Identifying RFI in voltage data from the Giant Metrewave Radio Telescope

Group 3, Radio Astronomy Winter School 2021

Abstract

Radio telescopes record the signal from the astronomical source. These signals are very weak, and large dishes or an array of antennas are used to increase the collecting area. Radio Frequency Interference (RFI) is man-made electromagnetic radiation, and often while recording the signal from an astronomical source, RFI (Radio Frequency Interference) gets inevitably recorded. RFI causes the distortion in the astronomical data, and hence it is needed to be identified and removed from data during the processing, which can be done using the differences in the statistical properties of RFI and astronomical signals. In this experiment, we were provided with data from antenna C11 and C12 of Giant Metrewave Radio Telescope (GMRT), observing at band-3 and band-5. We try to identify the channels containing RFI using the power spectrum.

Introduction

Radio telescopes are the receiver systems that collect and process electromagnetic waves in the frequency range of a few MHz to a few hundred GHz from celestial sources. Giant Metrewave Radio Telescope (GMRT) is an array of 40 parabolic dishes arranged roughly in 'Y' shape. The central group has 12 antennas with a short baseline. The rest are distributed along the three arms. Each dish had a diameter of 45m.

Radio Frequency Interference or RFI is the man-made signal that gets inevitably recorded with the astronomical signal, causing distortions. Hence it is desirable to remove this RFI during data processing. To do this, we use the differences in the statistical properties of the RFI and the astronomical signal.

RFI can be categorized as broadband and narrowband RFI. In broadband RFI, the interference is spread across the band, whereas in narrowband, it is confined to a particular part of the band. Sources of broadband RFI can be sparking on high-power transmission lines and distribution equipment (e. g., transformers), automobile sparking, and switching of inductive load. A few examples of sources of narrowband RFI are broadcasting transmitters, satellites, air traffic control systems, and radar.

To identify RFI, we first need power spectrums. The voltage-time series data can be analyzed in the frequency domain by using the Fourier transforms. Here we first divide the data into bins of, say size n, and Fourier transform each bin. If F() is the Fourier transform of the f(t), then FF* (F* is the complex conjugate of F) is the power spectrum.

The inverse Fourier transform of the power spectrum is the auto-correlation spectrum. To obtain the cross-correlation function, consider F(k) and G(k) as the Fourier transform of the f(t) and g(t), respectively. The Fourier transform H(k) of the cross-correlation function h(t) is given by:

$$H(k) = F(k)\, G(k)^*$$

This is referred to as the cross-correlation spectrum.

To identify the RFI, we use statistical properties. We know that the voltage data from the celestial source has a Gaussian distribution, whereas the power has exponential distribution. For exponential distribution, mean equals rms, and any deviation from this results from the non-celestial source. We can take the ratio of the expected rms, which will be proportional to the mean divided by the $\Delta\nu\tau$ and the observed rms, and use this distribution to locate the frequencies in the spectrum affected by RFI.

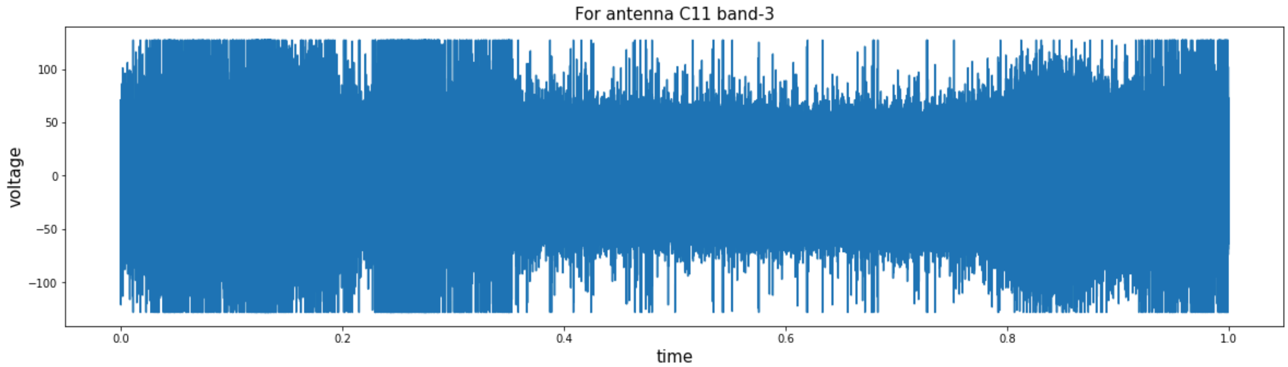
I. Data

We were given four data files corresponding to the voltage recorded by antennas C11 and C12 observed at band 3 (300 - 500 MHz) and band 5 (1050 - 1450 MHz) at each. Each of the data files has 4194304 voltage values. The data contains the voltage-time series recorded simultaneously. Each number is the digitized voltage recorded, and the consecutive numbers are recorded at a time interval of 2.5×10^{-9} s (1/400MHz).

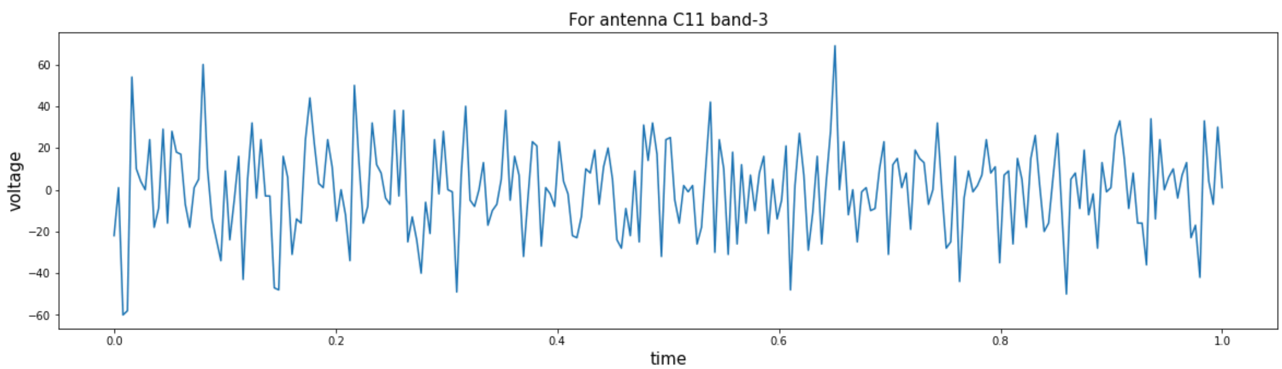
II. Data Analysis steps and plots

1. Properties of voltage-time series

To visualize the voltage-time series, we plot values in the data files with time for two antennas. Fig. 1 shows the plots for (a) the entire voltage-time series and (b) the first 250 voltage reading with arbitrary units for the antenna C11 observed in band-3.



(a)



(b)

Fig. 1: Voltage vs time plots for antenna C11 band-3 in arbitrary units. (a) entire Voltage-time series, (b) Voltage time series for the first 250 values in the file.

The histogram of the voltages in the data files was plotted to check the distribution properties of the voltage-time series data. It is found that the Gaussian curve can be fitted to this distribution by finding the mean and standard deviation. The histograms and fitted Gaussian curve with the parameter values are shown in Fig. 2

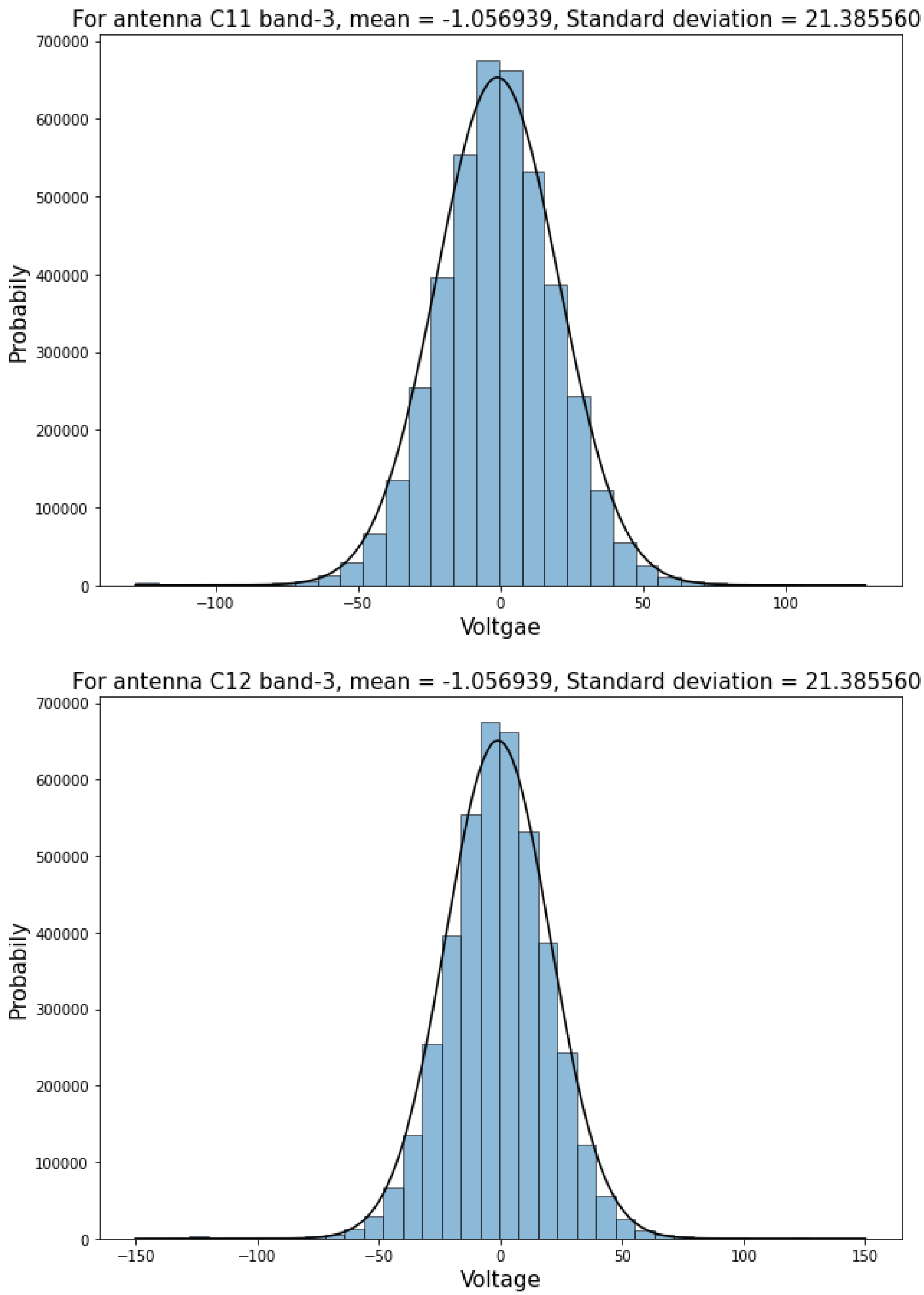
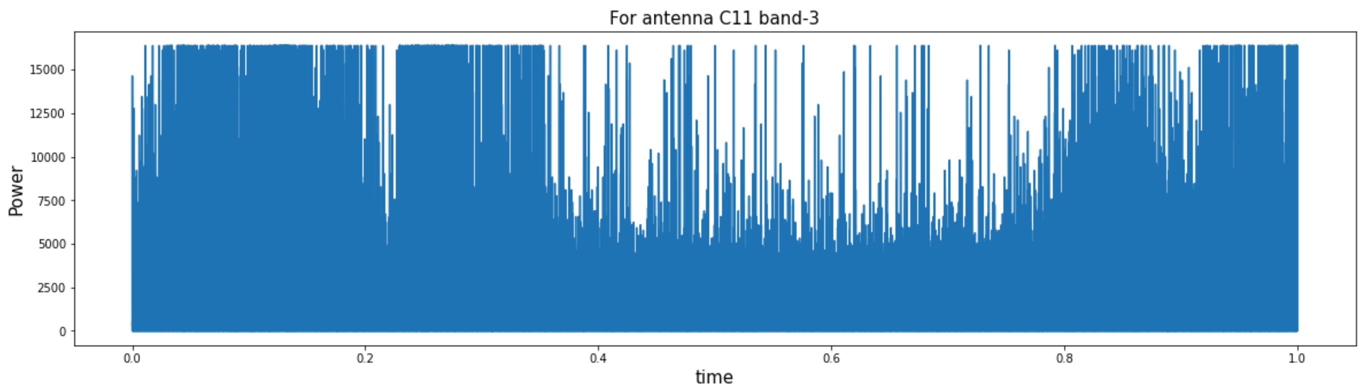


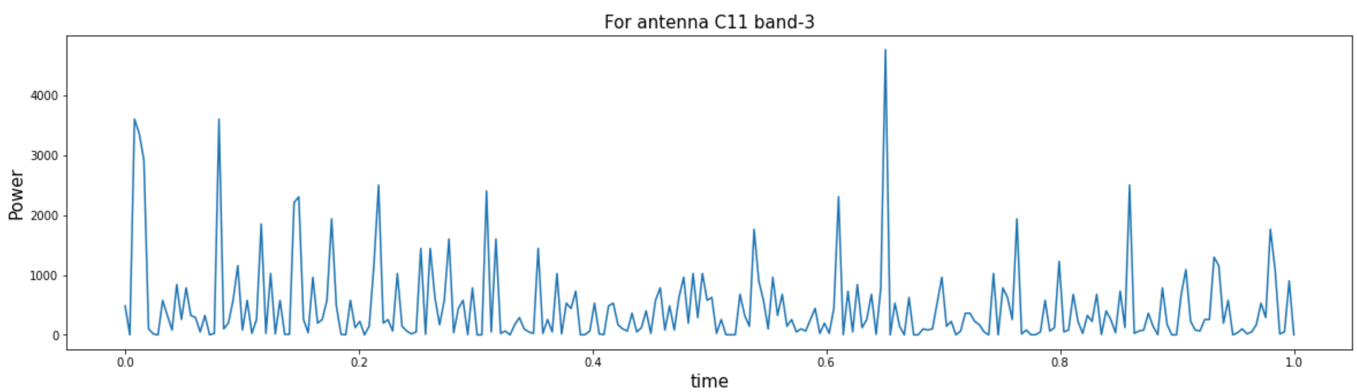
Fig. 2: Voltage distribution and fitted gaussian with the parameter indicated in the title of the plot.

2. Properties of the power

The power can be calculated by simply squaring the voltage values. To visualize the power-time series, this power was plotted against the time for the first 4096 values for showing the power variability (Fig.3)



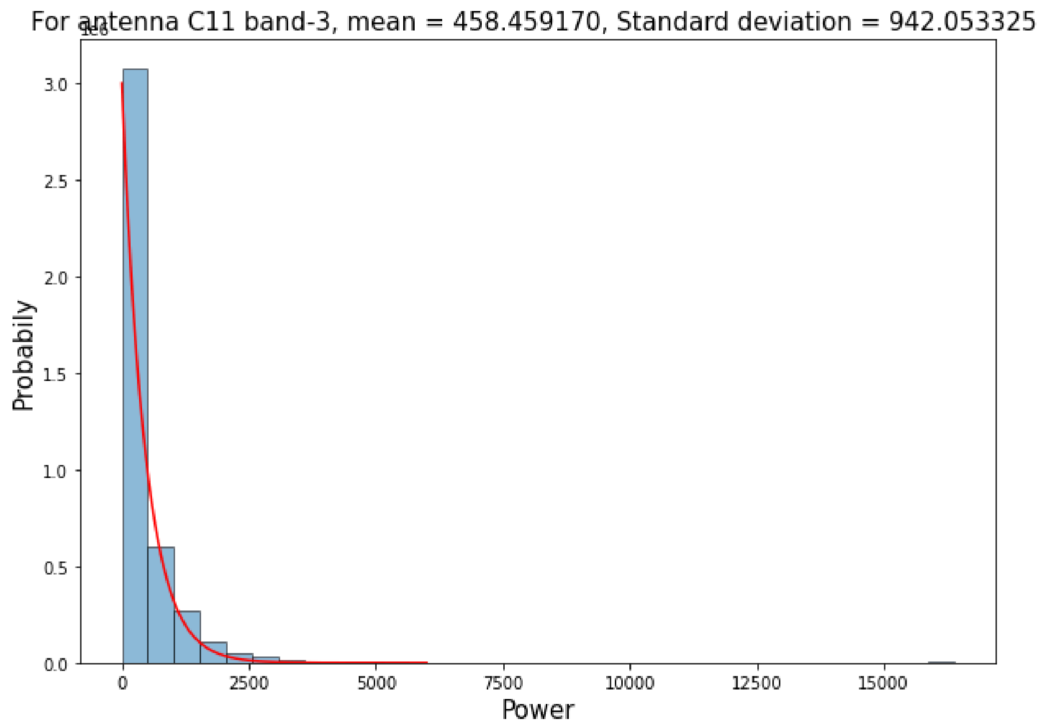
(a)



(b)

Fig. 3: Power-time series (a) for entire data, (b) for first 250 values in the data file.

Fig. 4 shows the distribution of power values. In this, the power is plotted on a histogram. An exponential function is fitted onto the data, and parameter values are shown in the title of the respective plot.



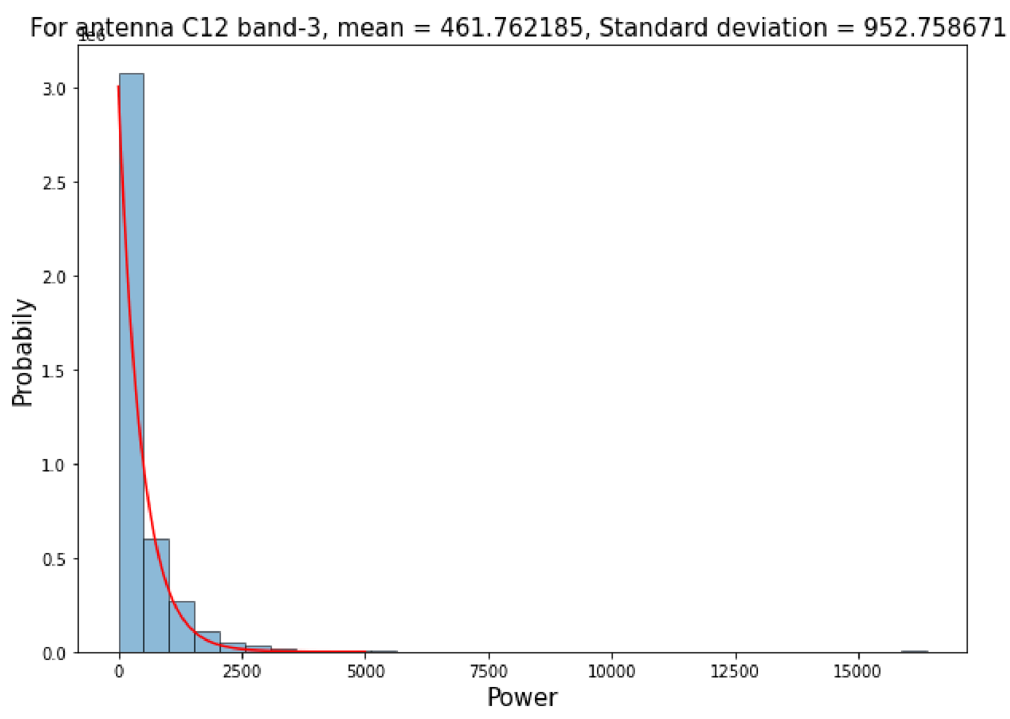


Fig. 4: Histograms and fitted chi-squared distribution function for powers (voltage squared $|V|^2$)

A similar analysis was repeated for the band-5 data from antennas C11 and C12.

3. Power spectrum and Dynamic Spectrum

To find the power spectrum for the given data, we need first to calculate the Fourier transform of the data to get the voltage values in the frequency domain. To do this, we divide the data into different bins of size 4096. The number 4069 = 2^{12} is chosen so that the fast Fourier transform requires less time for computation (An N-point FFT works considerably faster if the number of data points (N) is a power of 2). The total number of voltage values in each of the given data files is 4194304. Dividing this into bins of size 4096 results in 1024 bins. The Fourier transform was calculated for each of these frequency bins using **fft** function from **scipy module** in python. fft function returns 4096 values (half positive and half negative) in the frequency domain. However, due to the hermeticity of the data given, we know that the negative frequencies obtained carry no new information. Hence, we can discard the negative half of the Fourier transformed data. This results in 2048 frequency channels for each of the 1024 bins.

(a) For auto-correlation spectrum:

To obtain the power spectrum for an observing band of a given antenna, we take the absolute value of multiplication of the fft of bins and its complex conjugate. Plotting the power over each bin gives a 1024 power spectrum. We also average this 1024 power spectrum to get the auto-correlation mean spectrum by plotting averaged values with respect to the frequency channel (2048 frequencies). Fig. 5 shows the Auto-correlation mean spectrums for antenna C11 and C12 in band-3.

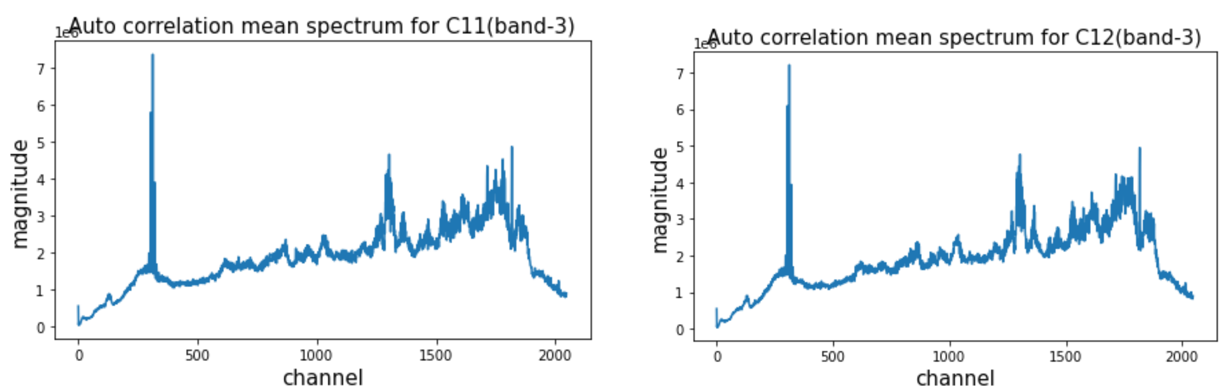


Fig. 5: Auto-correlation mean spectrum for C11 and C12 (band-3)

We can also plot 1024 bins across 2048 frequency channels (channel vs. time plot) with the intensity of spectrum indicated by the color map (Fig. 6). This is the dynamic spectrum for auto-correlation for the individual antenna at band-3.

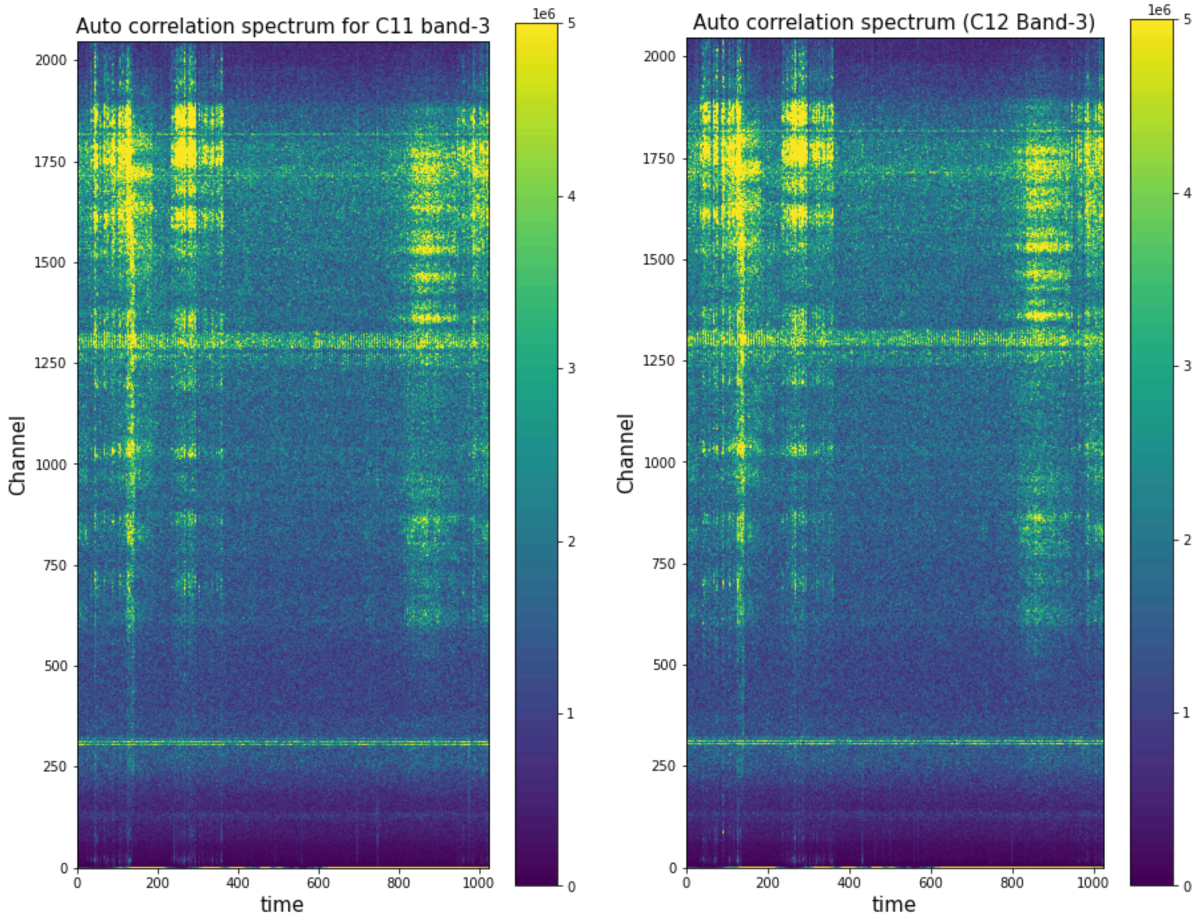


Fig. 6: Dynamic spectrum for antenna C11 and C12 at band-3.

(b) For cross-correlation spectrum

To find the cross-correlation power spectrum of antennas C11 and C12 for band-3, we take the absolute value of multiplication of Fourier transformed data of C11, and the complex conjugate of the Fourier transformed data of C12. Averaging these values over 1024 different spectra and plotting the averaged result across the frequency channels gives cross-correlation mean spectra for antennas C11 and C12 (band-3). We also plot the dynamic spectrum (Frequency vs. time plot with colored intensities) for cross-correlation. (Fig. 7)

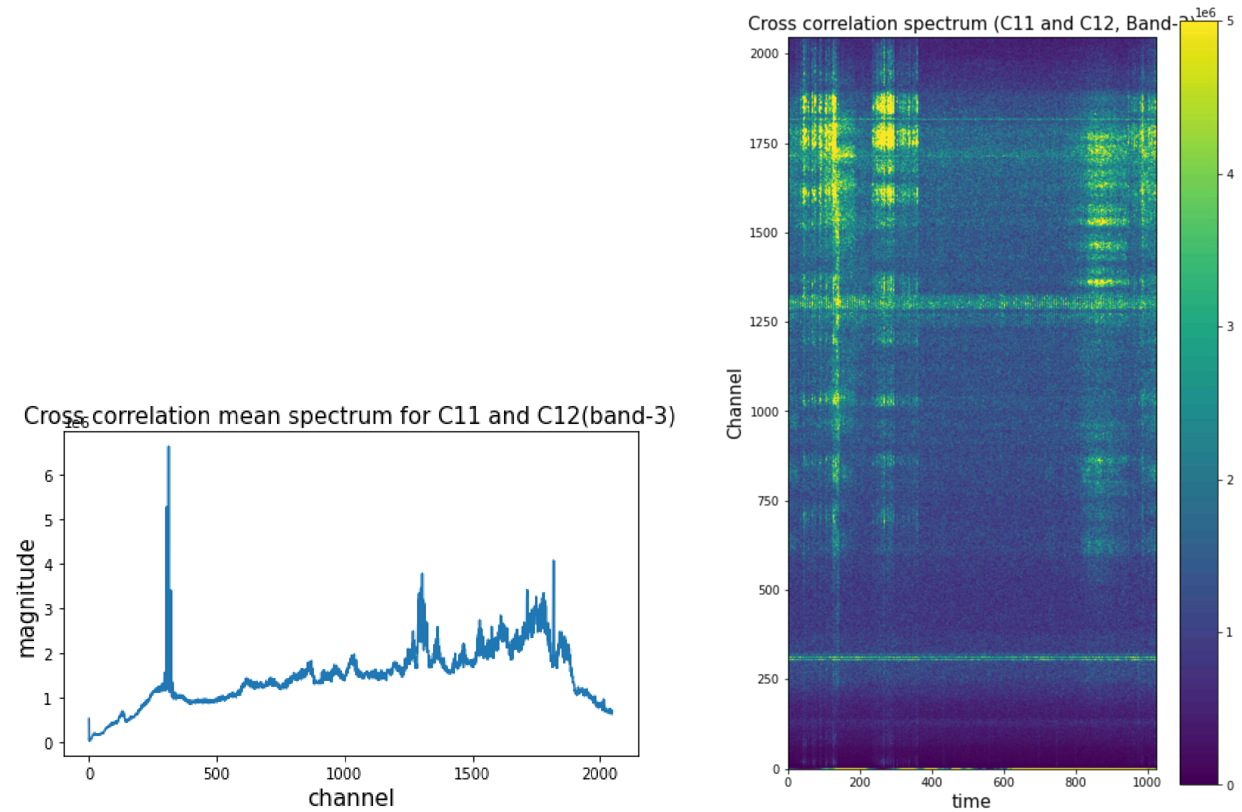


Fig. 7: Cross-correlation mean spectrum and dynamic spectrum for C11 and C12 antennas (band-3).

A similar analysis was repeated for the band-5 data from antennas C11 and C12. The graphs are shown in Fig. 8, 9 and 10

Graphs of power spectrum obtained for band-5, antenna C11, and C12:

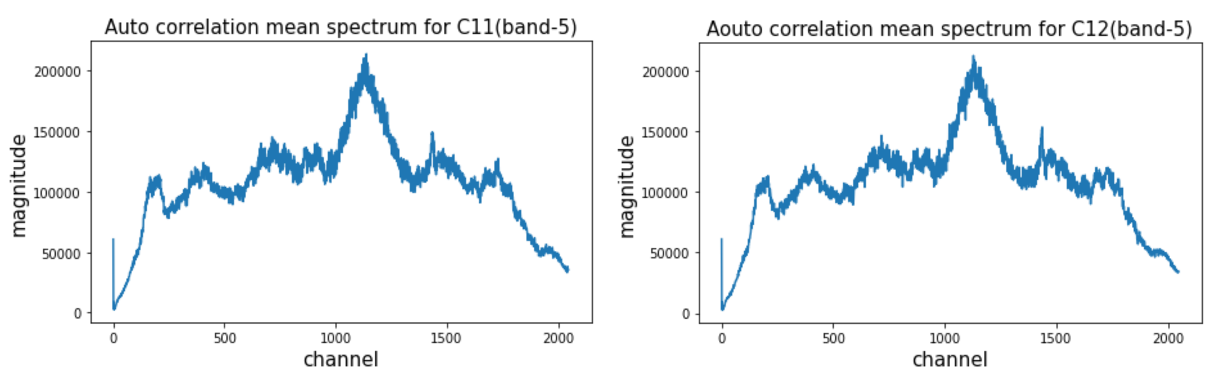


Fig. 8: Auto-correlation mean spectrum for C11 and C12 (band-5)

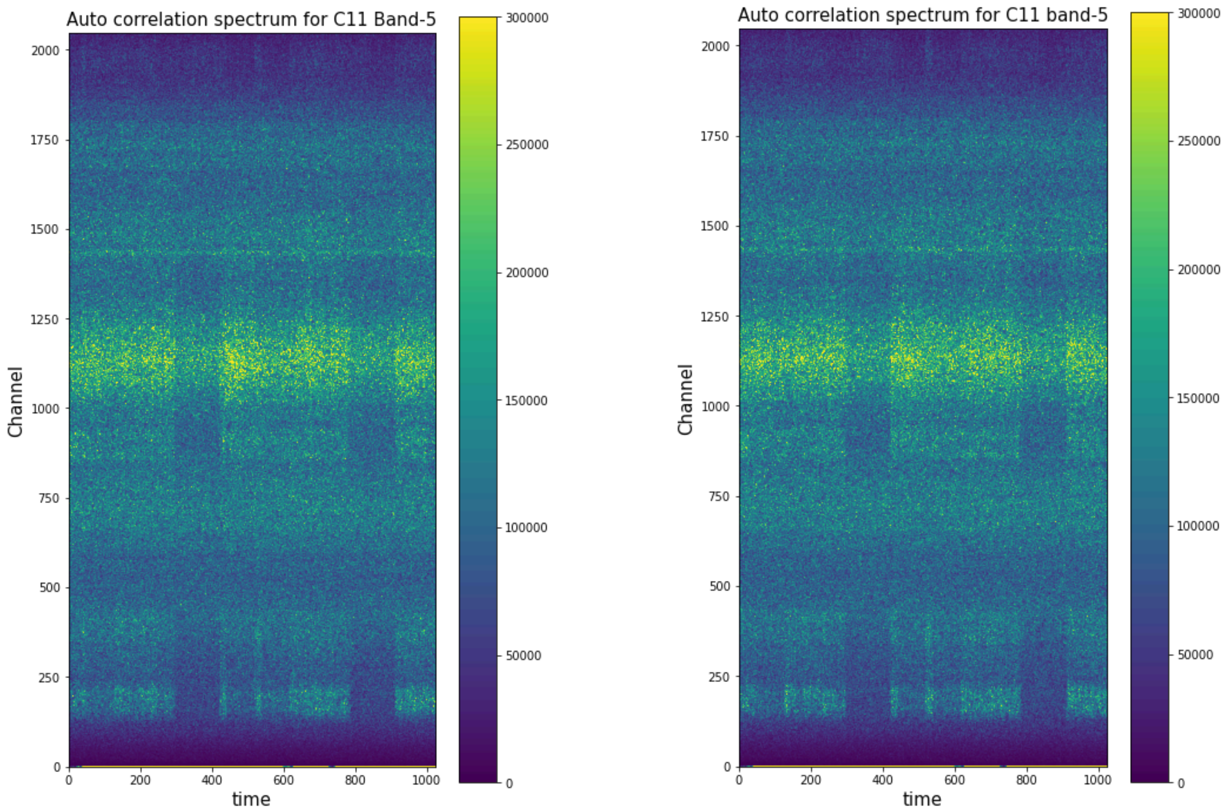


Fig. 9: Dynamic spectrum for auto-correlation for C11 and C12 (band-5)

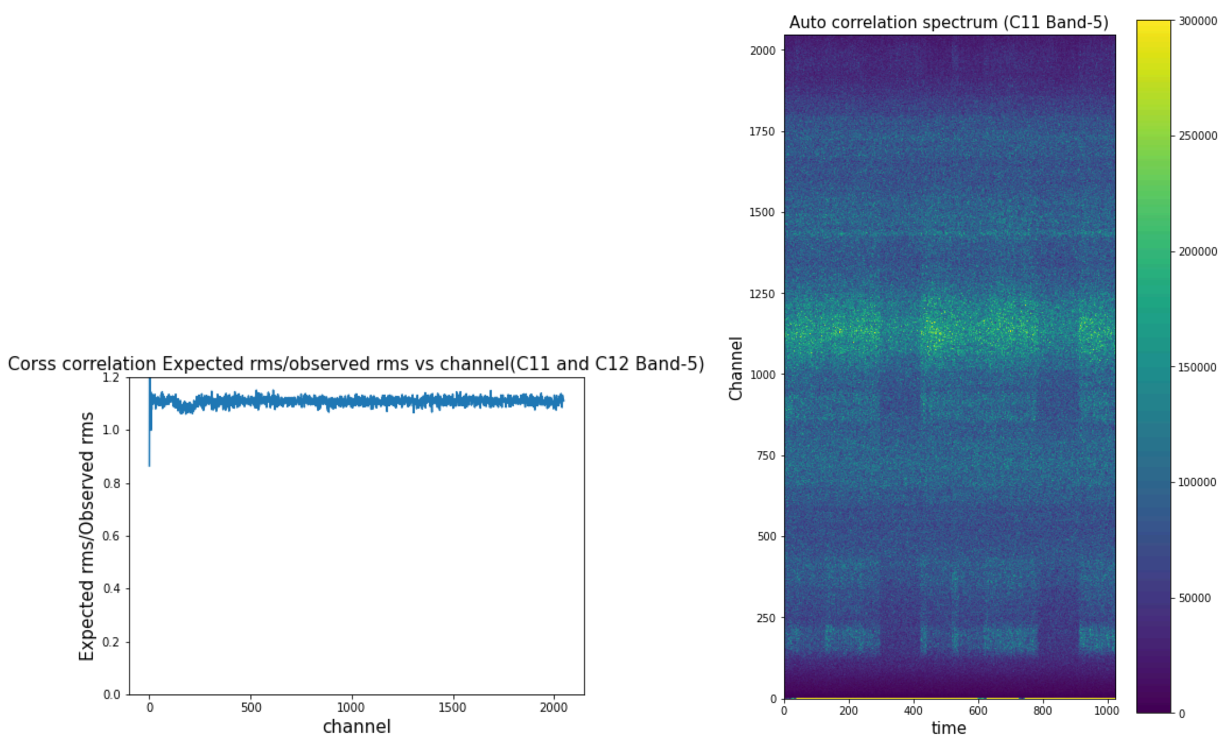


Fig. 10: Cross-correlation spectrum for band-5

4. Identification of RFI

The RFI is present in the signal from both antennas; hence we first identify the RFI in each antenna for the given observing band. To do this, we consider the Fourier transformed data (as obtained in the previous part of analysis for both auto-correlation and cross-correlation) having 2048 frequency channels and 1024 of such power spectra. We find the mean and the rms value for each frequency bin. We then find expected RMS for each channel using the following formula:

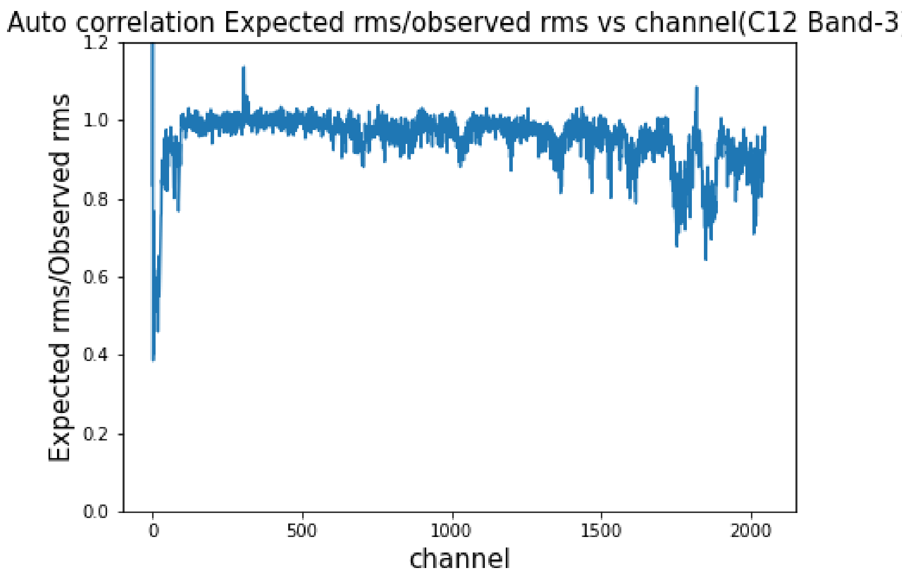
$$\text{Expected rms} = \frac{mean}{\sqrt{no.\, of\ samples}} = \frac{mean}{\sqrt{2\tau\Delta\nu}}$$

Where $\tau$is the time interval, and $\Delta\nu$is the bandwidth. We know that channel bandwidth ($\Delta\nu$) is inverse of time interval (τ) hence their product equals 1 ($\tau\Delta\nu\ =\ 1$). Therefore,

$$\text{Expected rms} = \frac{mean}{\sqrt{2}}$$

The ratio of the expected rms to the observed rms is plotted against the frequency channel (Fig.11 and Fig. 12).

Using the threshold of > 10%, the channels containing the RFI si identified.



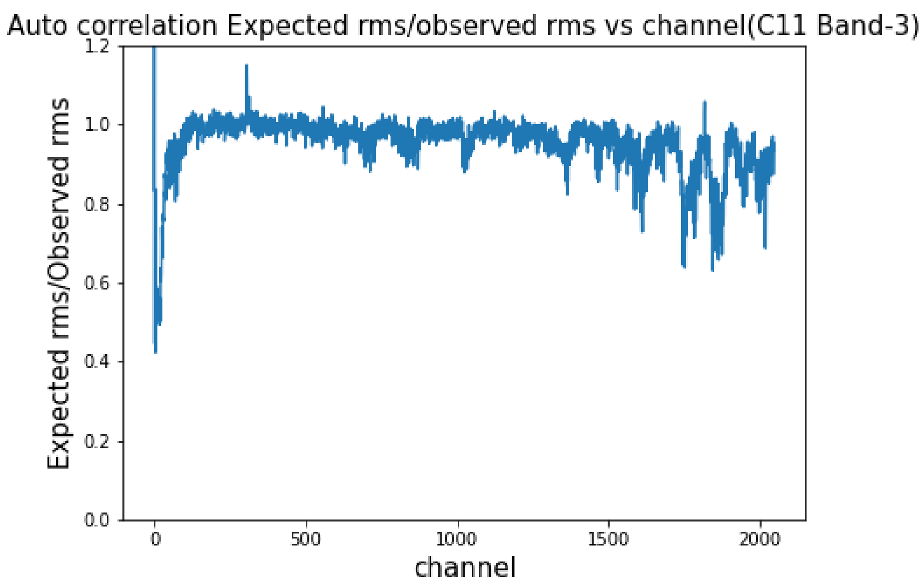
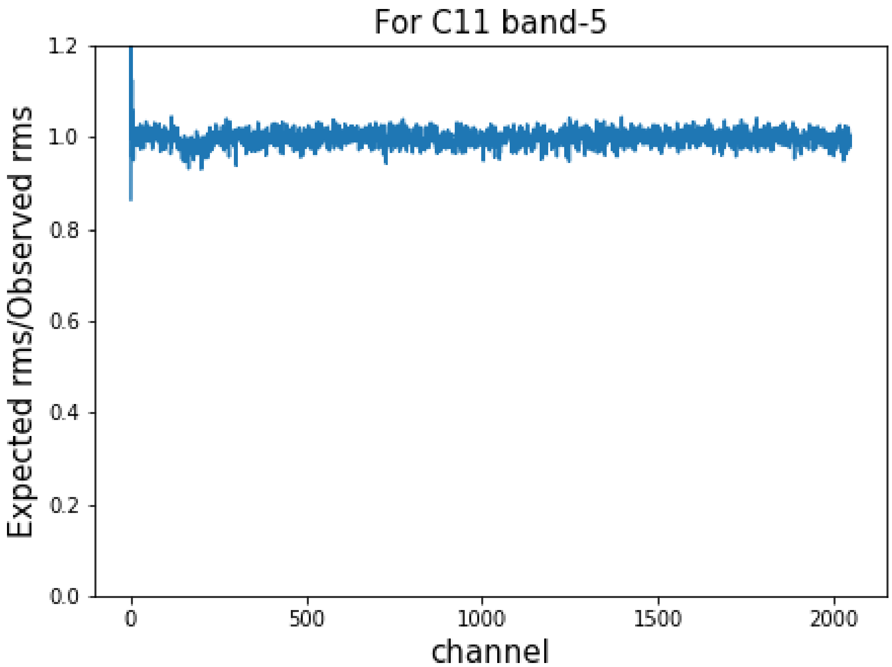


FIg. 11: Expected rms/mean rms for auto-correlation of antenna C11 and C12 at band 3.



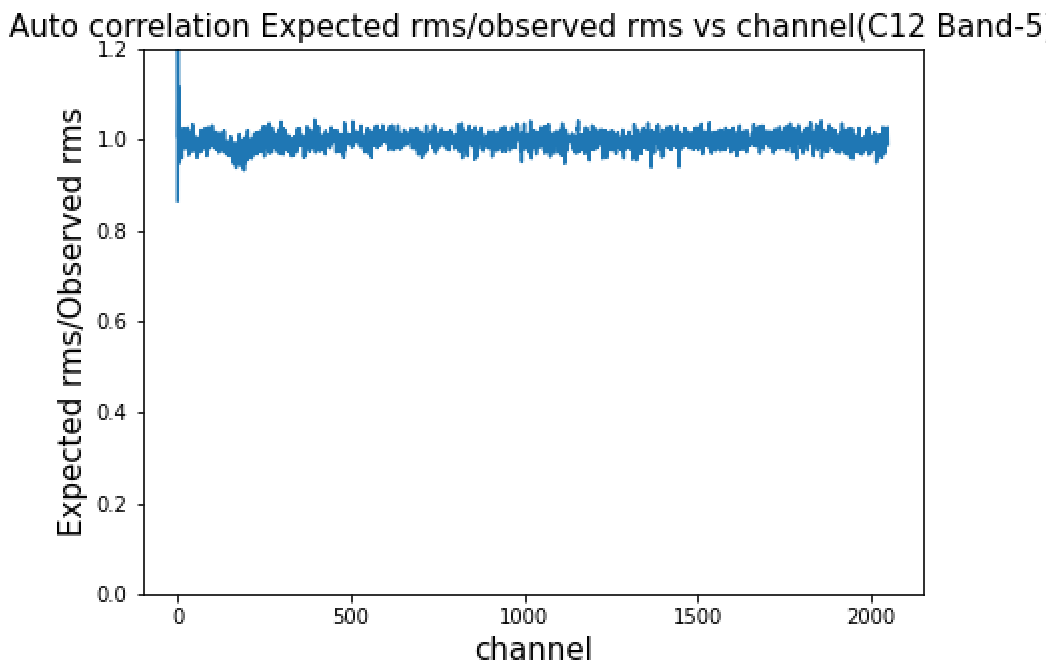


Fig. 12: Expected rms/mean rms for auto-correlation of antenna C11 and C12 at band 5.

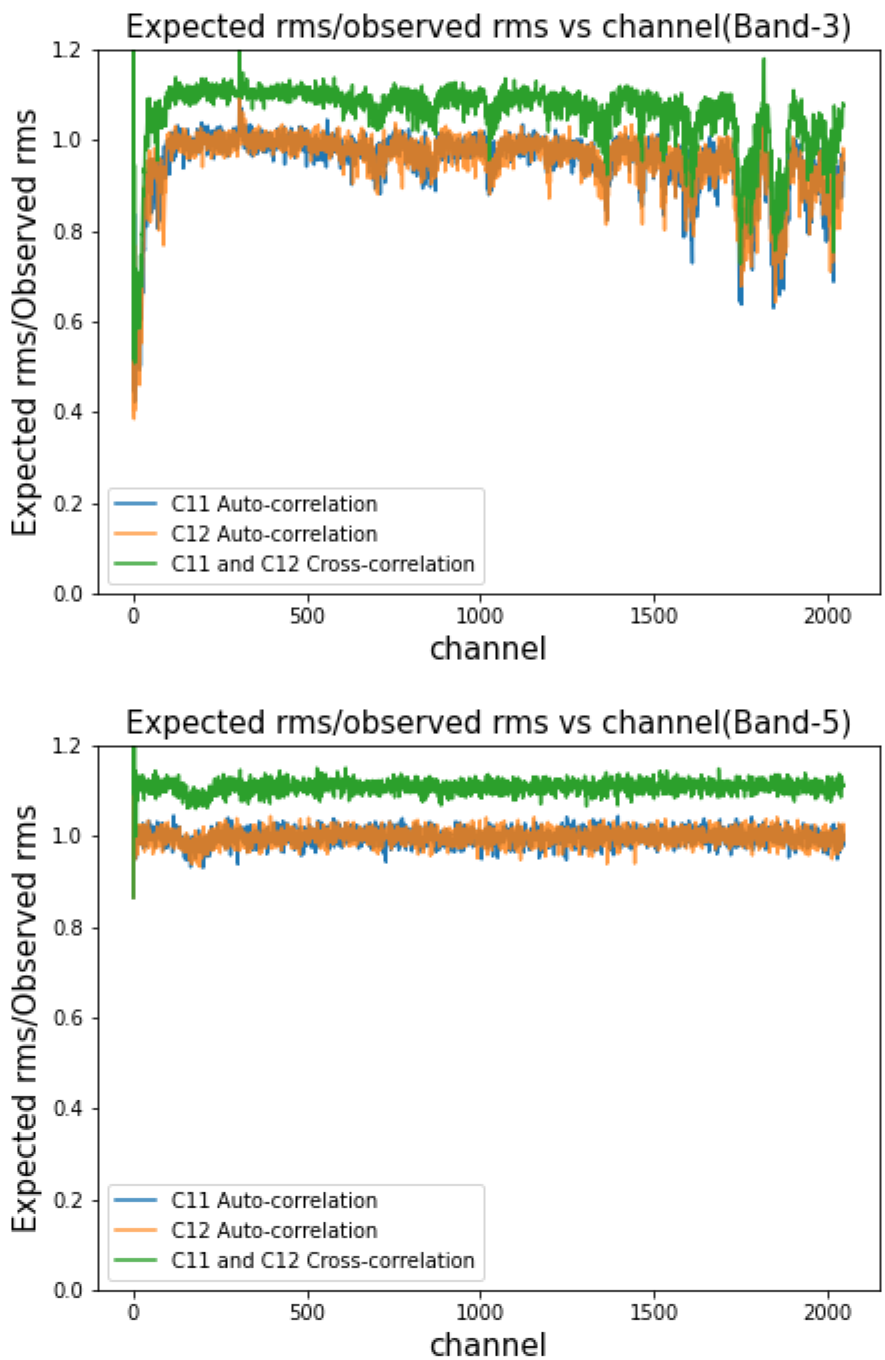


Fig. 13: Expected rms/mean rms plotted against frequency channels for auto-correlation and cross-correlation of antennas C11 and C12 for band-3 and band-5 separately.

III. Interpretation of graphs and discussion

The identified RFI channels for band-3 from auto-correlation of C11 antenna are 305, 712, 834, 1024, 1361, etc. (Many such channels are identified).

The identified RFI channels for band-3 from auto-correlation of C12 antenna are 306, 705, 1025, 1198, 1356, and so on. (Many such channels are identified)

We know that at the edges of the spectrum values will show large deviations and these are contributed by dominant quantization noise at the digitizer[1]

We can compare the bright horizontal lines in the dynamic spectrum for auto-correlation and see that the identified channels with RFI match these corresponding channels.

Fig. 13 shows the expected rms/mean rms for auto-correlation of C11 and C12 (both band-3 and band-5). We can see the difference between the two variations. The expected rms/mean rms for the cross-correlation is also shown in different colors in the same plot.

From Fig. 12, we can see no significant RFI in any of the channels for band-5 in both the antennas (C11 and C12). This is also confirmed by the expected rms/mean rms plot for the cross-correlation of the two antennas.

IV. Results

The channels containing RFI signals were identified using auto-correlation and cross-correlation expected rms/observed rms vs frequency plots. These RFI signals were also identified in the power spectrum for both bands.

V. Reference

1. Identifying RFI in voltage data from the GMRT Ruta Kale, Kaushal D. Buch, Bela S. Dixit

VI. Contributions of group members

The python code for analyzing the given data files and the report are done by Shreya Umesh Prabhu.