Statistical Noise Baseline in AC6 Data

Mykhaylo Shumko

2019-02-04

This report describes the procedure taken to calculate the statistical noise baseline in the Cumulative Distribution Function (CDF) plots made from AC6 microburst detections. The motivation for this analysis is to estimate how often (or fraction of the time) that a cross-correlation (CC) above a certain threshold (0.8 in this case) is due to random chance and is not physical. This will help identify where we should cut off the CDF to model the microburst scale size distribution.

**Count Binning Procedure**

The general approach here is to first bin AC6-A (since it has more 10 Hz data) dos1 counts at a similar magnetospheric location and condition. The binning was implemented by looping over every day with AC6A 10 Hz data, binning the counts by Lm\_OPQ, MLT\_OPQ, and AE. The AE data was appended to the file using minute cadence AE files. The binned counts were saved into its own file in the data/binned\_counts/ folder. Here is an example filename that designates what bin’s counts are in contained whithin “AC6\_counts\_4\_L\_5\_9\_MLT\_10\_300\_AE\_400.csv”. I defined bins in L, MLT, and AE. Currently, the bin edges are:

* L = {4, 5, 6, 7, 8},
* MLT = {0, 1, 2, 3,… 14}
* AE = {100, 200, 300, 400, 500}

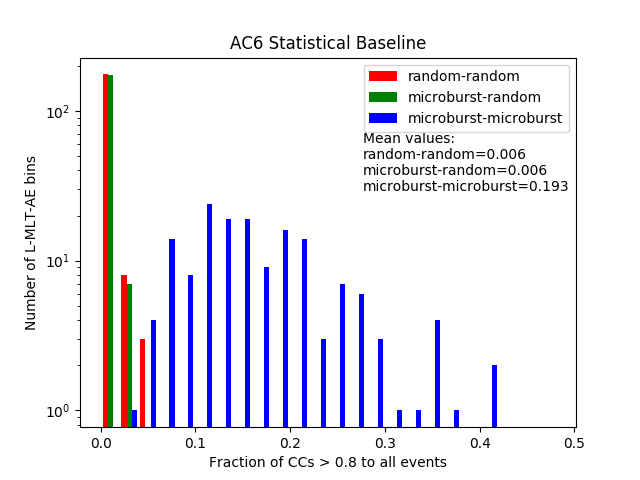
These files were created for convenience sake. It is much easier to loop over these smaller files than loop over all of the AC6A count files for each bin. I am currently redoing this binning for all MLTs and AE up to 800 and it should take a few days to complete.

**The Random Statistical Baseline**

The overall approach here is to calculate a range of baselines given as set of assumptions about what is CCd. One such baseline “**random-random**” is calculated with a CC of random counts to other random counts in each L-MLT-AE bin. The two CC windows are 1 s and 1.2 s with the *mode=’valid’* to account for timing and Poisson uncertainty. This assumption should result in the lowest fraction of valid events for the CDF (fraction of random CC realizations above 0.8).

Another baseline considered here is the “**microburst-random”** which is calculated in a similar manner to random-random, with CC calculated with randomly picked microburst detections against randomly picked data in each bin. This gives us knowledge of how often a microburst CCd with random data will give us a valid event. Note that the random data could include other microbursts!

The last baseline I’ve considered here is the “**microburst-microburst”** which is calculated in a similar manner to random-random and microburst-random, with CC calculated with two randomly picked microburst detections from each bin. Each bin must have at least 10 unique microbursts to CC or it is skipped to avoid CCing a handful of microbursts against themselves to bias the fraction towards higher values. This gives us information of how often two unrelated microburst detections are considered valid in my code.

The plot on the right shows a histogram made by binning all valid L-MLT-AE bins by the fraction of valid events in each L-MLT-AE bin.

It is interesting to note that random-random and microburst-random have a very similar distribution that has a mean value of ~ 1 % chance of a valid event.

The microburst-microburst distribution goes out to higher fractions as I would expect with a mean value of 0.2. Thus we can say that on average, 20% of the time two unrelated microbursts will have a CC > 0.8 and be considered valid by my CDF code.

**Interpretation**

I need to somehow relate these values to the statistical background that is present through the CDF. One option is I can take the mean of the microburst-microburst distribution and go from there. This would be easy and I can justify it as our highest possible baseline. I think we can do better. I am thinking that the true statistical baseline is a combination of the microburst-microburst and microbust-random fractions. For example, given a valid event in the CDF, it could have been two unrelated microbursts which had a 20% chance on average of being marked valid by my code. Also, a microburst CCd with random data has a 1% chance of being marked valid as well. Since both are probably happening, I think we need to consider the microburst duty cycle.

An idea to think through: I think the true statistical baseline is some combination of the microburst-random and microburst-microburst fractions. I think we need to think of this in terms of conditional probability. I am looking for a statistical baseline which I represent as the probability of a non-coincident microburst (NC), given a valid event (V) i.e. P(NC|V). Note that a coincident microburst (CM) probability is 1 – N. I have the probability of V given NC, i.e. P(V|NC) = 0.2 and the probability of V given a random event (R), P(V|R) = 0.01. I am starting to slowly piece this together and it smells like Bayes theorem. If that is the case, then we need to know P(NC), the microburst occurrence frequency. For this we could do this ourselves or lazily refer to other papers e.g. Paul’s 2003 paper on SAMPEX microbursts to get the microburst occurrence frequency. Thoughts?