Appendix H: Data Cleaning Guidelines

H.1 Introduction

This appendix describes guidelines for a process for cleaning data sets to be used in the AE9/AP9/SPM radiation belt climatology models. This process is designed to identify data that is corrupt, has significant background, is saturated, etc. This process does not address calibration and error bars. This appendix extends some aspects of the PRBEM Data Analysis Procedure V1.2 [*Bourdarie et al.*, 2012].

There are 5 recommended examinations, of which 1-4 should be considered as required:

1. Scatter plot (color-coded, e.g., by date or location) of the raw measurement (counts or counts/sec) against another correlated measurement. Ideally, multiple scatter plots would be made: at least one against a background species (e.g., protons in an "electron" channel), and at least one against the same species at a nearby energy (or look-direction or similar sensor)
2. Scatter plot (color coded) of the raw measurement against itself offset in time. This plot identifies transient data spikes.
3. A histogram the number of occurrences of every possible value of the raw measurement. This plot identifies saturation and bit errors, e.g., that can translate low count rates to high ones, or telemetry errors that store the wrong value in the intended data position.
4. A long-term plot of all the data versus time in sensible "bins", e.g., L bins, or equatorial pitch-angle and L bins.
5. A browse plot of each "pass" compared to prior and subsequent passes against a meaningful positional coordinate (e.g., L). For vehicles with only a few orbits per day, this plot can be a daily plot of all data for the day versus L.

By performing the first 4, and ideally all 5, of these examinations, rules can be generated that define measurement values, regions of space, regions of a scatter plot or periods of time during which the data should be flagged as suspect. The data flag should be a bitmap, so that the flag indicates which of the rules were violated, and the flag will likely need to be unique for each channel. For example, a 32-bit integer flag could uniquely identify which, if any, of 32 rules was violated, including all combinations of multiple rules being violated. Not all flags will rule out data's use in a climatology model, and so some indication should be given in the data cleaning report as to what bitmask should be used to exclude points from climatology.

It should be noted that a significant amount of detective work may be required to understand the origin of each type of bad data identified by this process. Therefore, the process should be done as soon as possible after launch, and repeatedly every few months thereafter, while memories are fresh.

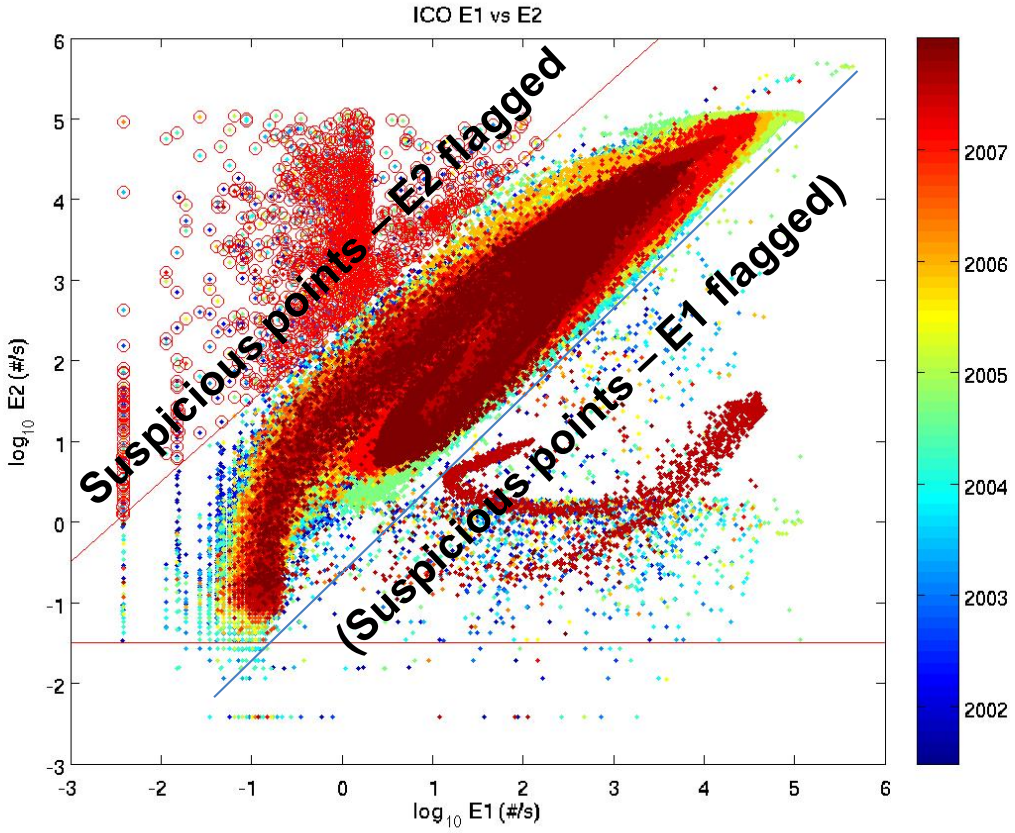

Figure 153. Example from ICO E1 vs. E2 shows bad E2 values above diagonal (a similar, transposed plot removes bad E1 values).

H.2 Scatter Plot Type I (Correlated Quantity or Potential Background)

In scatter plot Type I, the data channel is plotted against a correlated quantity or potential background. In the former case, the correlated quantity is used to define a region of acceptance: accept data only when it falls within a specified range of the correlated quantity. In the latter case, a potential background channel will show up as a diagonal upper or lower bound to the scatter plot. One can then flag points near the diagonal as possibly contaminated. In both cases, color-coding the points can provide a critical extra dimension: are deviations from the main cloud of points isolated in time, location, sensor temperature, etc.

Figure 153 shows a scatter plot of ICO Elec2 (E2) versus Elec1 (E1). The vast majority of points fall in the main cloud. Whenever E2 deviates too far from the main cloud, it should be flagged as suspicious. These suspicious points are much more common early in the mission (for unknown reason). It is also evident that there are similar bad points in E1, including an interval in 2007 (red structure in lower right). It is a judgment call whether to mark both data channels in these two outlier regions.

Figure 154 shows a potential background scatter plot. On HEO-F3, protons can register in the electron channels (e.g., Elec3), and electrons can register in the proton channels (e.g., Prot4). The diagonal line in the upper left defines a region where Elec3 should be flagged for possible proton contamination, and the nearly-vertical line on the right indicates a region where Prot4 should be flagged for possible electron contamination. The *L* color coding shows that electron contamination of Prot4 occurs primarily in the outer zone, where Prot4 is expected to have little or no signal anyway. The protons in Elec3 are primarily confined to the low *L* values of the inner zone, except during solar particle events, when they can occur at any *L* value.

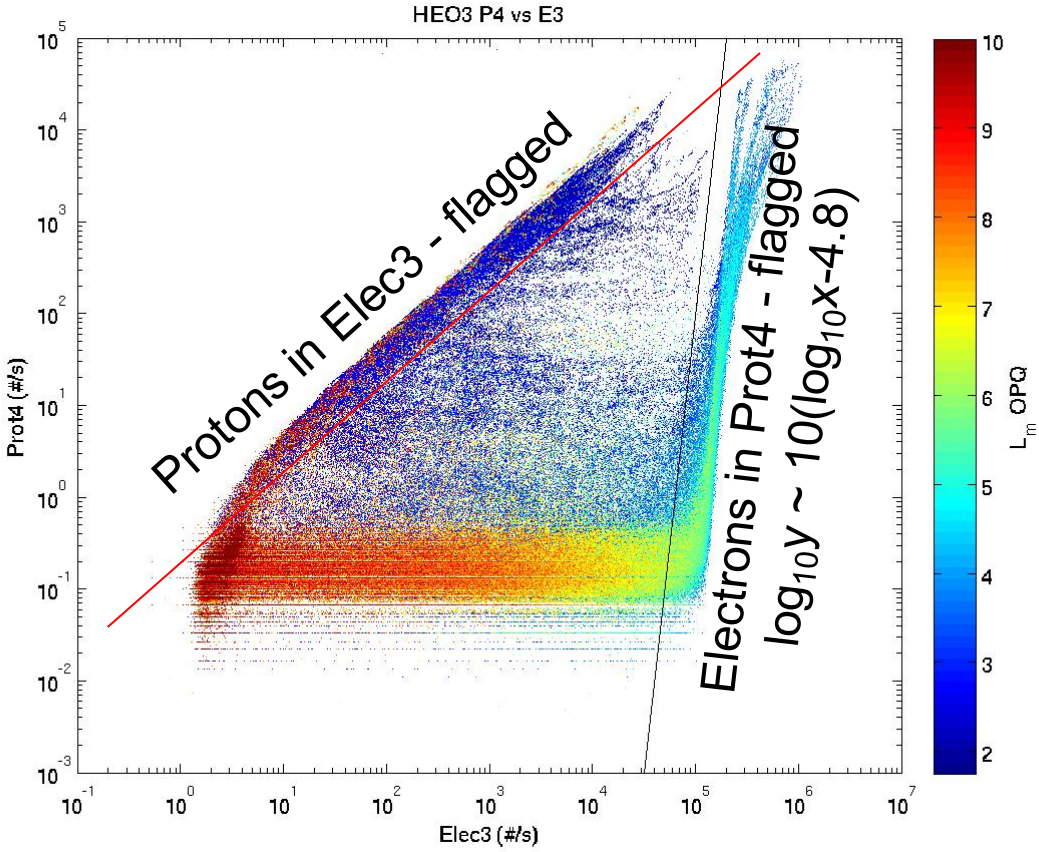

Figure 154. Example from HEO-F3 shows electron "background" in Prot4 channel. So long as Elec3 is to the left of the black line, the Prot4 can be used. Very few points in the inner belt are excluded.

H.3 Scatter Plot Type II (Time Offset)

An extension of Scatter Plot Type I is to plot a data channel against itself at a time lag. The time lag should be small to reveal data spikes. Figure 155 shows a self-scatter plot for HEO-F1 E4, where adjacent time points are plotted against each other, except when they are more than 30 seconds apart (the nominal cadence is 15 seconds). The simple limits defined by the red lines can remove many spurious data spikes. However, some data spikes may be real, so the use of the data flag associated with a self-scatter plot may be appropriate for climatology but not for other investigations.

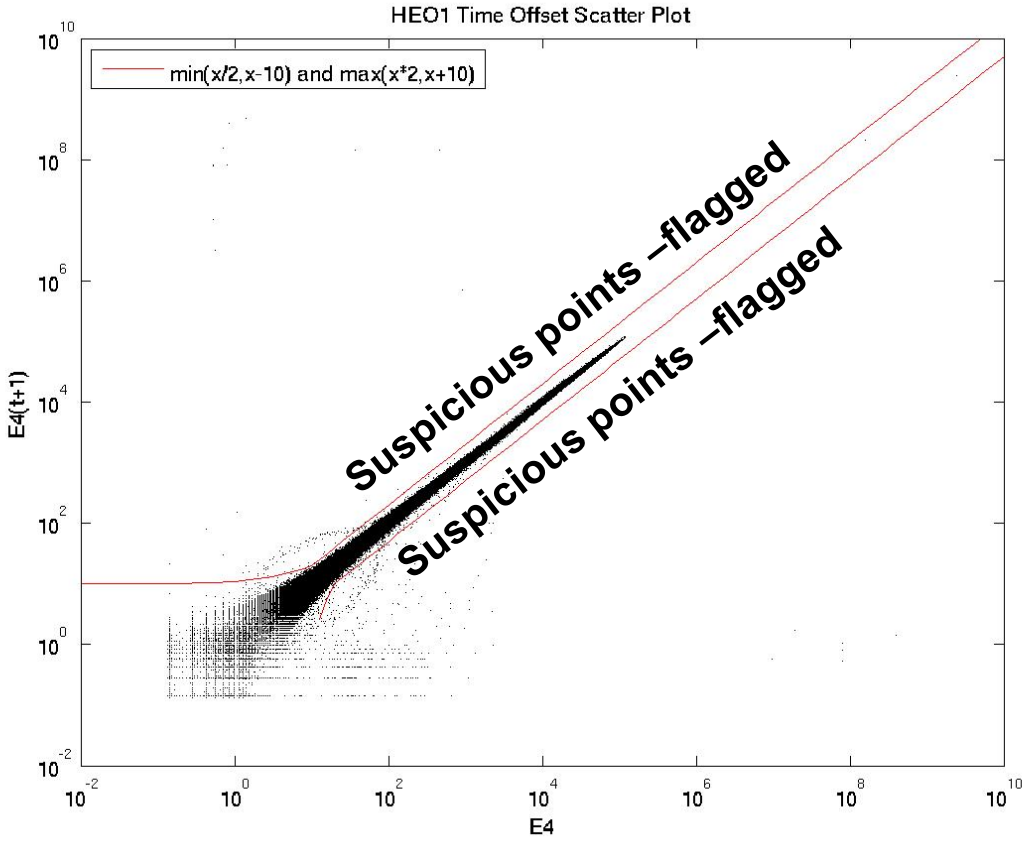

Figure 155. Example from HEO-F1 shows numerous artifacts manifest as large deviations in adjacent time points (<30 seconds apart).

H.4 Histogram

A histogram can reveal a variety of problems in a data set. It is best to make two kinds of histograms: one of raw digital values (e.g., original counts from the instrument) to detect bit errors, which show up as "echoes" of the main distribution at linear offsets that are powers of 2; the second histogram should be in counts per second to reveal dead-time issues. In Figure 156 below, we have shown counts per second for HEO-F3 Elec3, which is known to have dead-time saturation at high count rates. The figure shows this saturation feature as a bump on the tail of the distribution (the multiple traces evident at ~10^3 counts is an artifact of varying integration times versus digital compression and should not, in principle, be present in a plot of raw observed counts).

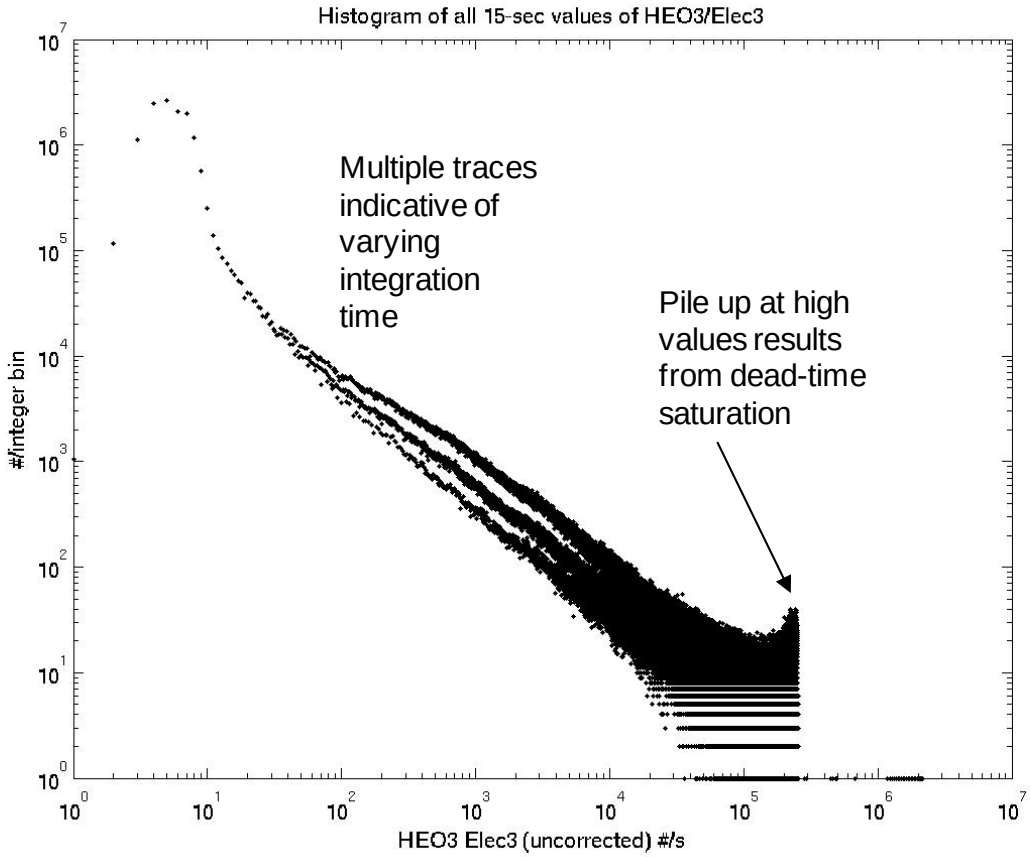

Figure 156. Example from HEO-F3/Elec3 shows saturation at high count rates.

H.5 Time Series

The final "required" plot type is a simple time series of data filtered or binned by reasonable criteria that should remove orbit effects and reveal only temporal radiation belt dynamics. These plots can reveal, for example, seasonal temperature effects, mode changes that were not processed correctly, or systematic degradation of the sensor.

Figure 157 shows an artifact in HEO-F3 Dose1 Rate and Dose2 Rate. Both data channels show suspicious high rates intermittently starting in 2003. While this plot does not lend itself to easily removing the suspicious points, it does reveal their presence, which can then be flagged out using one of the scatter plot types. Alternatively, one can compute a moving percentile (e.g., 95th) of the data channel and look for points that are more than some constant multiplier times that percentile. Because the tail of the distribution falls off very quickly, the 99th and 95th percentiles may be only a factor of 2 apart, and a factor of 2 beyond the 99th percentile may include the 99.99th percentile. Finally, one always has the option to define an exclusion boundary (such as the red one in Figure 157).

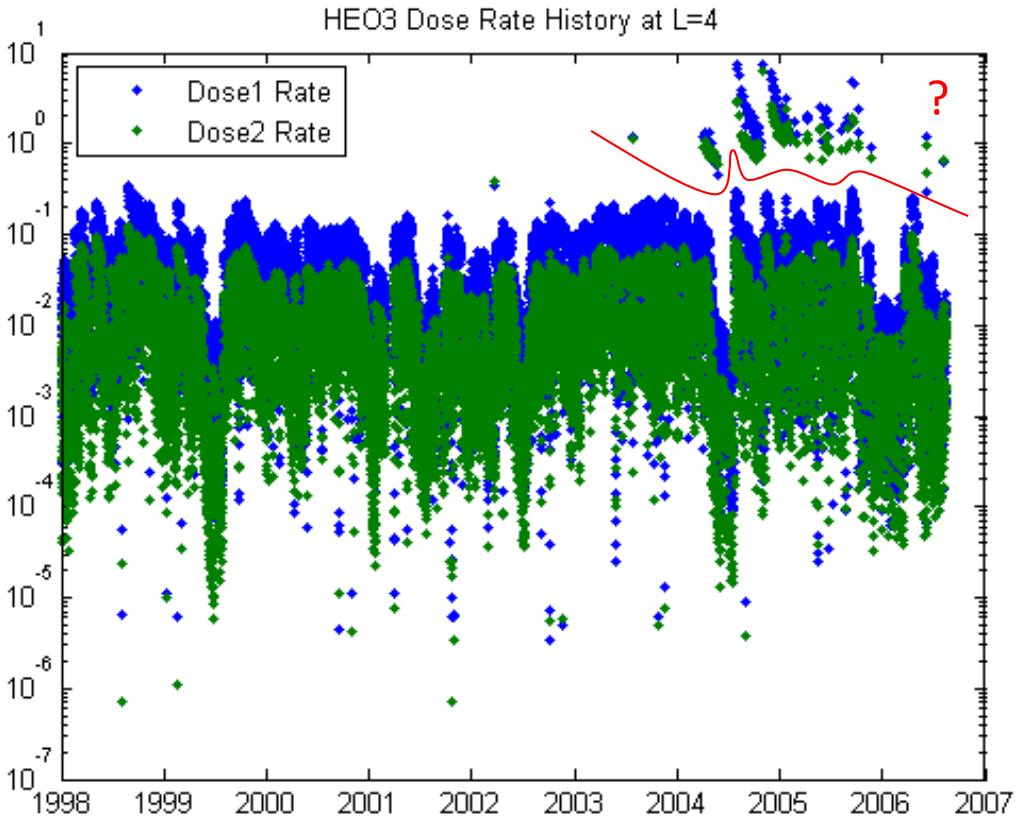

Figure 157. Example from HEO-F3 showing suspicious signature in a whole-mission time series plot.

H.6 Pass Comparison Browse Plot

To gain the highest confidence in the cleanliness of a data set, one should browse the entire data set. This is a tedious process, and so regarded as optional. An ideal browse plot allows one to compare a single pass through the radiation belts to its neighbors. Points that stand out can be individually flagged, and other artifacts in the data can be revealed. For example, Figure 158 reveals a systematic difference between inbound and outbound passes in ICO data (the signature is present in all energy channels). For inbound passes, the radiation belts are displaced to lower *L* than for outbound passes. Further investigation revealed that this artifact is likely a result of a timing error between the sensor data and the vehicle ephemeris. This data could be flagged, but it could also still be used for climatological studies: the timing error translates into extra variance in the binned data.

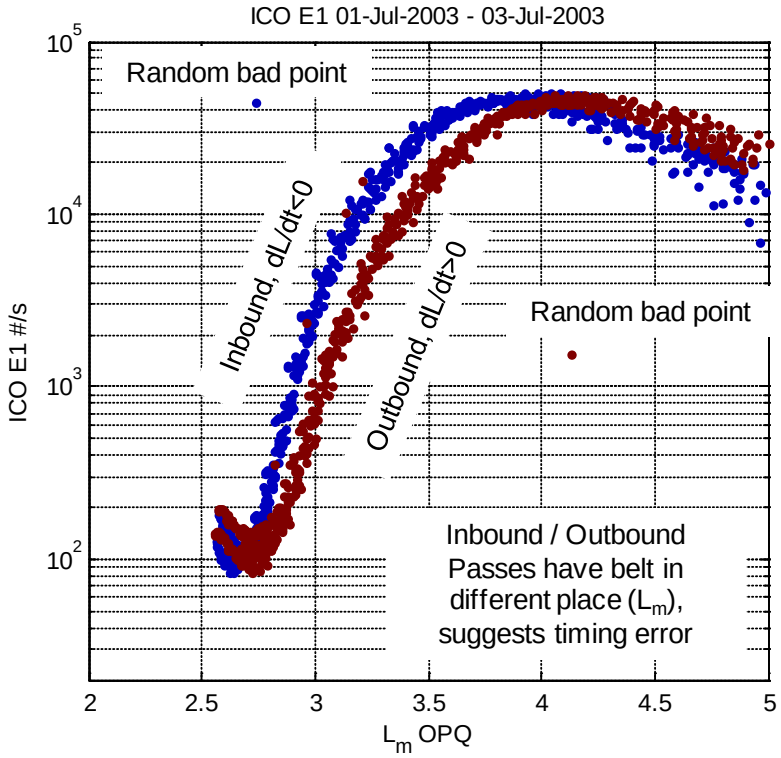

Figure 158. Example from ICO showing possible timing error between instrument and ephemeris.

H.7 Flag Description Table

Once the set of plots is made and all the artifacts are identified, the flagging rules should be collected into a table. The table should include a description of each rule, including any equation needed to apply the rule, the bit that is set in the data flag when the rule would exclude a data point, and an indication of what fraction of the data set is marked with that particular flag. Finally, an additional entry should be provided indicating the combined bit mask that should be used to exclude (by bitwise AND) points that should not be used in a climatology study. A sample table is given as Table 69.

Table 69. Example of data flag description table (numbers are for illustration only), for ICO Elec1 (E1).

Description	Set Bit	% Affected
Saturation: $E_1 > a$	1	5
Disgreement with E_2: $(E_1>10^b E\ 2^a)$ & $(E_1 > c)$ and $(E_2>d)$	2	1
Temporal discontinuity; $(E_1(t+1)>\max(2\ E_1(t), E(t)+10)$ | $(E_1(t)<\min(E_1\ (t-1)/2, E_1(t-1)-10)$	4	0.2
Mostly Proton Background: $E_1 < 10^b P_1{}^a$	8	7
Suspected Timing Error (by visual inspection)	16	40
Climatology Exclusion Mask	*15*	*13*

H.8 Summary

The process outlined in this section can be used to provide a fairly high degree of confidence in the quality of a resulting data set. To capture the results of that process into a data cleaning report, the following outline is recommended:

- Description of data set provenance
  - Where was the data downloaded from / who provided it
  - Interval covered (start/end dates)
  - What format was it in
  - Example file names
  - Summary of known preprocessing prior to data transfer
- Data description
  - Instrument description / method of measurement
  - Species and energy response, including background species
  - Angular response
  - Nominal geometric/energy/efficiency/livetime factors
- Catalogue of noteworthy plots
  - All plots leading to a flag rule
  - Other plots that help illustrate data flag
  - Omit (optional) data cleaning plots that do not reveal bad data or are not primary means of detecting and flagging bad data
- Summary
  - Table of data cleaning flags, with rule equations and comments, for each data channel
  - Flag bit mask recommended for excluding data from climatological analysis
  - Unresolved issues

Some aspects of the processed defined above can be automated: for example, the ONERA IPSAT tool facilitates some of the scatter plotting. Further, it may be desirable to develop a graphical user interface (GUI) to facilitate selecting bad points for flagging and to define exclusion regions in the various plots. Hosting such tools at a Virtual Observatory (e.g., ViRBO) or integrating them into a standard plotting tool (e.g., autoplot), could enable widespread, economical use of this process.