

Analyst Tools and Quality Control Software for the ARM Data System

Final Report for the period July 2004 through December 2007

Prepared March 2008

Sean Moore
Gary Hughes

Mission Research
ATK Mission Systems
6755 Hollister Avenue Suite 200
Goleta, CA 93117-5571

PREPARED FOR THE UNITED STATES
DEPARTMENT OF ENERGY
ENVIRONMENTAL SCIENCES DIVISION, SC-74
OFFICE OF BIOLOGICAL AND ENVIRONMENTAL RESEARCH

Work Performed Under DOE Grant DE-FG02-04ER63864

UNCLASSIFIED

Abstract

Mission Research develops analyst tools and automated quality control software in order to assist the Atmospheric Radiation Measurement (ARM) Data Quality Office with their data inspection tasks. We have developed web-based data analysis and visualization tools such as the interactive plotting program NCVweb, various diagnostic plot browsers, and a datastream processing status application. These tools allow even novice ARM researchers to be productive with ARM data with only minimal effort. We also contribute to the ARM Data Quality Office by analyzing ARM data streams, developing new quality control metrics, new diagnostic plots, and integrating this information into DQ HandS - the Data Quality Health and Status web-based explorer. We have developed several ways to detect outliers in ARM data streams and have written software to run in an automated fashion to flag these outliers.

We have also embarked on a system to comprehensively generate 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.

Table of Contents

Abstract.....	ii
Table of Contents.....	iii
1 Introduction.....	1
2 Coordination and Collaboration Wiki.....	3
3 Developing Better Quality Control Limits	5
3.1 ARM Statistical Analysis and Reporting System.....	5
3.1.1 Introduction.....	5
3.1.2 System Design and Method	5
3.1.2.1 Data Retrieval and Import.....	5
3.1.2.2 Data Acceptance and the Analysis Queue	6
3.1.2.3 Statistical Processing and Reports	7
3.1.2.4 Analysis Review and Feedback	7
3.1.3 Results.....	8
3.1.3.1 Improve Existing Limits	8
3.1.3.2 Implement Monthly Limits.....	9
3.1.3.3 Detect Abnormal Trends.....	10
3.2 Measurement Comparison Databases.....	11
4 New Automated Analysis Modules	17
4.1 Radar Wind Profiler (RWP).....	17
4.1.1 Background.....	17
4.1.2 Quality Analysis of RWP Data.....	19
4.1.2.1 RWP Analysis Code Output	20
4.1.2.1.1 QC Metrics.....	20
4.1.2.1.2 Diagnostic Plots	22
4.1.3 Quality Analysis of RASS Data.....	23
4.1.3.1 RASS Analysis Code Output.....	25
4.1.3.1.1 QC Metrics.....	25
4.1.3.1.2 Diagnostic Plots	26
4.2 Best Estimate Flux Value Added Product (BEFLUX VAP)	28
4.2.1 Background.....	28
4.2.2 Quality Analysis of BE Flux Data	28
4.2.2.1 BEFLUX Analysis Code Output	32
4.2.2.1.1 QC Metrics.....	32
4.2.2.1.2 Diagnostic Plots	33
4.3 Tower Water-Vapor Mixing Ratio Value Added Procedure (TWRMR VAP)	34
4.3.1 Background.....	34
4.3.2 Quality Analysis of TWRMR Data.....	35
4.3.2.1 Min / Max Checks.....	35
4.3.2.2 Sensor Cross-Comparison Checks.....	38
4.3.2.3 Sonde Cross-Comparison Checks.....	38
4.3.2.4 TWRMR Analysis Code Output.....	39

4.3.2.4.1	QC Metrics.....	39
4.3.2.4.2	Diagnostic Plots	41
5	Interactive Web-Based Tools.....	42
5.1	DQ HandS Plot Browser.....	42
5.2	NCVweb	43
5.3	Datastream Processing Status on DQ Computer	45
6	Automated Check and Alert System.....	46
7	Conclusion	48
8	Acronyms	49
9	References.....	50

1 Introduction

The Department of Energy (DOE) Atmospheric Radiation Measurement (ARM) program has acquired and continues to acquire an incredibly large quantity of data during its normal process of operation, all of which must be reviewed in some manner in order to ensure that the data is of "known and reasonable" quality¹. The thoroughness of this quality review will largely determine to what extent ARM data is used and how well the ARM program is respected within the research community. So it is with the utmost importance that sufficient software exists to quality-check data products in a timely and continuous fashion. There is also an ongoing need to find ways to increase the utility of ARM data among researchers and convey its usefulness to the public.

We believe we have made much progress in addressing these needs by providing easy-to-use, interactive analyst tools as well as several automated quality control software modules for the ARM Data Quality system. We have approached the problem by breaking the work into the following tasks:

1. Coordination with the Data Quality Office to identify data products requiring further quality control.
2. Developing better bounds for data values by: analyzing historical data, applying models, and consulting with experts.
3. Improving upon current min/max/delta quality criteria by implementing new rule-based, statistical based and cross-comparison based measures.
4. Developing new data analysis tools, specific to ARM data sets.
5. Developing a system that reassesses the quality of instrument level data streams based on information provided by value added products.

Coordination with the ARM data quality office has been accomplished by way of face-to-face meetings, frequent phone conversations, and various computer-assisted collaboration systems. We have participated in yearly science team meetings, developer meetings, and ARM working group meetings. We've also met with members of the instrument team in order to understand the peculiarities and performance characteristics of their specific instrumentation.

To facilitate coordination and collaboration among the data quality office personnel, ARM infrastructure, and instrument team, we have implemented and hosted a wiki system on a computer accessible to the public internet. Details on this system will be provided in Section 2.

In order to improve QC limits, we have reviewed and summarized many years of data to get a good understanding of what is typical. After manually compiling several climatological databases from sondes and other instruments, we set out to build a system that would methodically process the most commonly used datastreams, and keep tabs on the statistical results in a easy to access database. These climatologies and the system used to produce them will be described in Section 3.

Our new automated analysis modules developed during this period of performance have made good use of these statistical databases and will be further described in Section 4.

Ideally, all quality control checks would be automated and summarized for quick review by the data quality office each day. But until that dream is fully realized, it is important to have on hand a suite of interactive tools for analysts to use for inspection and visualization of potential data problems. We describe some of our latest tools that serve this purpose in Section 5.

Lastly, in Section 6 we describe the system we have to reassess instrument level data based on information provided by value added products.

2 Coordination and Collaboration Wiki

To facilitate sharing of information among Data Quality Office staff, Mission Research and other ARM infrastructure personnel, we have implemented and hosted a wiki system on one of our in-house computer systems². A Wiki is like a normal web site, except that the user can edit the content from within their web browser. Wiki is short for "wiki wiki", the Hawaiian word for "Quick". A Wiki is basically a shared, online, persistent whiteboard, in which anyone can add content, or change what is written, or change the organization of the content. Whatever the user enters using a simple web form is nicely presented, without use of any hypertext markup language know-how. The documentation is remembered, version-controlled, and never forgotten. Also, Web pages are linked to each other automatically -- no more "404 page not found" errors. Users create new pages by simply entering a topic name and filling out the details.

We based our wiki on an open-source, enterprise-class collaboration platform and knowledge management system known as TWiki³. We've customized the TWiki system to support the specific needs of the ARM data quality office (Figure 1). This system allows all data quality office personnel, instrument mentors, and other interested parties, to collectively edit documents related to instrument data quality. Each web page may be edited by any member, and may include graphics, tables, or arbitrary attachments. The pages are fully searchable and users may elect to be notified whenever there is a change.

The screenshot shows a wiki page titled "Atmospheric Profiling" on the ARM Data Quality Office wiki. The page layout includes a top navigation bar with "DQdocs > InstrumentGuidance > AtmosphericProfiling" and action links: "Edit this page", "Attach a file", "Printable version", and "More...". The main content area features a list of instrument links such as "G-Band Vapor Radiometer (GVR) (NSA)", "Profiling G-Band Vapor Radiometer (GVRP) (NSA)", "Microwave Radiometer Profiler (MWRP) (NSA & AMF)", "Microwave Radiometer (MWR) (AMF, SGP, TWP, NSA)", "High Frequency Microwave Radiometer (MWRHF) (NSA)", "Radar Wind Profiler (RWP) (NSA & SGP)", "Balloon-Borne Sounding System (SONDE) (NSA, SGP, TWP, AMF)", "Raman Lidar (RL) (SGP)", "Vaisala Ceilometer (VCEIL) (SGP, NSA, TWP, AMF)", "Micropulse Lidar (MPL) (SGP, NSA, TWP, AMF)", and "Surface Temperature and Humidity Reference System (SURTHREF) (SGP)". Below this is an "Instrument Page Shortcut" section with a long list of instrument abbreviations. At the bottom of the main content, it says "-- KenKehoe - 07 Sep 2005".

The left sidebar contains a search box for "Search DQdocs web", "Switch Webs" (DQ docs, User List, Sandbox), "DQdocs tools" (Recent changes, Topic list, Verbose topic list, Notify me of changes, Web preferences, Search Web, Advanced Search), and "TWiki" (Text formatting rules, TWiki documentation, Global preferences).

The footer of the page shows the revision history: "Revised: r1.6 - 18 Mar 2008 - 15:58 - KenKehoe" with links for "Edit", "Attach", "Printable", and "More...". Below that are links for "Referrers to this page", "View raw text", "Revision history", and a range of revisions "r1.6 | > | r1.5 | > | r1.4". In the bottom right corner, it says "You are logged in as: SeanMoore".

Figure 1: Sample page from the ARM Data Quality Office wiki. Notice user's login displayed in lower right corner, and links to view 'Revision History', 'Edit' current page, 'Attach' files, or show the page in a 'Printable' view. A search box is in the upper left, and quick access to various tools is available from the left side navigation menu.

TWiki also acts as an "application platform" to integrate a number of other functions. For example, we added a calendar plug-in to track and monitor DQ activities and travel.

Each page of the wiki is organized by a wiki word in CamelCase⁴, where the words are joined without spaces and each word is capitalized. This allows easy creation of new pages, or links to existing wiki pages without additional programming syntax.

The ARM Data Quality Office (DQO) processing system has grown to include many different programs and scripts to manage the large amount of processing and files. The Wiki pages have been an immense help as a way to quickly bring together documentation from a wide variety of sources. The discussion space has become an important area to ensure everyone is properly informed, and a notification system alerts users of any recent changes.

3 Developing Better Quality Control Limits

When possible, experts familiar with the instrumentation employed by the ARM 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.

3.1 ARM Statistical Analysis and Reporting System

3.1.1 Introduction

In order to comprehensively define limits for all produced data streams, we are developing a tool that systematically reviews the entire historical record of measurements⁵. This system is called the ARM Statistical Analysis and Reporting System, or ARM*STAR.

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.

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

3.1.2 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.

3.1.2.1 Data Retrieval and Import

ARM data is typically stored in daily NetCDF files and warehoused at the ARM Archive⁶. 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.

3.1.2.2 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. 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 2 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.

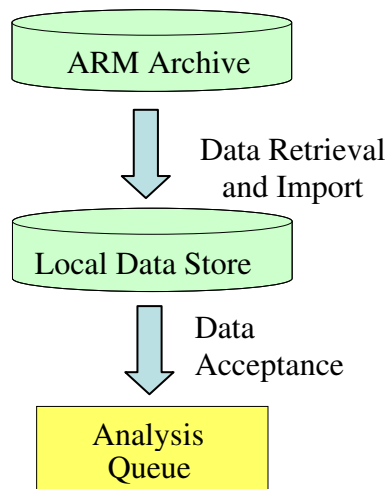


Figure 2: Data Flow into ARM*STAR from the Archive

3.1.2.3 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.

3.1.2.4 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 3 provides a concise summary of the processes involved that make up the statistical analysis system.

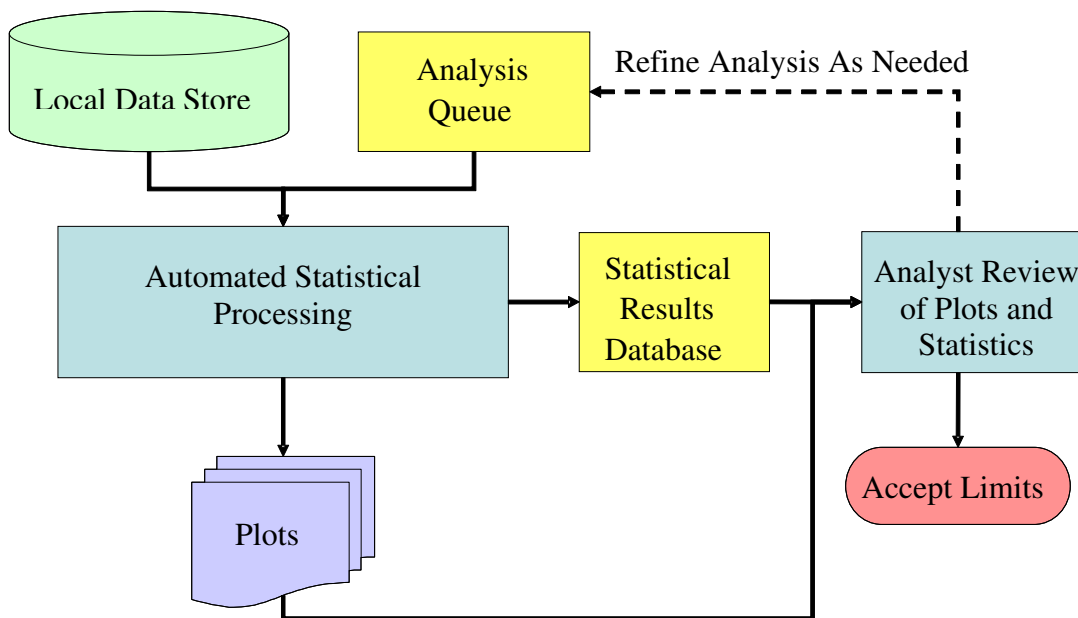


Figure 3: Automated Statistical Analysis System Block Diagram.

3.1.3 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.

3.1.3.1 Improve Existing Limits

Figure 4 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 the figure 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.

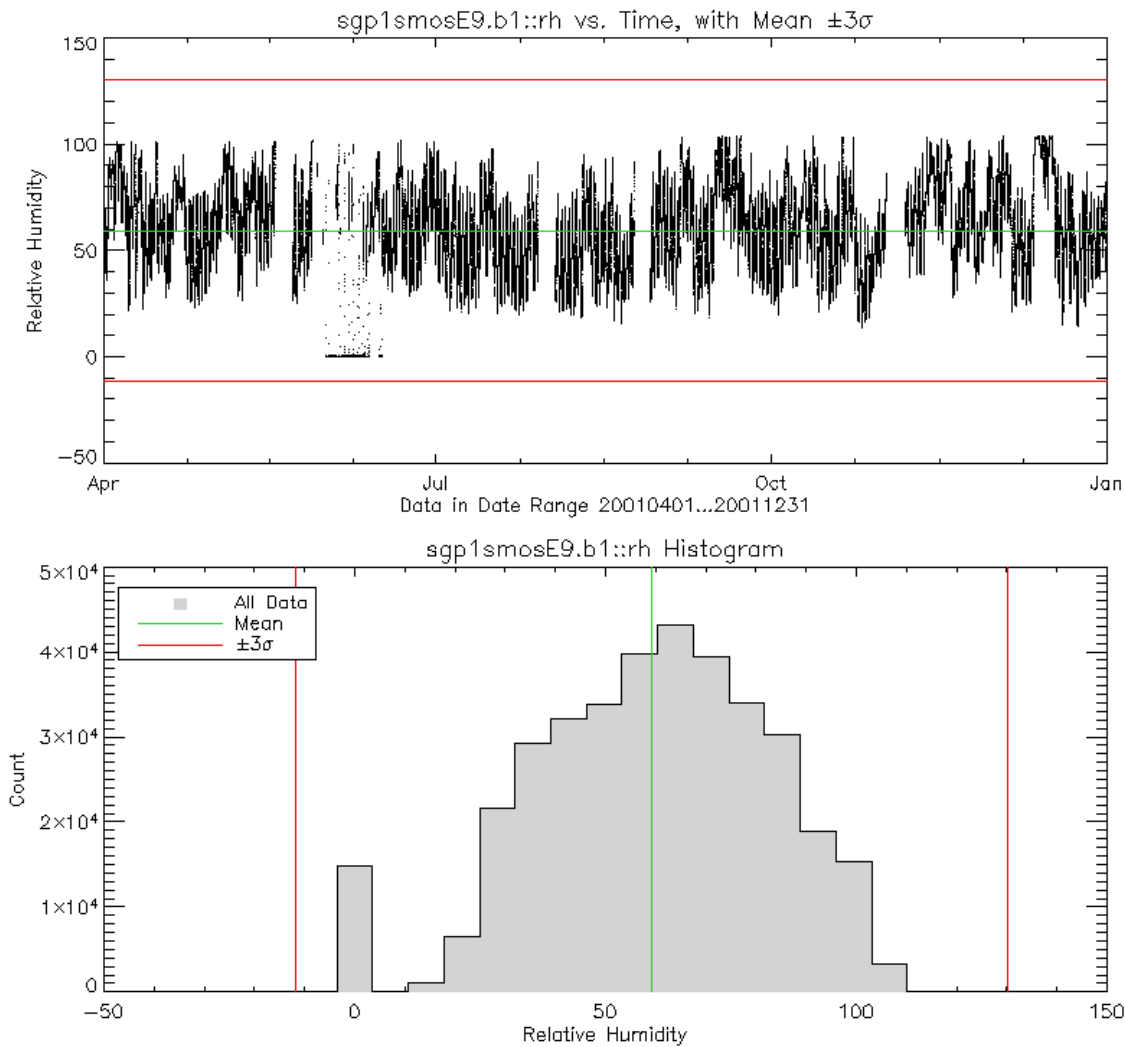


Figure 4: Application of analysis system to improve existing data quality limits. Data displayed is from the Surface Meteorological Observation System (SMOS) of the Southern Great Plains ARM Climate Research Facility.

3.1.3.2 Implement Monthly Limits

Many Value Added Products (VAPs) produced by ARM do not have valid data ranges defined. Figure 5 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.

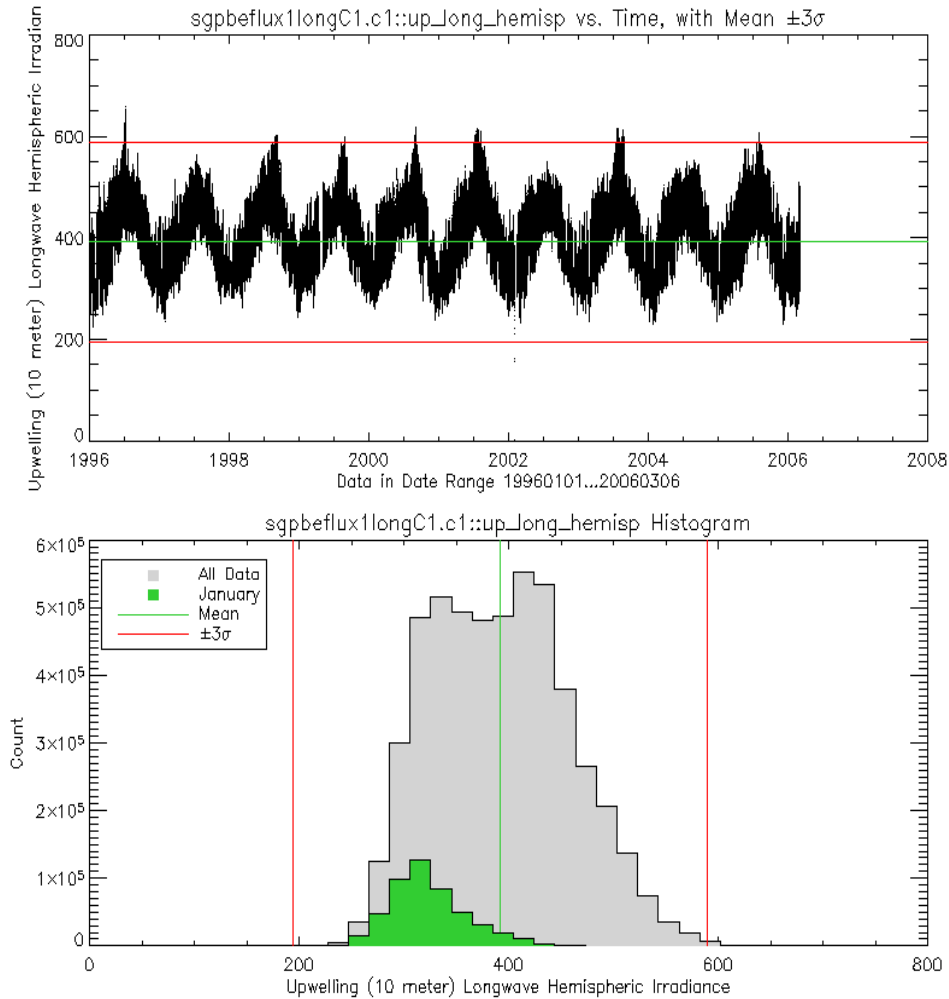


Figure 5: Upwelling longwave radiation exhibits a strong seasonal dependence, as seen in this time-series and frequency distribution graph. The distribution colored green represents values measured in January for each of the 10 years analyzed. The gray colored area represents all months.

3.1.3.3 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 6 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.

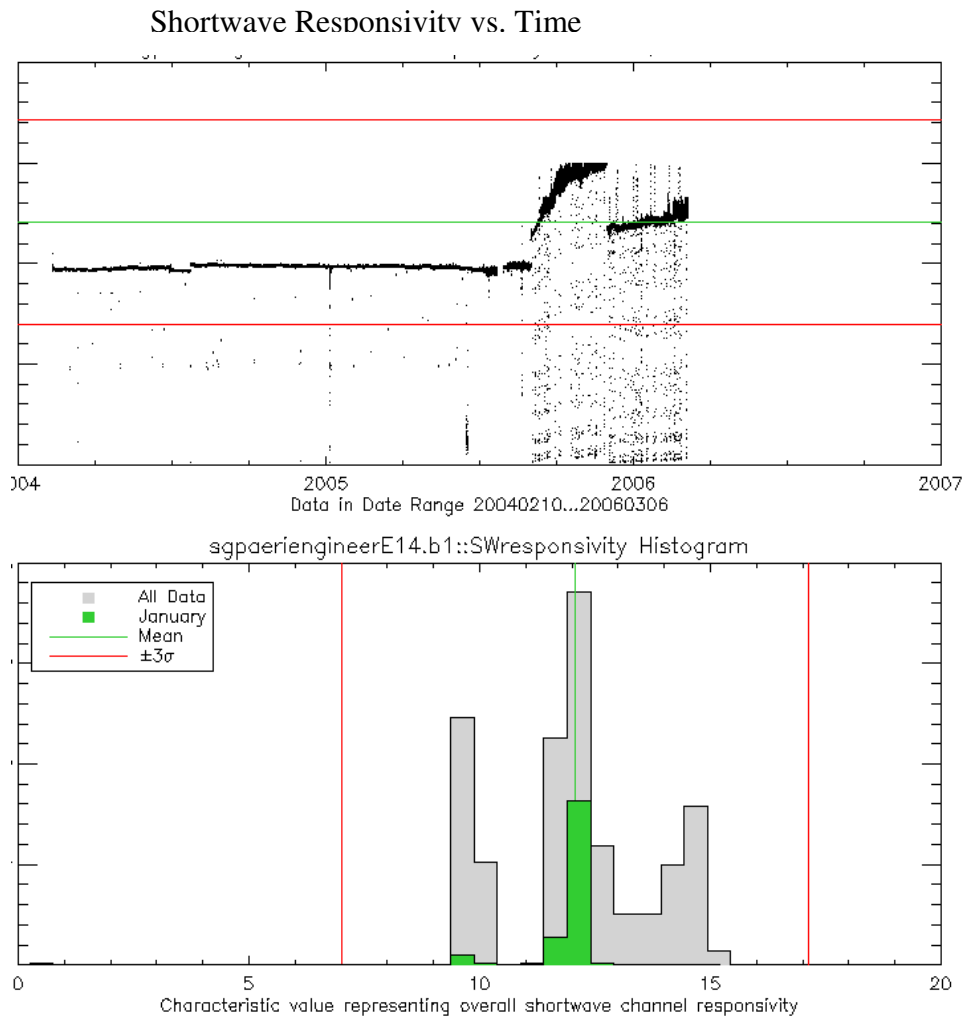


Figure 6: Multimodal distributions signify trend changes that may warrant investigation of possible data quality issues.

3.2 Measurement Comparison Databases

One technique for checking the quality of measurements is to compare those measurements to similar sources of data. Similar sources of data can be found from related instrumentation at nearby locations, models or retrievals of the measurement in question, or from past measurements taken under similar circumstances from the same instrument.

Within the Atmospheric Radiation Measurement instrument network, several different systems often measure the same quantity at the same site. For example, several ARM instruments measure time-series profiles of the atmosphere that were previously available

only from balloon-borne radiosonde systems. These instruments include the Radar Wind Profilers (RWP) with Radio-Acoustic Sounding Systems (RASS), the Atmospheric Emitted Radiance Interferometer (AERI), the Microwave Radiometer Profiler (MWRP), and the Raman Lidar (RL). ARM researchers have described methods for direct cross-comparison of time-series profiles and we have extended this concept to the development of methods for automated quality control (QC) of ARM datastreams.

The current data flow architecture at the Data Management Facility (DMF) does not allow direct cross comparisons to other data streams during the generation of quality checked (b-level) instrument data files. At the moment when QC checks are performed, the availability of comparison data streams cannot be guaranteed. In order to perform automated QC at the time of data ingest, another method is required. One method is to compare key values against historical ranges.

Using our techniques for processing and combining data over long time periods, we have built a sonde-based climatology database for each of the fixed ARM sites (NSA, SGP, TWP) plus the Barrow NWS site. These databases are organized in convenient NetCDF data files and contain monthly ranges, distributions, lapse rates, means and other statistical measures. We do this for: wind speed and direction (NSA and SGP), temperature, dew point, relative humidity, and calculated water vapor mixing ratio. The statistics for each quantity are computed and stored by month and by 50 meter altitude bin, up to about 25 km.

Since several high-level products attempt to retrieve these quantities, this database can be used to set reasonable bounds for the month and height of interest. Details on this approach will be presented in following sections.

Sample plots showing monthly averages for temperature and water vapor mixing ratio over the SGP are shown in Figure 7. Sample monthly statistics for wind speed are shown in Figure 8. Sample monthly temperature profiles for January and July are shown in Figure 9. Sample monthly profile statistics for water vapor mixing ratio are shown in Figure 10. Statistics for wind speed are based on 5,068 sonde balloon launches between 01 April 2001 and 15 October 2004. All other quantities are based on 10,040 sonde launches between 11 April 1994 and 07 June 2004.

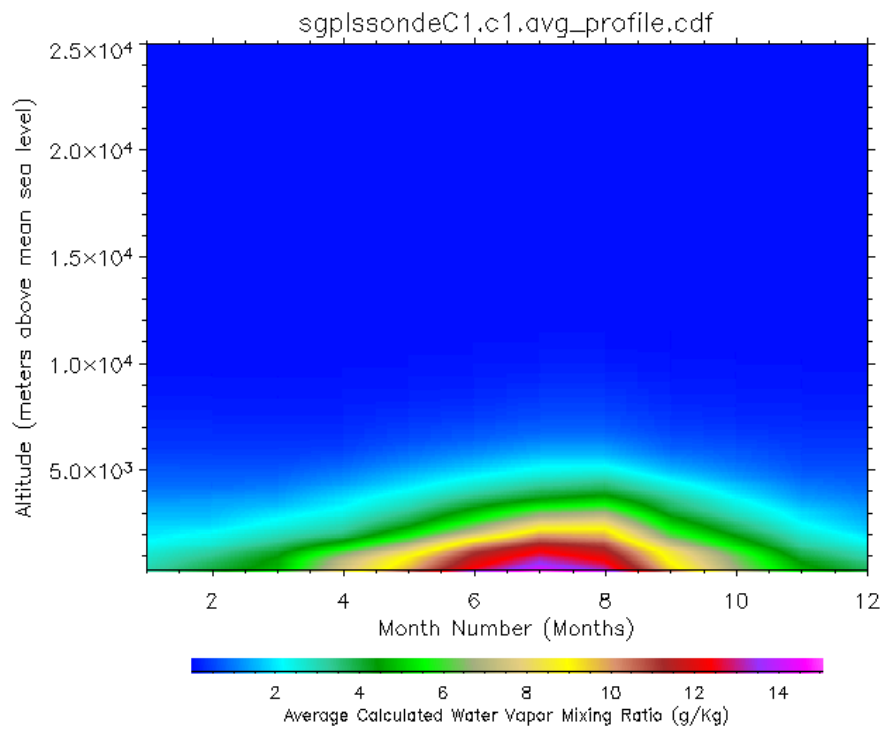
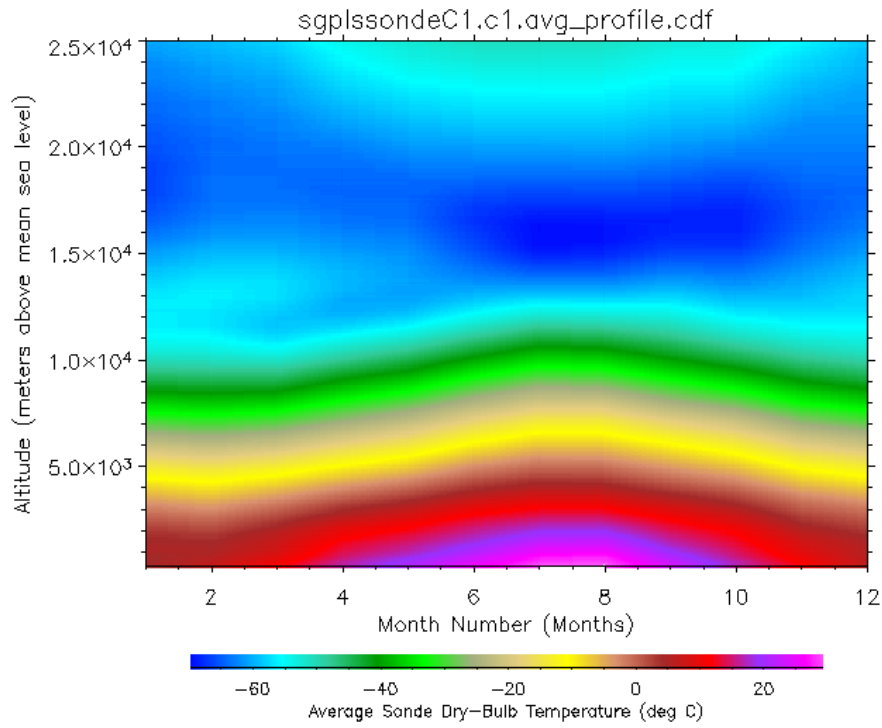


Figure 7: Average sonde dry-bulb temperature (top) and average calculated water vapor mixing ratio (bottom) for each month and each altitude bin over the SGP site. Results based on 10,040 sonde launches between 11 April 1994 and 07 June 2004.

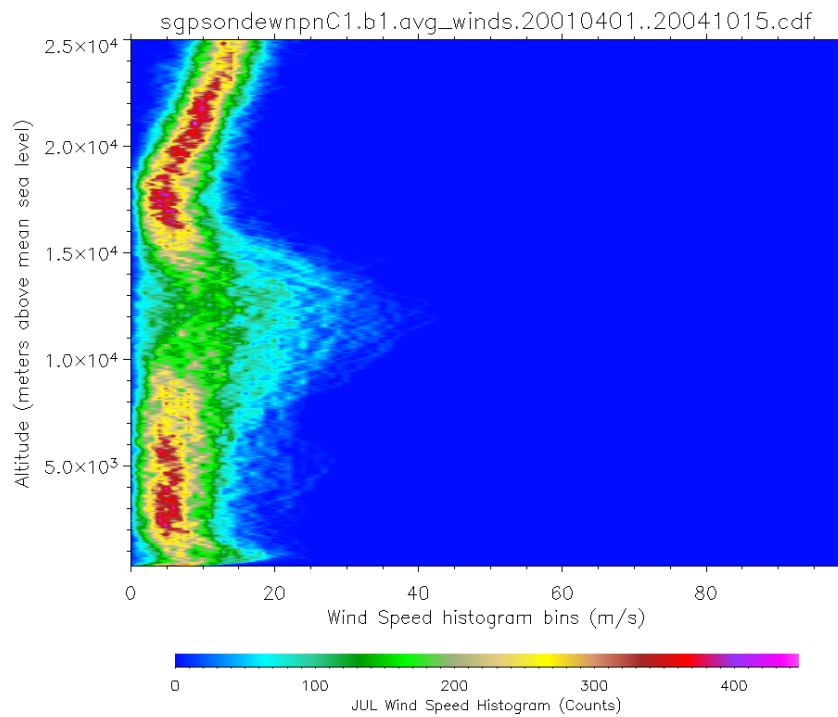
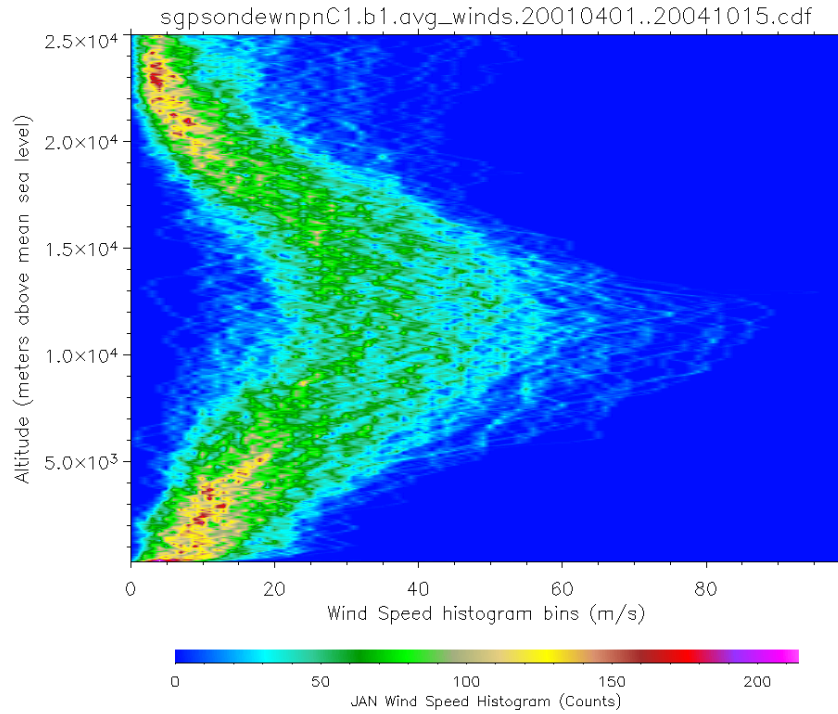


Figure 8: Sample monthly statistics for SGP wind speed profiles were calculated for 50 meter altitude bins based on 5,068 sonde balloon launches between 01 April 2001 and 15 October 2004. Displayed are profiles for the months of January (top) and July (bottom).

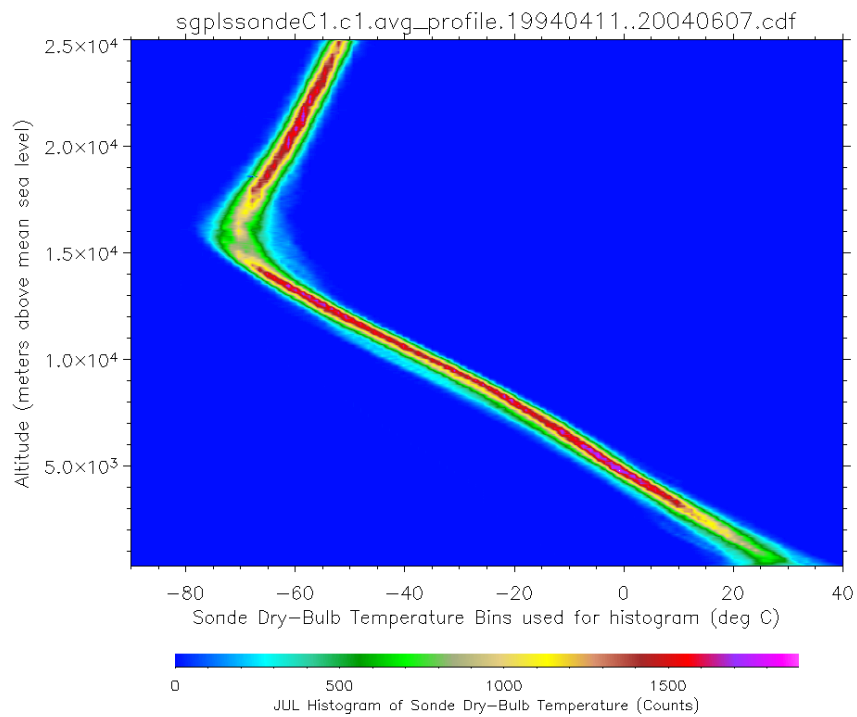
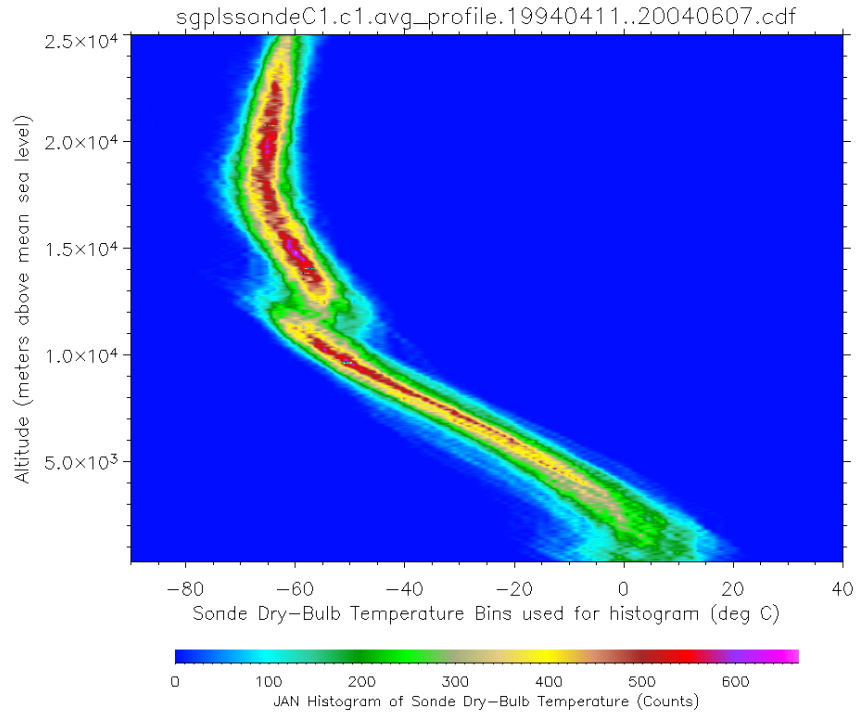


Figure 9: Monthly statistics for SGP temperature profiles were calculated for 50 meter altitude bins based on 10,038 sonde balloon launches between 11 April 1994 and 07 June 2004. Displayed are profiles for the months of January (top) and July (bottom).

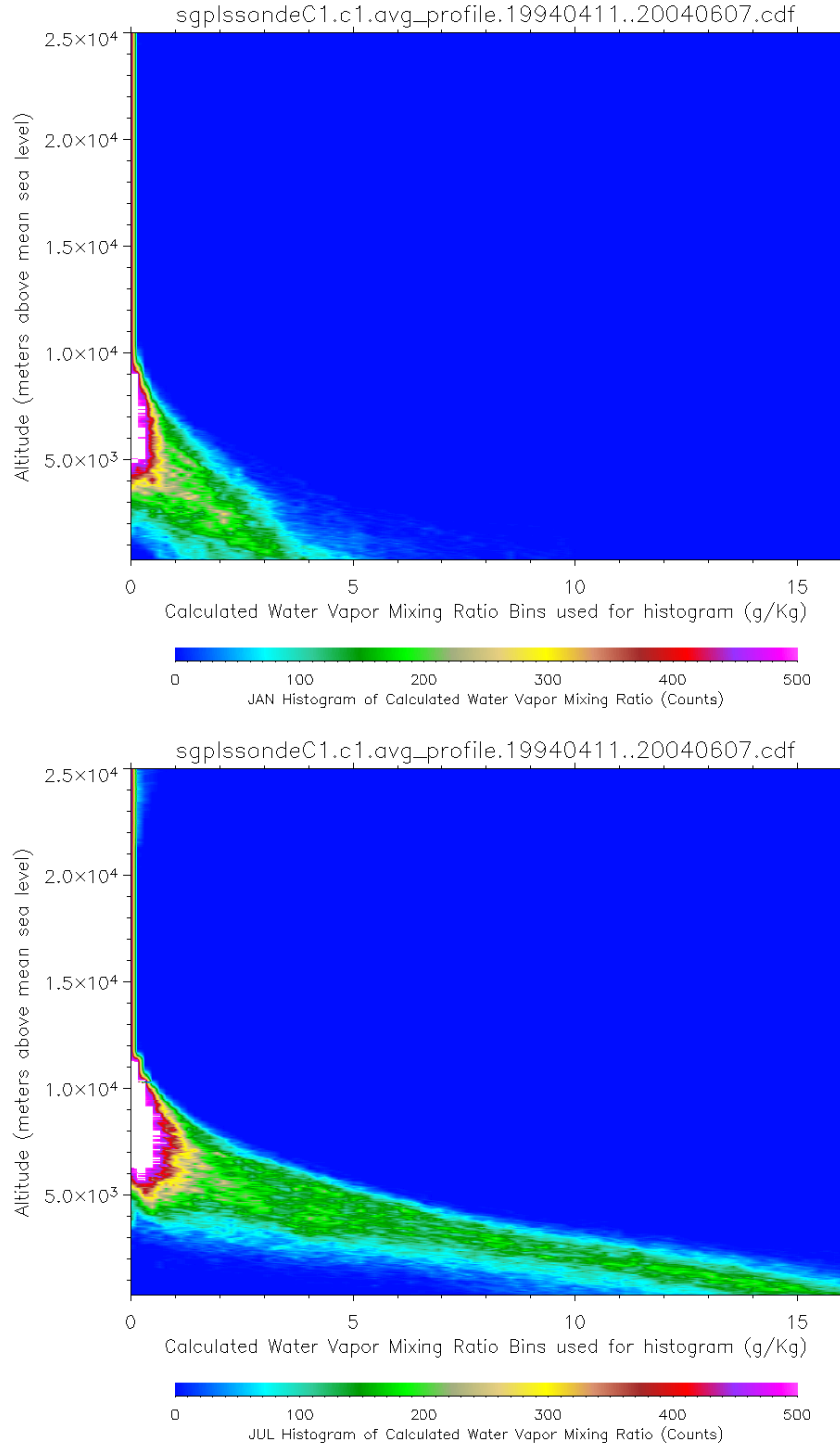


Figure 10: Monthly statistics for SGP water vapor mixing ratio profiles were calculated for 50 meter altitude bins based on 10,038 sonde balloon launches between 11 April 1994 and 07 June 2004. Displayed are profiles for the months of January (top) and July (bottom).

4 New Automated Analysis Modules

We developed automated Quality Assurance processing modules for the Radar Wind Profiler and the Radio-Acoustic Sounding System instruments. We also developed modules for the Tower Water-Vapor Mixing Ratio Value-Added Product (VAP), the Best Estimate Flux VAP, and the QC Flux VAP. The output from these modules is fed into DQ HandS for review by DQO analysts. None of these VAPS have defined valid data ranges, so we analyzed historical data to produce our own limits. A complete description of these processing modules is provided in the following sections. Hourly statistics for missing, failing and not available values are calculated for display in DQ HandS. Daily diagnostic plots of every field in the VAPS are generated for review using the DQ HandS plot browser.

4.1 Radar Wind Profiler (RWP)

4.1.1 Background

The Radar Wind Profilers (RWP50 and RWP915) are designed to provide a time series of vertical profiles of wind velocity (horizontal speed and direction). The RWPs operate by transmitting electromagnetic energy into the atmosphere and measuring the strength and frequency of backscattered energy from refractive index fluctuations that are moving with the mean wind. By sampling in the vertical direction and in two tilted planes, the wind components can be determined. The system consists of a single phased array antenna that transmits alternately along five pointing directions: one vertical, two in the north-south vertical plane (one south of vertical, one north of vertical), and two in the east-west vertical plane (one east of vertical, one west of vertical). The non-vertical beams are tilted at about 14 degrees from vertical.

The primary quantities measured with the system are the intensity and Doppler frequency shift of backscattered radiation. The wind speed is determined from the Doppler frequency shift of energy scattered from refractive index fluctuations (caused primarily by temperature fluctuations) embedded within the atmosphere. The wind speed is derived from measurements from the five beams. Since the individual components are not collocated in space, horizontal homogeneity is assumed in order to derive the wind vector as a function of height. The 50-MHz Radar Wind Profiler (RWP50) measures wind profiles from (nominally) 2 to 12 km. The 915-MHz Wind Profiler (RWP915) measures wind profiles from (nominally) 0.1 km to 5 km. Range gates for all RWP modes are shown in Table 1. The accuracies of these quantities, while dependent upon the accuracy of the frequency measurement, are also affected by atmospheric effects and vary considerably according to conditions. Nominal accuracy for wind speed is 1 m/s, and nominal accuracy for wind direction is 3 deg.

The Radio-Acoustic Sounding System (RASS) subsystem of each RWP provides time series of virtual temperature profiles and vertical wind velocity profiles. In the RASS

mode, virtual temperatures are recovered by transmitting an acoustic signal vertically and measuring the electromagnetic energy scattered from the acoustic wavefront.

In normal RWP and RASS operation, virtual temperature and vertical wind speed profiles are determined during the first 10 minutes of every hour and the wind profile is averaged over the remaining 50 minutes.

It takes, nominally, 30 - 45 seconds (dwell time) for the RWP to determine the radial components from a single pointing direction. The system cycles through five beams (South, North, East, West, and vertical) at low power, and then cycles the five beams again at a high power (longer pulse length) setting. Then the whole process is repeated. About five minutes elapse before the system returns to the beginning of its sequence. Within an averaging interval, the estimates from each beam-power combination are saved (11-12 in a 1-hr period) and these values are examined and compared at the end of the period to determine the consensus-averaged radial components of motion.

Consensus averaging consists of determining if a certain percentage (e.g., 50%) of the values fall within a certain range of each other (e.g., 2 m/s). If they do, those values are averaged to produce the radial wind estimate. The radial values are then combined to produce the wind profile.

The system transmits pulses at about 1-10 kHz rate into the atmosphere. The backscatter from each transmit pulse is sampled at, for example, a 1 MHz rate. This results in 1 sample every 150 m in range. The samples at each range gate are averaged together (time domain integration) over some number (e.g., 100) of pulses to produce a phase value for input into a FFT. After (e.g., 64) values are produced, the FFT is performed (one for each range gate). This process takes on the order of 1 sec. A number (about 30) of these spectra are then averaged together during the dwell time. At the end of the dwell time there is produced a single averaged spectrum from each range gate along the designated pointing direction.

The spectra are analyzed by the system before moving to the next pointing direction. This analysis produces estimates of the Signal to Noise Ratio (SNR), the noise, the mean velocity (proportional to frequency), and the first moment (spectral width) at each range gate.

RASS operation is essentially the same, except that the averaging time is about 10 minutes and only a single pointing direction (vertical) is used. Also, the atmosphere is "seeded" with a sound wave; the index of refraction changes created by the sound wave are the signal source. In order to sample both the sound wave (speed about 340 m/s) and the atmosphere (to remove air velocity from temperature estimates) a larger FFT is required (2048 points). This requires a smaller number of points for each time domain integration and increases the processor time required to calculate the FFT. Only a portion of the spectra are reported, namely a region near 0 Doppler shift to account for atmospheric motions and a region around the expected speed of sound. The "moments"

files consist of moments and widths from both the atmospheric portion of the spectrum and from the acoustic portion (the main contributor to the temperature calculation).

4.1.2 Quality Analysis of RWP Data

One approach for quality analysis of RWP a1-level data is to make an instrument cross-comparison check with sonde wind data for the main RWP data products. We use historical sonde files to establish nominal ranges and nominal lapse rates for wind velocity. The RWP profiles, which are based on remotely sensed data, are tested against the ranges derived from sonde data, which are based on direct measurements (a “ground-truth” assessment). If a RWP profile predicts a value that is outside the sonde range, then there is a defensible reason to raise suspicions about the RWP value⁷. The sonde ranges for wind velocity and wind velocity lapse rate are established by calculating the monthly means and standard deviations within 50m altitude bins. Example wind speed distributions for two different months over the SGP site are shown in Figure 8.

In addition to sonde cross-comparison checks, quality analysis of RWP a1-level data includes the following checks:

- (1) NOAA pressure-level dependent maximum velocity
- (2) ANL check for contamination from migrating birds
- (3) ANL check for 60Hz noise
- (4) ANL check for presence of rain
- (5) ANL check for validity of First Range Gate
- (6) ANL check for Mirrored Signal in Upper Profile

The RWP DQ Look-Up Table (LUT) is created from historical sonde files and historical RWP files. First, the sonde climatology is created and stored in a stand-alone NetCDF file. Then, the RWP LUT is made from a combination of the sonde climatology, and historical RWP data files in a separate step.

For automated quality analysis of RWP, each data value in the retrieved RWP profile is compared to the historical sonde range for the month of observation and corresponding altitude bin. If the RWP value falls outside of the sonde range $\bar{x}_{sonde} \pm 4.0 \cdot \sigma_{sonde}$ then the RWP value is flagged as failing max/min check (Figure 11). If the calculated RWP lapse rate falls outside the same range for average sonde lapse rates, then the RWP value is flagged as failing delta check.

Hourly statistics for missing, failing and not available values are calculated for display in DQ HandS. The total number of failing values for any given hour is divided by the total number of observations during that hour to give an hourly percentage of failing values. Statistics are also calculated for missing and not available values and reported to DQ HandS via the usual metrics file.

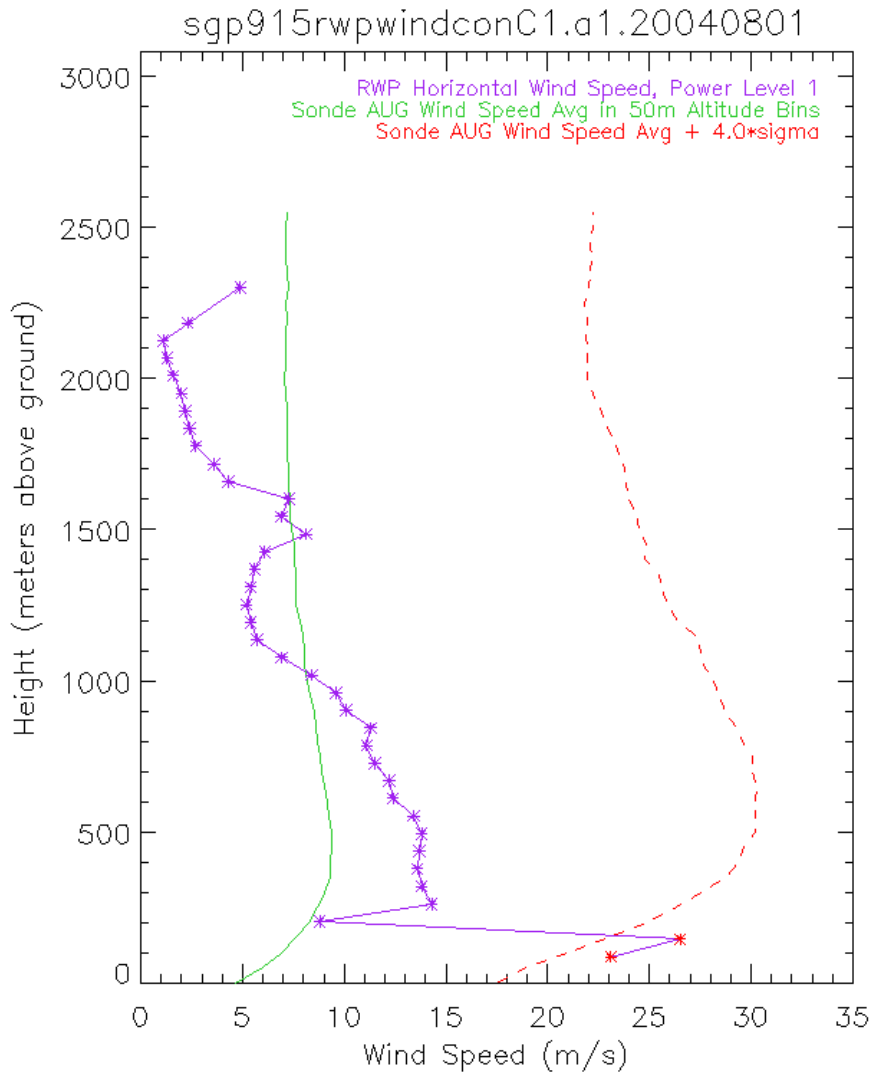


Figure 11: For August, the solid green line depicts the average wind speed profile as reported by the sondes at the SGP site. The red dashed line marks the spread of sonde values out to four standard deviations from the average. On 1 August 2004, RWP wind speed data (purple asterisks) were compared against the sonde climatology, and out of bound values were flagged (red asterisks).

4.1.2.1 RWP Analysis Code Output

4.1.2.1.1 QC Metrics

Statistics are compiled for each hour of the 24-hour period covered by the RWP NetCDF file. The statistics are written to a data file that is compatible with DQ HandS Explorer. This data file contains statistics for percent of values failing, percent of values missing, fail times for missing or failing values, and a quality flag indicating which quality criteria were violated during each hour. The percent failing statistic combines all values that are

outside the ‘good’ interval. The failure code indicates which quality criteria were violated within the hour, bit-packed with the following values:

- 0 => 'missing data value (-9999)',
- 1 => '< MIN Based on historical distribution for each variable',
- 2 => '> MAX Based on historical distribution for each variable',
- 3 => '> DELTA Based on historical distribution for each variable',
- 4 => 'NOT AVAILABLE',
- 5 => '< NOAA Minimum for pressure level of observation',
- 6 => '> NOAA Maximum for pressure level of observation',
- 7 => '< SONDE Minimum Wind for month and height of observation',
- 8 => '> SONDE Maximum Wind for month and height of observation',
- 9 => '> SONDE Maximum Wind DELTA for month and height of observation',
- 10 => 'Failed ANL check for contamination from migrating birds',
- 11 => 'Failed ANL check for 60Hz Noise',
- 12 => 'Failed ANL check for presence of Rain',
- 13 => 'Isolated Measurement (questionable value)',
- 14 => 'First Range Gate Value Questionable',
- 15 => 'Mirrored Signal in Upper Profile',

Summaries of the hourly statistics are then displayed in DQ HandS Explorer, as shown in Figure 12.

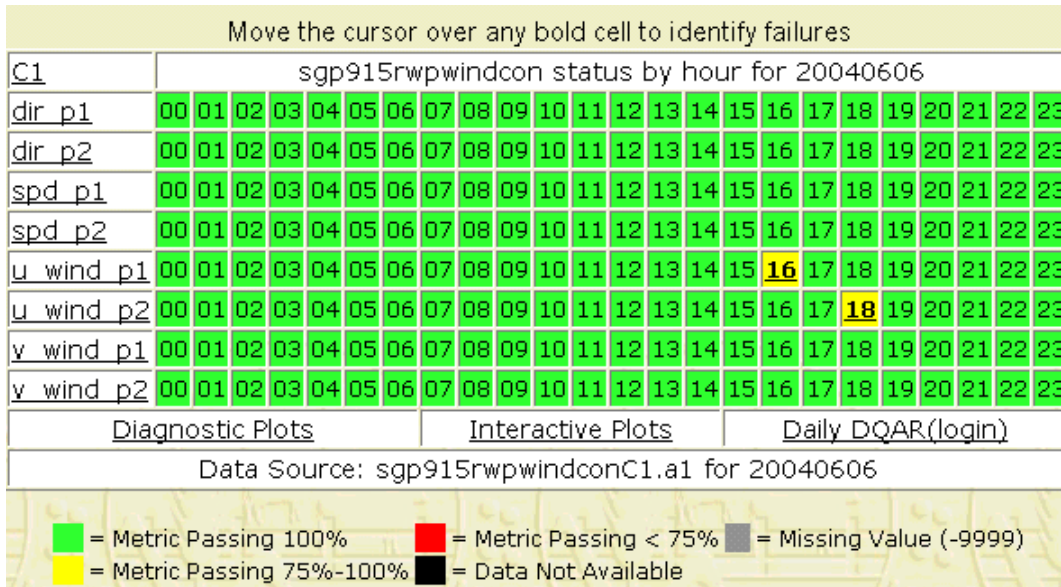


Figure 12: ARM DQ HandS Explorer display of hourly statistics for variables that were checked in a RWP file. Bold hour numbers indicate some data were not available during that hour. Data quality flags indicating which quality criteria were violated during that hour are displayed in the DQ HandS Explorer window when the cursor is moved over a cell with values that were either failing, missing or not available.

4.1.2.1.2 Diagnostic Plots

Plots are created showing the primary measurements over the 24-hour period covered by the NetCDF file. Sample wind velocity profiles are shown in Figure 13.

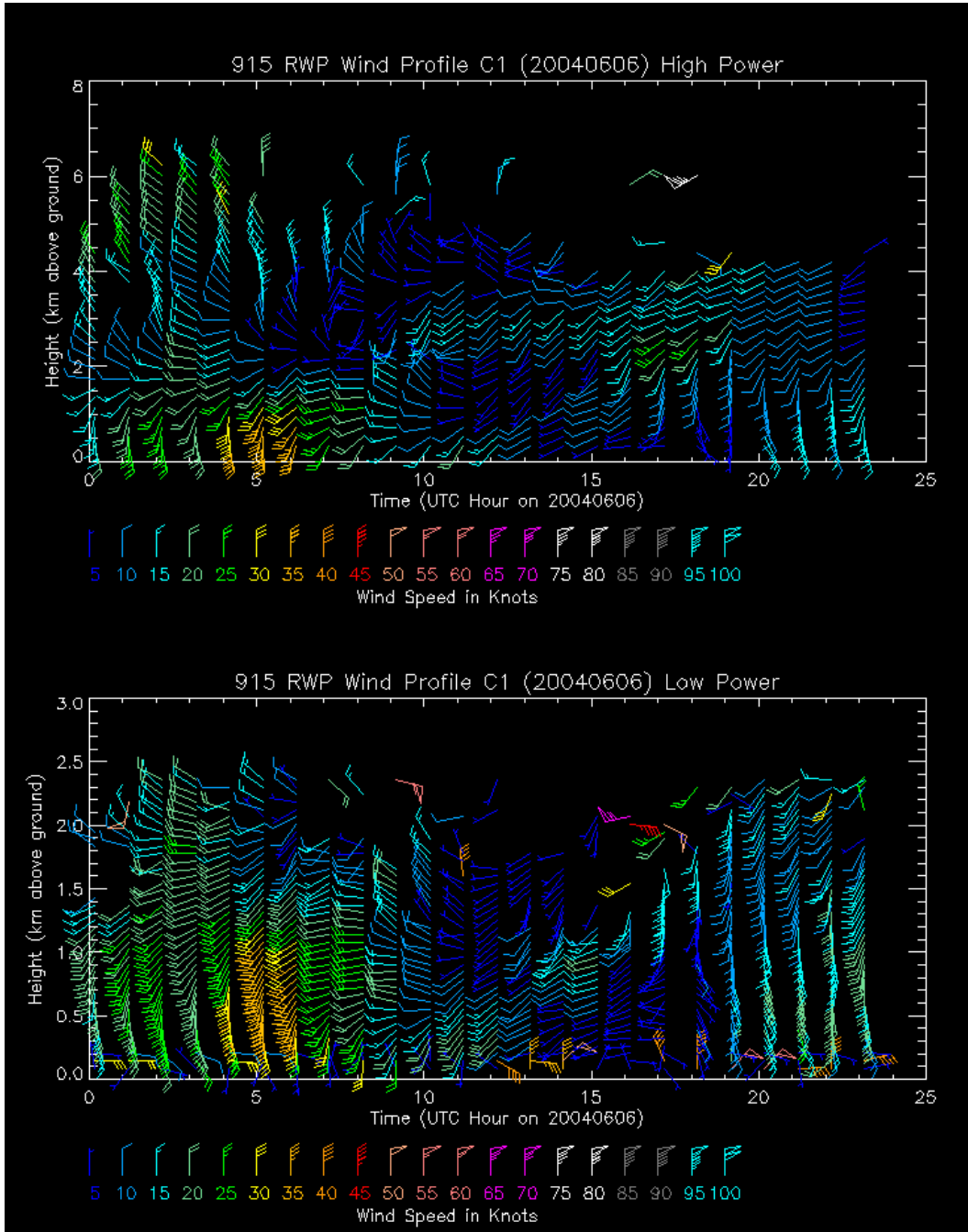


Figure 13: Sample RWP wind velocity profile plots at each power level.

4.1.3 Quality Analysis of RASS Data

Our approach for quality analysis of RASS a1-level data is to make an instrument cross-comparison check with our sonde climatology data for the RASS virtual temperature profile. Historical sonde files are used to establish nominal ranges and nominal lapse rates for virtual temperature. The RASS profiles, which are based on remotely sensed data, are tested against the ranges derived from sonde data, which are based on direct measurements (a “ground-truth” assessment). If a RASS profile reports a value that is outside the sonde range, then there is a defensible reason to raise suspicions about the RASS value. The sonde ranges for virtual temperature and virtual temperature lapse rate are established by calculating the monthly means and standard deviations within 50m altitude bins.

RASS temperature profiles are currently produced at SGP and NSA Central Facilities. Sonde data used to create the LUT for RASS QA came from `sgplssondeC1.c1` files for SGP, and `nsasondownpnC1.b1` files for NSA. These files contain the following values, all dimensioned by time:

pres	atmospheric pressure (hPa)
tdry	dry-bulb temperature (C)
rh	relative humidity (%)

The RASS QA procedure requires additional variables not available in the `sgplssondeC1.c1` or `nsasondownpnC1.b1` sonde data files, including dewpoint temperature and water vapor mixing ratio. These values are calculated from absolute temperature and relative humidity, as follows:

Dewpoint Temperature from Dry-Bulb Temperature and Relative Humidity:

$$\text{dewpoint temperature} = \frac{116.9 + 237.3 \cdot \ln(e_d)}{16.78 - \ln(e_d)} \text{ (}^\circ\text{C)}$$

e_d = vapor pressure (partial pressure of water vapor) (mb)

$$= \frac{\text{RH}}{100} \cdot \exp\left(\frac{16.78T_{dry} - 116.9}{T_{dry} + 237.3}\right)$$

T_{dry} = dry-bulb temperature ($^\circ\text{C}$)

RH = relative humidity (%)

Water Vapor Mixing Ratio from Dry-Bulb Temperature , Relative Humidity and Atmospheric Pressure:

$$\text{water vapor mixing ratio} = \frac{\text{mass of water vapor}}{\text{mass of dry air}} = \varepsilon \cdot \frac{e_d}{p - e_d} \text{ (g/Kg)}$$

$\varepsilon = 622.0$ (ratio of molecular weight of water vapor to dry air, x1000)

e_d = vapor pressure (partial pressure of water vapor) (mb)

$$= \frac{\text{RH}}{100} \cdot \exp\left(\frac{16.78T_{dry} - 116.9}{T_{dry} + 237.3}\right)$$

p = atmospheric pressure (mb)

For comparison to RASS virtual temperature profiles, the Sonde dry-bulb temperatures are converted to virtual temperatures using the Sonde dry-bulb temperature and water vapor mixing ratio, as follows:

$$\text{virtual temperature} = T \cdot \frac{1 + r_v}{1 + \varepsilon} \text{ (K)}$$

T = absolute temperature (sonde dry-bulb temperature, K)

$\varepsilon = 0.622$ (ratio of molecular weight of water vapor to dry air)

r_v = water vapor mixing ratio (g/Kg)

For automated quality analysis of RASS, each data value in the retrieved RASS virtual temperature profile is compared to the historical sonde range for the month of observation and corresponding altitude bin. If the RASS value falls outside of the sonde range $\bar{x}_{sonde} \pm 4.0 \cdot \sigma_{sonde}$ then the RASS value is flagged as failing max/min check (Figure 14). If the calculated RASS lapse rate falls outside the same range for average sonde lapse rates, then the RASS value is flagged as failing delta check. Hourly statistics for missing, failing and not available values are calculated for display in DQ HandS. The total number of failing values for any given hour is divided by the total number of observations during that hour to give an hourly percentage of failing values. Statistics are also calculated for missing and not available values and reported to DQ HandS via the usual metrics file.

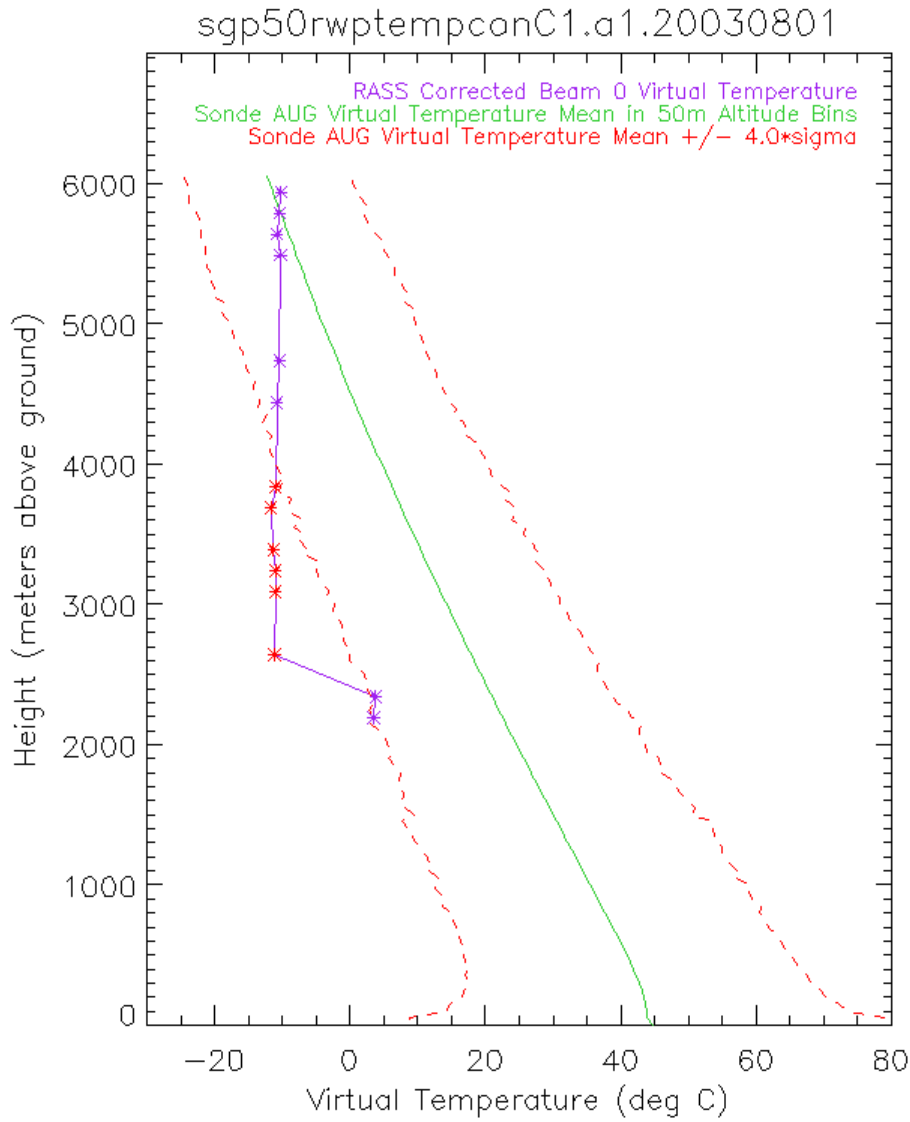


Figure 14: For the month of August, the solid green line depicts the average temperature profile as reported by the sonde at the SGP site. The red dashed lines mark the spread of sonde values out to four standard deviations from the average. On 1 August 2003, RASS temperature data (purple asterisks) were compared against the sonde climatology, and out of bound values were flagged (red asterisks).

4.1.3.1 RASS Analysis Code Output

4.1.3.1.1 QC Metrics

Statistics are compiled for each hour of the 24-hour period covered by the RASS NetCDF file. The statistics are written to a data file that is compatible with DQ HandS Explorer. This data file contains statistics for percent of values failing, percent of values missing, fail times for missing or failing values, and a quality flag indicating which quality criteria were violated during each hour. The percent failing statistic combines all values that are

outside the ‘good’ interval. The failure code indicates which quality criteria were violated within the hour, bit-packed with the following values:

- 0 => 'missing data value (-9999)',
- 1 => '< MIN Based on historical distribution for each variable',
- 2 => '> MAX Based on historical distribution for each variable',
- 3 => '< SONDE Minimum for month and height of observation',
- 4 => '> SONDE Maximum for month and height of observation',
- 5 => 'NOT AVAILABLE'

Summaries of the hourly statistics are then displayed in DQ HandS Explorer, as shown in Figure 15.

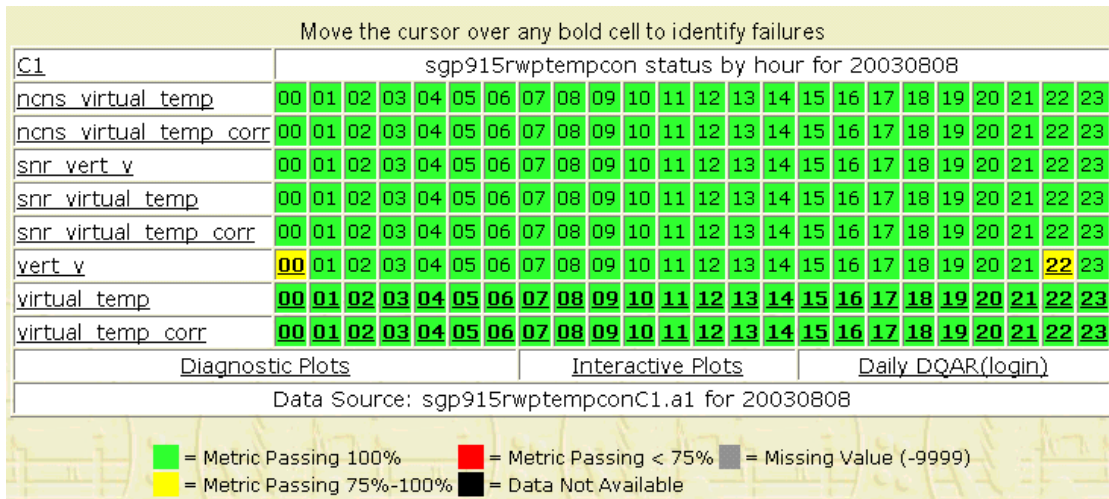


Figure 15: ARM DQ HandS Explorer display of hourly statistics for variables that were checked in a RASS file. Bold hour numbers indicate some data were not available during that hour. Data quality flags indicating which quality criteria were violated during that hour are displayed in the DQ HandS Explorer window when the cursor is moved over a cell with values that were either failing, missing or not available.

4.1.3.1.2 Diagnostic Plots

Plots are created showing primary measurements over the 24-hour period covered by the NetCDF file. Missing or not available values are not shown in the plot. Example showing virtual temperature profile and vertical wind velocity profile is displayed in Figure 16.

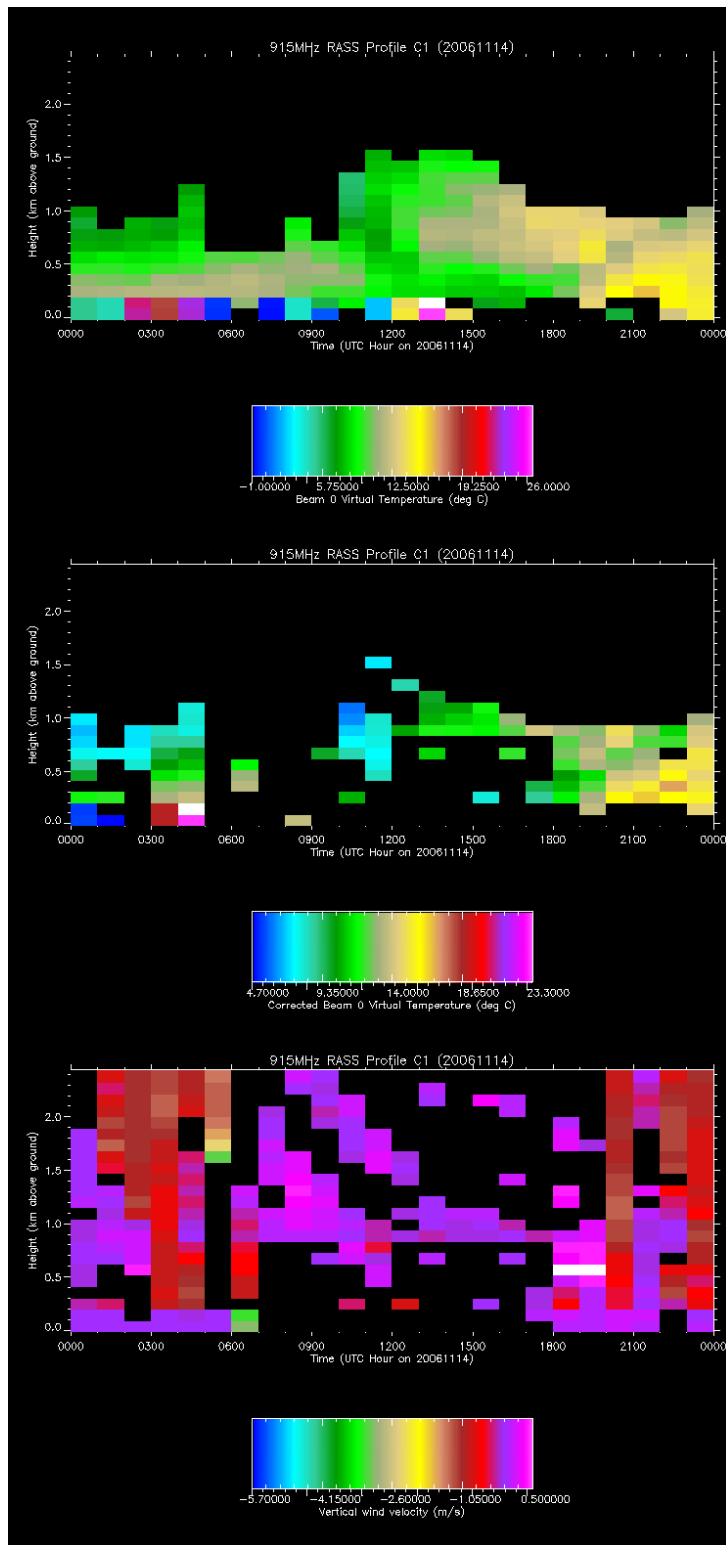


Figure 16: Sample RASS Virtual Temperature profile and Vertical Wind Velocity profile for each measurement mode.

4.2 Best Estimate Flux Value Added Product (BEFLUX VAP)

4.2.1 Background

This VAP uses data available from the three co-located surface radiometer platforms at the SGP Central Facility to automatically determine the best irradiance measurements available.

The Best Estimate Flux VAP processes data starting from March 22, 1997 when data from the three Central Facility radiometer systems, SIRS E13, C1, and "BSRN" (sgpsirs1duttE13.c1, sgpsirs1duttC1.c1, and sgpbsrn1duttC1.c1), are all available. In 2001, the diffuse SW instruments were switched to shaded B/W instruments, and the name "BSRN" was switched to BRS. Prior to that time, this VAP uses corrected diffuse SW from the DiffCorr1Dutt VAP as input. The 1-minute input data are compared to decide which will be used for averaging to get the best estimate. The output data are saved in two NetCDF files containing the best estimate values, QC flags, and the difference fields.

The input data 1-minute samples are compared to decide which will be used for averaging to get the best estimate. For diffuse and direct normal SW, and downwelling LW, three measurements from SIRS E13, C1, and BRS are used in the best estimate evaluation, while only two measurements from SIRS E13 and C1 are available for upwelling SW and LW evaluation.

The input data sources for BE Flux VAP are files containing measurements of downwelling SW direct, and diffuse irradiance, downwelling LW irradiance, and upwelling SW and LW irradiance measurements. The input files are the a1 level SIRS, SIROS, and BRS data streams. A more complete description of the input sources is available in sections 3 and 4 of [ARM Technical Report TR-008](#)⁸

4.2.2 Quality Analysis of BE Flux Data

The approach for quality analysis of BE Flux c1-level data is to apply min / max checks, with limits based on monthly histograms of historical data for each variable that is checked. Some of the variables in the BE Flux VAP are differences between measured values and the best estimate. For these variables, the same DQ checks are made, based on monthly histograms of each difference. Monthly histograms for each variable are generated from historical data (Figure 17 and Figure 18). The histograms are used to determine appropriate limits for min / max QC checks. For variables that are essentially normally distributed, QC limits are determined by adding a multiple of the monthly standard deviation to the monthly mean for each variable, e.g., $\bar{x}_{\text{var}} \pm n \cdot \sigma_{\text{var}}$. For BE Flux, the multiplier is currently set to 4.0. For variables that are not essentially normally distributed, floor and/or ceiling values are used as limits. These values are determined by

inspecting the histograms. A variable can have a floor value without a ceiling value, and vice versa. In these cases, the other limit is determined with $\bar{x}_{\text{var}} \pm n \cdot \sigma_{\text{var}}$. For those variables with designated floor and/or ceiling values, those values are used as the QC limit. For variables with no floor or ceiling value, the QC limit is determined with monthly statistics by $\bar{x}_{\text{var}} \pm n \cdot \sigma_{\text{var}}$. Histograms are generated only with values that lie between the var_min and var_max values as shown in Table 1.

Table 1: Min, Max, Floor, and Ceiling values used with BEFLUX quality analysis code.

	short_direct_normal	down_short_hemisp	down_short_diffuse_hemisp	up_short_hemisp
var_min	0	0	0	0
var_max	1250	1350	800	500
floor	-2	-2	-2	-2
ceiling	1100	1300	700	275

	short_direct_normal_diff	down_short_diffuse_hemisp_diff	up_short_hemisp_diff
var_min	-100	-50	-100
var_max	1000	200	100
floor	-2	-50	-65
ceiling	50	50	30

	down_long_hemisp	up_long_hemisp
var_min	0	0
var_max	550	700
floor	-9999	-9999
ceiling	-9999	-9999

	down_long_hemisp_diff	up_long_hemisp_diff
var_min	-25	-50
var_max	50	50
floor	-25	-15
ceiling	50	35

	net_surface_radiation	albedo	zenith
var_min	-500	0	0
var_max	1000	10	180
floor	-500	-2	-2
ceiling	1000	1	180

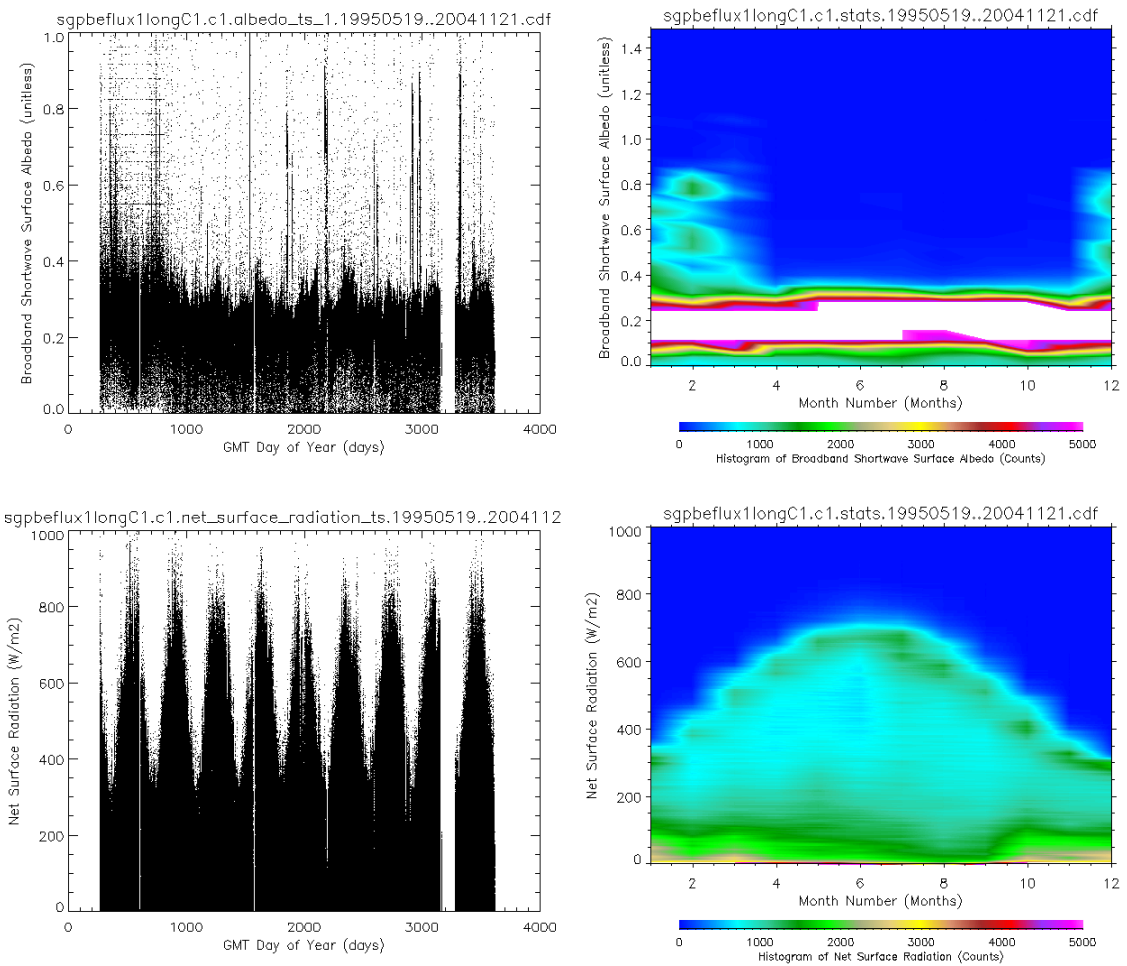


Figure 17: Long Time Series and Average Monthly Distributions of Albedo (top) and Net Surface Radiation (bottom). Statistics were calculated from 3,355 data files created between May 19th, 1995 and Nov 21, 2004.

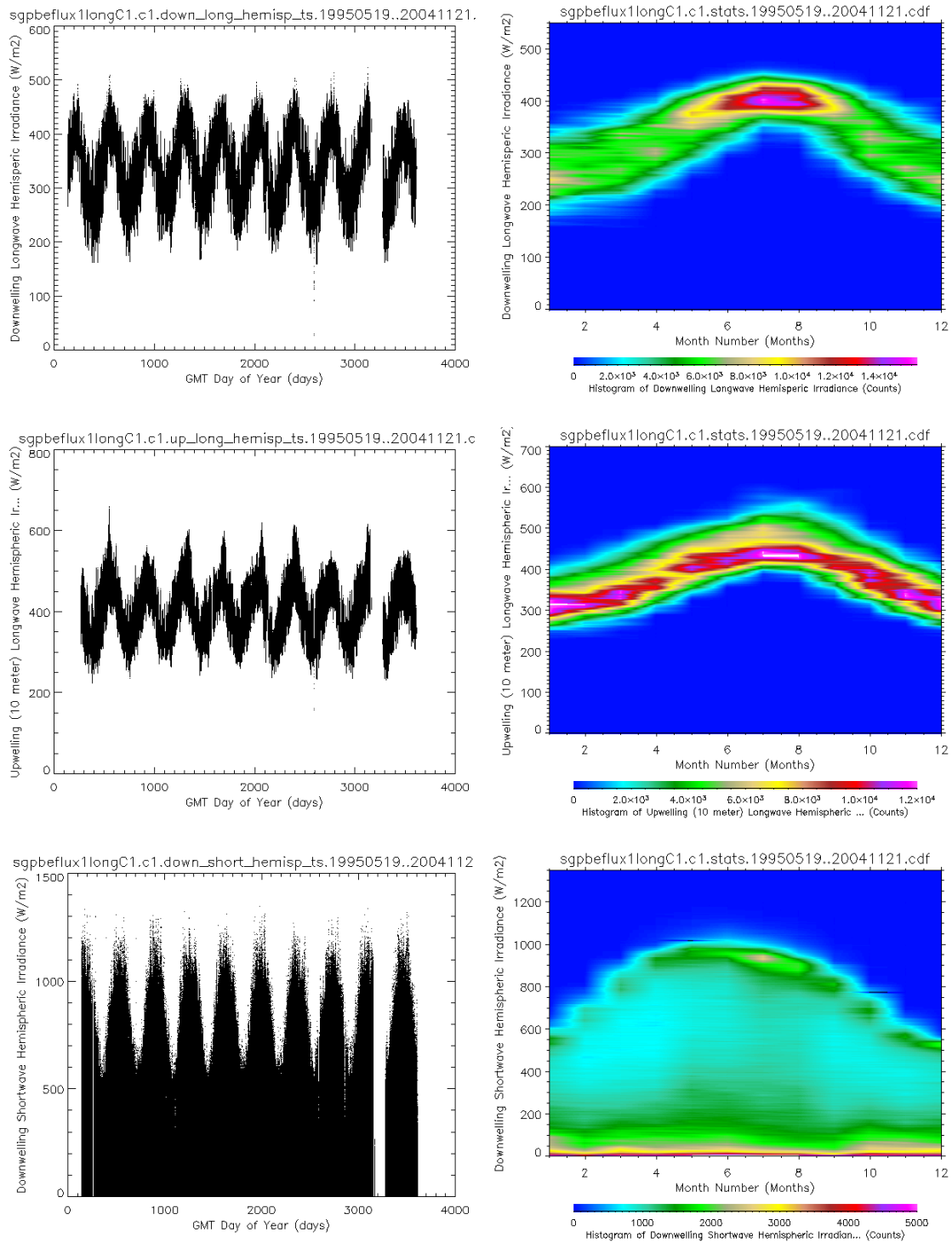


Figure 18: Long Time Series and Monthly Distributions of Hemispheric Irradiance. Downwelling Longwave (top), Upwelling Longwave (middle), and Downwelling Shortwave (bottom). Statistics were calculated from 3,355 data files created between 19 May 1995 and 21 Nov 2004.

4.2.2.1 BEFLUX Analysis Code Output

4.2.2.1.1 QC Metrics

Statistics are compiled for each hour of the 24-hour period covered by the BE Flux NetCDF file. The statistics are written to a data file that is compatible with DQ Hands Explorer. This data file contains statistics for percent of values failing, percent of values missing, fail times for missing or failing values, and a quality flag indicating which quality criteria were violated during each hour. The percent failing statistic combines all values that are outside the 'good' interval. The failure code statistic indicates which quality criteria were violated within the hour, bit-packed with the following values:

```
0 => 'missing data value (-9999)',
1 => '< MIN Based on monthly historical distribution for each variable',
2 => '> MAX Based on monthly historical distribution for each variable',
3 => 'NOT AVAILABLE',
```

Summaries of the hourly statistics are then displayed in DQ Hands Explorer, as shown in Figure 19.

Move the cursor over any bold cell to identify failures

C1	sgpbeflux1long status by hour for 20060715																							
albedo	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
down long hemisp	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
down long hemisp diff	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
down short diffuse hemisp	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
down short diffuse hemisp diff	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
down short hemisp	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
net surface radiation	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
short direct normal	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
short direct normal diff	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
up long hemisp	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
up long hemisp diff	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
up short hemisp	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
up short hemisp diff	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23
zenith	00	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23

[Diagnostic Plots](#) [NCVweb Interactive Plots](#)
 Data Source: sgpbeflux1longC1.c1 for 20060715

■ = Metric Passing 100% ■ = Metric Passing < 75% ■ = Missing Value (-9999)
■ = Metric Passing 75%-100% ■ = Data Not Available

Figure 19: ARM DQ Hands Explorer display of hourly statistics for variables that were checked in a BE Flux file. Bold hour numbers indicate some data were not available during that hour. Data quality flags indicating which quality criteria were violated during that hour are displayed in the DQ Hands Explorer window when the cursor is moved over a cell with values that were either failing, missing or not available.

4.2.2.1.2 Diagnostic Plots

Plots are created showing selected variables over the 24-hour period covered by the NetCDF file. Missing or not available values are not shown in the plot. If any variable in a plot contains values outside of the sonde range, these points are over plotted with red dots to indicate failing data. Example diagnostic plots are shown in Figure 20 and Figure 21.

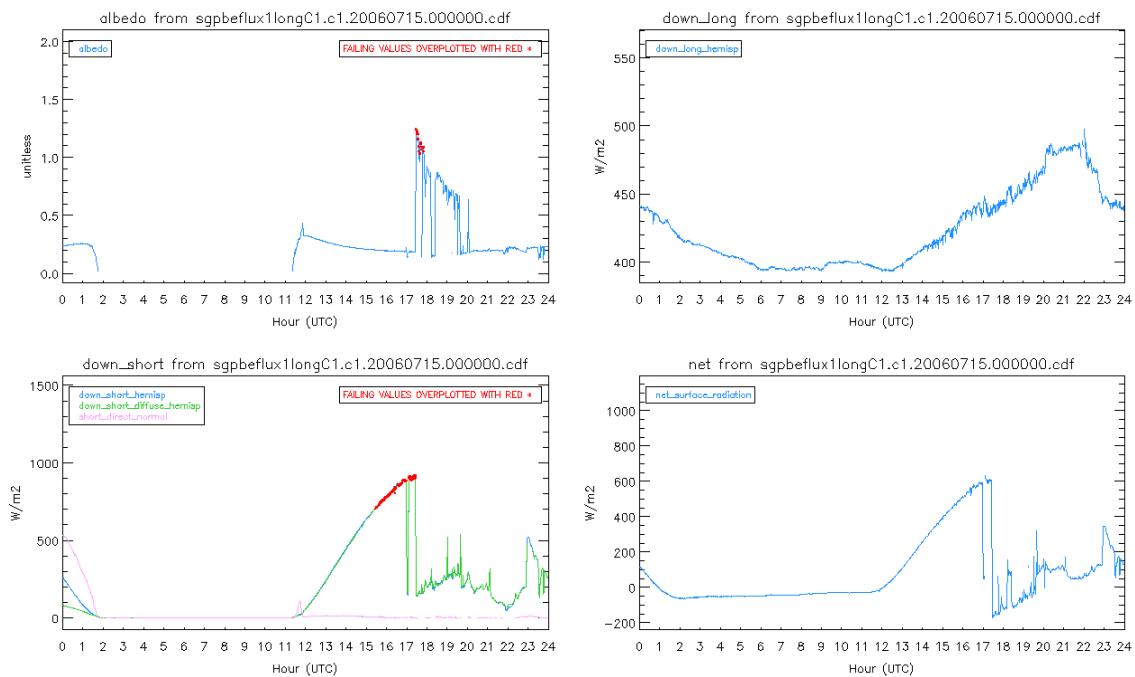


Figure 20: ARM DQ Hands Explorer example of diagnostic plots for BE Flux showing albedo (top left), downwelling longwave (top right), downwelling shortwave (bottom left), net radiation (bottom right).

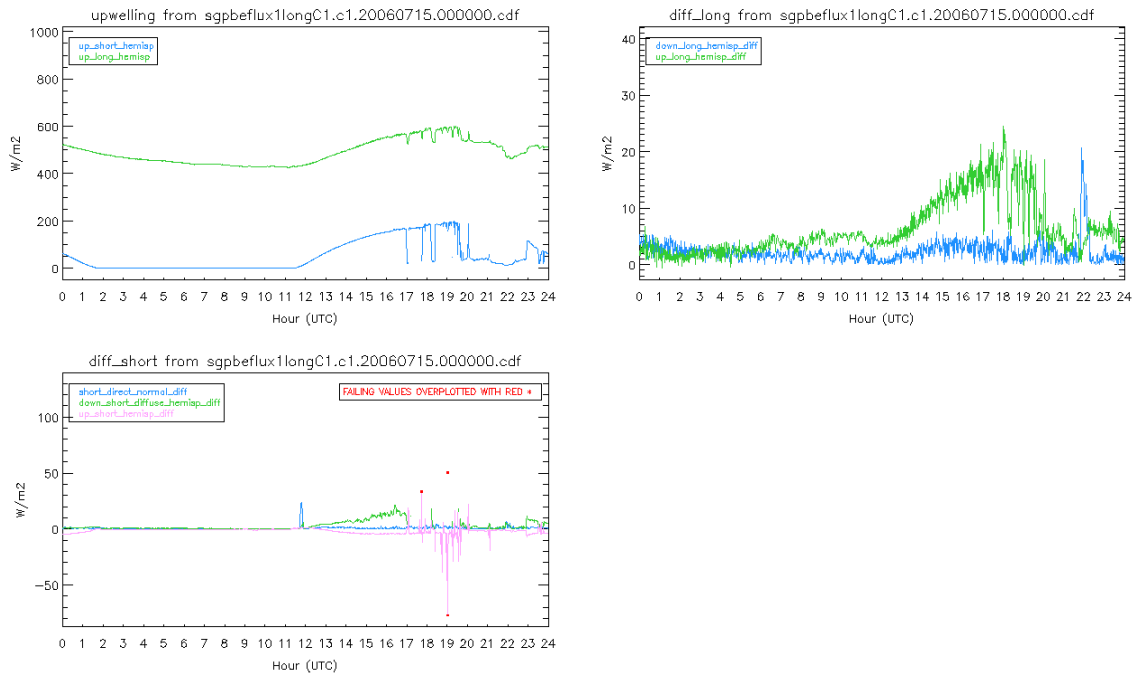


Figure 21: BE Flux analysis diagnostic plots showing long and shortwave upwelling (top-left), longwave difference plots (top-right), and shortwave difference plots (bottom).

4.3 Tower Water-Vapor Mixing Ratio Value Added Procedure (TWRMR VAP)

4.3.1 Background

The primary purpose of the TWRMR VAP is to calculate water vapor mixing ratio at the 25- and 60-meter levels of the tower at the SGP central facility. Since there are no barometric pressure sensors at those levels on the tower, the hypsometric equation is used along with surface pressure values from either the SMOS or the THWAPS to derive barometric pressures at the tower levels. After barometric pressure is derived for the 25- and 60-meter levels, water vapor mixing ratio at those levels can be calculated directly.

At the same time, the TWRMR VAP serves as a "best-estimate" product for temperature, relative humidity and pressure for the surface and the 25- and 60-meter levels. The primary input at the surface is the SMOS, while the primary inputs at the 25- and 60-meter levels come from the sensors in the southeast (SE) elevator. The alternate observations come from the THWAPS for the surface, and from the west (W) elevator for the 25- and 60-meter levels. The source of primary and alternate values can be changed manually with a command line switch at runtime when the TWRMR VAP is produced.

4.3.2 Quality Analysis of TWRMR Data

Three methods are used for quality analysis of TWRMR c1-level data: (1) min / max checks, with limits based on monthly histograms of historical data for each variable that is checked; (2) Sensor cross-comparison checks, based on monthly histograms of differences between two sensors measuring the same quantity; and (3) Instrument cross-comparison checks with sonde climatologies at the upper tower levels.

4.3.2.1 Min / Max Checks

Monthly histograms for each variable are generated from historical data (Figure 22). The histograms are used to determine appropriate limits for min / max QC checks. For variables that are essentially normally distributed, QC limits are determined by adding a multiple of the monthly standard deviation to the monthly mean for each variable, e.g., $\bar{x}_{\text{var}} \pm n \cdot \sigma_{\text{var}}$. For TWRMR, the multiplier is currently set to 4.0. For variables that are not essentially normally distributed, floor and/or ceiling values are used as limits. These values are determined by inspecting the histograms. A variable can have a floor value without a ceiling value, and vice versa. In these cases, the other limit is determined with $\bar{x}_{\text{var}} \pm n \cdot \sigma_{\text{var}}$. The current settings for DQ limits for TWRMR VAP are shown in Table 2. Histograms are generated only with values that lie between the var_min and var_max values in the table.

Table 2: Min, Max, Floor, and Ceiling values used with TWRMR quality analysis code.

	pres_02m	pres_25m	pres_60m
var_min	945	945	945
var_max	1015	1015	1015
floor	-9999	-9999	-9999
ceiling	-9999	-9999	-9999

	temp_02m	temp_25m	temp_60m
var_min	-30	-30	-30
var_max	45	45	45
floor	-9999	-9999	-9999
ceiling	-9999	-9999	-9999

	rh_02m	rh_25m	rh_60m
var_min	0	0	0
var_max	110	110	110
floor	0	0	0
ceiling	110	110	110

	vap_pres_02m	vap_pres_25m	vap_pres_60m
var_min	0	0	0
var_max	35	35	35
floor	-9999	-9999	-9999
ceiling	-9999	-9999	-9999

	mixing_ratio_02m	mixing_ratio_25m	mixing_ratio_60m
var_min	0	0	0
var_max	25	25	25
floor	-9999	-9999	-9999
ceiling	-9999	-9999	-9999

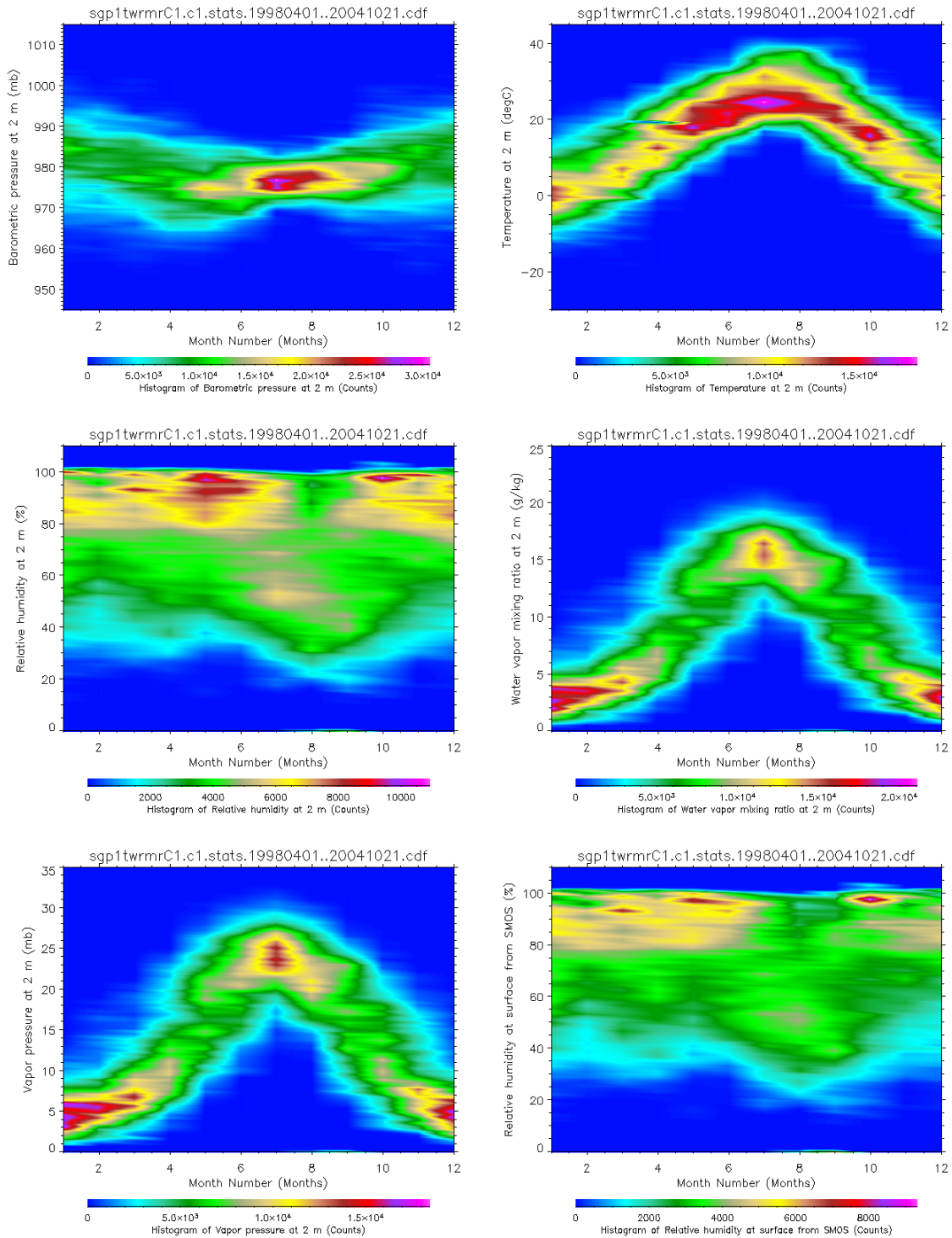


Figure 22: Historical monthly distributions of Barometric Pressure (top-left), Temperature (top-right), Relative humidity (middle-left), Water Vapor Mixing Ratio (middle-right), and Vapor Pressure (bottom-left) from the SGP Tower at 2m. Relative humidity from SMOS is also shown (bottom-right) as a secondary source. Statistics were calculated from 2,324 data files created between 01 Apr 1998 and 21 Oct 2004.

4.3.2.2 Sensor Cross-Comparison Checks

The tower instrumentation includes dual sensors (SE and W platforms) for measuring temperature and relative humidity at the 25- and 60-meter levels. Measurements of temperature and relative humidity are also made at the surface by THWAPS and SMOS. Since these calibrated sensor pairs measure the same quantity at the same place and time, one would expect the values they produce to be similar. If the measured values diverge, then there is a defensible reason to raise suspicions about the quality of one measurement or the other, or both of them. This condition may occur when both sensors report values within the designated min / max range, so making a quality assessment by comparing the two sensor values may reduce Type II errors. A specific example of this situation has been described, when the leads to the 25- and 60-meter sensors were switched.

Monthly histograms for the *difference* between each pair of sensors are generated from historical data. The histograms are used to determine appropriate limits for QC checks. One would expect the difference variables to be distributed essentially normally about zero. In this case, QC limits could be determined by adding a multiple of the monthly standard deviation to the monthly mean for each difference variable, e.g., $\bar{x}_{\text{var}} \pm n \cdot \sigma_{\text{var}}$. However, we found that most of the difference variables show systematic errors, and most distributions were not essentially normal. For implementing QC checks for TWRMR based on difference variable distributions, the limits have been extended to 6.0 to 8.0 times the monthly standard deviation.

4.3.2.3 Sonde Cross-Comparison Checks

Sonde data provide an opportunity for QC checks based on instrument cross-comparisons. The general approach for instrument cross-comparison is to establish nominal ranges and nominal lapse rates for measured quantities from historical sonde profiles. If a TWRMR value lies outside the sonde range, then there is a defensible reason to raise suspicions about the TWRMR value. The sonde ranges for each parameter and their lapse rates are established by calculating the monthly means and standard deviations within 50m altitude bins (see Section 3.2).

Sonde climatology data included in the TWRMR Lookup Table were derived from `sgplssondeC1.c1` files from the SGP site. The LSSONDE files contain the following values, all dimensioned by time:

pres	atmospheric pressure (hPa)
tdry	dry-bulb temperature (C)
rh	relative humidity (%)

The TWRMR QA procedure requires additional variables not available in the `sgplssondeC1.c1` sonde data files, including vapor pressure and water vapor mixing ratio. These values are calculated from absolute temperature and relative humidity, as follows:

$$\text{water vapor mixing ratio} = \frac{\text{mass of water vapor}}{\text{mass of dry air}} = \varepsilon \cdot \frac{e_d}{p - e_d} \text{ (g/Kg)}$$

$\varepsilon = 622.0$ (ratio of molecular weight of water vapor to dry air, x1000)

e_d = vapor pressure (partial pressure of water vapor) (mb)

$$= \frac{\text{RH}}{100} \cdot \exp\left(\frac{16.78T_{\text{dry}} - 116.9}{T_{\text{dry}} + 237.3}\right)$$

p = atmospheric pressure (mb)

T_{dry} = dry - bulb temperature (°C)

RH = relative humidity (%)

4.3.2.4 TWRMR Analysis Code Output

4.3.2.4.1 QC Metrics

Statistics are compiled for each hour of the 24-hour period covered by the TWRMR NetCDF file. The statistics are written to a data file that is compatible with DQ HandS Explorer. This data file contains statistics for percent of values failing, percent of values missing, fail times for missing or failing values, and a quality flag indicating which quality criteria were violated during each hour. The percent failing statistic combines all values that are outside the 'good' interval. The failure code statistic indicates which quality criteria were violated within the hour, bit-packed with the following values:

```

0 => 'missing data value (-9999)',
1 => '< MIN Based on monthly historical distribution for each variable',
2 => '> MAX Based on monthly historical distribution for each variable',
3 => 'NOT AVAILABLE',
4 => '< SONDE Minimum for month and height of observation',
5 => '> SONDE Maximum for month and height of observation',
6 => '> SONDE Maximum DELTA for month and height of observation',
7 => '> Maximum DIFFERENCE (2m - 25m) for month and height of observation',
8 => '< Minimum DIFFERENCE (2m - 25m) for month and height of observation',
9 => '> Maximum DIFFERENCE (2m - 60m) for month and height of observation',
10=> '< Minimum DIFFERENCE (2m - 60m) for month and height of observation',
11=> '> Maximum DIFFERENCE (25m - 60m) for month and height of observation',
12=> '< Minimum DIFFERENCE (25m - 60m) for month and height of observation',
13=> 'Failed Difference Check SMOS - THWAPS for month and height of
    observation',
14=> 'Failed Difference Check SW - W Sensors for month and height of
    observation'

```

Summaries of the hourly statistics are then displayed in DQ HandS Explorer, as shown in Figure 23.

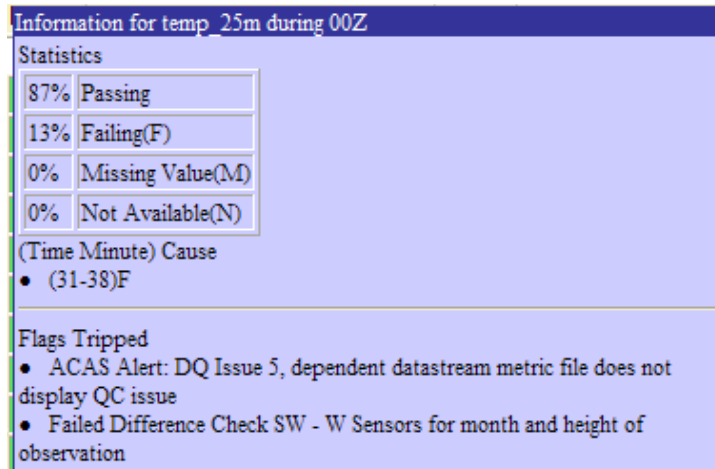
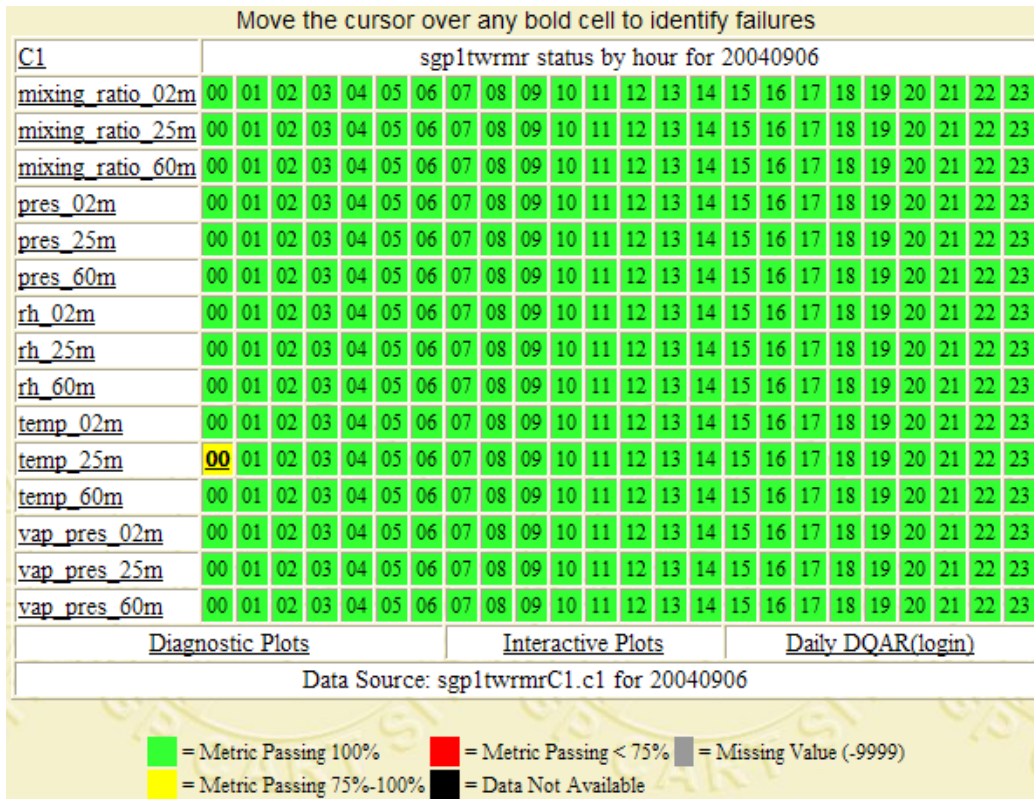


Figure 23: ARM DQ HandS Explorer display of hourly statistics for variables that were checked in a TWRMR file. Bold hour numbers indicate some data were not available during that hour. Data quality flags indicating which quality criteria were violated during that hour are displayed in the DQ HandS Explorer window when the cursor is moved over a cell with values that were either failing, missing or not available.

4.3.2.4.2 Diagnostic Plots

Plots are created showing selected variables over the 24-hour period covered by the NetCDF file. Missing or not available values are not shown in the plot. If any variable in a plot contains values outside of the valid range, these points are over-plotted with red dots to indicate failing data. Example diagnostic plots are shown in Figure 24.

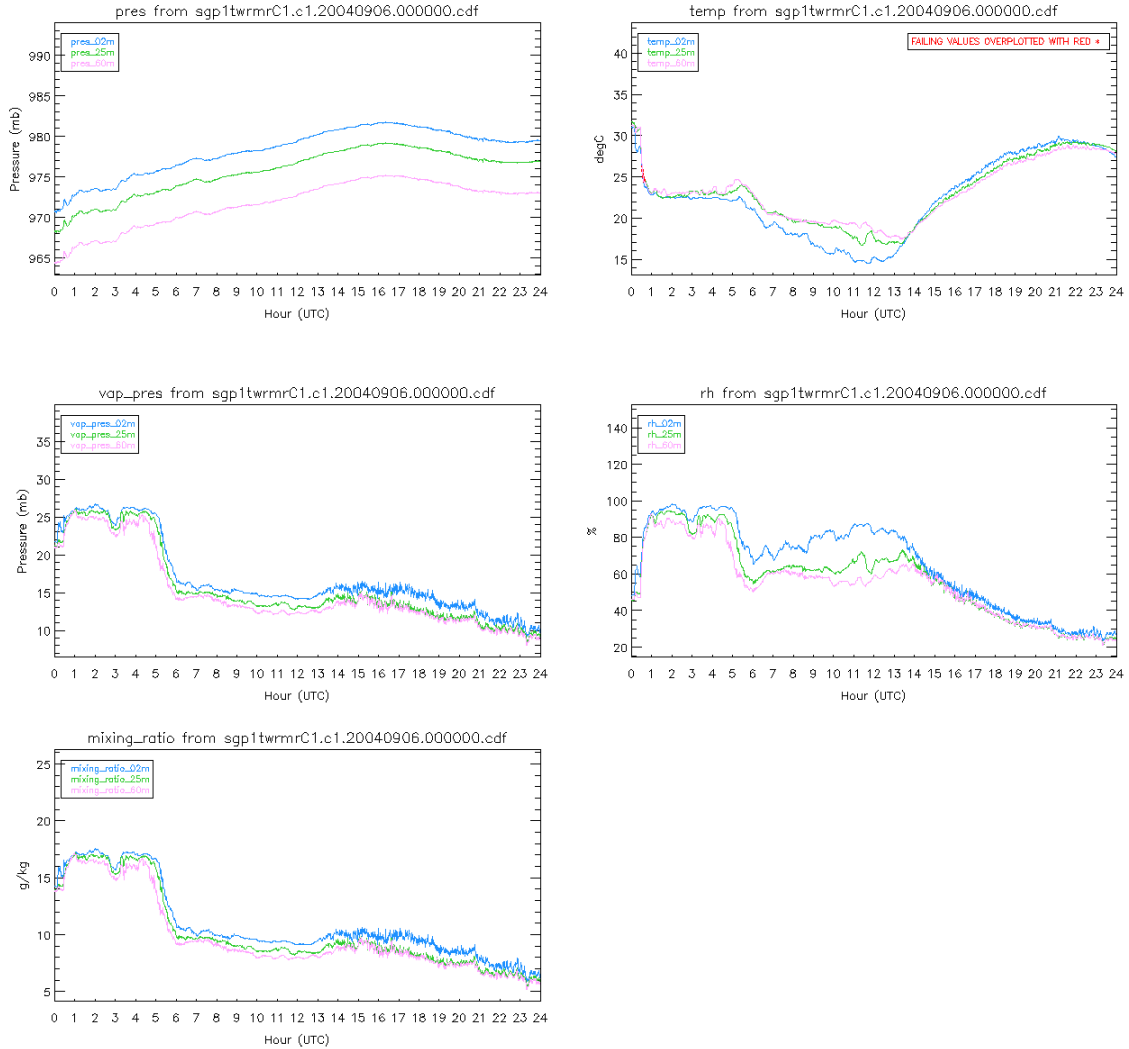


Figure 24: Diagnostic plots for TWRMR showing pressure, temperature, vapor pressure, relative humidity and calculated water vapor mixing ratio.

5 Interactive Web-Based Tools

5.1 DQ Hands Plot Browser

The data quality office produces daily diagnostic plots for key measurements from most ARM data streams. These files have been traditionally viewed on a daily basis, along with the data quality related metrics available in the Data Quality Health and Status (DQ Hands) Explorer.

During this period of performance, we have created a Web based tool to allow an analyst to view many plots at once and organize them by various criteria of importance to a particular user (Figure 25). Plots are available for all ACRF fixed and mobile sites. Ancillary processes work in the background to ensure that the Web based tool is always provided with the most up-to-date configuration parameters concerning sites, available data streams, instrument locations and plot availability for particular dates.

This technique benefits the user in that (1) the software does not need human intervention when new ARM data streams or plots are added, and (2) dead-end queries are eliminated. Since the application always knows valid ranges for each of the input parameters, the user is not allowed to formulate inappropriate queries. Users can browse by thumbnail or by list before viewing a full-size diagnostic plot. Multiple data streams and multiple types of plots can be selected concurrently for side by side comparisons.

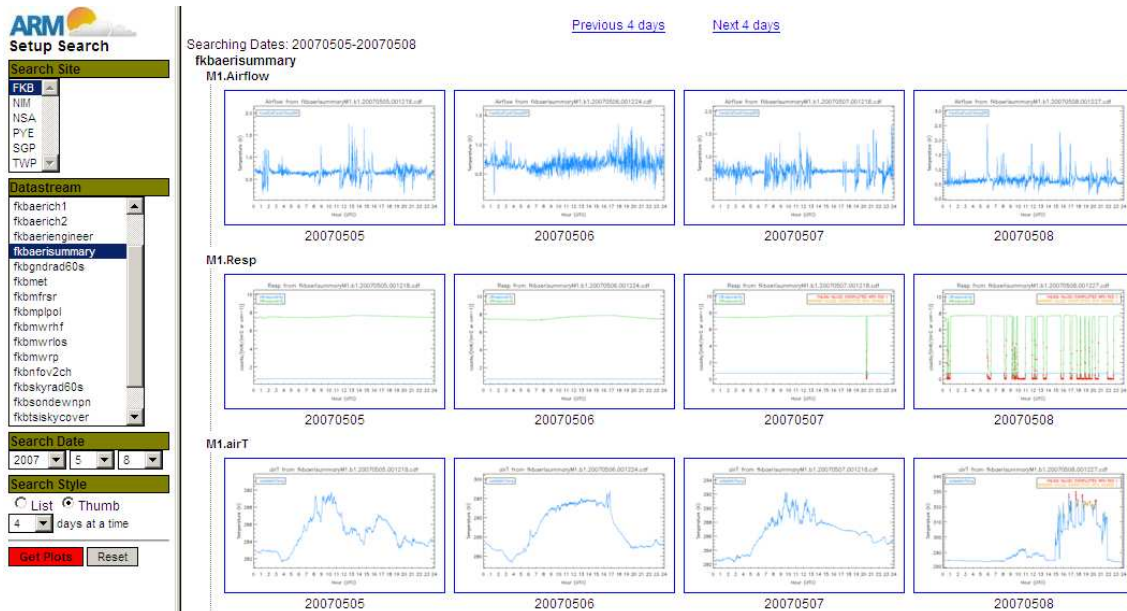


Figure 25: The DQ Hands Plotbrowser allows analysts to quickly scan thru multiple days of pre-generated diagnostic plots. Here we show housekeeping data from an AERI instrument. Displayed is Airflow (top), Responsivity (middle), and Air Temperature (bottom), over the course of 4 days.

5.2 NCVweb

NCVweb is a Web-based NetCDF data viewer and interactive plot tool we developed specifically for ARM data⁹. This application has been running successfully for several years on the data quality computer for inspecting recently acquired data, and at the ARM archive for inspecting 'standing order' data.

During this period of performance, NCVweb was integrated into the ARM archive's user request system. Once a data order for a user is fulfilled, the user receives a notification instructing them how to retrieve the data. Now, they are also provided a web URL for accessing their data via NCVweb.

NCVweb helps to eliminate the need of and problems associated with downloading large volumes of data, installing and configuring visualization software, or writing custom data exploration software. Since the tool is Web-based, ARM researchers and instrument mentors can visualize large and complex data sets without needing to be NetCDF savvy (see Figure 26).

The tool supports a wide number of visualizations as summarized in Table 3.

Table 3: NCVweb Plot Types

Plot Type	Example
X-Y	Time vs. Irradiance or Temperature vs. Altitude
X-Y-Y2	Ambient <i>and</i> Calibration Temperatures Vs. Time
Multi-Facility	Soil Temperature at several different Facilities Vs. Time
X-Y-Z Color Image	Continuous Temperature Profile (over height and time)
Horizontal Slices	Plots of Temperature vs. Time at several Altitudes
Vertical Slices	Plots of Temperature vs. Altitude at several, specific instances during the day

General plotting features include:

- Plots data across multiple files
- Detects and does not plot data points designated as "missing"
- Plots any compatible variable against any other
- Allows arbitrary zoom or pan to any data region
- Detects and does not connect data across significant time gaps
- Provides useful time conversions for the time axis
- Allows for various symbol and/or line combinations
- Supports multiple plot sizes to accommodate various screen sizes
- Auto-generates appropriate plot labels based on the data being viewed
- X/Y Axis Flip

Additional features:

- Choice of color table for false color three-dimensional visualizations
- Day/Night cycle background color coding (yellow/blue)
- Computes statistics for the aggregate data set
- Displays textual data values of the current zoom region
- Will convert NetCDF variables to ASCII formats for use with other tools
- Highlights data that have failed associated quality control flags
- Displays details of file contents, variables, and dimensions
- Thorough error checking and reporting
- Online help and directions

Other versions of NCVweb are being maintained on our project server to support specialized needs of the data quality office and of the external data center. As new ARM facilities were deployed, and as new datastreams emerged, we ensured that NCVweb would read and plot the new files.

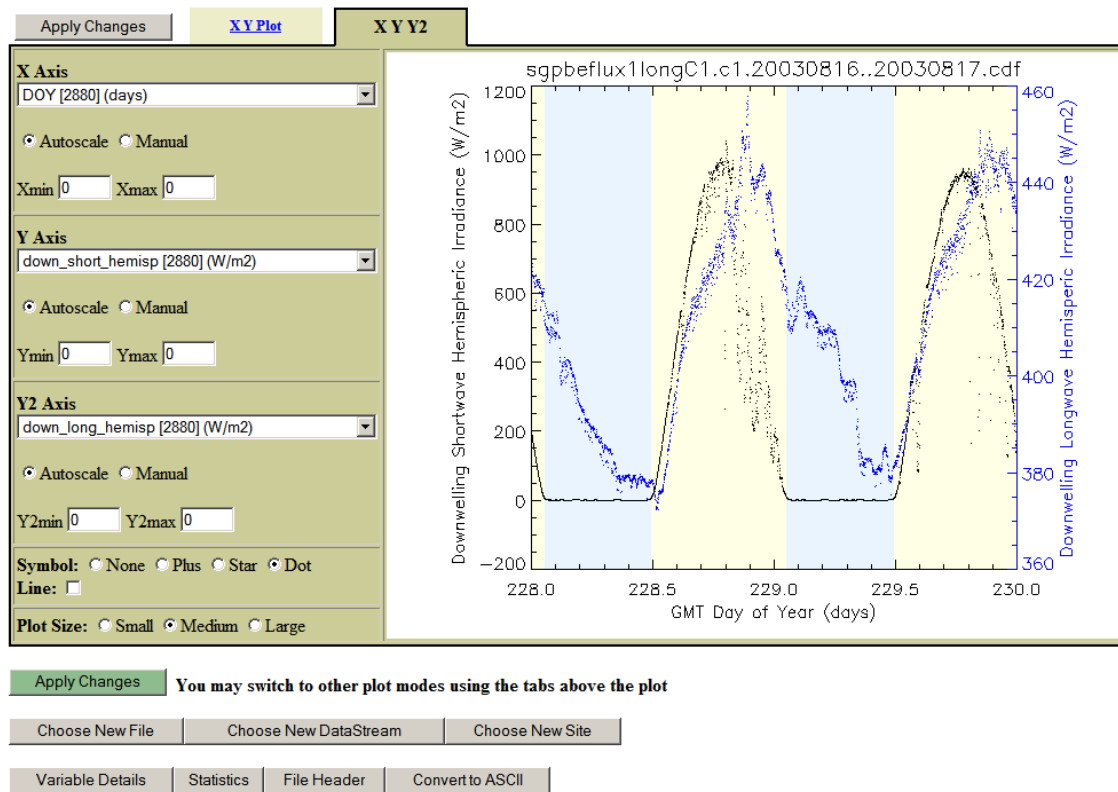


Figure 26: Example of using NCVweb to view Downwelling Shortwave and Downwelling Longwave radiation over the course of two days. The user can change the data to be plotted using the pull down menus on the left, as well as axis scaling and other plot characteristics. The background of the plot is colored blue during nighttime hours and yellow during daylight hours.

5.3 Datastream Processing Status on DQ Computer

A web based utility was developed to summarize and report on the processing status of ARM data streams at the DQ computer. This utility shows a tabulated list of the datastreams currently supplied to the data quality office. From this table, the user can easily determine which datastreams are delayed in delivery, which data streams have been processed into DQ metric files, which diagnostic plots have been generated, and whether the data quality office is delayed in producing any of these products. This information is updated daily (Figure 27).

Status of site NIM scanned on 20060622

Most recent files in each category have these dates

Datastream	Data	DQ metric	Plot
nim30ecorM1.b1	20060622	20060622	20060622
nimaerich1M1.b1	20060621	20060621	20060621
nimaerich2M1.b1	20060621	20060621	20060621
nimaeriengineerM1.b1	20060621	20060621	20060621
nimaerisummaryM1.b1	20060621	20060621	20060621
nimgndrad60sM1.b1	20060622	20060622	20060622
nimmM1.b1	20060622	20060622	20060622
nimmS1.b1	20060616	20060616	20060616
nimmfrsrM1.b1	20060622	20060622	20060622
nimmfrsrv2S1.b1	20060616		
nimmwrlsM1.b1	20060622	20060622	20060622
nimmwrrpM1.b1	20060622	20060622	20060622
nimradS1.b1	20060616	20060616	20060616
nimskyrad60sM1.b1	20060622	20060622	20060622
nimsondewnpnM1.b1	20060622		20060622
nimtsiskycoverM1.b1	20060622	20060622	20060622
nimvcei125kM1.b1	20060622	20060622	20060622
nimwacrM1.b1	20060622	20060622	20060622

Figure 27: Example of using the Datastream status tool to check on the datastreams from the mobile facility situated in Niamey, Niger. The green cells show datastreams and products that are up-to-date. The red colored cells show a delay in data delivery to the DQ host computer.

6 Automated Check and Alert System

We have developed a demonstration for an "Automated Check and Alert System", or ACAS. This system is intended to reassess the quality of instrument level data streams based on new information provided by value added products or quality measurement experiments.

The system works by consulting the VAP or QME metrics files for failures. It then checks a dependency table to see which instrument variables feed into the VAP or QME. If a problem is found in the higher level product, ACAS will look at the instrument metric file to see if the problem is also noted there. If not, an alert message is sent to interested parties.

In the ACAS file `mrc_acas.conf`, `%var_table` is a hash of hashes that stores variable dependencies for each datastream. Every hash key in `%var_table` represents a single datastream in the ACAS domain. For each datastream key, the associated value is another hash key representing a variable within the datastream that will be checked by ACAS. The variable key value is another hash consisting of "datastream" => "variable" key value pairs. These pairs denote datastream variables from which the original variable was derived or otherwise depends upon. An abbreviated example of a `%var_table` entry for `sgpqmeaeriprof` is shown in Figure 28.

In addition to these alerts, ACAS can generate quality flags for the instrument data stream when one of the higher-level checks are performed and have failed. As a test of this system, we currently insert ACAS triggered flags into the TWRMR metrics file to show how this might work.

We have solicited feedback from potential users in order to identify needed features. Early results with the TWRMR VAP suggest that the approach will in fact find additional issues with instrument data streams that have not been previously identified.

Once all the dependencies relating source data to value added products are known, we will be able to set up a production version of this system.

```

%var_table = (
    sgpqmeaeriprofC1 => {
        temperature_60m_resid => {
            "sgpaeriprofC1.c1" => "temperature",
            "sgp30twr60mC1.b1" => "qc_temp",
        },
        dewpoint_60m_resid => {
            "sgpaeriprofC1.c1" => "dewpointTemperature",
            "sgp30twr60mC1.b1" => ["qc_rh", "qc_temp"],
        },
    },
);

```

Figure 28: An example of an ACAS dependency tree for the sgpqmeaeriprofC1 datastream.

7 Conclusion

The approach used for this project has worked well for us over the years. The process of inspecting and analyzing ARM data requires not only a set of basic software tools, but tools that understand the peculiarities and unique characteristics of ARM datastreams. By concurrently developing new tools while we analyze and quality check ARM data, we find that these two tasks benefit each other and lead to results far superior to that which might be achieved if we had focused on either task alone. Better tools lead to better analysis. Experience with a great variety of data leads to better tools. We plan to continue this approach for future involvement with the ARM data quality office.

The data quality office now has a new arsenal of automated codes to help with their data inspection tasks. We have developed web-based and platform independent software tools in order to allow ARM researchers, instrument mentors and the data quality office to visualize large and complex data sets without requiring the user to be terribly computer savvy. Data visualization and quality assessments are the first step in the researcher's discovery process, and tools like NCVweb and DQ HandS remove some serious impediments that would normally interfere with this activity.

8 Acronyms

ACRF	ARM Climate Research Facility
AERI	Atmospheric Emitted Radiance Interferometer
ARM	Atmospheric Radiation Measurement (Program)
ARM*STAR	ARM Statistical Analysis and Reporting System
DMF	Data Management Facility
DOE	Department of Energy
DQ	Data Quality
DQ HandS	Data Quality Health and Status (System)
DQO	Data Quality Office
LUT	Look Up Table
NetCDF	Network Common Data Format
NSA	North Slope of Alaska
NCVweb	NetCDF Viewer on the Web
NOAA	National Oceanic and Atmospheric Administration
NWS	National Weather Service
QA	Quality Assurance
QC	Quality Control
QME	Quality Measurement Experiment
RASS	Radio Acoustic Sounding System
RWP	Radar Wind Profiler
SGP	Southern Great Plains
SMOS	Surface Meteorological Observing Station
THWAPS	Temperature, Humidity, Winds, and Pressure System
TWP	Tropical Western Pacific
VAP	Value-Added Product

9 References

- ¹ ARM Science Plan, Ch. 11: Data Management and Documentation Plan, accessible at <http://www.arm.gov/docs/documents/technical/sciplan/sp-11.pdf>
- ² Kehoe, KE, RA Peppler, KL Sonntag, and ST Moore. 2007. "Storing and organizing ARM Program measurements documentation for data quality purposes." Preprints, *14th Symposium on Meteorological Observation and Instrumentation*, January 14-18, San Antonio, Texas, American Meteorological Society, Boston Massachusetts, CD-ROM JP1.12. <http://ams.confex.com/ams/pdfpapers/118999.pdf>.
- ³ Thoeny, Peter. TWiki: Enterprise Collaboration Platform. <http://twiki.org/>.
- ⁴ CamelCase definition <http://en.wikipedia.org/wiki/CamelCase>
- ⁵ Moore, ST, RA Peppler, KE Kehoe, and KL Sonntag. 2007. "Analysis of historical ARM Measurements to detect trends and assess typical behavior." Preprints, *16th Conference on Applied Climatology*, January 14-18, San Antonio, Texas, American Meteorological Society, Boston, Massachusetts CD-ROM, P2.6. Available at <http://ams.confex.com/ams/pdfpapers/119946.pdf>.
- ⁶ Macduff, MC, and RC Eagan. 2005. "ACRF data collection and processing infrastructure." ARM TR-046 Preprints, *21st International Conference on Interactive Information Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology*, January 9-13, San Diego, California, American Meteorological Society, Boston, Massachusetts CD-ROM, J17.1. Available at http://www.arm.gov/publications/tech_reports/arm-tr-046.pdf
- ⁷ Coulter, R.L. and B.M. Lesht. "Results of an Automated Comparison Between Winds and Virtual Temperatures from Radiosonde and Profilers." Proceedings of the Sixth Atmospheric Radiation Measurement (ARM) Science Team Meeting, DOE CONF-9603149, March 1996.
- ⁸ Shi, Y. and C. N. Long, "Best Estimate Radiation Flux Value-Added Procedure: Algorithm Operational Details and Explanations". ARM TR-008, October 2002. http://www.arm.gov/publications/tech_reports/arm-tr-008.pdf
- ⁹ Moore, S.T. and S. Bottone, "NCVweb: an interactive Web-based tool for viewing Atmospheric Radiation Measurement (ARM) data". [19th Conference on IIPS](#), February 10th, 2003, San Diego, CA, American Meteorological Society.

ARM Website and Instrument Handbooks at <http://www.arm.gov/>