Automated Data Quality Assessment for the ARM Data System

M. Fisk
Mission Research Corporation
Newington, Virginia

D. H. Sowle and D. Terry
Mission Research Corporation
Santa Barbara, California

Introduction

The U.S. Department of Energy (DOE) Atmospheric Radiation Measurement (ARM) Data System has become a complex system for collecting, processing, archiving, and disseminating a wide range of data from the Southern Great Plains (SGP) Cloud and Radiation Testbed (CART) site, from instruments aboard manned and unmanned aircraft, and other ancillary sources (e.g., satellite imagery and associated data products). While efforts have been made to control and ensure high data quality, many data quality problems have been noted (e.g., Cess, Ellingson, Gautier, and Wiscombe 1997, private communication).

Many ARM Science Team members point out an existing, prevalent need to detect and treat artificial spikes in data, absurd non-zero values (e.g., of liquid water column in clear skies, or short-wave flux at night), and other artifacts that might be spurious outliers. They recommend that data that have obviously been corrupted be removed, flagged, or replaced by a universal bad-data flag. Citing the experience with the total ozone mapping experiment spectrometer (TOMS) ozone data over Antarctica, in which interesting data was removed as outliers, many scientists are reluctant to simply remove spurious data. Fortunately, the netCDF capabilities used to format the ARM data allow for a data quality indicator to be incorporated with the data, rather than removing or altering the data.

In this paper, we demonstrate the feasibility of a statistical outlier approach by analyzing multifilter rotating shadowband radiometer (MFRSR) data. Work at both the Pacific Northwest National Laboratory and the Atmospheric Sciences Research Center suggests that "standard lamp calibrations are not as reliable as originally thought" (Weseley 1997, private communication), hence, data quality is a concern. Our primary objective is to provide automated data quality assessment and flagging procedures, including related summary metrics. Our work complements existing ARM data quality efforts and is coordinated with the ARM science team, instrument mentors, site scientists, and the staff of the archive and experiment center. Our near-term focus is on radiation measurements, particularly spectral radiation and cloud properties.

Application

Using MFRSR broadband irradiance data taken over a 6-month period in 1997, we have 1) applied the usual minimum/maximum checks, 2) applied a statistical outlier test to flag additional anomalous data, and 3) applied an outlier test to further assess anomalous data under a clear-sky hypothesis. The outlier test uses corrections for the solar zenith angle corresponding to atmospheric path length. Training sets of apparently non-anomalous MFRSR data are used to test for outliers. To further refine this outlier set, we form a training set strongly dominated by clear days using inter-comparisons of whole sky imager (WSI) data and MFRSR narrowband data.

Figure 1 illustrates a data calibration issue associated with MFRSR data obtained from the ARM web site. Hemispheric irradiance data from the SGP/E1 site is plotted as a function of Julian day.

Allowing for the seasonal zenith change, it is obvious that the data calibration was altered after day 200. Figure 2 plots the same data versus the solar zenith secant, proportional to the atmospheric path length. Here, the upper curves represent data from selected days in June and July 1997, while the lower data are the obviously recalibrated values from October and November 1997. Note that using the minimum/maximum values of 0 and 1370, as obtained directly from the MFRSR netCDF file, flags only some of the anomalous data. The lower portion of the June/July data, which has an entirely different calibration standard than the October/November data, is not excluded using the simple minimum/maximum check.

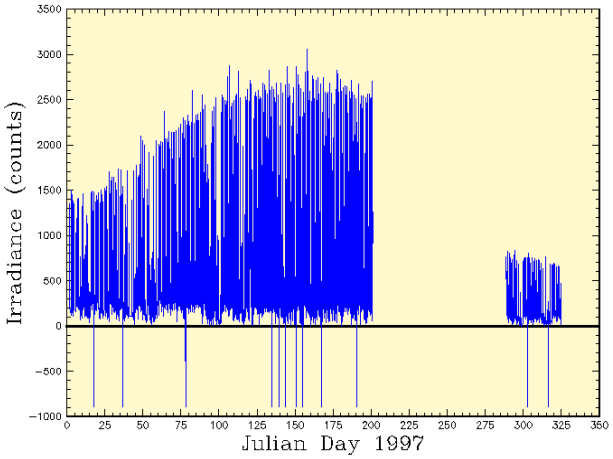

Figure 1. MFRSR hemispheric broadband irradiance data at SGP/E1 for 1997. Plotted here is irradiance, in 10-minute-averaged counts, versus Julian day. (For a color version of this figure, please see *http://www.arm.gov/docs/documents/technical/conf_9803/fisk-98.pdf.)*

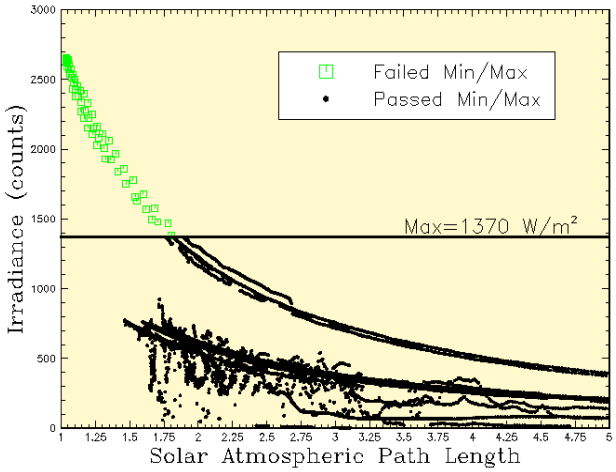

Figure 2. The simple minimum/maximum check detects only some of the anomalous data. Irradiance is plotted versus solar zenith secant, proportional to the solar atmospheric path length. (For a color version of this figure, please see *http://www.arm.gov/docs/documents/technical/conf_9803/fisk-98.pdf.)*

Figure 3 shows the result of the statistical outlier test, using data from the October/November time period as the non-anomalous training set. Here, the outlier test flagged all of the prior uncalibrated data, including that portion of data that had passed the minimum/maximum test. The fact that the recalibrated data from the October/November time period has not really been calibrated to give W/m^2 is not important; the point is that the outlier test has successfully discriminated among data with two completely different calibration standards.

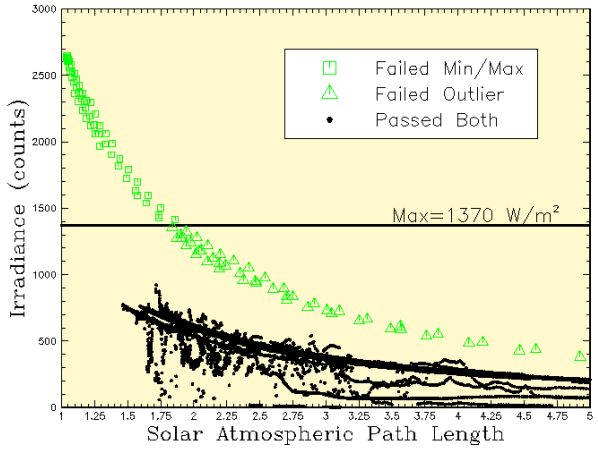

Figure 3. Using the same data, the statistical outlier test flags all of the obviously anomalous data. (For a color version of this figure, please see *http://www.arm.gov/docs/documents/technical/conf_9803/fisk-98.pdf.)*

Clear-Sky Analysis

A training set of clear-sky days is developed to further refine the outlier analysis. The WSI gives the opaque cloud fraction of the entire sky, which can be used to eliminate the heavily clouded days. Figure 4 (top) depicts the opaque cloud fraction for several days during the fall of 1997. The MFRSR hemispheric narrowband irradiance, available in six bands in the visible and near infrared (IR), can be used to further infer clear days. Figure 4 (middle) shows a plot of the narrowband irradiance in several of its bands. Clear days lack the high-frequency structure present in both the cloudy and partly cloudy days. As a direct comparison, Figure 4 (bottom) offers the broadband MFRSR hemispheric irradiance. The clear-sky training set, constructed from Figure 4 (top) and (middle), is then used to test all of the Figure 4 (bottom) MFRSR broadband data for outliers relative to expected observations under clear-sky conditions.

Figure 5 depicts the clear-sky training set. Some of the MFRSR broadband data that were thought to represent clear days, based upon the WSI, are actually flagged as outliers (plotted as triangles). This flagged data may represent thin cloud cover that went undetected by the WSI. Figure 6 shows the result of applying this training set to an expanded set of MFRSR broadband irradiance data from October and November 1997. The outlier analysis identifies datapoints inconsistent with the clear-sky hypothesis.

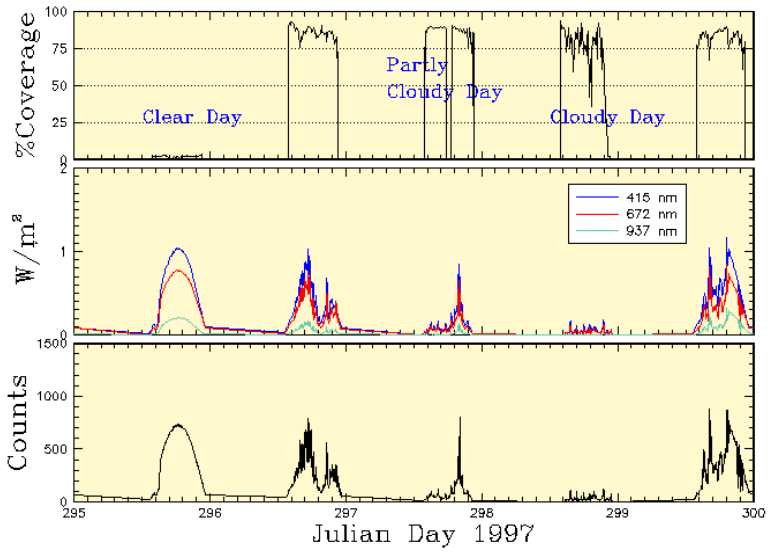

Figure 4. Clear-sky discrimination: (top) WSI opaque cloud fraction over the entire sky; (middle) MFRSR hemispheric narrowband irradiance in three selected visible and near IR bands; and (bottom) MFRSR hemispheric broadband irradiance, to be tested for outliers. (For a color version of this figure, please see *http://www.arm.gov/docs/documents/technical/conf_9803/fisk-98.pdf.)*

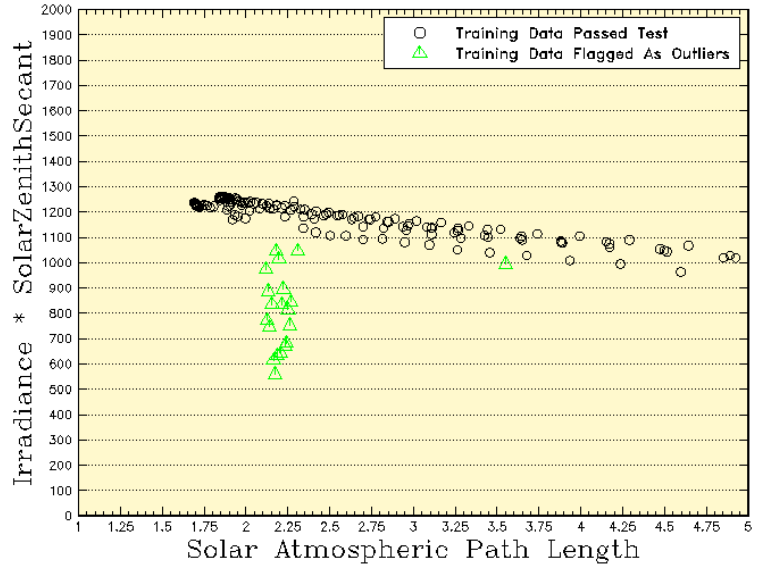

Figure 5. Training set of broadband MFRSR hemispheric irradiance data (counts multiplied by the solar zenith secant) that has been strongly correlated with clear-sky days. Some outliers (triangles) are detected, presumably corresponding to thin cloud cover days. (For a color version of this figure, please see *http://www.arm.gov/docs/documents/technical/conf_9803/fisk-98.pdf.)*

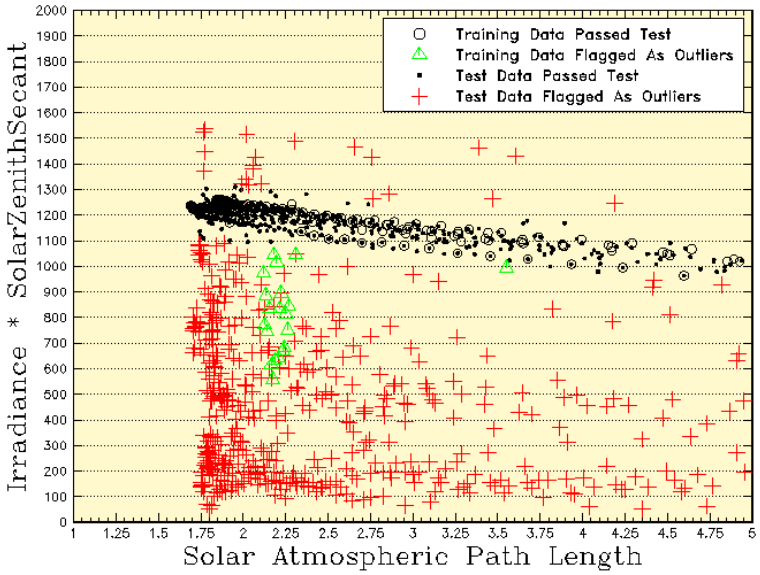

Figure 6. The clear-sky training set is used to test an expanded set of MFRSR broadband irradiance data from October and November 1997. Test data inconsistent with the clear-sky hypothesis are flagged as outliers, and superimposed with training data from the previous figure. (For a color version of this figure, please see *http://www.arm.gov/docs/documents/technical/conf_9803/fisk-98.pdf.)*

Results and Conclusions

Due to data availability of both MFRSR and WSI, we have performed a proof-of-concept study to assess the benefit of automated outlier detection using data from these instruments. While the simple minimum/maximum test, applied to MFRSR broadband irradiance data, will flag some anomalously calibrated data, the statistical outlier test – using validated MFRSR training data – flags all of the miscalibrated data for the time period analyzed. Using the WSI and MFRSR narrowband data to select clear-sky days, the outlier test is further able to flag MFRSR broadband data that is anomalous under the hypothesis of clear-sky conditions. Additional data and optimization are expected to improve these results.

The statistical outlier method employed is applicable to radiation data from other instruments. To complement existing efforts by the instrument mentors and the site scientists, we plan on applying similar data quality assessments to the following: shortwave spectroradiometer (SWS), with comparisons to the rotating shadowband spectrometer (RSS); ground-based radiometer autonomous measurement system (GRAMS), with eventual comparisons to other broadband and spectral measurement instruments; MFRSR and Cimel sunphotometer (CSPHOT) data comparisons, with regard to aerosol optical depth; WSI and micropulse lidar (MPL) comparisons, with regard to cloud fraction determination.

We, therefore, conclude that outlier detection can contribute significantly to the ARM data quality assessment. For applicable cases, the approach is objective and can be automated. Furthermore, the utility of the outlier approach is greatly enhanced by including intercomparisons to ancillary data. As such, the data quality approach must be tailored to the specific instruments.