Multi-dimensional radio frequency interference framework for characterizing radio astronomy observatories

Isaac Sihlangu,a,b,* Nadeem Oozeer,a,c and Bruce A. Bassett,a,b,c,d

aSouth African Radio Astronomy Observatory, Cape Town, South Africa
bUniversity of Cape Town, Department of Maths and Applied Maths, Cape Town, South Africa
cAfrican Institute for Mathematical Sciences, Muizenberg, South Africa
dSouth African Astronomical Observatory, Cape Town, South Africa

Abstract. Radio frequency interference (RFI) has historically plagued radio astronomy, worsening with the rapid spread of electronics and increasing telescope sensitivity. We present a multi-dimensional probabilistic framework for characterizing the RFI environment around a radio astronomy site that uses data from the telescope observations themselves. We apply the framework to about 1500 h of commissioning data (∼200 TB) from the MeerKAT radio telescope; producing a six-dimensional array that yields both average RFI occupancy as well as distributions around the mean as a function of key variables (time of day, frequency, baseline, azimuth, and elevation). This allows for automated alerting for significant new sources of RFI and monitoring of RFI trends. Our results provide the first detailed view of the MeerKAT RFI environment at high sensitivity. As expected, in the MeerKAT L-band, we find that the major RFI contributors are from the global positioning system satellites, flight distance measurement equipment, and the global system for mobile communications. Beyond characterizing RFI environments, our approach allows observers access to the prior probability of RFI in any combination of tracked variables, allowing for more efficient observation planning and data flagging.

© 2021 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JATIS.8.1.011003]

Keywords: radio frequency interference; radio astronomy; global positioning system; distance measurement equipment; global system for mobile communication.

1 Introduction

Radio signals from astronomical sources are extremely weak and easily overwhelmed by human-made radio signals from sources, such as cellphones, satellites, aircraft, and telescope electronics. Any radio signal other than the desired astronomical signal is called an unwanted signal or spurious radiation. Astronomers classify these signals as radio frequency interference (RFI) and they are increasingly threatening radio observatories due to our increasingly technological society.1

The MeerKAT radio telescope, referred to as MeerKAT onward, is among the most sensitive L-band radio telescope of its kind and is observing the radio sky with unprecedented depth and detail.2 Building such an instrument comes at the cost of picking up very faint RFI sources. Known RFI sources dominate the MeerKAT L-band frequency range from mobile (GSM) communications, flight distance measurement equipment (DME) from aircraft, and global position system (GPS) satellites.

Although radio astronomy has been carried out for decades, we have not seen much framework that collects and characterizes the RFI environment from the telescope observation measurements. If available, such frameworks are internal to the observatory and are rarely accessible. On the other hand, radio astronomers generally flag the outliers from their science data without

*Address all correspondence to Isaac Sihlangu, isihlangu@sarao.ac.za

2329-4124/2021/$28.00 © 2021 SPIE
caring much about their causes. These RFI flags are discarded and rarely fed back to the observatory. We, therefore, propose a framework that can allow any radio astronomy site to perform the task of keeping track of these RFI from the massive amount of data that has been collected.

Our proposed framework investigates the RFI occupancy surrounding the MeerKAT site using a probabilistic approach. For each observation file, we produce the probability of RFI as a function of various parameters. The paper is divided as follows. Section 2 describes the MeerKAT in-house RFI detection methodologies. Section 3 gives the framework used for analyzing the RFI occupancy, including the algorithm design and statistical methods. Section 4 provides the results, and discussions followed by a conclusion in Sec. 5.

2 MeerKAT SDP RFI Flagger

MeerKAT is an interferometry array of 64 radio dishes that is among the most sensitive radio telescope of its kind and will observe the sky with unprecedented depth and detail.\(^2\) The radio signals captured by the MeerKAT antenna are converted to voltages at the receiver. The receiver further amplifies and filters these voltages, which in turn get digitized and sent to the correlator beamformer (CBF).\(^3\) The MeerKAT correlator implements the FX/B signal processing.\(^2\) The F-engine coarsely aligns the voltages, corrects for both geometrical and instrumental delays, and splits the data into frequency channels. The aligned voltages from pairs of antennas can then undergo various processes such as the correlation of the signal by the X-engine or beamforming of the signal by the B-engine.\(^4\) The raw visibilities are further piped into the ingest at the 0.5-s dump period, which then processes the data to produce the visibility data product (which is usually averaged to 8 s or according to the user’s observing parameters as \(L_0\)). A calibration pipeline is run on the \(L_0\) visibilities to produce the calibrated visibility that is called the \(L_1\) data product.

RFI detection in the MeerKAT Science Data Processing pipeline happens at two stages, that is once at the ingest step and then at the calibration. The high-time-resolution RFI detection occurs during the ingest step, where the strong RFI is detected by checking for outliers in the frequency axis in individual correlator dumps. At this stage, averaging of data is carried out. The samples detected as RFI are excised, and only unflagged samples are averaged in time as per the observation requirement and further used in the data processing pipeline. The output of the ingest step is, therefore, an averaged RFI excised data-set with pertinent meta-data stored in telstate. [Telstate is a Redis database from telescope state (telstate) that contains meta-data.]

To take into account the data loss due to the ingest excision, each visibility data point has an associated weight (\(W_{\text{SDP}}\)) that allows us to calculate how many samples were averaged to produce the visibility. If we define \(N\) as the number of correlator samples, \(V_{\text{CBF}}\) as the visibility sample from the CBF, \(U\) as the set of indices of unflagged correlator visibilities, we can then calculate the SDP visibility sample \(V_{\text{SDP}}\) using the following equation:

\[
V_{\text{SDP}} = W_{\text{SDP}} \sum_{i \in U} V_{\text{CBF}[i]},
\]

where \(i\) represents correlator/beamformer sample index and \(W_{\text{SDP}} = \frac{1}{N_U}\), with \(N_U\) being defined as the number of unflagged samples by the ingest. The ingest flags become “true” when all \(N\) samples are flagged as RFI and as a result there is no excision of data. With partial flagging or excision of data, the ingest flags becomes “false.” The ingest RFI detection algorithm usually only detect narrow regions around the brightest RFI spikes in the data, and further flagging is required.

On the other hand, during the calibration step, the MeerKAT in-house developed RFI flagger (https://github.com/ska-sa/katsdpsigproc) (hereafter called the SDP flagger) is used. The SDP flagger is based on the variation of the VarThreshold method used in the classic AOFlagger algorithm.\(^5\) The SDP flagger works on a quasi-real-time model. It runs on a two-dimensional data array of time and frequency. As already mentioned in Sec. 1, the MeerKAT L-band frequency is corrupted mainly by RFI from the GSM, DME, and GPS sources. Therefore, a static mask was developed to flag the data for those frequency ranges. Due to the RFI dependence on
the baseline length, the mask is applied on short baselines (≤1000 m). A baseline is defined as a vector joining any two antenna pairs. For \( N \) antennas, the total number of baselines is \( N(N - 1)/2 \). The mask is used to determine the statistics from the smooth background which is eventually used to detect the RFI in the full unmasked data. The unmasked visibility data are loaded into the SDP flagger per scan, where a scan is defined as a collection of SDP visibilities over a specific time period. For MeerKAT, a scan is on average between 5 and 15 min. Furthermore, the algorithm treats each baseline in each scan independently, allowing the parallelization of the algorithm along the baseline axis.

A smooth background fit is applied to the unmasked data by convolving it with a 2D Gaussian with widths larger than standard RFI spike widths in both time and frequency. The convolved widths are smaller than any variations in the bandpass or changes in amplitude with time. This process ensures that the smooth background ignores any potential RFI spikes in the data but follows the actual shape of the background. Data already flagged from the ingest or from the static RFI mask are given zero weight and therefore do not contribute to the background estimation.

The fitted smooth background is subsequently subtracted from the data, and the standard deviation is measured from the masked residual. This standard deviation is used as the basis for the spike/outlier detection threshold. First, the data are averaged in time over the whole scan and outliers in the resulting 1D frequency spectrum are located; this is to find faint spikes in the time axis that would otherwise be missed. RFI channels in the 1D spectrum are flagged for all times in the scan. Finally, the entire data are flagged in the time and frequency dimensions.

3 KATHPRFI Framework

RFI is a nuisance for astronomers as it corrupts much weaker astronomical radio signals. These unwanted signals usually are marked/flagged by masking out the regions around what the astronomer decides as RFI. This decision is very often very subjective. Furthermore, there is a poor feedback mechanism from astronomers to the observatory about those flagged RFI. This information is crucial for the observatory finding ways to improve RFI mitigation strategies.

There is no way to keep track of RFI or quantify the site RFI environment as measured by the MeerKAT telescope itself. During the commissioning and testing phases of the MeerKAT telescope array, a massive amount of data were captured and stored in the archive. We decided to carry out a statistical analysis of the RFI environment as measured by the MeerKAT telescope using the Karoo Array Telescope Historical Probability Radio Frequency Interference (KATHPRFI) framework.

One of the goals of the KATHPRFI framework is to provide MeerKAT users (astronomers, telescope operators, RFI engineers, or anyone interested in the RFI health of the observatory) with a tool that will aid them to keep track of changes in the RFI statistics over a long period as measured by the telescope. For astronomers, having a better understanding of the RFI environment is vital in preparing observation proposals and carrying out scientific analysis of their experiment. On the other hand, the RFI statistics can also allow us to build an intelligent observation scheduler and monitoring system for the RFI on-site. The proposed tool can thus help the telescope operation team to understand the RFI environment on-site and also aid them to detect system failures and any other telescope electronics issues.

3.1 KATHPRFI: Design

We chose an evolutionary prototyping model in our design instead of a throw-away approach. The evolutionary prototyping is a life cycle model wherein the concept of the system is developed as the project progresses. The evolutionary prototyping approach allows easy modification of the system in response to the user’s inputs. Our motivation behind choosing an evolutionary prototyping approach instead of the throw-away approach is due to the complicated nature of RFI signals, as it is difficult to frame the specifications of our system from the word go.

The KATHPRFI framework retrieves the MeerKAT visibility dataset from the archive followed by doing subselection of the data and finally returns the RFI flags that we use to build our
statistics. Using the flags the KATHPRFI framework then constructs a “master” and a “counter” array, which contains the descriptive statistics about the RFI from each visibility dataset. The master array contains the number of RFI points per voxel, whereas the counter array contains the total number of observations per voxel. Both arrays are built around the concept of a 5D array. The dimensions are; time of the day ($T$), frequency ($F$), baseline length ($B$), elevation (El), and azimuth (Az). The shape of the 5D data array is $[24 \times 4096 \times 2016 \times 8 \times 24]$.

### 3.2 KATHPRFI: Algorithm

The KATHPRFI script starts by initializing the master and the counter arrays, as depicted by the algorithm in Fig. 1. Block 1 of the algorithm reads in the RDB file followed by the preprocessing stage whereby we remove any bad antennas from the data. An antenna is flagged as a bad if, during an observation, it fails for some reasons. The katdal (https://github.com/ska-sa/katdal) library is used to get the information about the antenna activities during the observation. If we find a STOP state on the activity lists of an antenna during an observing track, we would flag that specific antenna as a bad and remove it from the analysis.

In the second block, we choose the parameters of interest to produce a subset of the data. The katdal library allows us to do the selection. The cross-correlation product and the horizontal polarization product (HH) were chosen for the purpose of demonstration of our framework. However, our framework allows any possible combination depending on the user’s goals. For example, if one chooses the autocorrelation, some slight modification in the array will need to be done, since the baseline array size will be much less (64 instead of 2016).

Due to computational limitation, the data are binned. This is carried out in the final step of block 2. This causes the full resolution of certain attributes, such as the time of observation, AZ, and elevation, not to be stored. We have managed to maintain the full resolution of the frequency and the baseline axis. Time is hence binned per hour into 24 h of a day. The elevation is binned into 8 deg intervals and the AZ is binned in 15 deg intervals. Any mention of Az, elevation, and time hereafter will mean the respective bin value and not the actual value.

Blocks 3 to 5 are nested loops that go over timestamps, frequency channels, and baselines, respectively. This is done to update the master and the counter array based on the indices extracted from a particular observation file in question. If for example, an observation started...
at 10:07 and ran until 10:11, the KATHPRFI script will put all of the data for that time in the 10th h bin. It will also check which antennas were present during the observation and update the baseline array accordingly.

3.3 Resources and Limitations

The Python\textsuperscript{7} programming language and various python packages such as Numpy, Pandas, Dask, Numba, and Xarray have been used to build the KATHPRFI framework. The Xarray has a Zarr\textsuperscript{8} back-end that can be used to write and read compressed datasets to disk. Here the Zarr library is used to interface with Dask to support parallel reading and writing of data from disk.

Computer random access memory (RAM) was the greatest limitation of our framework because we could only store a limited amount of data in a finite amount of RAM. Thus RAM has influenced the resolution of our master and counter arrays. The data from these arrays were stored onto disk, and data access is known to be lazy. Each element, in our array, is a 16-bit unassigned integer which takes 8 bytes in memory. Taking these into consideration, the dimension of the array and the size of the data type, we computed the required RAM as follows:

\[
\text{required RAM per array} = \text{dimension} \times \text{data type size} = 24 \times 4096 \times 2016 \times 8 \times 24 \times 8 \text{ bytes} = 7.610 \times 10^{10} \text{ bytes.}
\]

The base two binary system\textsuperscript{9} (i.e., 1 megabyte = \(2^{20}\) bytes) is used as the unit of data measurement. Hence, as a result, the total memory required to store the array is

\[
\text{required RAM per array} = 7.610 \times 10^{10} \text{ bytes} \times \frac{1 \text{ megabyte}}{2^{20} \text{ bytes}} = 283.5 \text{ gigabytes (GB).}
\]

Thus for the two arrays, the total RAM required is 567 GB. As previously mentioned, the Zarr library is used to write and read the data from the disk. This library gives us an optimal compression of the Numpy array before they are stored onto disk by a factor of 100. The final Zarr file only takes 3 GB of disk space.

3.4 Statistical Methods to Calculate the Averages

Randomly chosen imaging observations were used to create the data set that we will call historical probability data release one. The observation dataset used is equivalent to 1500 h of observing time (~200 TB) which was collected from May 2018 to December 2018. These observations were carried out in the L-band, containing 4096 channels with a 208.984-kHz channel width. As mentioned previously, the datasets are run through the in-house RFI detection algorithm and the respective flag tables were created.

In order to calculate the probabilities, we adopt the following approach. Suppose that \(\alpha\) is the number of RFI samples as obtained from the master array and \(\beta\) is the number of non-RFI samples (i.e., total of counter array—total of master array), where counter array is the total number of observed samples. Then we can compute the probability estimate \(P(\text{RFI})\) in a voxel as follows:

\[
P(\text{RFI}|T, F, B, El, Az) = \frac{\alpha_{T,F,B,El,Az}}{\alpha_{T,F,B,El,Az} + \beta_{T,F,B,El,Az}}, \tag{2}
\]

where \(T, F, B, \text{El}, \text{Az}\) are the indices of time of the day, frequency, baseline length, elevation, and Az in a given voxel, respectively. In order for us to compute the probability of RFI for a given dimension, we need to marginalize over all other dimensions. For instance, if we want to compute the probability of observing RFI as a function of the frequency, we sum both master and counter array in all other axes except the frequency axis, and then we divide one by the other, and the resulting array will be the probability of observing RFI as a function of frequency. Mathematically, it can be written as
\[ P(\text{RFI}|F) = \frac{\sum_{T,B,\text{El},A} (\alpha_F)}{\sum_{T,B,\text{El},A} (\alpha_F + \beta_F)}. \]  

(3)

In order to calculate the average, we used two different methods.

- The first method was to update the master and counter array every time we get a new observation. At the end using Eq. (3), we computed the average RFI probability, effectively combining all observations into a single long observation. This is referred to as the combine average (CA).
- The other method consists of computing the average of each individual file using Eq. (3) and finally computing the average of those probabilities. We call this method the average of average (AoA).

These two averages will coincide if all the files have the same length or if the average of each file is the same, but in general the CA and AoA will differ. The AoA method is preferred, because it allows us to compute the RFI probabilities of each file and as a result we can also compute the errors associated with each voxel in our 5D array. Moreover, one requires sufficient computing resources to use the AoA method. On the contrary, the CA method, which requires less computing resources, will yield only one file, and hence statistical errors cannot be computed.

4 Results and Discussion

Once the framework set up and ran on all the datasets, we can now start looking into our findings. We would like to point out that our framework only looks at the outliers as detected by the algorithm. So any outlier that is above a certain threshold in the data will be flagged as an RFI, including sharp emissions lines such as HI. This work does not cover classification, grouping, and detailed analysis of known RFI sources.

4.1 RFI Occupancy Versus Time of the Day and Frequency

The overall RFI probability distribution picked up by MeerKAT in the HH polarization as a function of time of the day in coordinated universal time (UTC), and frequency in megahertz (MHz) is shown in Fig. 2. Our results show a clear pattern between the hour of the day and the RFI probability. We see a drop of RFI probability during the night time (i.e., 18:00 to 04:00 UTC) compared to the day time (i.e., 05:00 to 17:00 UTC). A maximum variation of 4% is observed between hours of the day in the RFI occupancy with an average of 23%. These results confirmed that the RFI probability is high during the day time compared to the night time. Our analysis, therefore, allowed us to validate some of the claims and hypothesis using MeerKAT commissioning imaging observations.

We noticed that at 05:00 UTC (corresponding to 07:00 South African Standard Time), the RFI occupancy goes up. This may be related to when activities begin in the nearby towns and cities and even on-site. At 10:00 UTC, we see a drop in the RFI occupancy, and similarly, at 14:00 UTC, we see another dip. These two times correspond to lunchtime and the end of the working day in South Africa, respectively. We cannot conclusively say that these human activities cause the observed increase in RFI occupancy; however, a correlation exists. One can also argue that there will be a dependency of RFI with time due to global navigation satellite system (GNSS), DME from aircraft, and other sources. These hypotheses will be covered with detailed analysis in a future work.

We also found that the RFI probability at the following frequencies: 1018, 1031, 1041, 1090, and 1103 MHz increases during the day and drops at night time. These frequencies are confined within the DME band allocated to the aircraft communication system. Therefore, these findings suggest that the observed increase in RFI probability during the day is most probably due to the aircraft passing over a region of the site.

Furthermore, there is a great deal of variation of RFI occupancy as a function of frequency at some frequency bands (e.g., 900 to 960 MHz) where we see 100% RFI whereas, at others
The RFI occupancy is down to <10%. We can see the three prominent frequency bands showing the highest probability of RFI in the MeerKAT site. Those are the GSM (900 to 960 MHz), aircraft transponders (1000 to 1200 MHz), and GPS satellites (1482 to 1600 MHz, and 1169 to 1280 MHz). From our analysis of ∼36.6% of the band at all the time, all the baseline is permanently flagged as RFI.

This paper is primarily interested in the RFI from known persistent sources such as GPS satellites, DMEs, and GSM. Emission from these sources is relatively constant, predictable, and regular. As a result, the variation in RFI probability from such sources is expected to be considerably small. To understand whether the observed fluctuations are statistically significant or are due to noise fluctuations, we computed the 68 percentile, which corresponds to the 1-sigma confidence interval for a Gaussian distribution. On the other hand, we suspect that the 95% confidence interval will include all sorts of outliers that may or may not be due to the radio signals. As a result, we found that the 68% confidence limits capture the RFI variability for the GPS satellites, DMEs, and GSM signals more accurately.

**4.2 Average RFI as a Function of Time of the Day**

We further investigated the statistical consistency of the RFI probabilities. Figure 3 shows the average RFI probability as a function of the time of the day, with the green region representing the 68% confidence interval. We used two methods to calculate the average as explained in Sec. 3.4. The blue line represents the CA, whereas the orange line represents the AoA. We observe a similar distribution of RFI from both methods.

As already mentioned in Sec. 4.1 at 10:00 UTC, we observed a drop in the RFI occupancy. We also found that the data are noisy for this time of the day, as shown by the 68% confidence interval. To understand this noisiness, we looked at the distribution of the RFI probabilities at a noisier and quieter hour of the day (Fig. 4). The huge variation of RFI probability that is observed could be explained by the long tail in the distribution of the RFI probabilities as shown in Fig. 4(a).
The considerable variation of RFI probability in noisier hours indicates some form of anomaly, which could result from several issues. Indeed, we found that some of the observation had zero probabilities, implying that no RFI was detected on any baseline and at any frequency by the algorithm. This is something impossible because of the permanent presence of RFI sources. Such non-detection of RFI may indicate a potential system problem. For instance if the correlator output zero value in the visibilities, the SDP flagger will not detect any RFI. This has a potential to bias our results and to avoid such biases, one need to build a well-defined data selection model that will reject all the observations that are not usable. One can check the visibilities of each observation and then select the flag. However, this will add extra processing and load to the framework.

4.3 Average RFI Occupancy as a Function of Frequency

We performed a similar analysis on the frequency axis; however, we decided to split the frequency spectrum into a known corrupted band and a clean band. We defined the corrupted band as the range of frequencies, in which the primary known RFI sources (GSM, DME, and GPS satellites) emit, whereas the clean band is the less corrupted part of the spectrum. We further

![Fig. 3 MeerKAT RFI occupancy as function of time of the day. The green region represents the 68% confidence interval computed over the historical observations. The blue line is the CA and the orange line is the AoA discussed in the text.](image1)

![Fig. 4 The histogram of RFI probabilities with a kernel density estimate (KDE) fit. The RFI probability distribution at (a) a noisy hour (10th h) and (b) quieter time of the day (7th h). The distribution of the probabilities in (a) has a long tail that is indicating some form of an anomaly. The RFI behavior is well defined by the average probability.](image2)
subdivided the clean band into lower and upper frequencies between 980 to 1070 MHz and 1310 to 1500 MHz. This subdivision was carried out by inspecting the RFI contribution as a function of frequency (Fig. 2).

Figure 5 shows the RFI averages with the blue and the orange line is computed from the CA and AoA method, respectively. Meanwhile, the green region represents the 68% confidence interval. We noticed a slight variation in the RFI occupancy in the corrupted band as shown by the 68% confidence interval limits, which are tightly constrained around the mean. However, we find frequencies for the lower and the clean upper bands (e.g., 1030, 1040, 1381, 1390, and 1492 MHz), in which the RFI occupancy is >10%, and spikes depict these in those regions. We observe a relatively high variation in RFI occupancy at these particular frequencies when looking at the 68% confidence interval.

We, therefore, looked at the distribution of probabilities of some of the clean band frequencies (Fig. 6). We expected the distribution of the RFI probabilities in the clean band to be close to zero, as there should not be any contamination. However, we see an extended tail distribution toward higher values of RFI probability. This long tail is a result of rare events appearing much more frequently than we expected. For example, the 1380-MHz L3 GPS band used for detecting

![Fig. 5](image)

**Fig. 5** The distribution of RFI probability as a function of frequency. The green region represents the 68% confidence interval. The blue line is the CA and the orange line is the AoA, the difference is explained in the last paragraph of Sec. 3.1. The average behavior of the known RFI source is constant and is well captured by the 68% confidence interval.

![Fig. 6](image)

**Fig. 6** Example histograms of average RFI probability distribution of some contaminated frequencies (a) from the clean band and (b) along with smooth KDE fits. The distribution of RFI in the clean band is mostly skewed towards zero probability, as expected since this band is supposed to be free of RFI. However, we do see some outliers at higher probability values.
nuclear activity on Earth seems to have been more active. The two frequencies shown are confined within the GPS L3 band.

4.4 Telescope Pointing Directions

Furthermore, we looked at how the RFI occupancy changes as the telescope points in various directions in the sky. The amount of RFI it measures is expected to change depending on the number of radio frequency transmitters in the field of view. It is anticipated that RFI due to terrestrial sources should be more dominant at low elevation. Figure 7 was used to examine this possibility. We noticed that between 20 deg and 50 deg elevation bin, the RFI probability is the highest, and it gradually drops as we go to higher elevations on both the CA and AoA methods. These results can explain that indeed at low elevation, we see more RFI than high elevations.

Likewise, we computed the 68% confidence interval for the elevation axis and the Az axis. We found that the 68% confidence interval limits on the elevation are tightly constrained around the mean; hence a small variation in RFI probability is observed. As for the Az plot, we found that some of the direction bins (30 deg and 140 deg) are too noisy. The observed variations are most probable due to an increased number of non-RFI outliers coupled with extra RFI from known towns in those regions.

The polar plot in Fig. 8 shows how much RFI is generated in the clean band as a function of Az (radial direction) and elevation (theta direction). MeerKAT uses a lower limit in the elevation, and we have chosen 20 deg as our lower limit (since most observations had this common lower limit). The white empty areas (Az bins: 225 deg to 240 deg and 345 deg to 360 deg) indicate a lack of data at these angles in our analysis.

The color scale ranging from purple through blue to yellow represents the probability of RFI occupancy, with yellow denoting the highest probability while purple is representing the lowest probability of RFI.

We noticed a hot-spot (maximum RFI occupancy) at lower elevations and Az bin of 135 deg that coincidentally points toward the town Beaufort West which is nearby the MeerKAT site. In addition, the RFI occupancy is relatively moderate across the Az angles at lower elevations. Looking at higher elevations (elevation bins >40 deg), the average RFI occupancy is about 2%. We require further investigations to confirm the sources of these RFI.

4.5 Baseline Length

Finally, we investigated the probability of RFI as a function of baseline length (Fig. 9). The baseline length that is referred to here is the physical distance between the antennas and not its $(u - v)$ projection in the sky. The blue and orange dots are the average RFI probabilities
from the two different averaging methods: CA and AoA. Meanwhile, the green region represents the 68% confidence interval. We notice that the RFI probability decreases as a baseline length from both methods.

To explain the observed decrease in RFI probability as a function of the baseline length, consider the complex visibility:

\[
V(u, v) = \int I(l, m) e^{-2\pi i (ul + vm)} dl dm. \tag{4}
\]

A single source is produced by multiplying the sky \(I(l, m)\) with the fringe pattern produced by the baseline integrated over a solid angle. The angular distance between two consecutive

---

**Fig. 8** RFI occupancy as a function of the telescope pointing direction for the clean band. We can notice a hot spot at low elevation and Az bin of 135 deg which is pointing toward nearby towns.

**Fig. 9** RFI occupancy for the full MeerKAT telescope in the L-band as a function of Baseline length (m). The blue and the orange dots are the mean RFI probability from the CA and AoA averages, respectively, whereas the green region represents the 68% confidence interval. The decrease of the RFI probability with increase of baseline length is due to moving RFI sources with respect to the static sky, which causes the phase of the RFI to oscillate rapidly on long baselines compared to short baselines which then tend to progressively average out on longer baselines when the visibilities are averaged over typical timescales (0.5 to 8 s for MeerKAT).
5 Conclusions

Radio astronomers usually flag RFI and outliers from their data without caring much about the origin and source of the contamination. On the other hand, radio observatories are very interested in accurately characterizing and understanding the RFI environment around the observatory to ensure the best quality of data possible from the telescope.

We have presented a framework that allows the capture of RFI as measured from the telescope itself. This framework also provides a multi-dimensional statistical view of the RFI environment. Using around 1500 h of MeerKAT telescope array data as a demonstration, we produce the RFI and outlier occupation probabilities over several months as a function of time of the day ($T$), frequency channels ($F$), baseline length ($B$), elevation (El), and Az. Overall, we can say that the RFI environment is dynamic from these preliminary findings. The clean band is supposed to be as clean as possible from RFI. Still, collectively, our results show evidence of ongoing activities that are worth investigating in the future.

The KATHPRFI framework can be applied to any archival data from observatories to understand the evolution and nature of the RFI. Beyond its use for alerting to new sources of RFI and understanding trends in the RFI environment, our results can provide useful prior probabilities for RFI flagging.

Acknowledgments

We would like to thank Christopher Finlay, Tom Mauch, the SARAO data science team, and the MeerKAT RFI working group for discussions and support during this project. The MeerKAT telescope is operated by the South African Radio Astronomy Observatory, which is a facility of the National Research Foundation, an agency of the Department of Science and Innovation. This research has been conducted using resources provided by the United Kingdom Science and Technology Facilities Council through the Newton Fund and the South African Radio Astronomy Observatory.

References

1. P. Fridman and W. Baan, “RFI mitigation methods in radio astronomy,” Astron. Astrophys. 378(1), 327–344 (2001).
2. F. Camilo et al., “Revival of the magnetar PSR j1622-4950: observations with Meerkat, Parkes, XMM-Newton, swift, Chandra, and Nustar,” Astrophys. J. 856(2), 180 (2018).
3. K. Asad et al., “Primary beam effects of radio astronomy antennas: II. Modelling the Meerkat l-band beam,” MNRAS 502(2), 2970–2983 (2019).
4. T. Mauch et al., “The 1.28 GHz Meerkat deep2 image,” Astrophys. J. 888(2), 61 (2020).
5. A. Offringa et al., “Post-correlation radio frequency interference classification methods,” Mon. Not. R. Astron. Soc. 405(1), 155–167 (2010).
6. R. A. Carter et al., “Evolving beyond requirements creep: a risk-based evolutionary prototyping model,” in Proc. Fifth IEEE Int. Symp. Req. Eng., IEEE, pp. 94–101 (2001).
7. Python Software Foundation Developers, “Python,” https://www.python.org/.
8. Zarr Developers, “Zarr,” https://zarr.readthedocs.io/en/stable/api/storage.html.
9. GbmbOrg Developers, “Bytes to MB Conversion,” https://www.gbmb.org/bytes-to-mb.
10. A. Thompson, “The response of a radio-astronomy synthesis array to interfering signals,” *IEEE Trans. Antennas Propag.* 30(3), 450–456 (1982).

**Isaac Sihlangu** received his bachelor’s degree in physics, his honors degree in astrophysics, and his master’s degree specializing in astronomy and space science from the University of Cape Town. He is a junior science software developer at the South African Radio Astronomy Observatory (SARAO) under the Square Kilometre Array Africa (SKA-Africa) project. His current research interests involve mining of the existing enormous amount of MeerKAT observations for transient variables using shallow machine learning methods and deep learning.

**Nadeem Oozeer** received his PhD from the University of Mauritius in 2008. Formerly, he worked as a senior commissioning scientist at the SARAO under the SKA-Africa project before moving to a data scientist position. He is also a visiting researcher at African Institute for Mathematical Sciences South Africa. His interests include big data analytics, machine learning, and artificial intelligence.

**Bruce A. Bassett** is a professor of applied mathematics at the University of Cape Town, a head of cosmology at the African Institute for Mathematical Sciences, and a research astronomer at the South African Astronomical Observatory. His interests include cosmology, applied AI, and Bayesian inference.