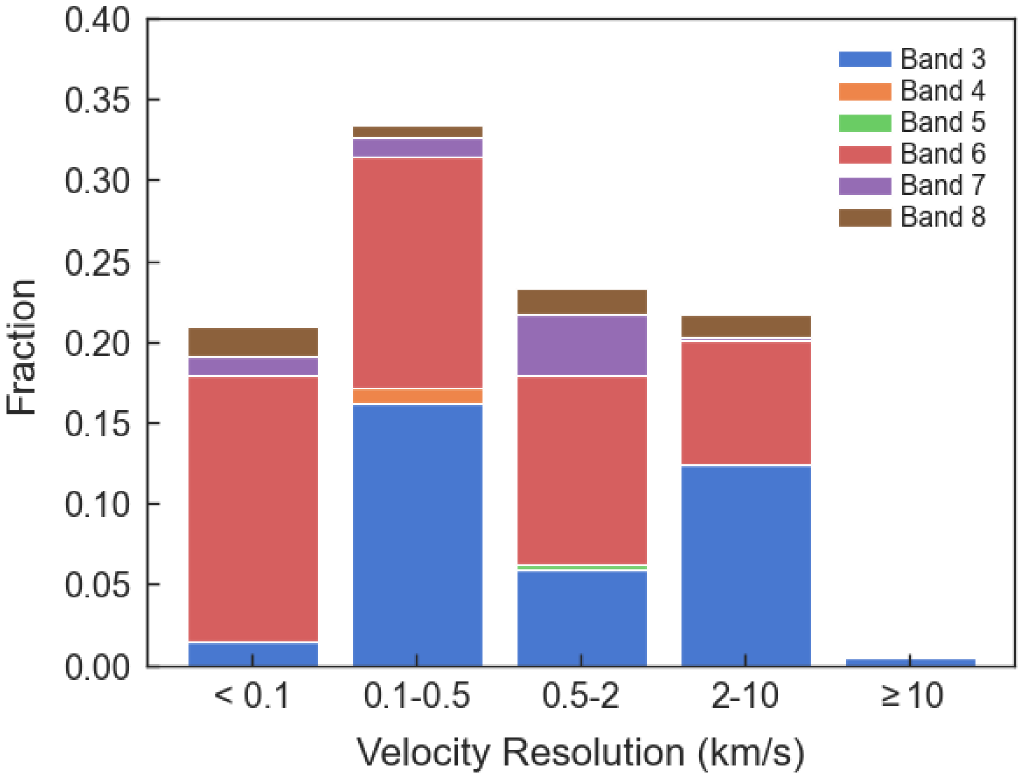


Figure 15. Distribution of velocity resolution per MOUS in the sample of the Cycle 7 and 8 data sets for Total Power Array.

Table 13. Data Rate Properties for TP Array

		Cycle 7 & 8	Early WSU	Later WSU
Data Rate	Median (GB/s)	0.001	0.007	0.016
	Time Weighted Average (GB/s)	0.003	0.016	0.033
	Maximum (GB/s)	0.017	0.132	0.263
Number of Channels	Median	14,272	118,516	296,288
	Time Weighted Average	37,103	296,288	415,688
	Maximum	135,234	666,222	1,332,443

Table 14. Data Volume Properties for TP Array

		Cycle 7 & 8	Early WSU	Later WSU
Raw Data Volume	Median (TB)	0.03 ± 0.26	0.11 ± 1.19	0.25 ± 2.38
	Time Weighted Average (TB)	0.34 ± 0.75	1.62 ± 2.77	3.36 ± 5.55
	Maximum (TB)	4.03	19.42	38.83
	Total (TB)	–	245.43 ± 9.03	503.85 ± 18.02

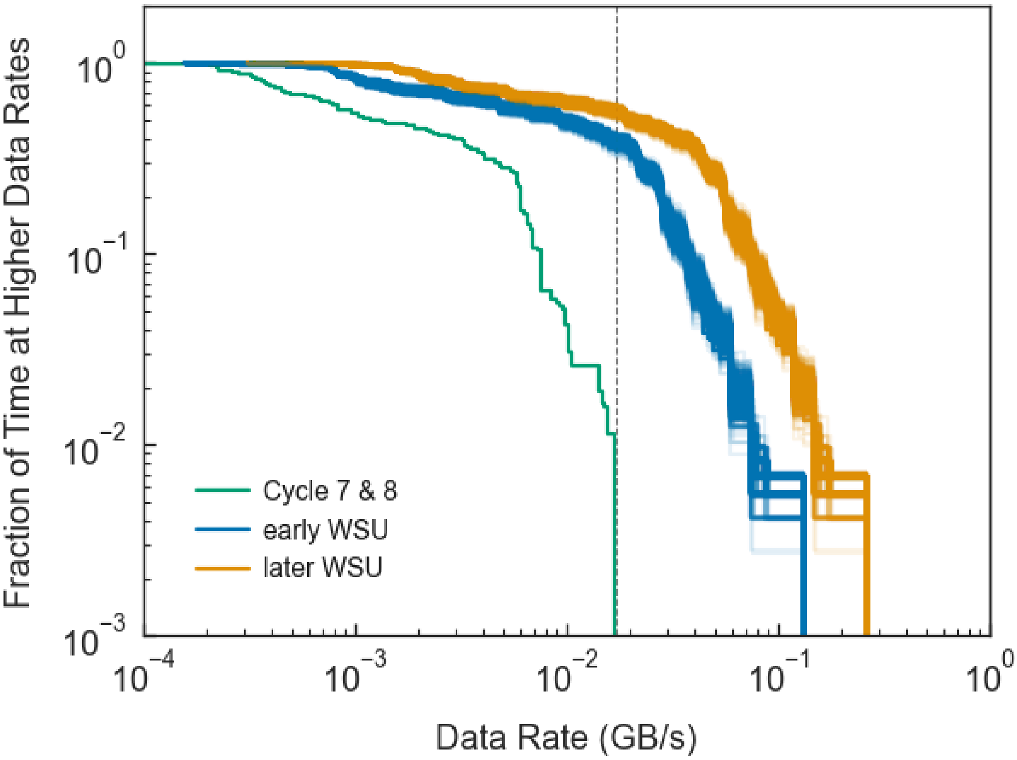


Figure 16. Complementary cumulative distribution of data rates weighted based on observing time for Cycle 7 & 8, early and later WSU. The vertical dashed line represents the maximum data rate of 17 MB s^{-1} in the specification of the current ACA Spectrometer.