Using Historical Atmospheric Radiation Measurement Program Measurements to Systematically Improve Quality Control Limits

S.T. Moore
ATK Mission Research
Santa Barbara, California

K. Kehoe, R. Peppler, and K. Sonntag
Atmospheric Radiation Measurement Program Data Quality Office
University of Oklahoma
Norman, Oklahoma

Introduction

Some Atmospheric Radiation Measurement (ARM) data streams either have no valid data ranges defined for the quantities contained within, or have limits too broad to catch instrument problems. In the absence of any expert judgment as to what appropriate limits should be, we look at historical data in order to determine typical data ranges. The ARM Program has amassed more than ten years of continuous data for some measurements, providing a wealth of samples just ripe for statistical analysis. We are building a system that will systematically process this historical data in order to improve Quality Control limits.

Objectives

- Set quality control limits for data streams without any valid data ranges defined.
- Suggest better limits for b-level and c-level products based upon the historical record.
- Determine which measurements might benefit from imposing monthly and site specific varying limits.
- Provide ARM Climate Research Facility with new tools for visualizing measurements.

System Design and Method

A block diagram of the system is shown in Figure 1.

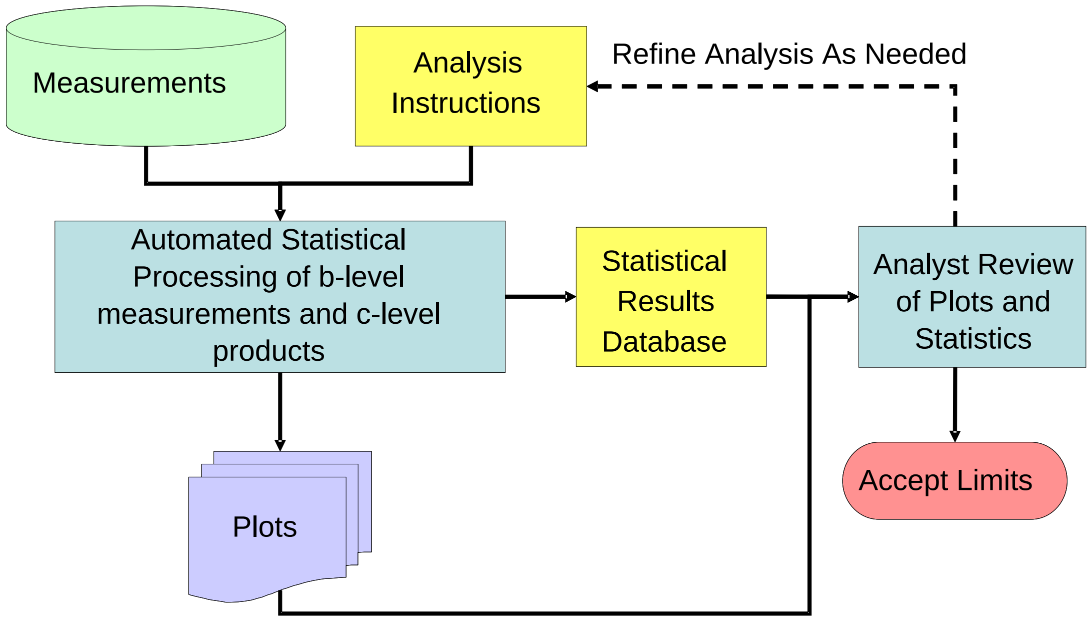

Figure 1. Statistical Analysis System Block Diagram.

The ARM Archive will supply us b-level and c-level data from data streams currently in production. A project server with 500 GB of storage has been configured to host roughly 10 years of this data. Only a handful of the very largest datasets will be excluded due to space limitations on the server.

The operator decides which quantities to analyze using a measurement selection interface or simply chooses to process all variables in a given datastream. Data can be analyzed over custom time ranges, by year, by month, or by the entire archive. The statistical analysis system reads each file, excludes data marked as missing or bad and concatenates the remaining data into an array for analysis. Various statistics are computed, such as mean, min, max, daily min, daily max, and percentage of samples passing existing range checks. Two plots are generated for each run. The first is a time-series of the data and the second is a frequency distribution. Each plot has relevant statistical measures overlaid to assist analysis. This information is saved in a database for subsequent review.

If analysis determines that new limits are appropriate for a given datastream quantity, the analyst will be able to suggest and store new limits using this tool. A database will keep track of the new limits along with pointers to the details of the analysis, including any plots generated. A Web-based application may be used to peruse the generated plots, statistics and proposed limits. Instrument mentors or other interested parties will be able to review and refine proposed limits before the data quality office or the data management facility includes them as part of their daily automated processing.

Examples

Figure 2 shows one example of how our tool might be used to improve Quality Control limits. Surface Meteorological Observing Station b-level data files currently define valid relative humidity range to be greater than -2% and less than 104%. The bottom plot of Figure 2 clearly shows an abnormal spike at 0 that is outside of the normal distribution. Analysis of this data suggests the existing range could be tighter to catch more problems.

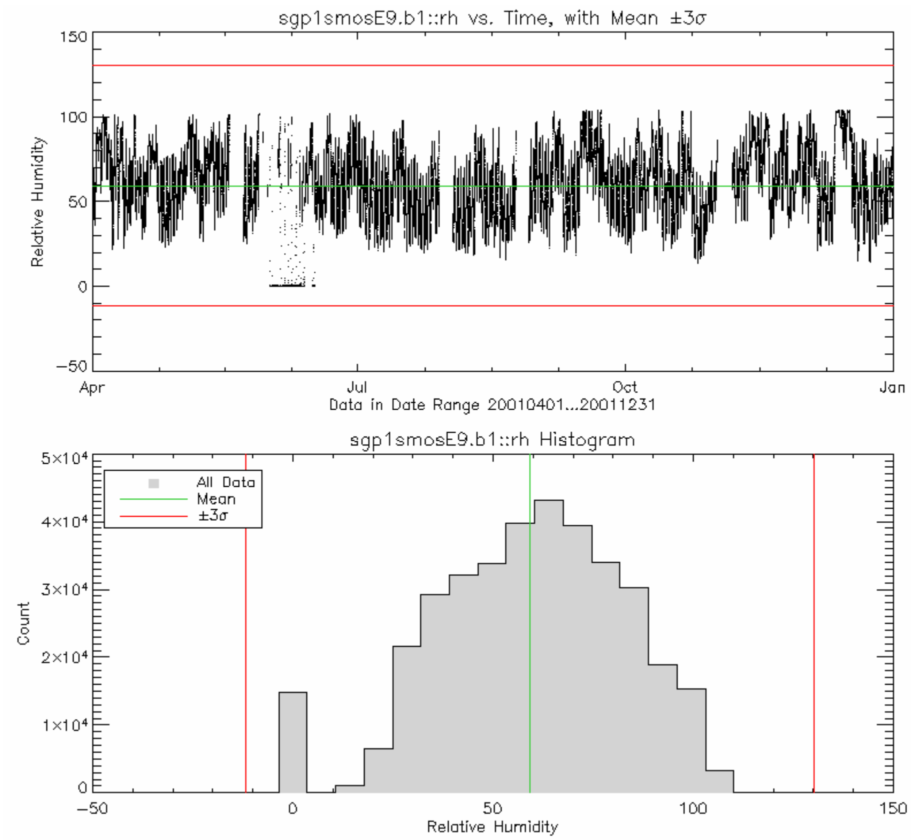

Figure 2. A preliminary version of our analysis system processed Surface Meteorological Observing Station relative humidity data from the 2001 year time frame. The existing limits are defined to be -2% to 104%. The frequency distribution plot shows an abnormal spike outside of the normal distribution.

Many Value Added Products produced by ARM do not have valid data ranges defined. Figure 3 shows upwelling longwave radiation data from one of these Value Added Products. Using the plots generated by this analysis system, the analyst can quickly suggest some appropriate limits, based on more than 10 years of measurements. In this case, the analyst has decided to also process the data by month. Data from January (for all years) is colored green in the frequency distribution plot, versus gray for all months. In this fashion, January specific limits can be determined at a glance in addition to global limits.

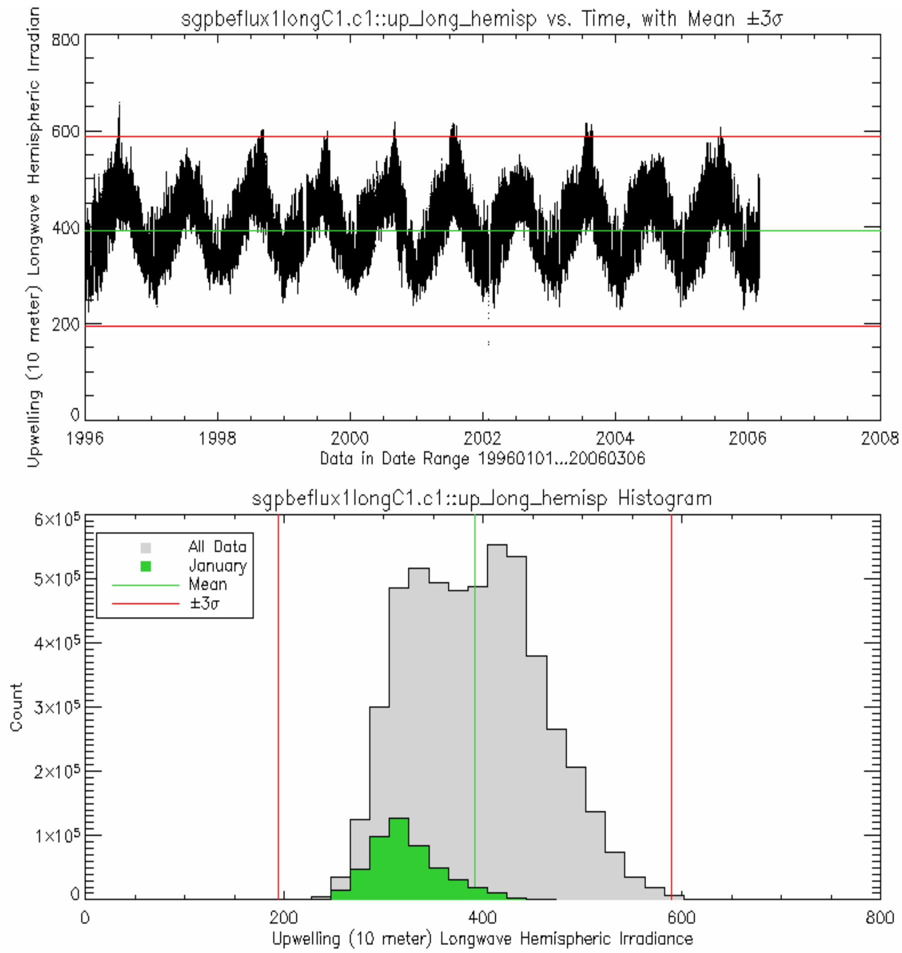

Figure 3. Time-series and frequency distribution for upwelling longwave radiation from an ARM value added product. More than 10 years of data is represented. January data from each year is colored green.

Conclusion

This analysis system will generate long time-series plots, frequency distributions, and other relevant statistics for scientific and engineering data in each b-level or c-level data stream. Furthermore, frequency distributions categorized by month or by season are useful to set valid data ranges specific to those time domains. These statistics can be used to set limits that when checked, will improve upon the reporting of suspicious data and the detection of instrument malfunction early. The statistics and proposed limits will be stored in a database for easy reporting, refining, and for use by other processes. Web-based applications to view the results will also be made available.

Acknowledgements

This research was supported by the Office of Science, U.S. Department of Energy, Grant Number DE-FG02-04ER63864.

Point of Contact

Sean Moore, sean.moore@arm.gov, (805) 963-8761 ext 243