

# RFI Mitigation for the ngVLA : A Cost-Benefit Analysis

## ngVLA Memo #70

Urvashi Rau, Rob Selina, Alan Erickson

December 10, 2019

### Executive Summary

This report analyzes the characteristics of anticipated RFI in the ngVLA era and evaluates the cost versus benefit of integrating RFI mitigation solutions within the architecture of the end-to-end data acquisition and processing system. ngVLA Memo #71 discusses the proposed system architecture design.

**Analysis :** RFI characteristics discussed in ngVLA Memo #48 are matched with appropriate choices of RFI mitigation algorithms at one or more stages of the data acquisition and signal processing chain. Each type of RFI is characterized by the fraction of the array that is likely to be affected, the number of hours per day the interferer is likely to be seen and typical signal duty cycles, with the assumption that the allocated bands are entirely filled. Mitigation options include post-processing flagging, in-correlator flagging of visibilities at high time and frequency resolution, RFI modeling and subtraction prior to averaging for archival, real-time antenna based RFI detection and flagging prior to correlation, and the natural effect of decorrelation on long baselines. Metrics are defined to quantify the fraction of data lost due to RFI with and without active mitigation and to relate this to the amount of extra observing time required to achieve imaging sensitivity assumed with no RFI. Compute loads are also estimated for each of the proposed solutions, taking into account data rates at which the algorithms must operate to be numerically effective. Scenarios are then analysed in terms of input RFI characteristics as well as mitigation solutions.

**Results :** Our broad conclusions are consistent with collective wisdom, but this exercise has resulted in quantitative estimates of the cost versus benefit. With post-processing flagging alone (current practice) the ngVLA is at risk of losing 20-35% of data across the entire ngVLA frequency range. Continuum science would still be feasible with increased observing time, but some key spectral line science is compromised. Real-time antenna based solutions are expected to have a low impact but they are the only option for nano-sec pulsed RFI and are important for gathering diagnostic information. In-correlator high time-resolution flagging is targeted to communication transmissions that are expected to have on/off duty cycles at  $\mu s$  to  $ms$  timescales and if this is the case, outlier detection/flagging algorithms could improve data losses from 35% to <10%. For continuous data transmissions seen by large fractions of the array, RFI modeling and subtraction algorithms may be the only option for reducing losses in otherwise unusable bands, but these are currently experimental.

**Recommendations :** A series of experiments have been defined to answer key questions that must be carried out prior to the detailed system design. In general, real-time antenna based RFI detectors and in-correlator high resolution flagging options should be implemented, along with the appropriate transmission of flags or weights throughout the data acquisition system. The ability to support modeling and subtraction algorithms must be designed-in, (in conjunction with fast-imaging modes) but need not be implemented in initial phase. In addition to mitigation options, RFI avoidance is the obvious first line of defense and includes monitoring and smart (predictive and reactive) scheduling as well as negotiated spectrum sharing with key RFI generators (where possible).

# Contents

<b>1</b>	<b>RFI Characteristics</b>	<b>3</b>
1.1	Frequent Short-duration RFI	3
1.2	Infrequent Short-duration Local RFI	3
1.3	Continuous RFI	3
<b>2</b>	<b>Mitigation Approaches</b>	<b>4</b>
2.1	Detection and Flagging	4
2.2	Modeling and Subtraction	5
2.3	Integrated System Design	6
<b>3</b>	<b>Evaluating Cost vs Benefit</b>	<b>7</b>
3.1	Defining the Parameter Space	8
3.1.1	Array Configuration	8
3.1.2	RFI Characteristics	8
3.1.3	Data Processing Stages and Algorithms	10
3.2	Metrics to Quantify the Impact of RFI and Mitigation Options	10
3.2.1	Fraction of Data Loss	11
3.2.2	Extra Observing Time Required to Compensate for Data Loss	12
3.2.3	Compute Load	12
<b>4</b>	<b>Analysis and Results</b>	<b>13</b>
4.1	Effects of RFI and Various Mitigation Options	13
4.2	Impact on ngVLA Science	20
4.3	Impact on ngVLA System Design (and Regulation)	21
<b>5</b>	<b>Future Work: Tests and Prototypes</b>	<b>23</b>
5.1	Pre-correlation RFI Detection and Blanking/Filtering	23
5.2	In-correlator Flagging at $\mu s$ and $ms$ Resolution	23
5.3	Real-time Subspace Projection Methods	23
5.4	Real-time Calibration, Imaging and Peeling	24
5.5	Understanding the RFI Environment	24
5.6	Scheduling Around Satellite Orbits	24
5.7	RFI Attenuation Due to Signal Decorrelation	24
5.8	RFI Monitoring and Database Generation/Use	24

# 1 RFI Characteristics

Radio Frequency Interference (RFI) affects data from radio interferometers in several ways. The choice of algorithmic options and the optimal processing stage at which to employ them depend strongly on the characteristics of the RFI as seen by the array. The discussion below builds on the RFI forecast from ngVLA Memo #48 to classify RFI sources based on their time signature (to identify at what time resolution they must be excised), the fraction of the array affected (to justify global versus local solutions) and the fraction of data that is potentially lost (to justify the use of extra measures).

## 1.1 Frequent Short-duration RFI

Some commercial and military radio communication protocols consist of high power, repeating, short duration transmissions, often with separate frequency channels for uplinks and downlinks and on/off duty cycles on the micro-sec to milli-sec timescales. One current example is aircraft distance measuring equipment (DMEs) transmissions between 960 and 1150 MHz which operate with 12 micro-second pulses spaced by 50 micro-seconds, in 100 kHz wide channels spaced by 1 MHz. The time-frequency occupancy of each such signals is low when viewed at (for example) a 10 micro-second time resolution and 50kHz frequency resolution, such as that available within the real-time data acquisition system or the correlator. However, when viewed at the post-processing stage where the data are typically at a 1 second / 1 MHz time/freq resolution the RFI signal appears continuous and forces a 100% data loss. As these RFI sources are usually airborne or in orbit, they will be seen by a large fraction of the array (for the current VLA as well as the ngVLA). A calculation of projected occupancies for signals such as aircraft DMEs suggests a 2% to 18% data loss if flagging is applied at the appropriate high time resolution compared to a 20% to 40% data loss if dealt with only in post-processing.

## 1.2 Infrequent Short-duration Local RFI

RFI is expected from numerous ground-based sources (humans with electronic devices, car radars, RFIDs, on-site equipment, etc). Some may be extremely high power transmissions (e.g. car radar) which will necessitate the design of robust receiver front-ends or observing strategies to avoid pointing individual antennas in particular directions. In general though, the time-frequency-antenna occupancy of such RFI is expected to be very low for transient sources and only moderately high for fixed ground-based transmitters. Flagging of affected visibilities will result in only a small fraction of data loss. Pre-correlation, antenna-based detection of such RFI is useful primarily from a diagnostic point of view and for generating meta-data to be used later on to label time/frequency ranges containing RFI that may be only partially correlated across the array. Only RFI that is intermittent on a timescale shorter than the correlator multiply-accumulate step will benefit from real-time excision although the exact details of whether to blank the data points or replace them with lower-power noise are as yet unclear.

## 1.3 Continuous RFI

All data transfer protocols such as cell phone signals, satellite radio, digital TV and (future) satellite internet are expected to be continuous in time, present over a wide range of frequencies and visible to a large fraction of the array. From bands already allocated to such emissions (shown in Table 2), it is likely that such emissions will affect a significant fraction of many of the ngVLA observing bands. Cell 5G towers may be present in towns near all outlier antennas and a few plains spiral antennas. As long as no cell 5G tower is transmitting near the central dense core of the ngVLA, only a small fraction of mid-to-long baselines will be affected by each tower. It is critical that no cell 5G towers be installed within sight of the ngVLA dense core. At present, the only mitigation option for such RFI is flagging to discard these data or in the case of predictable satellite transmissions to avoid the frequency ranges during the observation. Experimental algorithms exist to model and subtract such signals but none have been proven to be robust enough for routine use.

## 2 Mitigation Approaches

RFI mitigation may be classified into outlier detection algorithms that generate flags to mask RFI-affected data from further processing, modeling and subtraction algorithms that attempt to recover sky signals from beneath the RFI, and smart RFI avoidance strategies that adapt observing schedules to predictable and current RFI patterns. The following sections introduce and compare these general approaches.

### 2.1 Detection and Flagging

**Algorithms :** Algorithms for robust outlier detection in radio interferometry include 1D and 2D statistical and model-based methods. Spectral Kurtosis filters are recently gaining popularity for 1-D real-time data streams with analyses conducted in both time-series and spectral domains. Autoflag algorithms such as 'rflag' and 'tfcrop' within the CASA/AIPS [3]/[6] offline data reduction software (and similar approaches at other observatories [11]) operate on complex visibilities and exploit a-priori knowledge of RFI characteristics compared to the sky signal. For example, 'rflag' uses hierarchical robust statistics to build up a model of the underlying noise level in the data against which it calculates flags in a second pass through the data. The 'tfcrop' algorithm calculates robust piecewise polynomial model fits to the smooth frequency and time variation of the sky visibilities underlying RFI spikes and then calculates flags against this average model. All these algorithms must be tuned for data with different RFI characteristics. Experiments such as RealFast and Fast-VLITE that detect astronomical transients have had success in employing spectral de-dispersion checks to separate sky signals from terrestrial RFI signals at  $\mu s$  to  $ms$  time resolutions. The output from all these algorithms is a Boolean flag (or float weight) per complex visibility data point.

**Data Partitioning :** Automatic flagging algorithms with specific parameter tunings operate best when applied to subsets of the data that contain similar RFI characteristics and sky signals. Various data partitioning schemes are therefore possible, allowing the algorithms to be run in parallel on smaller subsets of the input data. For example, for a pointed observation, RFI characteristics usually vary across spectral windows (frequency ranges) but remain similar across all timesteps and scans. For a fast mosaicing observation, the RFI and the sky signal may vary with position on the sky and these data may require a sky-position based partitioning. For the ngVLA configurations being evaluated, partitioning will also involve different subsets of antennas. The 100 antennas in the central dense core will see more correlated RFI than the outlier antennas and might require algorithm tunings that differ from the rest of the baselines. Outlier antennas are also likely to see different RFI characteristics and may be best treated as separate subsets. Flag extension protocols will be required for baselines that cross these partitions and would have to be applied after flags have been detected for all baseline subsets. A basic requirement for any outlier detector is that the outliers form a small fraction of the input data. For RFI autoflagging, this means choosing time, frequency and baseline subsets such that the RFI-affected data fill less than half of the data. This choice depends strongly on the RFI characteristics as well as the time and frequency resolution at which the data are viewed.

**Compute Load :** The compute requirements of autoflag algorithms scale linearly with the number of visibilities, typically requiring one or two passes through the data (or subset). Mathematical operations include  $O(10)$  iterations of statistics calculations (mean, median, mad, stddev), averages along the time/frequency axis and 1D linear least squares polynomial fits. For the purpose of analyzing compute costs and data rates at which the algorithms must operate, a reasonable partitioning scheme would group together baselines from about 30 antennas, frequency ranges matched to observed spectral windows, and time ranges of a few minutes to allow the algorithms to make use of aggregate RFI-to-noise ratios. Although a few synchronization points may still be required in between algorithm stages, they are not expected to amortize out the gains expected from data parallelization with such partitioning. This estimate is based on current best practices with the VLA, but will need to be re-evaluated with high time resolution data and the ngVLA baseline distribution.

## 2.2 Modeling and Subtraction

When recorded at sufficient time and frequency resolution such that no phase decorrelation occurs, an RFI signal may be modeled in much the same way as an astrophysical source and its contribution subtracted out from the affected visibilities. Such algorithms may operate either in the image domain or via subspace projection applied to visibilities arranged in matrix form [8, 14, 4, 1, 7, 9].

**Subspace projection :** These algorithms typically operate on complex visibilities that are arranged as antenna correlation matrices (or other arrangements that lend themselves to easy factorization of the RFI signal). Principal component analyses (the simplest being an Eigen value decomposition) combined with projection operations are able to separate the RFI signal from the sky signal. These methods rely on the RFI being 'orthogonal' to the sky signal in some way, either by vastly differing brightness levels or being far apart on the sky. The concept of self-calibration is implicit in the way dominant eigen modes are solved for and these methods can operate on uncalibrated data. ngVLA Memo #38 describes one experimental method that works around partially decorrelated RFI by partitioning the array into subsets within which RFI phase coherence is preserved along with protocols to manage baselines that cross subset boundaries. Other experimental ideas exist, but so far none have been proven to be robust enough for routine use. Note that all these methods require that autocorrelation visibility data be recorded.

**Imaging and Peeling :** All-sky imaging in real time is another potential approach towards detecting and modeling RFI as simply another source on the sky (or near the horizon for terrestrial sources). This approach uses sky position as a strong distinguishing factor between sky signal and RFI (a stronger constraint than general orthogonality) but requires adequate pre- (or self-) calibration in the direction of the RFI signal. Algorithms for direction-dependent calibration and source peeling exist but none have yet been evaluated for the purpose of RFI modeling and subtraction. Experiments such as RealFast and Fast-VLITE on the VLA have prototyped frameworks meant for real-time imaging for the purpose of detecting astronomical transient signals (and also distinguishing them from time varying RFI signals) and such frameworks can be re-used to evaluate RFI mitigation approaches.

**Data Partitioning :** Modeling and subtraction algorithms can (in principle) operate separately per timestep and frequency channel but will require data from all baselines to be viewed together. Also, the requirement of phase coherence of the signal being intact places constraints on the maximum baseline length, channel width, time resolution and angular separation between the interferometer phase center and the RFI source. For example, for baselines of 2km or less, phase coherence is preserved at time and frequency resolutions of 10ms and 50kHz for a source up to 20deg away from the interferometer phase center. For the ngVLA, it may suffice to consider such algorithms only for baselines formed from the central dense core (and the plains spiral), both from the point of view of numerical and computational feasibility as well as the relatively small impact of simply flagging longer baselines that see correlated RFI.

**Compute Load :** Computing for subspace projection methods scales with the square of the number of antennas (complex visibilities are replicated to fill in the missing half of the correlation matrices, prior to matrix decomposition). The algorithm itself involves a sequence of standard matrix math operations and one  $N_{ant} \times N_{ant}$  eigen value decomposition per timestep and channel. Off the shelf custom solutions for FPGAs or GPUs may be worth evaluating for a real-time implementation. Computing for imaging based methods will scale with the number of visibilities and the angular resolution (and therefore number of pixels in the image) required to accurately distinguish RFI sources from sky sources. Cost estimates may be extrapolated from current RealFast and Fast-VLITE experiments that have imaged RFI at high time resolution, as well as ngVLA imaging computing costs (for only the ngVLA core).

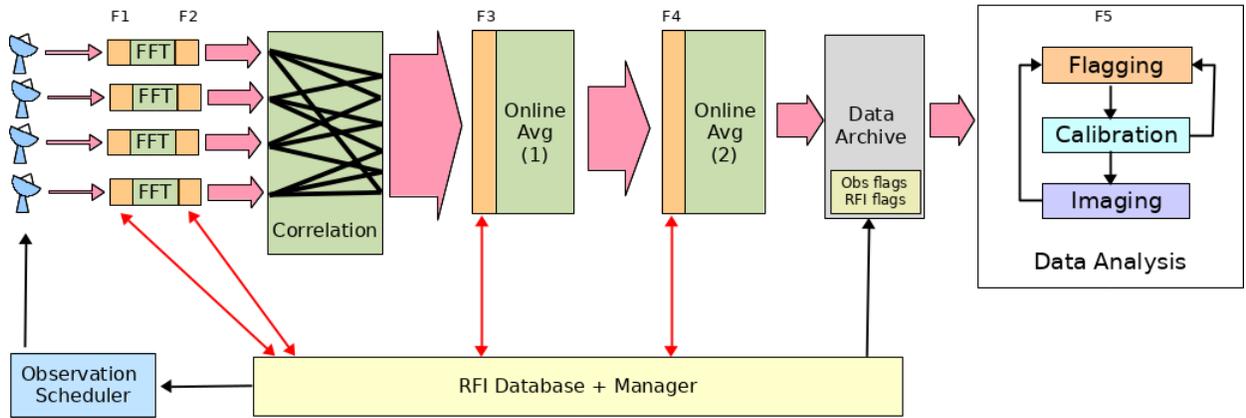


Figure 1: A schematic diagram of an online data acquisition system, archiving and post-processing. Yellow boxes labeled as F1 through F5 represent different signal stages. F1 and F2 represent pre-correlation voltage streams and their short-term Fourier transforms. F3 indicates  $\mu s$  timescales suitable for flagging and recovering data in between intermittent communication signals. F4 indicates  $ms$  timescales more amenable to modeling and subtraction algorithms. F5 indicates solutions applied at archival and post-processing data resolutions. An RFI database is to be maintained across the entire system with useful information derived to be stored in the archive along with the data.

### 2.3 Integrated System Design

With a wide variety of RFI, multiple detection and mitigation options and a range of time resolutions of both the RFI and the observed data spanning nine orders of magnitude between a real time acquisition and the archive, it is important to include RFI mitigation as part of the end-to-end signal processing chain.

**Signal Processing Stages :** Overall, three signal processing stages are being considered for the application of the mitigation approaches described in the previous sections. The first is pre-correlation real-time antenna-based voltage streams (time series and short-term spectra). The second is correlated visibilities viewed at timescales ranging from  $\mu s$  to  $ms$  and 10s of kHz to MHz channel bandwidths. The third is post-processing at archival time/frequency resolution of 1sec/1MHz. Fig.1 illustrates the concept.

**Monitoring / Mining / Avoidance / Removal** There are many other strategies that may be employed to either prevent RFI from entering the system or to use information about RFI early in the signal chain to tune algorithms that will be applied downstream.

1. Store a-priori information about RFI characteristics in a database. Use real-time RFI detection results to dynamically generate meta-data to add to the database.
2. Identify RFI characteristics by mining the current RFI database and use this information to choose and tune outlier detection and RFI modeling algorithms downstream.
3. Implement RFI avoidance via smart scheduling in-between known intermittent RFI emissions (in direction, frequency and time), or in response to a real-time RFI detection system. For frequency bands with a few satellites moving in known orbits, experiments at other observatories suggest that this is feasible.
4. The use of a reference array specifically for RFI monitoring and finding and removing the source of RFI when possible.

ngVLA Memo #71 discusses an integrated system design that could implement multiple mitigation strategies within a real-time system and enable them to interact usefully via a dynamically created and utilized RFI database.

### 3 Evaluating Cost vs Benefit

Information from the ngVLA RFI forecast memo (Ericksen 2018) is used to classify potential RFI based on the expected time-frequency occupancy, with ground-based RFI sources being visible to only a small fraction of antennas, airborne RFI being visible to a large fraction of the array, and satellite transmissions affecting the entire array. The signal processing stages considered are real-time pre-correlation voltage streams per antenna, correlated visibilities at high time (and frequency) resolution within the correlator backend and time-averaged visibilities saved to the archive and used for post-processing. Algorithms for real-time statistical outlier detection and flagging, visibility based flagging and RFI modeling and excision are matched to RFI characteristics that they are likely to be optimal for, and their compute load and resource requirements are evaluated for the most appropriate stage in the signal stage for their implementation.

Implications on imaging performance and the required observing time to reach target sensitivities (due to data lost to RFI) are evaluated for methods currently in use, improvements expected using additional production-ready concepts, and finally potential further improvements requiring algorithm R&D.

**RFI Impact Simulator :** Given the vast number of uncertainties about actual RFI characteristics, the suitability of chosen RFI mitigation options, the optimal data resolution at which to view the RFI and estimates of algorithmic efficiency, an RFI Impact Simulator was written to effectively navigate the parameter space for different RFI scenarios. All RFI characteristics shown in Table 2 are editable, each of four RFI mitigation options may be chosen separately or together, and the effect of RFI decorrelation can be optionally added in for a few specific target image dynamic ranges. A sample screenshot is shown in Fig.2. In the example shown below, the bottom left plot shows the fraction of data lost to RFI (assuming perfect flagging at post-processing data resolutions) separately for each of 7 RFI types. The atmospheric Oxygen band is shown in grey and the proposed ngVLA observing bands are marked at the top of the plot. The bottom right panel shows the effective extra observing time required to reach target sensitivity (a scale factor relative to 1 for the ideal situation of no RFI). The two traces on the plot indicate best and worst case scenarios of multiple RFI types affecting the same subset of data versus distinct subsets of data. Details of all the calculations are documented in the following sections and results are presented in Sec.4.



Figure 2: <https://gitlab.nrao.edu/rurvashi/ngvla-rfi-impact-simulator>

### 3.1 Defining the Parameter Space

#### 3.1.1 Array Configuration

Name	Number of antennas	Diameter	Max Baseline	Classification for RFI
Main Array : Core	94	18m	<1.25 km	remote location
Main Array : Plains	74	18m	<36 km	mostly remote
Main Array : Mid-Baselines	46	18m	<1000 km	some remote, some near humans
Short Baseline Array	19	6m	<0.06 km	remote
Long Baseline Array	30	18m	<8856 km	near humans

Table 1: Proposed ngVLA antenna configuration [2]

From the point of view of RFI impact assessment, three categories of antennas were chosen from Table 1.

1. *outlier* : The 30 long baseline and 46 mid-baseline antennas are likely to be near humans (cities/towns), with baseline lengths of a few km to several 1000s of km. Local or disjoint subsets of the outer Plains antennas may also be considered in this category. We assume that at any given time, about 30 antennas see 'local' RFI.
2. *core* : We consider the core as 113 (94+19) antennas of the main core and the short baseline array with a maximum baseline of about 1.25km, plus about 30 inner Plains antennas within a 5km radius from the core. RFI from the telescope site or nearby highways is likely to be seen by all these antennas.
3. *full* : All 263 antennas. Satellite RFI is assumed to be visible over the entire main and short baseline array (with 1000km footprints [13]).

This categorization is used in our calculations as follows. RFI characteristics are described in terms of what fraction of a day an interferer is seen over one or more of these subsets of the array, the numbers of antennas and baselines for each set are used in calculations of data loss fractions, and the distribution of baseline lengths within each subset are used in calculations of the effect of decorrelation.

#### 3.1.2 RFI Characteristics

The RFI forecast from ngVLA Memo # 48 was used to construct Table 2 in which RFI sources are classified into several categories based on similarities in their RFI characteristics.

1. *People* : Intermittent and mostly non-repeating terrestrial RFI seen always on outlier antennas and for part of the day near the core.
2. *Ultra-Wide Band Signals* : A new commercial protocol for short-distance device communication with nano-second pulses and extremely wide band footprints.
3. *Cell 5G* : Mostly terrestrial continuous data transmissions affecting all outlier antennas at all times and visible over the core and the full array for a small fraction of each day. Low- and mid-band (< 6GHz) transmissions are expected to be long-range with high-band transmissions being more localized.
4. *LEO Satellites* : Low Earth Orbit Satellite Internet transmissions expected to be continuous for nearly the entire day/night and visible to the entire array.
5. *Aircraft and Ground Radar* : Terrestrial and airborne transmitters using communication protocols with clear on/off duty cycles on the 10s of  $\mu s$  timescale. These transmissions are likely to be seen by outlier antennas most of the time and about half the time at the core.
6. *Satellite communication* : Similar signal characteristics as aircraft and ground radar, but visible to the entire array for a significant fraction of the day.

Several assumptions have been made in this classification of RFI types as well as in estimates of the subsets of the array that can see the RFI and the fraction of time such transmitters are on. Using the RFI impact simulator tool described above, all these values may be edited and the results recomputed.

## RFI Characteristics

Type	Freq Range	Freq Res	Time Res	Time Frac	Array Frac
Cell 5G - High ['cell phones', 'local p2p', 'cars on Hwy 60', 'IoT']	[[24.25, 27.5], [31.8, 33.4], [37.0, 43.5], [45.5, 50.2], [50.4, 52.6], [66.0, 76.0], [81.0, 86.0]]	200.0 kHz	0.0001 s	1.0 0.2	outlier core
LEO Sat ['spacex', 'oneweb', 'boe- ing', 'Iridium Data', 'sirius XM']	[[1.61, 1.63], [2.2, 2.33], [10.7, 12.7], [13.85, 14.5], [17.8, 18.6], [18.8, 19.3], [27.5, 29.1], [29.5, 30.0], [17.8, 19.3], [27.5, 29.1], [29.5, 30.0], [37.5, 42.0], [47.2, 50.2], [50.4, 51.4]]	200.0 kHz	0.0001 s	0.9	full
People ['wifi', 'blue- tooth', 'wireless usb', 'wireless video', 'car radar']	[[0.9, 0.93], [2.4, 2.5], [3.6, 3.7], [4.9, 5.0], [5.1, 5.9], [61, 61.5], [3.0, 11.0], [5.7, 5.9], [76.0, 81.0]]	20000.0 kHz	0.001 s	1.0 0.4	outlier core
Aircraft Comm ['ground radar', 'aircraft comm', 'aircraft data', 'air- craft nav', 'aircraft radar', 'drones']	[[1.24, 1.37], [2.7, 3.0], [5.6, 5.65], [9.0, 9.2], [15.7, 16.2], [0.978, 1.213], [4.2, 4.4], [5.3, 5.5], [8.7, 8.9], [9.3, 9.5], [8.0, 12.0], [12.0, 18.0], [13.2, 13.4]]	100.0 kHz	2e-05 s	0.5 0.9	core outlier
Sat Comm ['GNSS', 'SAR', 'Iridium Comm', 'Military', 'Cloud- Sat']	[[1.1, 1.8], [2.0, 2.3], [2.5, 2.7], [3.4, 4.2], [7.25, 7.75], [8.0, 9.0], [10.7, 12.8], [18.0, 20.0], [23.0, 27.0], [94.01, 94.09]]	100.0 kHz	2e-05 s	0.4	full
UWB ['wifi', 'blue- tooth', 'wireless usb', 'wireless video', 'car radar', 'surveillance', 'handheld comm']	[[1.99, 10.6], [7.5, 8.5], [22.0, 29.0]]	900000.0 kHz	1e-09 s	0.9 0.2	outlier core
Cell 5G - Low/Mid ['cell phones', 'cars on hwy60', 'aircraft 5G']	[[1.427, 1.518], [3.3, 3.8], [5.15, 5.925]]	200.0 kHz	0.0001 s	1.0 0.25 0.1	outlier full core

Table 2: Potential RFI that the ngVLA will see, based on ngVLA memo #48. RFI types are listed along with allocated frequencies, signal bandwidth and duty cycle, and fraction of a day that the RFI is likely to be present over different subsets of the array. *Note* : All our calculations assume that entire allocated bands are filled when a transmission is seen. This (we hope) is an over-estimate that represents the worst-case scenario.

### 3.1.3 Data Processing Stages and Algorithms

Three data processing stages are identified, each with a wide range of time and frequency resolutions at which to view the data. Table.3 lists these options and matches them to algorithm choices.

Signal Processing Stage	Time resolutions	Channel resolution	Algorithms	Types of RFI
Antenna-based flagging	0.1ns, 1ns, 10ns	100 kHz	1D filtering	Sparks and UWB pulses
Baseline-based flagging	10 $\mu$ s, 0.1ms, 1 ms, 10ms	100 kHz	1D and 2D outlier detection and flagging	Intermittent RFI with on/off duty cycles.
High time resolution modeling and subtraction	1 $\mu$ s, 10 $\mu$ s, 0.1ms, 1ms	10 kHz	Subspace Projection and Imaging	Continuous signals that mimic astrophysical sources
Post Processing	0.1s, 1s, 10s	1 MHz	Outlier detection and flagging	Long-duration and weak RFI

Table 3: Data processing stages matched with algorithms and targeted RFI types.

The purpose of enumerating these data resolution options is to calculate accurate RFI filling fractions for different types of RFI that might have on/off duty cycles at different timescales. Based on the input RFI characteristics and the algorithm efficiency it would also be possible to assess whether there are a few preferred data resolutions that must be optimized in the system design or whether there is a critical need for flexibility (and added complexity) to dynamically adapt to different types of RFI.

## 3.2 Metrics to Quantify the Impact of RFI and Mitigation Options

A series of metrics were defined to quantify the impact of RFI, to assess improvements after applying one or more RFI mitigation options and to analyse the operational cost of doing so.

1. **Fraction of Data Loss :** The first is the fraction of data lost and it depends on the subsets of the array that see correlated, uncorrelated and decorrelated RFI from multiple RFI source types, the true RFI filling fraction at the time-resolution of on/off duty cycles compared to the time resolution at which the data are viewed, and the expected efficiency of the mitigation algorithms that have been applied.
2. **Increased Observing Time :** The above fraction of data loss can be converted to an estimate of the total observing time required to compensate for the data loss (a factor relative to the situation of no RFI).
3. **Compute Load :** Finally, the compute load per solution may be estimated from the data rate at which the algorithm must operate and the number of operations to be done per visibility datum (relative to the simplest case of only averaging the data down to archival time and frequency resolutions).

Analyses may then be done to compare the improvement in data loss fractions with the operational cost of applying the mitigation solutions that produced the improvement.

### 3.2.1 Fraction of Data Loss

Several factors play a role in calculating an effective fraction of data loss.

**Baselines with correlated RFI :** Each RFI type in Table.2 is associated with a subset of the array over which the RFI is seen (outlier, core, full). The number of affected baselines (both antennas see the RFI) is  $N_{base_{rfi}} = N_{ant}(N_{ant} - 1)/2$ .

**Effect of decorrelation :** Depending on the angular separation between the RFI and the phase reference location, the channel bandwidth, baseline length and time range per integration, RFI can de-correlate and be partially suppressed on long baselines. A key assumption is that the RFI is continuous over the time range being considered for decorrelation and that its amplitude and phase structure is not otherwise corrupted.

Specifically, for a source located at an angle  $\theta$  away from the phase reference position, the correlation coefficient will be attenuated by  $sinc(\phi)$  where  $\phi = 2\pi\delta t\delta\nu$  and  $d$  is the physical baseline length. The delay between the signals as seen by two antennas gives  $\delta t = \frac{d}{c}sin(\theta)$  and the observation channel bandwidth decides  $\delta\nu$ . This is the attenuation due to the channel width (from ngVLA Memo #38). Attenuation by decorrelation is also achieved by increasing the integration timestep where  $\delta t$  now represents the timestep and  $\delta\nu = \frac{c}{d sin(\theta)}$  is the fringe frequency. The total attenuation is calculated as the product of the two  $sinc$  functions for the specific combination of time and frequency binning.

The above calculation applies to a single baseline length. An approximate distribution of baseline lengths for the ngVLA is derived from Table.1 based on the number of baselines below 5km, between 5km and 1000 km, and longer than 1000 km. An overall decorrelation attenuation is computed as an average across the full range of affected baseline lengths, weighted by the relative number of baselines in each baseline-length bin.

An average attenuation of at least 40dB would be required to suppress an RFI signal such that it may be considered insignificant for expected ngVLA imaging dynamic ranges of  $10^4$ . For any given source direction  $\theta$ , we can therefore calculate what fraction of baselines will be effectively RFI-free and subtract this number of baselines from  $N_{base_{rfi}}$ . In the RFI impact simulator, relative source directions of 20 deg and 90 deg are options that represent satellite and terrestrial RFI respectively. Attenuation thresholds of 20dB, 40dB and 60dB are also options to represent target imaging dynamic ranges of  $10^2$ ,  $10^4$ ,  $10^6$  respectively.

**Baselines with uncorrelated RFI :** Interference seen by only one antenna contributing to a baseline will not correlate. This is a natural means of RFI suppression in an interferometer. The resulting mean visibility from that baseline will be devoid of RFI but the measurement will be associated with a higher level of noise. Our calculations assumed a 20% increase in noise. Baselines with correlated RFI that has undergone attenuation due to decorrelation are also treated similarly with a 20% increase in measurement noise. An increase in noise is translated to an effective baseline loss fraction (assuming Gaussian statistics) that is then added to  $N_{base_{rfi}}$ .

**RFI filling fractions :** An RFI filling fraction is calculated as  $\frac{\max(\delta t, t_{on})}{t_{on} + t_{off}}$  where  $t_{on}$  and  $t_{off}$  define the on/off duty cycle of each type of RFI and  $\delta t$  is the time resolution at which the data are viewed. Flagging efficiency is highest (filling fraction is lowest) when only the time/frequency ranges truly affected by RFI are flagged and all surrounding non-corrupted data are left intact. For example, if the data resolution is too low (larger than the on/off duty cycle), then intermittent RFI that may be present only for 50% of the time will appear continuous and result in 100% data loss. An optimal match between flagging efficiency and compute load is achieved when the data resolution is matched with the duty cycle of the RFI signal itself.

**Algorithm Efficiency :** In our calculations, outlier detection and flagging algorithms are assumed to operate perfectly with the fraction of RFI-affected data exactly matching the RFI filling fraction calculated at the data resolution being considered. Modeling and subtraction algorithms are assumed to work as intended only over 50% of the affected baselines, for situations where the number of interferers is low and their positions are likely to be known. For LEO satellites, we assume that the algorithms will have very little effect reducing the RFI data loss by only 10%.

**Overall fraction of data loss** For each RFI type in Table.2,RFI mitigation option and data resolution combination from Table.3, the fraction of data loss is computed as

$$f_{loss} = \frac{Nbase_{rfi}}{Nbase_{total}} \times \frac{Nhrs_{day}}{24hrs} \times \frac{\max(\delta t, t_{on})}{t_{on} + t_{off}} \quad (1)$$

The data time resolution that produces the lowest  $f_{loss}$  is chosen as the optimal stage of data processing. When multiple RFI mitigation algorithms are applied, the option that results in the lowest  $f_{loss}$  is picked as the optimal choice of algorithm for the RFI type. The resulting fractional data loss per RFI type is what is displayed in all the figures in Sec.4. Finally, when there are multiple RFI emitters in the same frequency range, we calculate two metrics from the list of data loss fractions for all overlapping RFI. The first is the maximum, which assumes that all types of RFI overlap in the data. The second is the sum, which assumes that each type of RFI (unfortunately) affects different subsets of the data. Practical reality will be somewhere in between.

### 3.2.2 Extra Observing Time Required to Compensate for Data Loss

For the above two estimates of fractional data loss, we calculate the extra amount of observing time that would be required to compensate for this data loss and achieve the same theoretical imaging sensitivity as an observation with no RFI. This calculation assumes Gaussian statistics and the fact that the same fractional data loss will apply to extra observing time as well. Plots in Sec.4 show the ratio of the total time required to compensate for losses due to RFI to the time that would be required if there were no RFI. A value of 1.8 implies a an 80% increase in observing time.

### 3.2.3 Compute Load

The following factors influence the compute load of each RFI mitigation option. More detailed versions of the calculations below will be required to assess options for compute infrastructure at each data processing stage, based on what combination of data resolution and algorithm has the most beneficial effect (in terms of reducing the fraction of data lost to RFI).

**Data Rate :** The dominant factor that will influence system design is the data rate at which algorithms need to operate. This is given by the total number of samples or visibilities to be operated upon per second (per subband). Parallelization will be routine and the algorithms are amenable to straightforward data partitioning. For example, outlier detection algorithms need to view ranges in time and frequency but can operate independently on different baselines. Modeling and subtraction algorithms can operate separately per timestep and channel but need to see all baselines together.

**Number of operations :** The number of operations per visibility data point is another key factor that will vary with the choice of algorithm. For an initial analysis we consider 'only averaging', 'auto-flagging + averaging', and 'modeling/subtraction + averaging' and use relative scale factors of 1, 20, 200 to indicate the difference between the number of operations per datum.

**Number of operations per second :** For a given combination of data resolution and algorithm choice, the product of the data rate and the number of operations per datum gives the number of operations to be performed per second. Table.4 quantifies this for a few data resolutions and algorithm choices (for one 100 MHz sub-band). Outlier detection is assumed to run on all baselines whereas modeling and subtraction algorithms would apply only to core antennas within a 1km diameter where RFI phase coherence is expected to be preserved. Each type of algorithm can make use of different parallelization axes and this is encoded in the last column.

Data Processing Stage	Time Res	Chan Res	Data Rate (Nvis/sec/band)	Number of operations per second (per band)
Baseline-based flagging	5 $\mu$ s	100 kHz	6.8e+12 for 263 antennas	1.3e+14 serial (4e+9 for each of 3.4e+4 baselines)
Modeling and Subtraction	1ms	20 kHz	3.1e+10 for 113 core antennas	6.3e+12 serial (1.2e+6 for each of 5e+6 timesteps and channels)
Post-processing flagging	1s	1 MHz	3.4e+6 for 263 antennas	6.8e+7 serial ( 2e+3 for each of 3.4e+4 baselines)

Table 4: Approximate data rates and relative number of operations per second (per subband) for three mitigation options. *These initial calculations are meant to be interpreted only as order-of-magnitude estimates to compare the relative compute load between RFI mitigation options, and not as absolute compute loads. Detailed calculations must be performed prior to formal system design.*

## 4 Analysis and Results

This section presents results of the above calculations and summarizes the estimated impact on ngVLA science (continuum, spectral line, time-domain), system design and resource sizing, and key areas where regulation might be essential.

### 4.1 Effects of RFI and Various Mitigation Options

Figures 3 through 6 show results for the RFI characteristics listed in Table.2, with different RFI mitigation options added on one by one. For each type of RFI, the fraction of data loss is calculated for all frequencies within the *allocated* ranges. It assumes the worst case of commercial bands being entirely filled, even though in practice at any given time only some fraction of each allocated band may be assigned and used by a nearby transmitter. All calculations include the effect of RFI decorrelation, assuming a 20deg separation between the observation phase center and the RFI and a 40db attenuation threshold to indicate when a baseline is effectively free of RFI. For each situation, plots are shown for the fraction of data loss and the estimates extra observing time required to compensate. Fig.7 shows the estimated effect of decorrelation on long baselines for multiple attenuation thresholds and Fig.8 shows the effect of varying some key RFI characteristics such as the presence of 5G Cell towers near plains-spiral antennas and reduced off-duty-cycle gaps for satellite communication signals.

**Comparison with current reality :** For frequencies between 1GHz and 10GHz, data and experience from the VLA may be used to validate the estimates. Typical data loss fractions are about 40% at L-Band (1-2 GHz), 30% at S-Band (2-4 GHz), and at about 20% at X-Band (4-8 GHz) and observation subbands are typically placed to avoid known satellite transmission bands. Fig.3 shows calculations with comparable loss fractions although data from the current VLA suggest that entire allocated satellite communication bands are not simultaneously filled.

### Post-processing Only



Figure 3: When the data are viewed at post-processing resolution, all RFI that is intermittent at timescales shorter than 1sec appear continuous. Data loss fractions range from 20% to up to 80% for satellite RFI that is continuous and seen by the entire array. RFI from people or objects moving around the array is expected to have low impact because of the short amount of time it is present. In this estimate, it is assumed that cell 5G transmissions are seen only at most 50 outlier antennas. Between 1 and 10 GHz, these fractions roughly match current reality. *This figure represents the lowest-cost option with no real-time RFI mitigation, but most frequency bands would require almost double the observing time to compensate for data lost due to RFI and all LEO bands would be practically unusable.*

## Post-processing + Antenna-based flagging

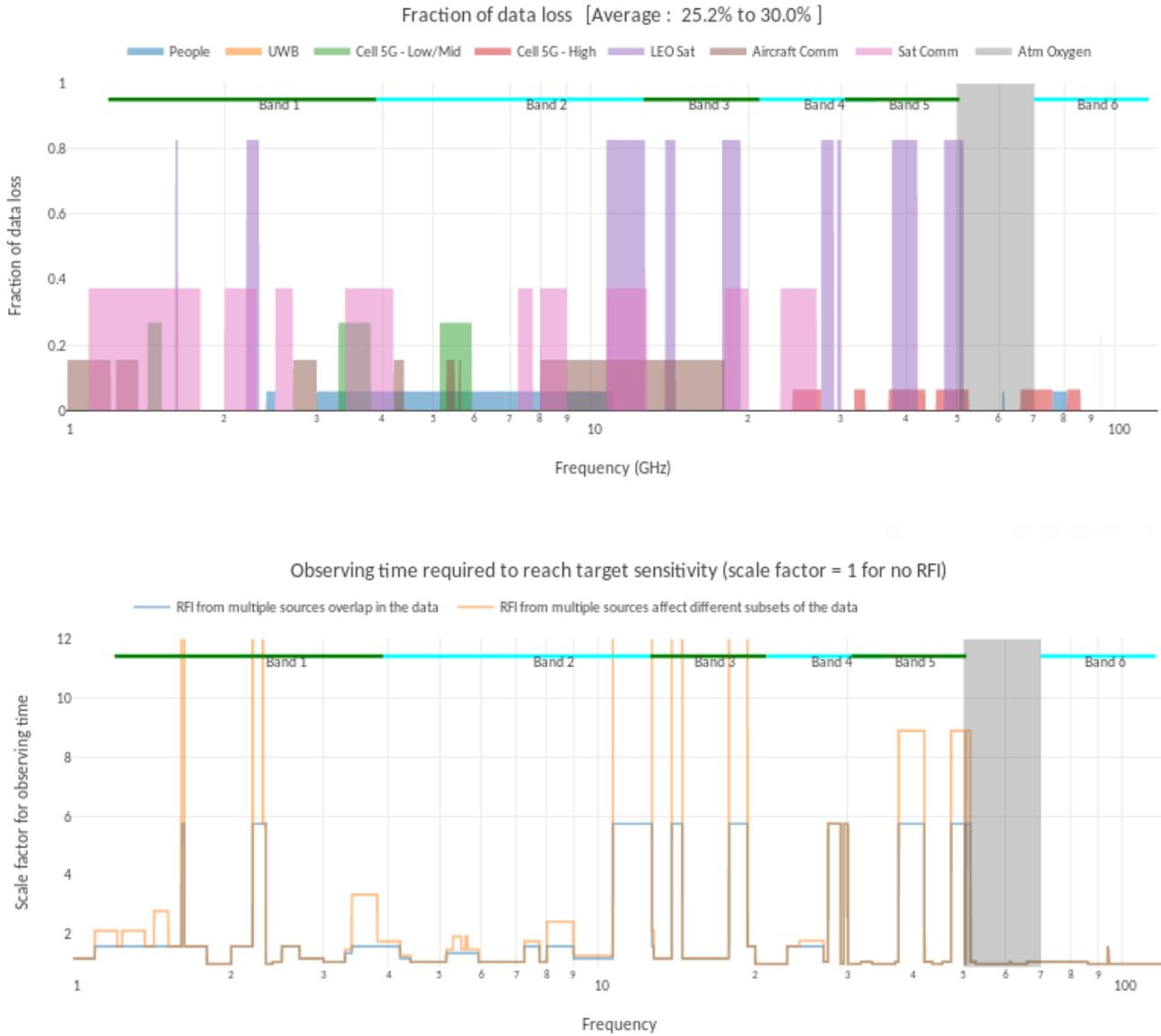


Figure 4: When real-time antenna-based detection on voltage time series is included, RFI that it is intermittent on timescales shorter than about  $1\mu s$  may be eliminated (the simplest approach being blanking). Ultra-wideband signals, nano-sec scale sparks and some DME pulses are most suited to this stage of processing. It is important to note though that actual mitigation at this stage is relevant only for signals with on/off duty cycles much smaller than  $1\mu s$ . Detections of infrequent UWB pulses or strong tones (continuous up to  $\mu s$  timescales or more) could be used to guide subsequent post-correlation flagging operations done at a lower data rate to match the true on/off duty cycle of the signal. In these calculations, excision (blanking) appears to have a low overall impact primarily due to UWB signals being local and infrequent, but this could change in the future. In any case, the information gathered at this stage would be of significant use downstream.

## Post-processing + Antenna-based flagging + In-correlator flagging

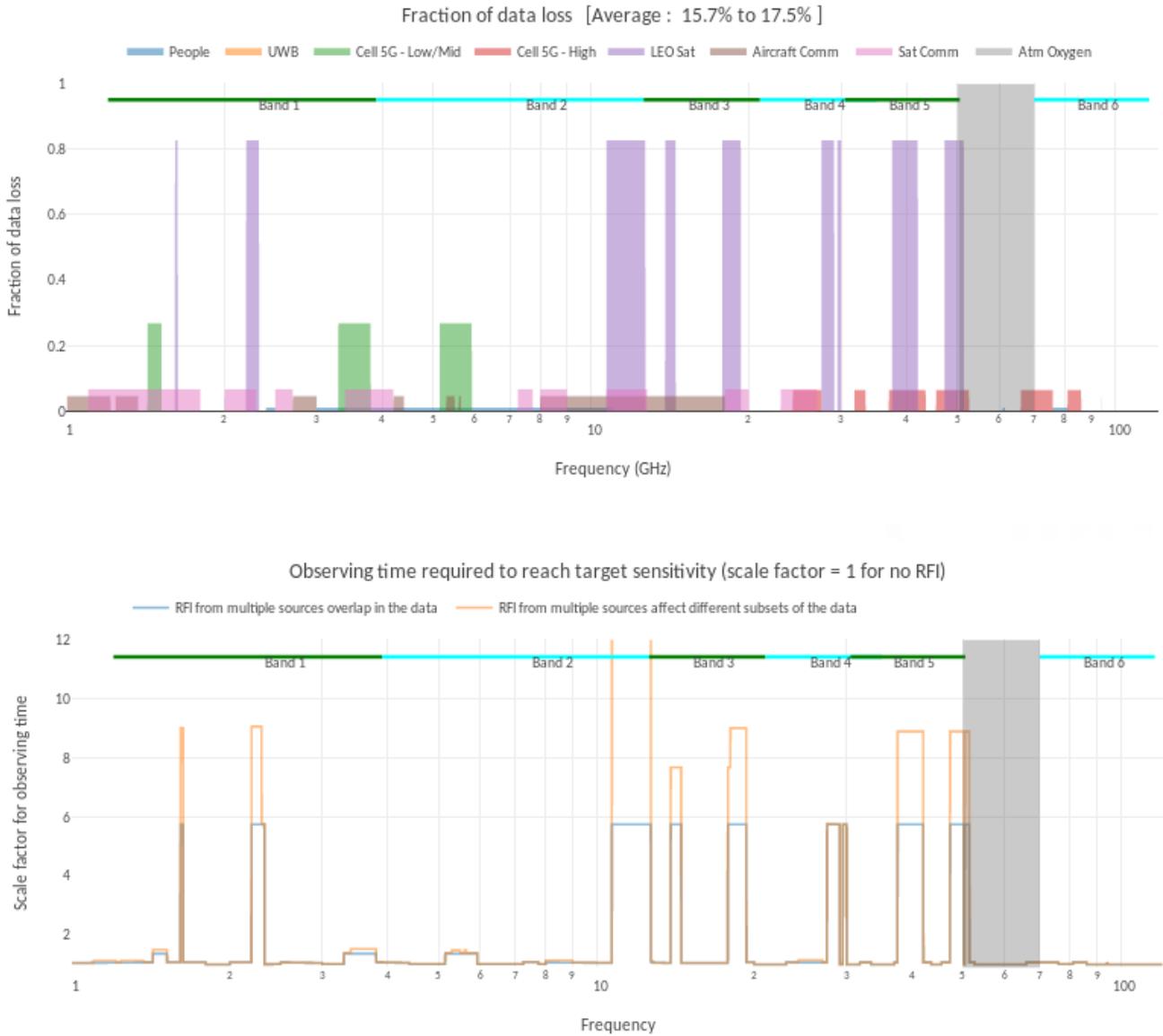


Figure 5: In-correlator flagging is expected to be effective when there are usable gaps in time and frequency between commercial transmissions, even within allocated bands being shared by multiple operators. All communication protocols (in theory) have clear on/off duty cycles (for example, few 10s of  $\mu s$  gaps in between (say)  $5\mu s$  pulses) and employ frequency hopping and separate bands for uplinks and downlinks. Flagging accuracy is maximized when the data are viewed at or at a higher resolution than the signal duty cycle, and compute load is lower when the data are viewed at lower time/freq resolution. Therefore an optimal balance is achieved by matching the flagging data resolution with the duty cycle and signal bandwidth of these types of RFI (for example, a few  $\mu s$  and 10 kHz). These calculations assume perfect flagging at a timescale matched to the signal such that the data loss exactly equals the filling fraction of the RFI. If successful, the second plot shows that in-correlator flagging is likely to have a significant positive impact on operations. But, a big assumption here is the presence of usable gaps at time and frequency resolutions of  $\mu s$  to  $ms$  10-100 kHz and it is imperative that measurements be carried out to verify (or invalidate) this detail.

## Post-processing + Antenna-based flagging + In-correlator flagging + Modeling and subtraction

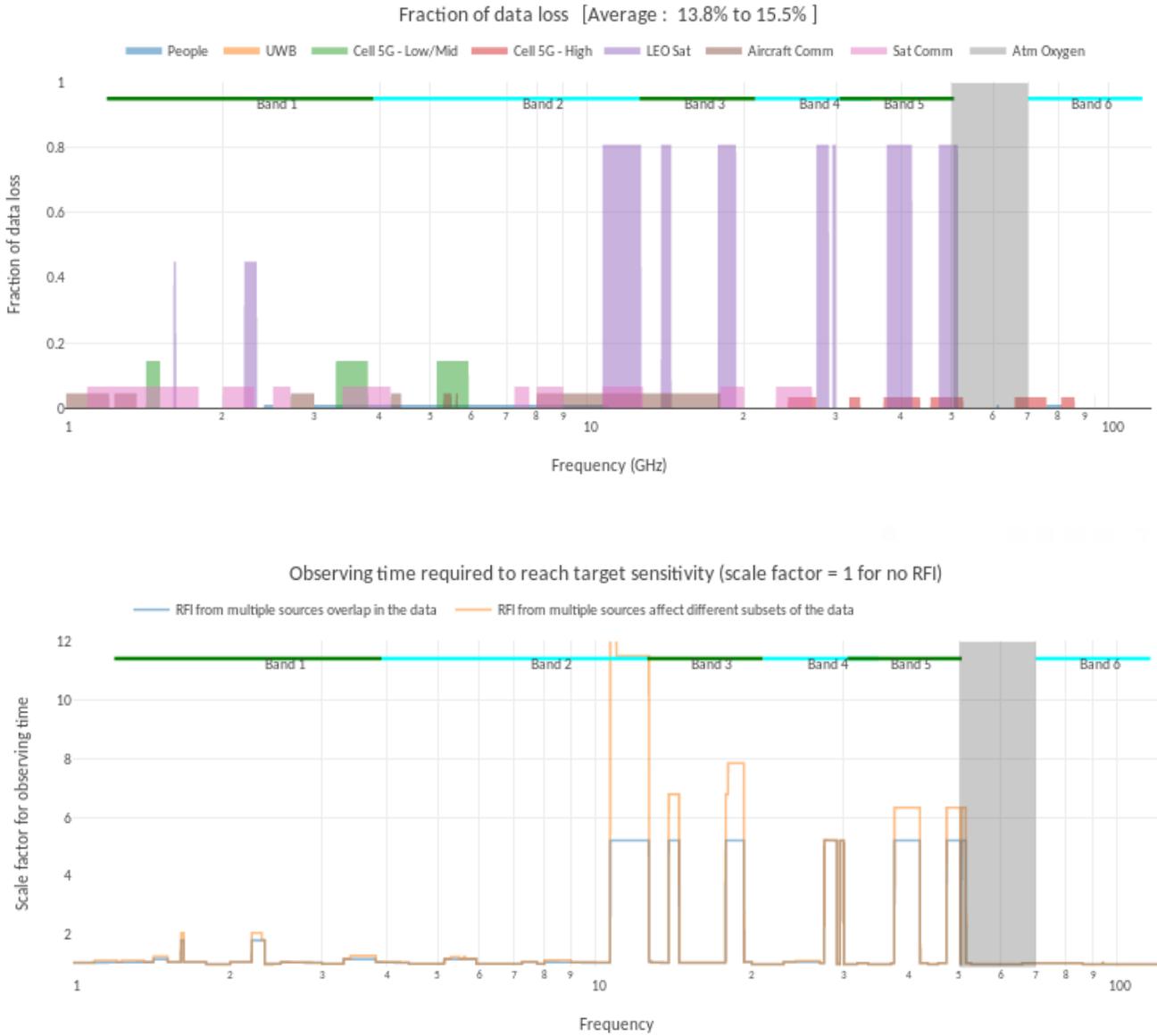


Figure 6: RFI modeling and subtraction algorithms are expected to be useful for signals that are continuous up to  $\mu s$ ,  $ms$  or longer timescales, thus mimicking an astrophysical source that can be imaged. These calculations assume that subtraction is accurate and sufficient for up to 50% of baselines that see the RFI, but only when a few interferers are visible at once and whose locations are likely to be known a-priori. 5G Cell towers and geo-stationary or GPS satellites are the primary RFI sources to target with such methods. It is important to note, however, that all modeling and subtractions methods are currently experimental. Several prototypes exist but none so far have proven to be robust enough for practical use. *This calculation is therefore an entirely theoretical and optimistic estimate.* Commercial LEO satellite constellations are likely to produce too many simultaneous interferers for our current algorithms and data recovery expectations in the satellite internet bands are minimal.

## Effect of decorrelation (for 4 different situations) + Post Processing Flagging

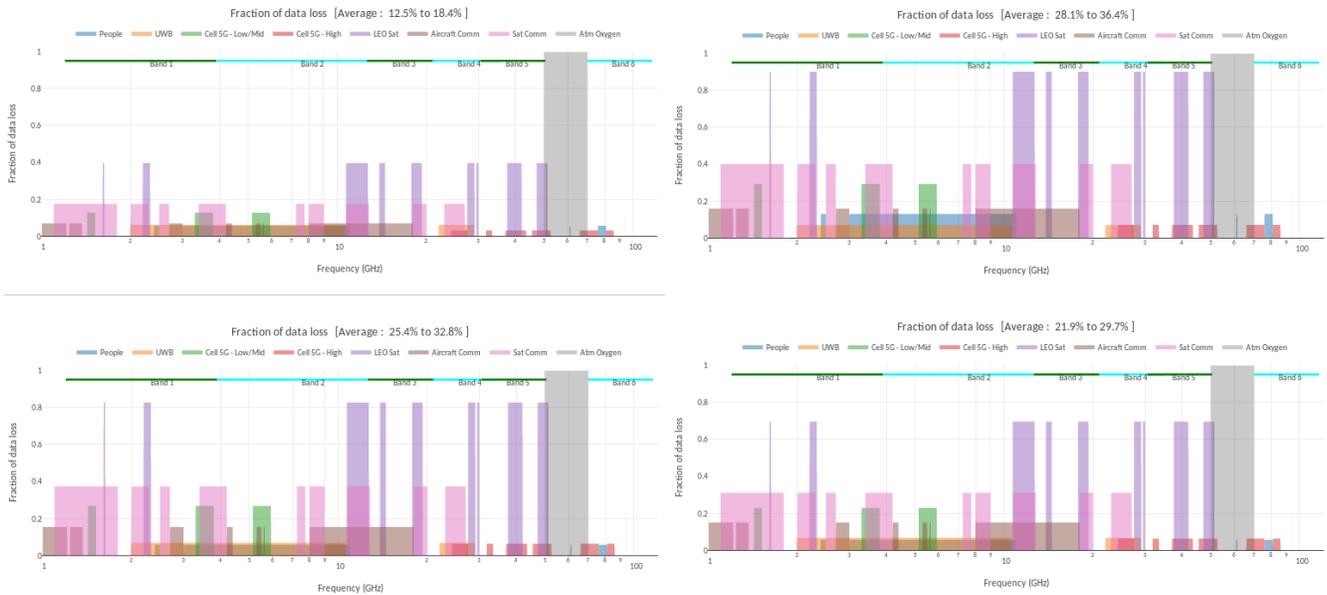


Figure 7: The estimated effect of decorrelation is shown for 4 different situations. Two parameters are varied. The first is the angular distance of the interferer from the observation phase center (which governs the projected baseline length in the direction of the interferer). Two cases are considered with 20 degrees representing satellite RFI and 90 degrees representing terrestrial RFI. The second parameter is an attenuation threshold below which the RFI may be considered non-existent, given by the target imaging dynamic range.

1. TOP LEFT : Angular distance of 20 deg and an attenuation threshold of 20 dB.
2. TOP RIGHT : Angular distance of 20 deg and an attenuation threshold of 60 dB.
3. BOTTOM LEFT : Angular distance of 20 deg and an attenuation threshold of 40 dB (same as figures 3 to 6).
4. BOTTOM RIGHT : Angular distance of 90 deg and an attenuation threshold of 40 dB.

The top row of plots illustrate the difference between dynamic range limits of  $10^2$  and  $10^6$  and suggest that the effect of decorrelation may be useful at very low imaging dynamic ranges but there is likely to be no effect for high dynamic range imaging. Even for imaging dynamic ranges of  $10^4$  that the ngVLA is being designed for, there is only a minimal effect.

The bottom row illustrates the (only minor) effect of the angle of arrival of the RFI.

This calculation has many assumptions of stationarity, continuity, and phase-coherence of the input RFI signal over a duration of 1 sec and assumes maximal decorrelation by all baselines seeing the RFI end-on. It is therefore important to verify (or invalidate) these calculations by experimental measurements, perhaps using a controlled RFI source located along one of the VLA arms.

## Effect of Varying the RFI characteristics

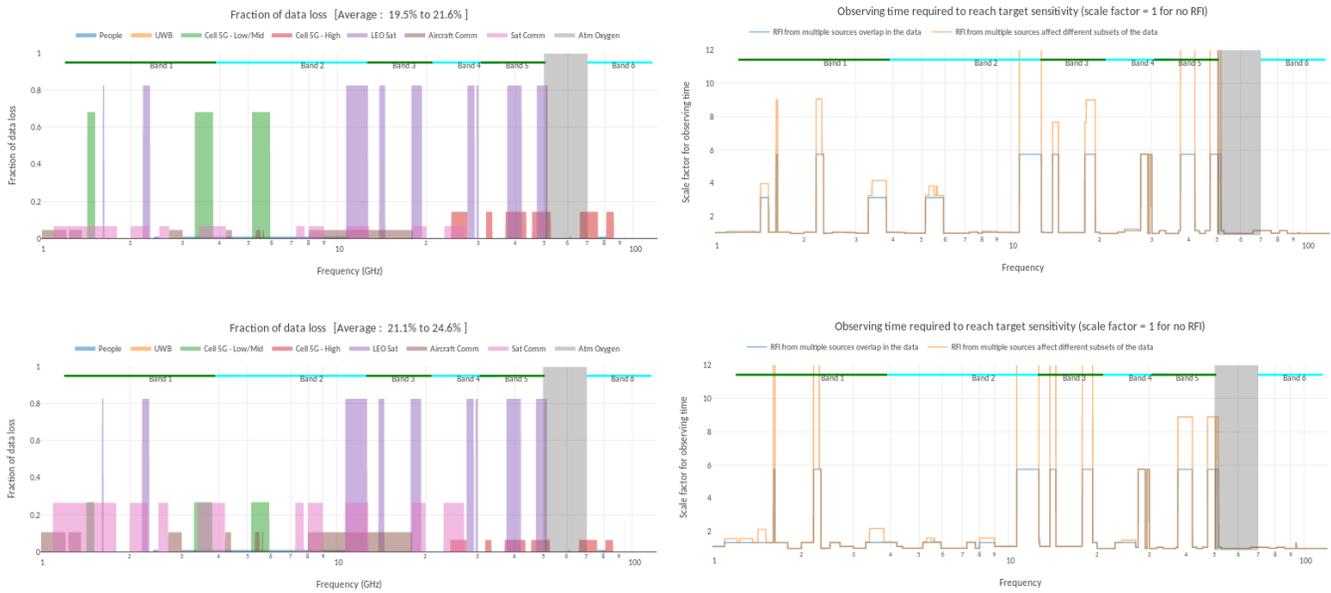


Figure 8: The RFI characteristics used in this analysis represent an educated guess/forecast based on current experience and information about emergent technology and FCC frequency allocations. An obvious question then is how robust our technology and algorithm choices for RFI mitigation for the ngVLA are to deviations from the predicted RFI environment. The RFI impact simulator may be used to evaluate this, and two illustrative examples are shown below. RFI mitigation options chosen were post-processing flagging, antenna-based flagging, and in-correlator high time resolution flagging and these figures can be compared with Fig.5 which shows results from the same choice of RFI mitigation options with the original RFI characteristics.

2. TOP : If long-range cell 5G transmissions are seen by the entire array (including the core), and present for 80% of the time duration of an observation. Such a scenario might arise if a cell 5G tower is placed in view of the ngVLA core and plains-spiral antennas. In Table.2 and Fig.5, it is already assumed that all outlier antennas (outside of the plains spiral) will see short-range cell 5G transmissions. By their nature, such signals are considered continuous when on and modeling and subtraction algorithms are the only viable option (if they can be proven to be robust for fixed ground-based interferers). Therefore, with current options, data losses in low-band 5G bands might approach LEO satellites in terms of fractional data loss. *It is therefore important to ensure that the main core and inner plains spiral antennas are shielded from 5G cell towers.*
2. BOTTOM : In Table.2, satellite and aircraft communication signals are predicted to have on/off duty cycles of  $20\mu s$  and  $100\mu s$ . Here, the off duration was changed to  $10\mu s$  to represent the situation where there may not be many usable time/freq gaps at in-correlator data resolutions. In-correlator flagging becomes less effective and the data loss of 30% is not much less that achieved by post-processing alone (Fig.3). *In this situation, the case for in-correlator flagging becomes weak and the high time resolution RFI characteristics must be measured and evaluated prior to detailed system design.*

## 4.2 Impact on ngVLA Science

**Continuum Imaging :** Continuum imaging typically combines data from multiple observing frequencies to form a single image of the average intensity distribution as well as a model of the broad-band spectral structure. Multi-frequency data contribute to lowering the aggregate image noise level, increasing image reconstruction accuracy by providing additional measurements of the sky visibility function, and providing spectral information. Spectral structure is typically modeled by low-order polynomials across a broad frequency range. It is therefore possible to use RFI-free or low-RFI regions only to build up continuum imaging sensitivity and to still derive an accurate spectral model. The effect of RFI on continuum imaging is therefore primarily data loss that can be compensated by more time spent on source. Extra care is required mainly to ensure a roughly uniform distribution of sample weights across the frequency range with the explicit inclusion of RFI-free (or low RFI) regions on both ends of the range to minimize extrapolation errors during wideband image model reconstruction. *Based on the VLA's current ability to perform continuum science between 1 and 10 GHz and the similarity in the expected RFI profile between 10 and 100 GHz, it is expected that given sufficient extra observing time, the ngVLA will be able to do similar levels of continuum science even without additional RFI mitigation strategies.*

**Spectral Line Science :** All spectral line science that overlaps with commercial transmissions will be negatively affected. Table 10 of ngVLA Memo #48 [5] lists the frequency ranges of interest for a series of molecules listed in ngVLA Memo # 19 [10] under the Key Science Project of charting the assembly, structure and evolution of galaxies from the first billion years to the present. Based on the overlap in frequencies, we can make the following conclusions (although we also note that it is insufficient to consider rest frequencies of spectral line emissions as they may red-shift into any part of the spectrum).

1. *Lines such as Formaldehyde, Ammonia and Red-Shifted HI overlap with satellite communication bands and will likely require in-correlation outlier detection and flagging to be able to work with high and low resolution gaps between transmissions.* Satellite and aircraft RFI currently affect red-shifted HI observations with the VLA and provide an ideal test bed for experiments and prototypes that must be done prior to finalizing detailed ngVLA system design.
2. Deuterated water, Ammonia Formylium, Hydrogen Cyanide, Formylium, Diazenlium, and Carbon Monoxide all reside between 70GHz and 114 GHz. Cell 5G bands and car radar are expected to be the primary emitters in this region and reach up to 86 GHz. These observations are likely to be feasible (at angular resolutions offered by the ngVLA core and plains spiral) as long as 5G transmissions are limited to only outlier antennas. Car radar is expected to be intermittent and therefore largely ignorable from the point of view of fractional data loss. *As of now, frequencies higher than 88 GHz are mostly clear, but it remains to be seen how FCC allocations in this band proceed in the next several years.*

**Time-domain Science :** Transient astrophysical sources (repeating or non-repeating), pulsars, nano-second pulsed RFI and transmissions from moving satellites all share signal characteristics. *The false positive rate will continue to be a significant concern for transient science, especially when multiple telescopes are scheduled to respond to sky triggers.* Multiple efforts are under way at various observatories to employ smart detection and source classification algorithms and such systems would have to be developed for the ngVLA both in real time on voltage streams as well as with fast imaging. Experiments such as RealFAST, Fast-VLITE and HERA+EPIC are ideal prototyping platforms. It is also worth noting also that the system design and infrastructure that will be required for time-domain imaging science is identical to that required for high time resolution RFI modeling and subtraction algorithms.

### 4.3 Impact on ngVLA System Design (and Regulation)

The RFI environment between 10 and 100 GHz is expected to be similar to what is currently seen at the VLA between 1 and 4 GHz. We also expect that antennas away from the ngVLA and plains spiral are likely to see more RFI than the current VLA. Post processing alone could result in data losses between 20% and 35% across the ngVLA-observable spectrum, with losses of more than 80% in the LEO satellite bands. The implications for continuum science are increased observing time in low-RFI or RFI-free bands, and spectral line science in commercial (especially LEO) bands is likely to be inaccessible. Although most continuum and line science will still be possible (especially for sources where the new science depends mostly on spatial-frequency sensitivity distribution offered by the ngVLA), there is a strong case for recovering observing time and spectrum through multiple RFI mitigation schemes designed into the end-to-end data acquisition and processing systems for the ngVLA. The general approach is summarized in Sec.2.3 and described in detail in ngVLA Memo #71 [15]). The rationale is summarized below.

**Real time antenna based detection :** Extremely short-duration interference such as sparks and commercial UWB transmissions have a very broad-band footprint and must be masked at the highest possible time resolution it is practical to do so, especially if the signals repeat and have an on/off duty cycle shorter than 1 sec. If all such signals remain local to a few antennas and are transient, the impact of such RFI is likely to be low. However, RFI detection at this stage will still be useful to gather meta-data to inform subsequent processing stages. While it would be ideal to apply RFI mitigation on per-antenna voltage streams for some types of RFI detected at this stage, the act of dropping samples and applying arbitrary pre-correlation filtering will cause complications downstream that must be thoroughly understood and accounted for prior to using the data for target science.

**In-correlator high resolution flagging :** Options for outlier detection and flagging at  $\mu s$  to  $ms$  timescales and 10-50kHz channel resolutions are likely to reduce data losses due to satellite and aircraft communications from 40% (assuming the interferer is visible for 6hrs/day) down to a few % to match the true RFI filling fraction (a communication protocol with a  $90\mu s$  gap between  $10\mu s$  signal pulses will have a 10% RFI filling fraction when the interferer is visible). Visibilities from each baseline may be treated separately and viewed either as 1D time or frequency streams or as 2D views of time-frequency structure. This allows for massive parallelization along the baseline axis (more than 30,000 for 263 antennas). Such algorithms are amenable to implementation on FPGAs to accommodate these high data rates. *A system requirement shared by RFI mitigation as well as schemes such as baseline-based averaging is for flag/weight streams to be transported along with the data.* It is still unknown whether the RFI environment at these time/frequency resolutions offers enough usable free space in between transmissions to make the above investment worthwhile, and experiments must be conducted to properly quantify this prior to detailed design and costing.

**High time resolution modeling and subtraction :** Options for manipulating visibilities at  $ms$  and 10kHz time/freq resolution should be included in the system design, but may not critical to implement during initial construction. At these time/freq resolutions, phase coherence of the RFI signal will be preserved across the central dense ngVLA core, and RFI modeling and subtraction algorithms may be effective when there are few localized sources such as some satellites, aircraft containing 5G, or automobiles on Hwy 60 as they pass through the ngVLA core. *Subspace projection algorithms additionally require that autocorrelations be recorded along with cross-correlation visibilities.* If modeling and subtraction algorithms are effective, the RFI loss fractions in these commercial bands could potentially be reduced from 90% down to 50% or less depending on what fraction of the array does not suffer from decorrelation. However two major caveats are (a) LEO satellite internet bands are still likely to be inaccessible when transmissions are present simply because of the large number of interferers and (b) modeling and subtraction algorithms have not yet been proven robust enough for routine use, even for isolated emitters. The LEO satellite case is described in further detail below. Experiments and algorithm development on this topic are therefore essential

before detailed design and implementation. *However, the system design and infrastructure required for these algorithms is identical to that for the science case of real-time transient imaging and detection. Therefore, an investment in either one area will benefit and enable the other as well.*

**Post-processing flagging :** Post-processing outlier detection and pipeline heuristics will have to handle the rest of the RFI similar to current operations. Enhancements in post-processing could involve smart tuning of algorithms and feedback systems that integrate post-processing algorithms and infrastructure with real-time RFI mitigation algorithms. Further, some RFI may be too weak to be visible in individual high-resolution antenna or baseline streams and will require significant data averaging before the RFI appears above the detection threshold. Such RFI would require post-processing that interleaves RFI excision with image reconstruction (as is already done on occasion by running outlier detection and flagging algorithms on residual data prior to resuming imaging).

**RFI database and smart scheduling :** A-priori information about predictable RFI or real-time information about incoming RFI may be used in multiple ways through the data acquisition and analysis chain. For example, knowledge about satellite orbits could feed into a smart scheduler that observes at affected frequencies only when the satellite is not visible. Statistics gathered about RFI at high time resolution may be used to automatically tune outlier detection algorithms later in the signal chain. A reactive system may also be designed to alter observation parameters based on currently seen RFI (such as band shifting). ngVLA Memo #71 [15] contains more design and implementation details.

**The role of regulation :** If all predicted technologies that will generate RFI for the ngVLA become reality, it will be worth exploring options to (a) shield the ngVLA core and plains spiral antennas from 5G cell towers, (b) prevent system saturation and reduce the impact of LEO satellite RFI by a cooperative agreement to not point high power beams directly at ngVLA antennas and to confine the RFI to the allocated bands alone, and (c) possibly negotiate band sharing by working in between other transmissions, especially for specific spectral line observations which would otherwise be inaccessible. The concept of a National Radio Dynamic Zone is another interesting avenue to explore. *ngVLA Bands 3, 4 and 5 are particularly at risk to RFI scenarios that would be significantly abated by such measures.*

**LEO satellite constellations :** The recent proliferation of LEO satellite constellations could permanently alter our ability to perform radio astronomy observations in bands allocated for satellite internet. We expect that these transmissions will be continuously on when present with 10s of satellites visible to the array at once. Modeling and subtraction algorithms might be our primary mitigation option, but present significant challenges.

Even if subspace projection or imaging based algorithms work perfectly, the ideal implementation could require the equivalent of all-sky high-angular-resolution imaging with direction-dependent self-calibration, performed (in real time) at time resolutions of  $\mu s$  to  $ms$  and channel bandwidths of 10 kHz, and followed by algorithms to use multiple frames to separate true sky emission from satellite traces. The practical cost and compute complexity of such a system is a few orders of magnitude larger than the ngVLA data processing costs for Band 1 (which is itself considered an edge case from the point of view of compute feasibility, even at post-processing data resolutions).

The only feasible regime for such algorithms to operate is at a low angular resolution, using baselines formed only from the central core of the ngVLA plus the inner plains spiral antennas. This is the only regime where computing is likely to be tractable for the data rates at which phase coherence is preserved across these short baselines, and where all antennas are likely to see the same RFI environment (both key criteria for modeling and subtraction algorithms). RFI decorrelation on longer baselines might prove useful, at least at low dynamic range, but detailed experimental analyses of such situations must be carried out to determine feasibility and inform the mitigation strategy.

## 5 Future Work: Tests and Prototypes

The VLA is an ideal test-bed for many of the ideas presented in this report. Currently, only post-processing flagging solutions are in routine use and all estimates of the efficacy of real-time solutions are based on a theoretical analysis of the RFI environment and the expected numerical accuracy of various algorithms. Some ideas have been explored via other existing projects and telescopes and our analysis has extrapolated from those results. Prior to any detailed design decisions for the ngVLA, it is essential that each type of solution be very clearly demonstrated and prototyped. Below is a list of topics that require very specific research and development effort over the next few years in order to prepare for the ngVLA. A large fraction of these efforts will benefit current telescopes such as the EVLA as well.

### 5.1 Pre-correlation RFI Detection and Blanking/Filtering

Test existing solutions such as threshold-based-blanking within WIDAR and the VLA. When a sharp power increase is detected, data are dropped for a predetermined length of time (from both antennas in a baseline pair). Initial tests (ref. P.Demorest) have shown promise in eliminating very short duration pulsed RFI. In its simplest form, this may be usable for known emitters with predictable blanking intervals. Implement prototypes of spectral kurtosis filters and evaluate the results including what types of RFI are actually removed. A clear understanding of the effect of editing the data at this early stage must be part of any design decision. This includes either masking samples along the time or frequency axis or by replacing values with simulated noise. Existing work (ref. O.Ojeda) on simulated data show promise.

### 5.2 In-correlator Flagging at $\mu s$ and $ms$ Resolution

Evaluate the efficacy of existing automatic flagging algorithms (available in post-processing software) and other generic outlier detectors on visibility data viewed/recorded at high time and frequency resolution. Evaluate whether assumptions about RFI filling factors at this cadence are valid or not and if the algorithms are tunable for such RFI. An optimum balance of RFI excision efficiency versus compute load is achieved when the data are viewed at a time/frequency resolution matched to the duty cycle and bandwidth of the signal itself. It would therefore be instructive to learn if there is a preferred time/frequency resolution that work for most intermittent RFI or if the system design critically needs to support a wide range of averaging options between (say) the  $\mu s$  and  $s$  timescales. Extrapolations based on autoflag algorithms run on intermittent RFI at post-processing time/frequency resolutions suggest that if there are usable gaps at these high time/frequency resolutions, the algorithms will be able to operate accurately and be tunable (ref. U.Rau). The compute load and data rate at which such flagging is most efficient will also need evaluation before attempting an accurate sizing or design of compute infrastructure.

### 5.3 Real-time Subspace Projection Methods

Demonstrate success on real VLA data or with data from telescopes such as MeerKAT whose array configuration is a closer approximation of the ngVLA dense core. Work is currently ongoing on this topic with preliminary semi-positive results for the controlled situation of one or two satellites in the field of view. For the VLA, data recorded at 10 ms and 50 kHz resolution suffice to preserve phase coherence for baselines up to 3km (VLA C-config) at S-Band and algorithm tests have been initiated on observations of the Sirius XM satellites (ref. B.Jeffs at BYU, and J.W.Steeb at NRAO). Many questions remain, ranging from how details about the correlation and recording are translated to the algorithm formulation, how effective these methods may be in the presence of a large number of emitters, how critically they depend on a-priori knowledge of the positions of emitters and how accurate the resulting subtracted data are from the point of view of science-recovery.

## 5.4 Real-time Calibration, Imaging and Peeling

Leverage existing frameworks for fast-transient imaging (RealFast, Fast-VLITE, LWA/HERA with EPIC) that are already showing the ability to image RFI on low-resolution all-sky and full-beam maps. Demonstrate accurate imaging of the RFI as a (moving) point source and demonstrate RFI cancellation by visibility prediction and subtraction. This is a technique used in post-processing to remove the effect of unwanted bright sources far from the region of interest but which the instrument is still sensitive to. Algorithms to image and subtract stationary RFI imaged at the North Pole may also be worth evaluating. Algorithms for direction-dependent self-calibration are in use at various observatories in post-processing for the purpose of foreground source removal for Epoch of Reionization experiments and it may be of use to evaluate such algorithms for RFI excision as well.

## 5.5 Understanding the RFI Environment

Record measurements of voltage streams per antenna and high time and frequency resolution visibilities and examine them to determine the true nature of different types of RFI and to assess how prevalent each type is. For example, communication protocol specifications may show on/off duty cycles with clear gaps in time and frequency between transmissions from the same emitter, but it is unclear how much of this we may see in practice when there are multiple emitters dynamically sharing the same time/frequency allocation. Another example is that since post-correlation masking and excision is easier to analyse from the point of view of post-processing, do we see enough short-duration wide bandwidth interference to warrant effort on pre-correlation editing of the voltage streams? Also, if for example, most intermittent RFI turns out to have duty cycles comparable to 1sec, the case for real-time (in-correlator) flagging itself becomes weaker.

## 5.6 Scheduling Around Satellite Orbits

Use existing information about RFI patterns (temporal, spectral, spatial) to schedule observations, based on the data characteristics required for the target science. One obvious application is to create dynamic schedules to work around known satellite orbits. Recent experiments at the GMRT [12] have shown promise with such a system being evaluated for science observations.

## 5.7 RFI Attenuation Due to Signal Decorrelation

Theoretically analyse the expected attenuation due to decorrelation across ngVLA long baselines for ground-based and airborne emitters. Craft an observational experiment to evaluate this effect using the VLA and compare against theoretical predictions. A rough estimate present in this document suggests that about 10% of the total number of baselines may see a viable amount of attenuation (i.e. 40 dB for a target imaging dynamic range of  $10^4$ ) but this must be verified (or falsified) by actual measurements.

## 5.8 RFI Monitoring and Database Generation/Use

Construct a real-time RFI monitoring framework (with simple detection algorithms) and generate a database that could potentially be mined and used for smart scheduling as well as for automatic tuning of downstream algorithm choices. A prototype exists at the VLA (ref. A.Erickson) that generates a database consisting of RFI metadata. Further experiments would include coordinating an observation along with database generation and then exploring ways to use the gathered information to tune (say) post-processing autoflag algorithms or to generate explicit manual flag commands to be applied either in real-time or in post-processing. Another dimension of this effort is to apply machine learning algorithms to auto-classify types of RFI and detect predictability that could be then used for smart scheduling.

## References

- [1] R. Athreya. A new approach to mitigation of radio frequency interference in interferometric data. *The Astrophysical Journal*, Volume 696, Number 1, 2009.
- [2] C. Carilli, V. Rosero, A. Erickson, B. Mason, E. Murphy. ngVLA array configuration : Reference design description. *ngVLA Reference Design, Vol I*, 2019.
- [3] CASAdocs. Flagging visibilities and calibration tables. <https://casa.nrao.edu/casadocs/casa-5.6.0/calibration-and-visibility-data/data-examination-and-editing/flagging-measurement-sets-and-calibration-tables>, 2019.
- [4] W. D. Cotton. Removal of interference from low frequency VLA data. *Obit Memo Series, Memo #16*, 2009.
- [5] A. Erickson. ngVLA radio frequency interference forecast. *ngVLA Memo Series, Memo #48*, 2018.
- [6] E. Greisen. RFLAG. <http://www.aips.nrao.edu/cgi-bin/ZXHLP2.PL?RFLAG>, 2019.
- [7] J. W. Steeb, D. B. Davidson, S. J. Wijnholds. Mitigation of non-narrowband radio frequency interference incorporating array imperfections. *JAI Special Issue : Interference Mitigation Techniques in Radio Astronomy - Volume 08, Issue 01*, 2019.
- [8] M. C. Burnett, B. D. Jeffs, R. A. Black, K. F. Warnick. Subarray processing for projection-based RFI mitigation in radio astronomical interferometers. *ngVLA Memo Series, Memo #38*, 2018.
- [9] M. Lambert, B. D. Jeffs. Radio frequency interference cancellation, appraisal, detection, and correction. *ngVLA Memo Series, Memo #56*, 2019.
- [10] ngVLA Science Advisory Council. Key science goals for the next generation very large array (ngVLA). *ngVLA Memo Series, Memo #19*, 2017.
- [11] A. Offringa. AOFlagger. <https://ui.adsabs.harvard.edu/abs/2010ascl.soft100170/abstract>, 2010.
- [12] P. Raybole, S. Sureshkumar, S. Katore, S. Rai. Real time prediction, detection and co-existing with satellite interference at GMRT. *IEEE RFI Conference Proceedings*, 2016.
- [13] I. Portillo. A technical comparison of three low earth orbit satellite constellation systems to provide global broadband. <http://www.mit.edu/~portillo/files/Comparison-LEO-IAC-2018-slides.pdf>, 2018.
- [14] R. Perley, T. J. Cornwell. Removing RFI through astronomical image processing. *EVLA Memo Series, Memo #61*, 2013.
- [15] R. Selina, U. Rau, A. Erickson. RFI mitigation in the ngVLA system architecture. *ngVLA Memo Series, Memo #71*, 2019.