P2.6

ANALYSIS OF HISTORICAL ARM MEASUREMENTS TO DETECT TRENDS AND ASSESS TYPICAL BEHAVIOR

Sean T. Moore*
Mission Research and Technical Services,
Santa Barbara, California

Kenneth Kehoe, Randy Peppler, Karen Sonntag
ARM Data Quality Office, CIMMS,
University of Oklahoma, Norman, Oklahoma

1. INTRODUCTION

The Data Quality Office of the Department Of Energy's Atmospheric Radiation Measurement (ARM) Climate Research Facility (ACRF) is charged with ensuring that released data is of a known and reasonable quality suitable for scientific research (Sonntag 2003). A number of automated and manual inspection procedures are used to inspect approximately 5000 data fields each day. The automated procedures typically rely on comparing measurements to a set of valid data ranges, while the human analysts sort out the false alarms as well as catch abnormalities that are not simple to detect by any computer algorithm.

When possible, experts familiar with the instrumentation employed by the program are consulted in order to define valid data ranges. However, some ARM data streams have never been assigned valid ranges, or have limits too broad to catch serious instrument problems.

In order to comprehensively define limits for all produced data streams, we are developing a tool that systematically reviews the entire historical record of measurements. The ARM Program has amassed more than ten years of continuous data for some instruments, providing a wealth of samples just ripe for statistical analysis. Our tool produces statistical summaries, frequency distributions, diagnostic plots, suggested quality control limits and a feedback mechanism to help keep instrument mentors and the data quality office in agreement regarding validation checks.

* *Corresponding author address:* Sean T. Moore, Mission Research and Technical Services, P.O. Drawer 719, Santa Barbara, CA 93102; e-mail: sean.moore@arm.gov.

Visualization tools developed as part of this effort will help analysts detect abnormal trends early, leading to quick problem resolution and an overall higher level of data quality.

2. SYSTEM DESIGN AND METHOD

The system can be broken down into the following processes: data retrieval, data import, data acceptance, statistical processing, report generation, analysis review, and feedback.

2.1 Data Retrieval and Import

ARM data is typically stored in daily NetCDF files and warehoused at the ARM Archive (Macduff 2005). Measurements from any given instrument are usually grouped into a small number of data streams (collection of similarly structured files).

The standard way to obtain ARM data is via a Web-based request for later retrieval by FTP. When data is ordered in this fashion, our data mirroring process will automatically retrieve the data and import it into our system. Our import process sorts through the received files and moves them into our local data store.

The ARM Archive has supplied to us b-level and c-level data from most data streams currently in production. A data server with 500 GB of disk storage has been configured to host roughly 10 years of this data. Only a handful of the very largest datasets have been excluded due to space limitations on the server. Once the system is fully functional, additional storage can be added as needed.

2.2 Data Acceptance and the Analysis Queue

Files in our data store are automatically inspected for fields appropriate for statistical analysis. Only time-varying and floating point NetCDF fields are

accepted. Multidimensional fields are accepted as long as one dimension is temporal.

An operator may decide which of the accepted fields should be queued up for analysis. An interactive data selection process displays a choice of variables, the dates of availability and a choice of desired analyses. Alternatively, the system simply queues up all appropriate fields and schedules them for each of the supported analyses.

A queue record consists of an identifier uniquely representing the field of interest, an analysis start and end date, an analysis type, an optional flag to pre-filter suspected outliers, an optional analysis month of interest, a run status flag, and a message field used only if the run flag indicates an error has occurred. The queue is implemented as a MySQL database table. Figure 1 depicts the flow of data from the archive that ultimately results in the population of the queue.

The use of a queue ensures that machine resources are efficiently managed while computing the statistics. Even if multiple simultaneous users are requesting an analysis, the queue will prevent the jobs from stepping on each other. Also, the queue provides convenient documentation of the parameters required to perform each analysis. As the analysis or plotting code is improved, it is a simple matter to reset the run flag for each entry and process the queue again to update the results.

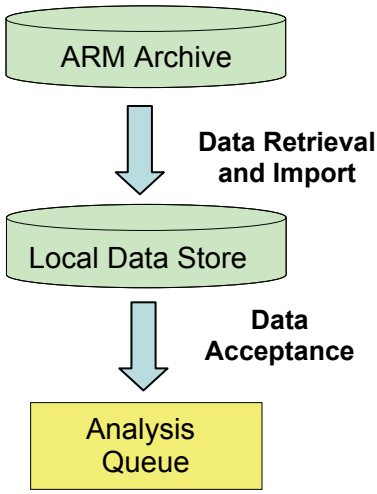

Figure 1. Data Flow into the Analysis System.

2.3 Statistical Processing and Reports

In preparation for statistical analysis, an entry is pulled from the analysis queue, and all data and metadata for the entry is read for the time range requested. Metadata for the data field may specify existing valid range limits. If so, these limits are optionally used to filter out extreme outliers before any statistics are tabulated. The system excludes data marked as missing or bad and concatenates the remaining data into an array for analysis.

Various statistics are computed, such as mean, minimum, maximum, standard deviation, daily minimum, daily maximum, and percentage of samples passing existing range checks. If the analysis is by month, only data collected during the month specified is used, but over the entire selected time range. For example, the analysis may be to look at data recorded in just the months of January for all years between 2001 and 2005.

The statistics, along with all parameters required to repeat the analysis are stored or referenced in another MySQL database table, one record per run. The analyst name and a link to all associated graphics are also included in the record. Two plots are generated for each variable and for each analysis run. The first is a time-series over the time period specified, and the second is a frequency distribution. Each plot has relevant statistical measures overlaid to assist review.

2.4 Analysis Review and Feedback

A web-based application is available for users to peruse the generated plots, statistics, and the proposed limits. A web-based front-end to the database tables is also available.

If analysis determines that new limits are appropriate for a given quantity, the analyst will be able to suggest and store new limits using this tool. The database will keep track of the new limits along with pointers to the details of the analysis and generated plots. 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.

Figure 2 provides a concise summary of the processes involved that make up the statistical analysis system.

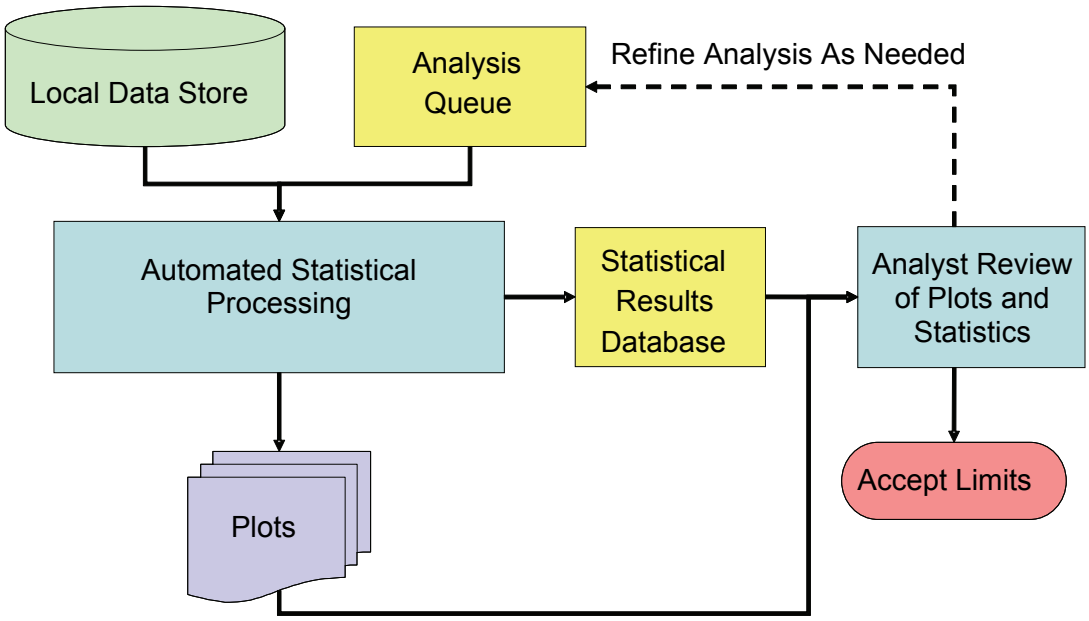

Figure 2. Automated Statistical Analysis System Block Diagram.

3. EARLY RESULTS

After analyzing a few of the ARM data sets with this tool, we believe that all of the objectives we hoped to achieve will be met. Namely, we want to use the system to improve upon existing quality control (QC) limits, we want to set reasonable limits for datastreams without limits, and we want to define monthly limits where appropriate. We also wish to quickly detect any abnormal trends.

3.1 Improve Existing Limits

Figure 3 shows one example of how our tool might be used to improve existing Quality Control (QC) limits. Metadata for a suite of meteorological sensors currently define valid relative humidity range to be between -2% and 104%. The bottom plot of Figure 3 clearly shows an abnormal spike at zero that is outside of the normal distribution for the year. Analysis of this data suggests the existing range could be tighter to catch more problems.

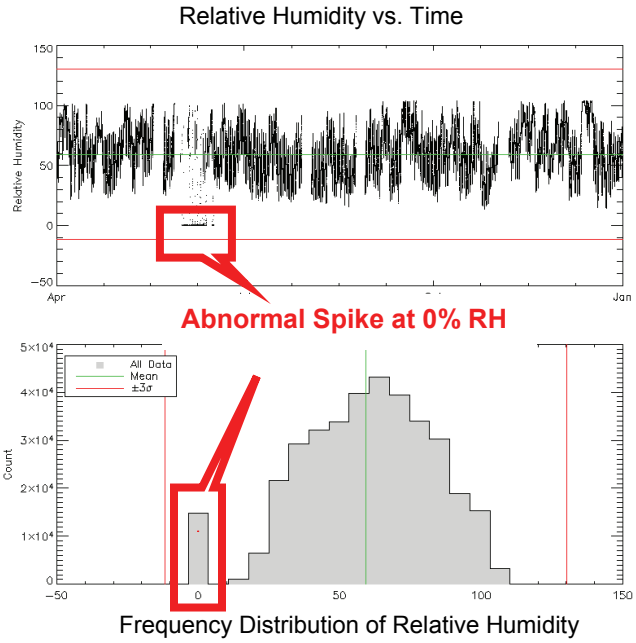

Figure 3. Application of analysis system to improve existing data quality limits. Data displayed is from the Surface Meteorological Observation System (SMOS) of the Southern Great Plains ARM Climate Research Facility.

3.2 *Implement Monthly Limits*

Many Value Added Products (VAPs) produced by ARM do not have valid data ranges defined. Figure 4 shows a time-series and frequency distribution graph for upwelling longwave radiation from an ARM VAP. More than ten years of data is represented. The distribution colored green represents values measured in January for each of the ten years analyzed. The gray colored area represents all months. Using the plots generated by our analysis system, the analyst can quickly suggest some appropriate monthly or global limits for a data product such as this.

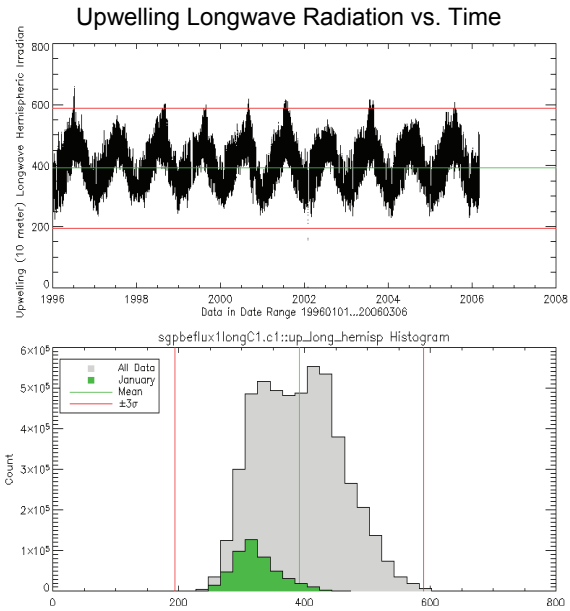

Figure 4. Upwelling longwave radiation exhibits a strong seasonal dependence, as seen in this time-series and frequency distribution graph. The distribution colored green represents values measured in January for each of the 10 years analyzed. The gray colored area represents all months.

3.3 *Detect Abnormal Trends*

Ideally, we wish to detect instrument problems long before those problems begin to affect the quality of the primary scientific measurements. Many data streams include housekeeping, calibration or engineering measurements. Since our analysis system comprehensively processes all time varying fields in the data streams, we can use the statistics gathered on the ancillary data to spot looming instrument problems. Figure 5 is an example of processing many years worth of shortwave responsivity from an infrared detector through our system. The multimodal frequency distribution is a dead giveaway that something is not quite right. The current daily inspections of such data do not always flag such subtle changes in the housekeeping data. By maintaining a record of the long-term trends and typical data ranges, the analyst looking at daily fluctuations will have a much easier time telling whether or not those fluctuations are indicative of a data quality issue.

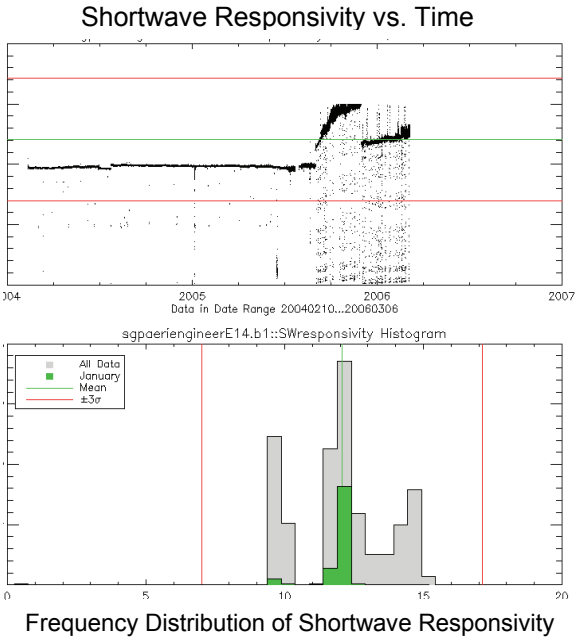

Figure 5. Multimodal distributions signify trend changes that may warrant investigation of possible data quality issues.

4. CONCLUSION

The analysis system described in this paper generates long time-series plots, frequency distributions, and other relevant statistics for scientific and engineering data in most high-level, publicly available ARM data streams. Furthermore, frequency distributions categorized by month or by season are made available to help define 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 early detection of instrument malfunction. The statistics and proposed limits are stored in a database for easy reporting, refining, and for use by other processes. Web-based applications to view the results are also available.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

Macduff, M.C., 2005: ACRF Data Collection and Processing. *21st International Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*. **17.1**

Sonntag, K.L., 2003: Automated Quality Control of Atmospheric Radiation Measurement Program (ARM) Data from the Southern Great Plains (SGP), North Slope Alaska (NSA), and Tropical Western Pacific (TWP) Cloud and Radiation Testbed (CART) Sites. *12th Symposium on Meteorological Observations and Instrumentation.* **6.2**