Data and Information Management in the Ecological Sciences: A Resource Guide

Edited by

William K. Michener, John H. Porter, and Susan G. Stafford

This publication should be cited as:

Michener, W.K., J.H. Porter, and S.G. Stafford. 1998. Data and information management in the ecological sciences: a resource guide. LTER Network Office, University of New Mexico, Albuquerque, NM.

Publisher: LTER Network Office, University of New Mexico, Albuquerque, NM

Preface

More than 100 individuals attended a two-day workshop (August 8-9, 1997) entitled "Data and Information Management in the Ecological Sciences" that was held at the University of New Mexico in Albuquerque. Objectives of the workshop were to: effect technology transfer, especially at biological field stations and marine laboratories; facilitate people networking; communicate training needs and opportunities; identify future needs for data management at field stations; and produce hard copy and digital versions of the proceedings. Workshop instructors provided comprehensive overviews of the technological infrastructure for a data management system (e.g., hardware, software, communications, and networking), data entry, quality assurance, database management systems, metadata, archival, the World Wide Web, and scientific visualization. Additional roundtable discussions focused specifically on software for field stations, challenges and opportunities at field stations, and site-specific data management implementation.

The editors are grateful to:

- Kathleen Parkhurst, Ellie Trotter, Paula Houhoulis, and Paula Johnson for coordinating onsite logistics;
- Paula Johnson, Patty Sprott, Paula Houhoulis, and Jean Turn for technical editing and production of the hardcopy version of the workshop proceedings;
- Jean Turn for design of the Web version of the workshop proceedings;
- the instructors who all gave 150% effort, particularly Hilary Swain whose caricatures of insects and synthetic abilities were sincerely appreciated; and
- the attendees for their attentiveness and enthusiasm.

To all of the above, we offer a heart-felt THANKS!

THE H. J. ANDREWS CLIMATOLOGICAL FIELD MEASUREMENT PROGRAM

Donald L. Henshaw

U.S. Forest Service Pacific Northwest Research Station,

3200 SW Jefferson, Corvallis, OR 97331

Frederick A. Bierlmaier

Forest Science Department, Oregon State University, Corvallis, OR 97331

Hazel E. Hammond

U.S. Forest Service Pacific Northwest Research Station,

3200 SW Jefferson, Corvallis, OR 97331

Abstract. Research modeling in ecology and hydrology demands the collection of short time-step, spatially distributed climatic measurements. Standardized sets of measured parameters with standard collection methods are critical for comparability of measurements. Standard data archival formats, quality assurance procedures, and sound mechanisms for method documentation are essential for efficient handling of large quantities of electronic data. Web access to near real-time climatological data as well as long-term archives has proven invaluable to researchers.

INTRODUCTION

Research modeling and other spatial studies in hydrology and ecology are placing increasing demands on long-term measurement collection systems. The advancement of remote sensing and Geographical Information System (GIS) technology has spurred the development of spatially distributed ecosystem models requiring short time-step, spatially-distributed measurements to understand processes and test hypotheses at different scales. Coordinated climatological field measurement programs with standardized collection methods, quality assurance procedures, and data archival formats are necessary to meet the needs of multidisciplinary programs. Long-term measurements of climatological variables at various temporal and spatial scales provide a necessary foundation for understanding ecosystem processes and documenting environmental changes (Greenland 1993).

Technical advances in climatological instrumentation such as meteorological sensors, data loggers, and telecommunications allow for routine measurement of critical climatological parameters at fine spatial and temporal scales. Collection of vast amounts of data with improved accuracy is now possible, but quality assurance of these data is not guaranteed without significant changes in field measurement programs. Field collection techniques, quality assurance procedures, and data processing and archival systems must evolve to meet the

demands of rapidly growing data collection efforts. Documentation of the methodologies and archival formats are necessary for the long-term survival and utility of this information.

Researchers also wish to take advantage of new web and telemetry technologies to gain near real-time access to critical stream and weather conditions. Access to climatic summary data is also requested shortly after significant weather events. Computer systems employing telemetry and web technology provide researchers remote access to raw weather data and a basis to determine the necessity of a site visit (e.g., for purposes of storm sampling). Telemetry also provides technicians with a way to monitor the operation and accuracy of these remote stations.

HISTORICAL PERSPECTIVE/BACKGROUND

The hydroclimatological measurement program of the H. J. Andrews Experimental Forest LTER site started in 1952. The existing system evolved as an accumulation of measurement programs with various designs and objectives in response to individual projects and scientist interests (Rothacher et al. 1967, Emmingham and Lundburg 1977, Waring et al. 1978). Until recetly, most meteorological data collection had been driven by very site-specific questions. Climatic stations as well as precipitation and temperature networks were established independently. Field techniques, collection methods, data handling procedures, and the parameters measured were highly variable among measurement sites. In the early 1990's, it became clear that new objectives for the hydroclimatological measurement program had superseded many of the original objectives and would require a more comprehensive network design and new instrumentation (Henshaw et al. 1995).

Much of the Andrews Forest (*http://www.fsl.orst.edu/lter*) occupies rugged terrain with steep mountain slopes and extensive conifer stands. The elevation range of 420 to 1630 m makes most of the forest accessible only by snowcat during most winters. This has traditionally limited the ability to find suitable, accessible sites for meteorological measurement, and especially precipitation measurement. To overcome the limited environmental representativeness of our measurement system, two new additional weather stations were built to give the forest four comprehensive "benchmark" meteorological stations (See site map). Additionally, a system of secondary precipitation and temperature measurement stations has been established to augment the benchmark stations. Secondary stations are strategically located to examine spatial variability of precipitation, air, soil, and stream temperature along environmental gradients. Radiotelemetry was established at the "benchmark" stations and selected stream gaging stations to create "virtual" accessibility to these remote sites.

Modifications were made to both new and existing stations to standardize the variables measured, temporal resolution, collection methodology, sensor heights, and instrumentation. Quality assurance/quality control procedures as well as data archival formats and procedures were also modified and standardized. To track changes in instrumentation and methodology, collection time resolution, and sensor heights, and to fully document problems encountered with meteorological measurement, a database schema for metadata was established (See Figure 1).

Figure 1. Andrews Forest LTER meteorological metadata schema.
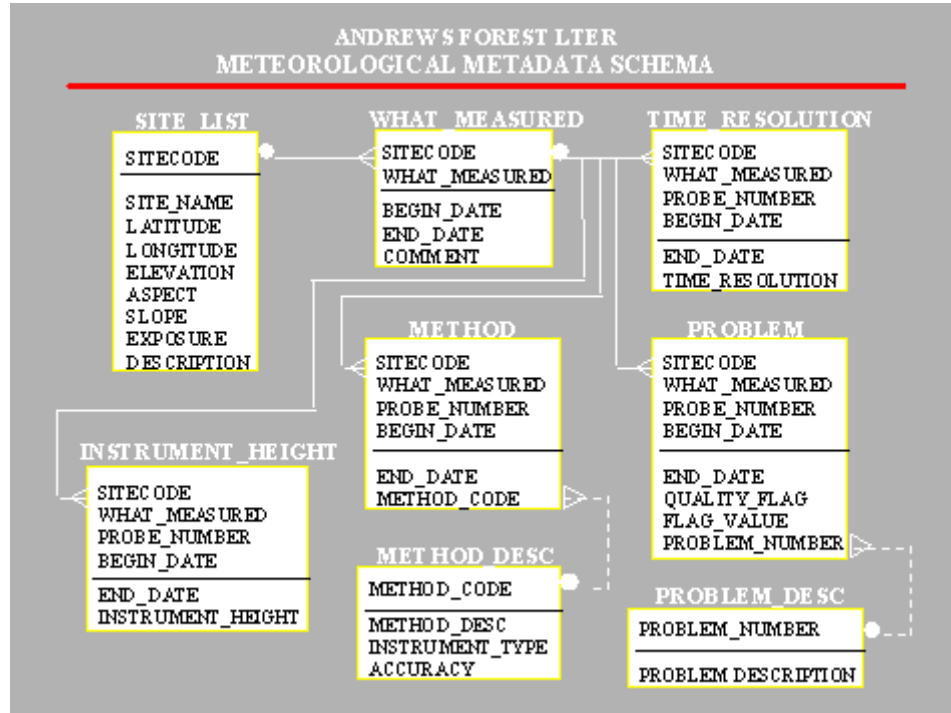
METEOROLOGICAL DATABASES AND METADATA

The Forest Science Data Bank (FSDB) has been actively involved with managing hydrological, meteorological, and other related data for the Andrews LTER for over twenty years (Stafford et al. 1986). Due to the piecemeal evolution of the measurement program, separate databases and data formats were established for these independent efforts. For example, climatic stations collected different sets of variables and were structured in separate databases. The precipitation network and thermograph networks also were independent databases. To prepare for the increasing influx of electronically collected data, databases were standardized for all meteorological data collected on the Andrews. This standardization has greatly simplified processing methods by allowing the use of standard programs for quality assurance checking of all meteorological data. An added benefit is the ability to automatically create web pages for access to these data.

Within the Forest Science Data Bank (FSDB) every data set table has an associated metadata file that: 1) describes the table's variables (attributes) including each variable name, definition, units, and domain (data type and format); 2) indicates whether the variable is coded; and 3) describes attribute information critical for generic quality assurance checking, such as whether a variable a) can be null, b) is a primary key, and c) shows the common variable range (Spycher et al. 1996). When attributes are listed as coded, a second metadata file lists all coded variables, all valid codes, and code definitions. The meteorological data take advantage of these metadata files to document the common database structures and properly define the measurement variables. Table 1 lists the independent tables that are established for the measurement parameters. Separate tables are maintained for daily and higher temporal resolution data.

All documentation or metadata regarding the hydroclimatological measurement system is stored and maintained by the FSDB. Records of all changes in field instrumentation, sensors, locations, sensor heights, and temporal resolutions are maintained for every measurement variable in a relational database schema (See Figure 1). Tables within the schema include specific measurement site descriptions and the parameters measured at each station with spanning dates. Histories of changes in collection resolution, sensor heights, instrumentation, and encountered problems are maintained. To avoid redundant descriptions, independent tables are established to describe methods (with instrumentation) and problems that were encountered.

QUALITY ASSURANCE (QA)

Electronic data collection provides a mechanism to collect immense quantities of data with very sophisticated measurement sensors and data loggers. Meteorological data are commonly collected at 15-minute or hourly time intervals and sometimes with multiple sensors for each measurement parameter. This is a tremendous change from the historical data collection efforts which consisted primarily of digitized and manually read records from various recording charts and punch tapes, and hand-recorded field notes. However, the transition to electronic data collection does not assure that the collected records are reliable and error-free, and the data require extensive quality assurance (QA) checking. Also lost is the instantaneous feedback the field technician had when examining a chart in the field. This bird's-eye view of seeing the overall graphical picture is replaced with a worm's eye view of the data logger's single discrete digital values.

In October 1991, the LTER Data Managers conducted a survey on the use of electronic data collection instrumentation at the LTER sites (Ingersoll and Chapal 1992). It was discovered that most meteorological data are collected, stored, and transferred by electronic means at LTER sites. The resultant survey responses provided an assessment of common problems and solutions encountered in electronic data collection and describe the QA procedures at the LTER sites. It is clear that while new electronic technologies have greatly reduced the need for manual data entry and digitizing, field techniques and QA protocols need to evolve to reduce the increasing chances of undetected problems residing in final data archives.

QA procedures for electronically collected data must involve field technicians, the site data manager, and the principle investigators to be most effective. Additionally, the ability to detect problems quickly provides more timely feedback to field technicians. Field procedures used at the Andrews and other LTER sites to help provide quality assurance include:

- Routine equipment inspection: Pre-designed check sheets that can lead the technician through standard protocols are very useful.

- Routine instrument calibration: Local site or manufacturer recalibration or instrument replacement is critical.

- Independent sensor checks: Make independent measurements side-by-side with the field instrument to ascertain field sensor reliability. The reliable lifetime for some electronic sensors can be relatively short (~1 year).

- Redundant measurements: Provide an alternate sensor as a backup. For example, a non-electronic recording chart might be used alongside an electronic data logger.

- Telemetry: Radio or satellite telemetry is an invaluable tool in knowing whether a station is functioning properly and can quickly alert technicians to measurement problems.

- Mechanisms for noting problems or making comments: Field recorders, field notebooks or check sheets, and tape recorders provide a means to describe observations.

- Communication: Strong communication between the field technicians, data manager and principle investigators.

QA continues once the data reach the laboratory. The Andrews LTER performs three types of programmatic quality assurance checks of electronically collected meteorological data.

Generic checking is performed on every data set in the FSDB by referencing the metadata to find and report potential problems in the data. Problems might include data values out of normal range, data values missing when "not null" is specified, use of improper codes (e.g., for coded variables such as data quality flags), and improper data structure or data format of ASCII files.

Special rule checks specific to climate data include: 1) checks for date and time validity and proper sequencing; 2) checks that minimum<=mean<=maximum when applicable; and 3) checks for proper placement of missing flags in the quality code fields. Computer code for implementing these special rules is maintained with the climatic metadata. Visual checking of data values is performed by: 1) graphing a single parameter over time; 2) graphing redundant or multiple sensors for comparison, and 3) graphing related or associated parameters together. Examples include: 1) displaying a station's mean, maximum, and minimum values over time in a moving time graph; 2) creating scatter plots of the data of one site versus another; and 3) displaying wind data in polar graph form. Models have also provided valuable insight into discovering inconsistencies in measurement sensors, e.g., snow hydrology models (Duan 1996).

Once the quality assurance checking has been completed, there remain substantial questions in determining how problem data should be documented. The Andrews LTER data structures employ data quality flags for every measured value. Determining whether a value is tagged "questionable" or "missing" is very subjective. Inaccurate values showing a consistent bias may still have redeeming value and are typically coded "questionable" and retained. Very questionable data may be ultimately coded "missing" and discarded if no interpretation is possible. Since researchers and modelers often require that no values be "missing" over key time periods, these missing values will frequently be estimated. Most estimation techniques and equations are saved in the metadata, and values are tagged "estimated." Field technician notebooks or logs with comments can be invaluable in evaluating problem data. A "Comments" database table is maintained to describe periods of flagged data and provides a useful log of problems and solutions.

DATA ACCESS

Historical climatic and stream flow data were collected on strip charts, circular charts, punch tapes and hand-written check sheets. In the early 1980's, weather station technology allowed capture of our primary station's data on a nearby computer at the Andrews headquarters (Bierlmaier and McKee 1989). However, all data were hand-delivered to the Corvallis research facility, data processing took weeks and in some cases years, and researchers had no access to any real-time weather condition data. In 1993, a dedicated 56Kb direct network link was established, greatly improving the connectivity of Andrews computers with the Corvallis research group's local area network.

In 1995, radio telemetry capability was added to the remote benchmark stations and two stream gauging stations providing instantaneous access to remote weather stations (See Plate 7). During a storm in February of 1996 when access to remote sites was impossible due to washed out roads, fallen trees, and flooding, radio telemetry provided data to researchers that would not otherwise have been available.

In 1996, a suite of Microsoft™ software, including Foxpro™ were used in conjunction with Campbell Scientific™ software and data loggers to develop programs for automatically accessing raw telemetry weather station data and processing the data into web pages accessible over the World Wide Web (*http://fredb2.fsl.orst.edu/*). This near real-time weather web page is updated hourly in the winter and daily in the summer. All of the Andrews Forest LTER historical and current climate databases can be found on the WWW at *http://www.fsl.orst.edu/lter/data/studies/ms01/ms01fmt.htm*. All metadata files associated with this as well as the data sets themselves are written to HTML and ASCII files for web access.

CONCLUSION

As research modeling in ecology and hydrology demands the collection of short time- step, spatially distributed climatic measurements, coordination of climatological field measurement programs becomes very important. Standardized sets of measured parameters with standard collection methods are critical for comparability of measurements. Standard data archival formats, quality assurance procedures, and good mechanisms for method documentation are essential for efficient handling of large quantities of electronic data and their long-term usefulness. Web access to near real-time data has also proven invaluable to researchers in assessing remote site conditions.

ACKNOWLEDGMENTS

LITERATURE CITED

Bierlmaier, F. A., and W. A. McKee. 1989. Climatic summaries and documentation for the primary meteorological station, H.J. Andrews Experimental Forest, 1972 to 1984. Gen. Tech. Rep. PNW-242. US Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, OR.

Duan, J. 1996. A coupled hydrologic-geomorphic model for evaluating effects of vegetation change on watersheds. Dissertation. Oregon State University. Corvallis, OR

Emmingham, W.H., and G.A. Lundburg. 1977. Climatic and physiological data summaries for the H.J. Andrews reference stand network. Coniferous forest biome internal report 166. University of Washington, Seattle, WA.

Greenland, D. 1993. The climate of the H.J. Andrews Experimental Forest, Oregon, and its regional synthesis. USDA Forest Service, Pacific Northwest Research Station. Cooperative Agreement No. PNW 92-0221.

Henshaw, D.L., W.A. McKee, and A. Sikka. 1995. Program for hydroclimatological measurement at the H.J. Andrews Experimental Forest. *http://www.fsl.orst.edu/lter/research/compplns/clima/hjamet.htm*

Ingersoll, R., and S. Chapal. 1992. Management of electronically collected data within LTER. *http://www.fsl.orst.edu/lter/im/ingersol.htm*

Rothacher, J., C.T. Dyrness, and R.L. Fredriksen. 1967. Hydrologic and related characteristics of three small watersheds in the Oregon Cascades. U.S. Department of Agriculture, Forest Service, Pacific Northwest Forest and Range Experiment Station. Portland, OR.

Spycher, G., J.B. Cushing, D.L. Henshaw, S.G. Stafford, and N. Nadkarni. 1996. Solving problems for validation, federation, and migration of ecological databases. Proceedings of the Eco-Informa Workshop, Global Networks for Environmental Information, November 4-7 1996, Lake Buena Vista, FL, 11:695-700. Environmental Research Institute of Michigan(ERIM), Ann Arbor, MI.

Stafford, S.G., P.B. Alaback, K.L. Waddell, and R.L. Slagle. 1986. Data management procedures in ecological research. Pages 93-113 in W. K. Michener, editor. Research data management in the ecological sciences. The Belle W. Baruch Library in Marine Science No. 16. University of South Carolina Press, Columbia, SC.

Waring, R.H., R.H. Holbo, R.P. Bueb, and R.L. Fredriksen. 1978. Documentation of meteorological data from the coniferous forest biome primary station in Oregon. Gen. Tech. Rep. PNW-73. US Department of Agriculture, Forest Service, Pacific Northwest Forest and Range Experiment Station Portland, OR.

Plate 7. H.J. Andrews Experimental Forest Telemetry System. (From Chapter 19, Henshaw et al., The H.J. Andrews Climatological Field Measurement Program.)