7-2-2012

Data access and visualization benefits from implementation of a hydrologic information system

Stephen Brown

Stephen Wesley Brown
Candidate

Civil Engineering
Department

This thesis is approved, and it is acceptable in quality and form for publication:

Approved by the Thesis Committee:

Julie Coonrod, Chairperson

Karl Benedict

Mark Stone

DATA ACCESS AND VISUALIZATION BENEFITS FROM IMPLEMENTATION OF A HYDROLOGIC INFORMATION SYSTEM

BY

STEPHEN WESLEY BROWN

Bachelor of Science
Earth & Planetary Sciences
University of New Mexico, 2010

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Science
Civil Engineering

The University of New Mexico
Albuquerque, New Mexico

May 2012

ACKNOWLEDGEMENTS

My Excellent Committee:

Julie Coonrod, Karl Benedict, Mark Stone

Research Scientists:

James Thibault, Chi Bui

Funding and Technical Support:

NMEPSCoR, EDAC, CUAHSI

Unwavering Support from Family:

Brown and Thomson

Sustenance on Long Days:

Espresso, Breakfast Burritos, Rice Crispy Treats

DATA ACCESS AND VISUALIZATION BENEFITS FROM IMPLEMENTATION OF A HYDROLOGIC INFORMATION SYSTEM

BY

STEPHEN WESLEY BROWN

B.S. EARTH AND PLANETARY SCIENCES, UNIVERSITY OF NEW MEXICO 2010

M.S. CIVIL ENGINEERING, UNIVERSITY OF NEW MEXICO 2012

ABSTRACT

In 2010, the National Science Foundation (NSF) implemented new guidelines for all scientists applying for grants. A Data Management Plan (DMP) is now required for all proposals in which data are created or gathered while working under the grant. Several organizations have produced templates and applications to assist with the construction of DMPs. The data plans provide a good overview of data processing and storage but do not provide any guidance for managing data during the research process.

Large temporal hydrologic data sets can provide a rich insight to complex hydrologic and ecological systems. Complications arise when attempting to query and present the data in ways that are useful for exploring and validating research hypotheses. Common tools, such as Excel or Matlab, may be helpful if you know the exact sequence of data you want to analyze. Frequently, this is not the case. Looking at long term trends, adding and removing additional variables, or comparing local results to external national datasets are difficult or impossible with these tools.

To overcome the limitations of current data management methods, a Consortium of Universities for the Advancement of Hydrologic Science Inc. - Hydrologic Information

System (CUAHSI-HIS) server was deployed in collaboration with Earth Data Analysis Center (EDAC) and the New Mexico Experimental Program to Stimulate Competitive Research (NMEPSCoR). Data products on the server are stored in a relational database using WaterML, an XML based language introducing standardization to the hydrologic community and facilitating distribution and aggregation of hydrologic data.

Four project types from different agencies have been selected to explore the process of obtaining and ingesting data into an HIS. Three of the projects are university based with different stakeholders and the fourth is a state funded project carried out by a contractor.

Tools developed by CUAHSI for ingesting measurements into the database made processing the raw data straightforward. After the data were formatted properly, automated processes allowed millions of measurements to migrate from Excel files into the HIS. Aggregating the data and metadata without support from the principal investigator proved difficult. Deciphering the provenance of derived data proved exceptionally difficult from a data manager perspective with little experience in specialized disciplines.

Datasets that previously required hours to download, aggregate, and visualize are can now be processed in minutes. Repetitive analysis tasks can be automated within the HIS, integrating local regional, and national datasets by spatial and temporal extent and delivered to the research team in a variety of formats. The CUAHSI-HIS components make data discovery and analysis streamlined in addition to satisfying the NSF DMP requirements.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

SECTION A: DATA MANAGEMENT

CHAPTER 1: INTRODUCTION AND BACKGROUND

1.1 Research Motivation

In 2010, the National Science Foundation (NSF) implemented new guidelines for all scientists applying for grants. A Data Management Plan (DMP) is now required for all proposals in which data are created or gathered while working under the grant. The initial focus of this research was to analyze the methods to satisfy the new DMP requirements and examine the effectiveness of the open source Consortium of Universities for the Advancement of Hydrologic Science, Inc - Hydrologic Information System (CUAHSI-HIS) on the data management and accessibility portion of DMP. Upon speaking with many Principal Investigators (PIs), research scientists, and college students the scope broadened to encompass the general issue of hydrologic discovery.

Three key elements encompass hydrologic discovery:

1. Local File Management
2. Hydrologic Data Reconnaissance
3. Researcher Workflow Tools

This manuscript explores issues that arise when trying to migrate archive datasets from multiple agencies in to a public HIS, the impact of implementing a CUAHSI-HIS on the three key elements of hydrologic discovery, and the effectiveness of the CUAHSI-HIS on satisfying NSF DMP requirements.

1.2 Importance of Research

Corporations have long known the value of efficient relational databases for tracking inventory, employee productivity, and consumer spending (IMT Strategies, 1999). In 1971, Terrence O'Brien developed a model to analyze consumer spending by determining the relationships between specific behavioral variables (O'Brien, 1971). Even though relational database systems were still in the semi-theoretical phase (Baxendale and Codd, 1970) in the early 1970's, O'Brien was already using the foundations of relational database technology to help corporations sell more products. Fast forward forty years and most researchers in the hydrologic sciences are still storing data collected from the field or laboratory in flat file spreadsheets.

Over the course of this research project, considerable time was spent discussing data collection, processing, and management with scientists in a wide array of disciplines. A majority said they spent more time getting data ready to analyze than actually conducting the analysis. Many times, the final analysis is conducted in Excel, limiting the scope of the investigation to small temporal and spatial slices manageable in a spreadsheet.

One researcher had half a dozen years of data from several locations and instrument clusters stored in spreadsheets on his computer. When asked how regional trend analysis for the project area was conducted, he replied "Excel". Tens of thousands of research dollars and thousands of hours of labor went into collecting that data without a viable method for extracting valuable knowledge from the dataset. In addition, a regular backup regime is not in place, risking loss of the entire project history.

Another research team had graduate students from multiple universities sharing data from a well instrumented long term research site. One student would visit the site, download the data, and email a copy to the student at the other university. The individual researchers were diligent about keeping project notebooks but there was no tracking system to make sure all analysis was conducted on the same unaltered dataset. The provenance of the data was unavailable and conducting detailed peer review of any collaborative papers complicated.

The National Science Foundation is aware of the problems associated with data collection, distribution, and storage, responding with the DMP requirement in 2010. Mandating research scientists prepare DMPs, is a giant step forward in data accountability. The next step, hopefully coming in the future, is a Data Workflow Plan (DWP) that records the data collection and analysis process in detail. Full provenance of the data would be recorded from programming of the device to final QA/QC in a public database.

When discussing data management and processing with research scientists, many are unaware of products and services available to streamline their workflows. Open source databases that store measurements in a standardized, portable format with well developed processing and visualization tools are available now.

The objective of this research is to chart a path for PIs, researchers, and students to:

1. Easily satisfy the NSF DMP requirements
2. Save time managing data from instruments and models
3. Maximize research team access to project data
4. Standardize data for discovery, visualization, and analysis
5. Conduct rapid reconnaissance of large hydrologic datasets

Even though hydroinformatics was referenced as early as 1991 (Abbott, 1991), incorporating information technology in the hydrologic sciences has been slow. Hydroinformatics is still, after twenty years, an emerging science.

1.3 Data Management Plan

In 2010, the National Science Foundation (NSF) implemented new guidelines for all scientists applying for grants. A Data Management Plan (DMP) is now required for all proposals in which data are created or gathered while working under the grant. University libraries, software developers, and NGOs have prepared documentation to assist with creation of DMPs (CSDMS, 2012, DataONE, 2012, Brunt, 2012, Olendorf, et al., 2012). Useful tools, like the CUAHSI-HIS, are being developed to satisfy the basic requirements of the NSF leaving the methods of managing the data to the researcher's discretion.

The NSF Proposal and Award Policies and Procedures Guide states:

> *Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data. . .created or gathered in the course of work under NSF grants. (NSF, 2010)*

The NSF Proposal Preparation Instructions, provide guidelines to ensure a proposal will conform to NSF policies for research data distribution and sharing (NSF, 2012):

1. *the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project;*

2. *the standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies);*

3. *policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements;*

4. *policies and provisions for re-use, re-distribution, and the production of derivatives; and*

5. *plans for archiving data, samples, and other research products, and for preservation of access to them.*

When deploying a new instrument to the field, a scientist will likely use the tools at their immediate disposal to access, visualize, and store datasets. The first line of defense is frequently the software that was delivered with the device or Microsoft Excel. Researchers conducting analysis of large hydrologic systems frequently are required to process and visualize thousands of data points over large time scales to capture the intricacies of complex system dynamics. Processing and viewing these large datasets in Excel is inefficient and may leave scientists without the opportunity to explore correlations between disparate datasets. An effective method of satisfying the NSF requirement and facilitating efficient processing and analysis of large datasets is to store all measurements and resulting products in a hydrologic database.

This paper aims to provide insight to the process of deploying a university based Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) HydroServer

hydrologic database system to satisfy the key NSF requirements of data and metadata standards, data access, sharing, and long term archiving. Several projects were chosen for analysis, showcasing acquisition of hydrologic data from researchers and ingesting large datasets as well as complications and special issues that arise when working with researchers from a data manager perspective.

1.4 Data Management Plan Implementation

Template DMPs are available from many universities and organizations that rely upon NSF funding. These templates vary by discipline and provide general guidelines to satisfying the requirements of the NSF. October of 2011, the DMPTool was released by a consortium of universities and organizations, providing guidance and resources for creating data management plans (UC Curation Center, 2012). The DMPTool is a web based application that steps a researcher through the sections of a DMP customized for their field of study. After an account is created, DMPs for various projects can be created, shared, and saved. Universities are able to customize the DMPTool for their specific data management and warehousing. For example, the University of New Mexico (UNM) may develop an NSF-EAR (Earth Sciences) template for hydrologic science that pre-fills fields with details pertaining to data being stored at the Earth Data Analysis Center (EDAC) data farm and archived in LoboVault (UNM library research archive) at the completion of the project. Details specific to standards of practice developed at UNM may include server configuration, backup regime, database structure, metadata format, etc (DataONE, 2012). Standardization will streamline DMP preparation and assist with budgeting for data and workflow management services.

1.5 Data overload

Development of a data management plan does not outline the actual measurement processing steps or methodologies used during active research. Data visualization is composed of two parts. The physical location of the data in a directory structure and the information contained within each file. Figure 1 shows a Sunray tree, produced by Treevis (Randelshofer, 2012), of the physical location of the data and metadata collected for a multi-year multi-station evapotranspiration and well project on the Rio Grande. Each rectangle in the chart is a folder or file totaling 19,630 folders or files and 9.3GB of data, resulting in over 35 million hydrologic measurements. After the physical data files are processed and understood, the measurements contained within can be paired with metadata and ingested in a database for analysis of the actual hydrologic information.

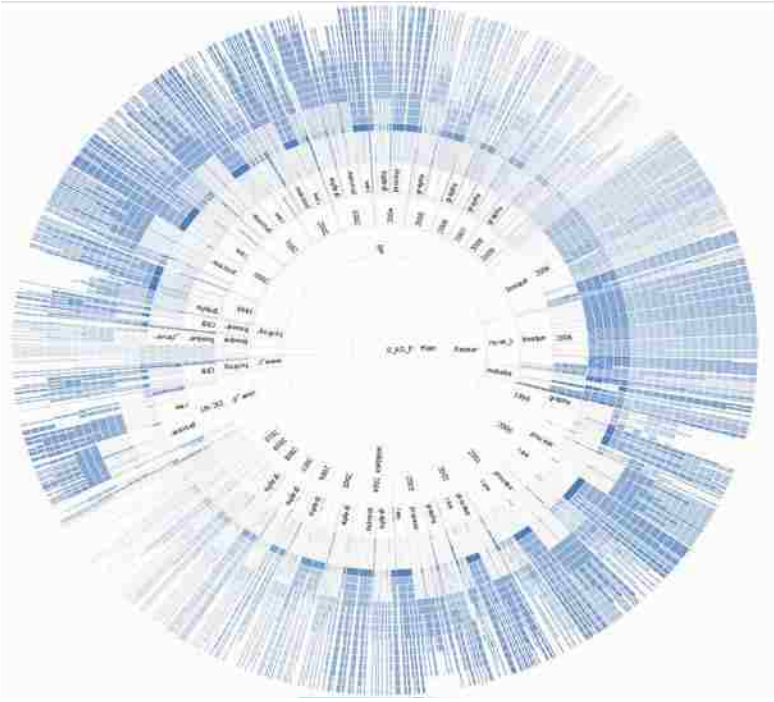


Figure 1: Visualization of Rio Grande ET file structure on disk.

Data management quickly becomes the primary task for a researcher when presented with an abundance of measurements. Conducting analysis on the data is severely limited by the technologic skill of the researcher. An ecologist or biologist must now become an information technologist, skilled in programming simply to visualize the dynamic system captured by the instrumentation.

A proliferation of instrument deployments has led to an exponential growth in the amount of data collected in recent years. As of 2008, the United States Geologic Survey (USGS) National Water Information System (NWIS) and Environmental Protection Agency (EPA) STOrage and RETreiveal (STORET) databases held 350 million data points (Beran and Piasecki, 2008), factoring all other government agencies, universities, and research bodies into the total count, there are more data available today than at any time in history. More measurements are being collected every second, adding to the knowledgebase. Pressure transducers and water quality sondes have recently dropped in price from a range that only government and large corporations could afford to reasonable amounts for small municipalities and educational institutions.

Researchers must process, error check, and preview millions of measurements to publish peer reviewable results. Much of the analysis time is spent pushing data around and not doing any real problem solving. The situation is exasperated when attempting to compare locally collected data with regional or national datasets.

Familiar analysis tools collapse under the weight of millions of rows of data. Microsoft has addressed growing datasets by increasing Excels data handling over the last 20 years from 16,384 to 1,048,576 rows (Office Watch, 2012). Performing a calculation on an Excel sheet

with more than 250,000 rows still requires a robust computer workstation. Elegant and faster solutions are available in the form of custom programming: C++, Java, Python, Matlab, etc. Although programming offers a solution for the computer savvy hydrologist, it does not provide an easy path to continue the research when the programming hydrologist moves to another assignment. Incorporating undergraduate and graduate students with limited computer and programming skills into the data collection and analysis process is hindered as well. A better solution to address core data acquisition, management, and visualization is necessary.

1.6 Paradigm Shift in Hydrologic Data Management

The *"Committee on Opportunities in the Hydrologic Sciences, Water Science and Technology Board"*, was created by the National Research Council in 1991 to address pressing issues in the hydrologic sciences. Key data requirements outlined in the "*Opportunities in the Hydrologic Sciences / Committee on Opportunities in the Hydrologic Sciences, Water Science and Technology Board, Commission on Geosciences, Environment, and Resources, National Research Council"* publication, here forward referred to by its colloquial name, the Blue Book, include maintenance of long term data sets, improved information management, dissemination of data from multidisciplinary experiments, and extensive student interaction with the field and laboratory research process. Since most hydrologic science is multidisciplinary, the Committee suggested open access to products of observation and experimentations to the scientific community at large. Critical to the advancement of hydrologic science, datasets need to include comprehensive metadata including purpose,

location, instruments, spatial and temporal range, etc. The data then need to be cataloged and archived allowing efficient mining by future scientists (National Research Council, 1991).

In 1992, Jeff Dozier developed a 'Data base centric' model for interacting hydrologic data. The key elements were bi-directional communication between a database management system and the user by way of recipe management, graphical query language, intelligent search, and visualization. Several issues have slowed the development of a robust hydrologic database system: storage speed and capacity, network bandwidth, relational database structures, and visualization software (Dozier, 1992). Technology has now matured enough to assemble functional data management systems. Moore's Law has held true, essentially doubling computer speeds every eighteen months (Miller, et al., 2009). Hard drives are now affordable in 1TB+ sizes with 6Gb/s transfer rates, 1.5Mbs wide area network speeds are common even in homes, the Internet has pushed database optimization forward, and photo and video editing are possible on basic home computers.

1.7 CUAHSI Hydrologic Information System (CUAHSI-HIS)

The Consortium of Universities for Advancement of Hydrologic Science, Inc. (CUAHSI) has developed an open source Hydrologic Information System (HIS) to manage temporal instrument and model data. The data are stored in an Observation Data Model (ODM) specifically developed in a relational database structure to efficiently store hydrologic data and metadata. The combination of these tools resulted in the development of three key components of the HIS (CUAHSI, 2012).

The CUAHSI-HIS is composed of three key components (Figure 2):

HydroServer: Data storage, portability, and distribution

HIS Central: Metadata repository for HydroServers and national datasets

HydroDesktop: Data discovery, visualization, and analysis

The three separate but interlinked components have been built attempting to solve the issues raised in the Blue Book. The HydroServer was designed to store hydrologic measurements, facilitate data ingest, streamline quality assurance, and distribute data to local and remote researchers. HIS Central is a master metadata library, storing data about the data located on the network of HydroServers and national hydrologic databases. HydroDesktop is a desktop application allowing researchers to search through dozens of hydrologic databases at once by spatial and temporal extent, graph the measurements, and conduct statistical and modeling analysis via integrated and custom plug-ins (CUAHSI, 2012).

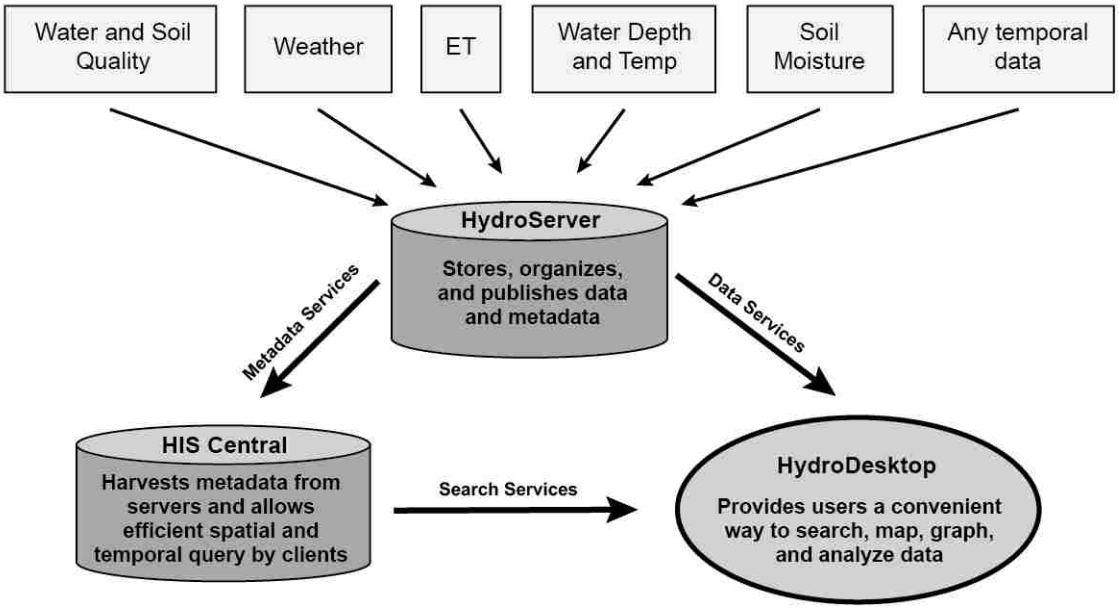


Figure 2: CUAHSI-HIS components, HydroServer, HIS Central, and HydroDesktop

The trinity of tools for storing, sharing, and querying data within the CUAHSI-HIS is very powerful but one problem still remains; national organizations each store data in custom formats. These databases cannot communicate with each other. Two solutions were developed to address these issues, WaterML and Controlled Vocabularies.

WaterML is a new language developed by CUAHSI to facilitate communication of hydrologic data. Based on eXtensible Markup Language (XML), WaterML standardizes variable names and units allowing communication between disparate hydrologic databases with the goal of having a universally accepted communication structure within the international hydrologic community (Open Geospatial Consortium, 2008). In 2011, OGC started the WaterML 2.0 Standards Working Group (SWG) to develop a hydrologic data standard consistent with the OGC Standards Baseline, building on the foundation of WaterML 1.0 and 1.1. International support for the WaterML 2.0 SWG is illustrated by the current members of the working group; CUAHSI, KISTERS, Australian Bureau of Meteorology, National Oceanic and Atmospheric Administration (NOAA), Geological Survey of Canada, Deltares, 52°North, USGS, and German Federal Institute of Hydrology (Open Geospatial Consortium, 2012).

Controlled Vocabularies (CVs) are standardized terms used to describe hydrologic concepts (Table 1). Creating a standard set of terms allowed mapping of unique databases to one common description, facilitating data sharing between systems. The master CVs are published as a set of XML web services from CUAHSI, that are locally stored in the HydroServer database (CUAHSI, 2012). Periodic updates of the CV can be implemented locally via CUAHSI's ODM Tools at the discretion of the server manager. Requests for additions are managed by CUAHSI.

Table 1: CUAHSI Controlled Vocabularies

CensorCodeCV:	Used to populate the CensorCode field of the DataValues table
DataTypeCV:	Used to populate the DataType field of the Variables table
GeneralCategoryCV:	Used to populate the GeneralCategory field in the Variables table
SampleMediumCV:	Used to populate the SampleMedium field in the Variables table
SampleTypeCV:	Used to populate the SampleType field in the Samples table
SiteTypeCV:	Used to populate the SiteType field in the Sites table
SpatialReferences:	Defines the coordinate systems used in the Sites table
SpeciationCV:	Used to poputate the Speciation field in the Variables table
TopicCategoryCV:	Used to populate the TopicCategory field in the ISOMetadata table
Units:	Defines the units used in the Variables and Offset types tables
ValueTypeCV:	Used to populate the ValueType field in the Variables table
VariableNameCV:	Used to populate the VariableName field in the Variables table
VerticalDatumCV:	Used to populate the VerticalDatum field im the Sites table

1.8 Standardization

Critical to understanding large scale hydrologic systems is standardization. Multiple agencies and research groups have equipment deployed, each with a different methodology for variable designation, unit identification, and data warehousing hindering interoperability. WaterML 2.0 is currently under a working committee with the Open Geospatial Consortium (OGC) to be defined as the international standard for sharing hydrologic data. Wide acceptance of an international hydrologic data naming and storage standard will allow independent researchers to easily compare national datasets with locally gathered measurements. Third party open source and commercial software developers will be motivated to write WaterML conduits, streamlining data import and export, expanding the opportunities to process, model, and visualize data in yet to be discovered ways.

The CUAHSI HydroServer currently delivers data in WaterML 1.x format with development underway to support migration to WaterML 2.0, proving an excellent foundation for a small

research group looking to streamline their data management. Additionally, the databases can easily be moved from one HydroServer to another in the event of a hardware failure or increase in demand for the data. One HydroServer can house multiple distinct databases with different permissions and local or global accessibility. Data from different projects are not mingled.

In research environments it is common for data to outlive the Principal Investigators (PI) involvement on the project. Researchers that use custom programming to process and store project data leave the next coordinator to figure out what data management methods were used. The new team member may be forced to learn an arcane programming language simply to process current data streams. Using a HydroServer to store raw incoming and QA/QC data provides a standard platform for data ingest. The metadata describing the processing steps is incorporated and available for future researchers to verify that no errors were introduced were verifying the measurements. Full documentation for setup and management of the HydroServer is available from CUAHSI.

1.9 HydroServer Deployment

A CUAHSI-HIS server system has been deployed in collaboration with EDAC located at the University of New Mexico (UNM), Albuquerque and the New Mexico Experimental Program to Stimulate Competitive Research (NMEPSCoR). EDAC provided a virtual server in EDAC's data center with a high speed connection to the Internet allowing international access to the regional datasets. NMEPSCoR provided funding for this study. The virtual server was deployed to meet the specifications outlined in the HydroServer Setup and Prerequisites guide (Valentine, 2012).

The CUAHSI development team designed HydroServer as a complete research data presentation tool. This includes the core database structure, tools for loading small time series datasets and real-time streaming datasets, QA/QC visualization systems, and end user website and mapping interfaces. Our installation is primarily focused on data ingest and distribution.

Data products on the server are stored in an Observation Data Model (ODM) specifically developed to manage time series data in a relational database (Horsburgh, et al., 2008). The formal structure of the database or the Database Schema (Appendix C) was developed from the ground up to incorporate metadata and data. Integrated metadata management ensures future researchers will have access to the provenance of the data.

WaterML, an XML based language specifically designed to facilitate distribution of hydrologic data and metadata, is used to transfer data streams. WaterML also acts as a translator between the CUAHSI-HIS and external hydrologic databases, like the USGS's NWIS database. Service requests are made with CUAHSI's WaterOneFlow web service, using HTTP based REST requests. REpresentational State Transfer (REST) architecture involves a client computer sending a HyperText Transfer Protocol (HTTP) request containing a detailed description of the requested information to a waiting server. Upon receiving the request, the server responds with a parcel of data and resumes waiting (Vinoski, 2007). REST interfaces are uniform and can be incorporated easily into software with external data access capability. HydroServer WaterOneFlow REST services are available from several popular analysis packages, Excel, Matlab, and HydroDesktop.

Microsoft SQL Server provides the database foundation for the CUAHSI ODM. SQL Server provides a robust, scalable environment for large hydrologic datasets. A version of the

CUAHSI ODM has been customized for mySQL, an open source relational database. Having a free option to storing and distributing hydrologic data using the standard CUAHSI data model will help many small schools and institutions with limited budgets.

ESRI's ArcServer is an optional component installed on the server for online publishing of dynamic maps. Map integration with the HIS can provide land use/cover, soil type, and more for the project region. Instrumentation locations are automatically updated on web maps from the geo-referenced locations stored in the HIS.

Key to the flexibility of the CUAHSI-HIS is the integration of metadata and controlled vocabularies to the data model. The controlled vocabularies are standardized descriptions for hydrologic variables. Linking measurement types from each new dataset to the standardized descriptions allows data collections with different names and units to be queried through HIS Central. To facilitate metadata and variable setup in the HydroServer, an Excel spreadsheet was created with dynamic dropdown menus to select the controlled vocabularies. Using this spreadsheet has assisted in gathering all required metadata from the researchers in the most expedient manner.

See Appendix A for a detailed overview of the server configuration.

CHAPTER 2: DISCOVERY

2.1 Taming the Tree

As illustrated by the 19,000+ folder and file, directory tree in the physical data storage of the Rio Grande Evapotranspiration (Rio-ET) project (Figure 1), conducting analysis is often

encumbered by inefficient, archaic storage methods. Using directory tree file management is the first line of defense for any computer user trying to organize important documents and data. When a dataset grows to multiple years with multiple stations contained in dozens of separate folders and files, the ability to conduct data discovery is severely limited. Finding the station of interest may be simple but aggregating several years for that station is time consuming. Similarly, finding one year may be straight forward but merging several sites is daunting. A scientist may be limited to studying a small subset of collected variables or reduce the spatial extent, potentially excluding critical influencing factors from the analysis.

Evaluation of dynamic environmental systems requires data from more than source. A scientist frequently needs data from several data repositories, both regional and national. The most common national data source for hydrologic information is the National Water Information System (NWIS), provided by the USGS. This Internet based data repository contains historic and real-time measurements from more than 1.5 million water-data collection sites in the US and Puerto Rico (2002). Additional common national data repositories for hydrologic analysis include the National Weather Service (NWS), and the Environmental Protection Agency (EPA).

Traditionally, obtaining data from these agencies required visiting each agency website, finding the location of the data, the station of interest and temporal range then downloading the data. This process would be repeated for every station at each agency. Assuming the spatial extent of the study area is known, obtaining the necessary data is simple albeit time consuming. If the research area or temporal range is expanded the entire search process must be conducted again. Further complications arise when trying to integrate locally collected

data with other regional or national datasets. A research scientist can be left with dozens of tables in different formats and time ranges to aggregate before any data analysis can begin.

2.2 Case Study: Location of national datasets near Rio-ET sites

CUAHSI's HydroServer was used to ingest ten years of evapotranspiration and groundwater data from a series of towers and wells running along the Rio Grande from Albuquerque to Bosque del Apache. Configuration of the HydroServer was straight forward and well documented by CUAHSI. Familiarity with Microsoft Windows Server and SQL Server were helpful but not necessary to setup a stable HIS platform. Technical assistance from the EDAC at UNM was valuable and a key component to future researchers developing a data ingest and QA/QC workflow. HydroDesktop was used to determine availability of national datasets near the Rio Grande ET towers.

2.2.1 Query regional HydroServer

The Rio-ET tower locations were initially queried directly from the regional HydroServer hosted at EDAC. A HydroServer does not need to be registered with HIS Central to access the data. When a project is in the data collection phase, it may be beneficial to keep the data private until the results have gone to press.

Several online basemaps, to assist finding the region of interest, are built-in to HydroDesktop. Within a couple minutes a map of the towers is available with custom icons for the project sponsor, NMEPSCoR (Figure 3). The label engine in HydroDesktop is set to avoid collisions resulting in several stations being unlabeled. Zooming in to the map reveals more station names (not shown).

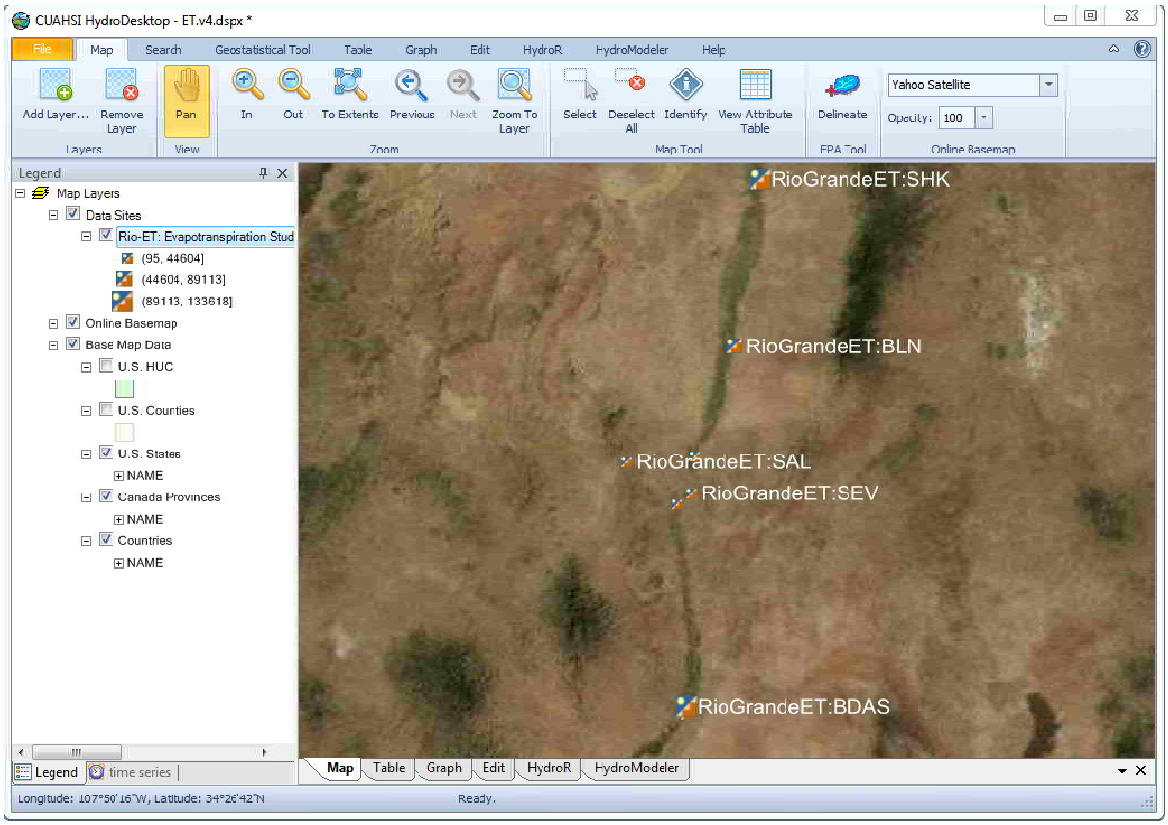


Figure 3: HydroDesktop: Query results for local project data

2.2.2 Query HIS Central for nearby data

In Figure 4, a regional search is conducted of the HIS Central hydrosphere near the Sevilleta National Wildlife Refuge, returning over 3500 data series from the NWS, EPA and NWIS surface and ground water sites. Labeling was set to show the agency and station number for easy identification. Detailed metadata from HIS Central contains key station features including: station name, station number, sample frequency, quality control level, and XY location. A shapefile with all queried station locations and associated metadata can be exported for use in other mapping applications.

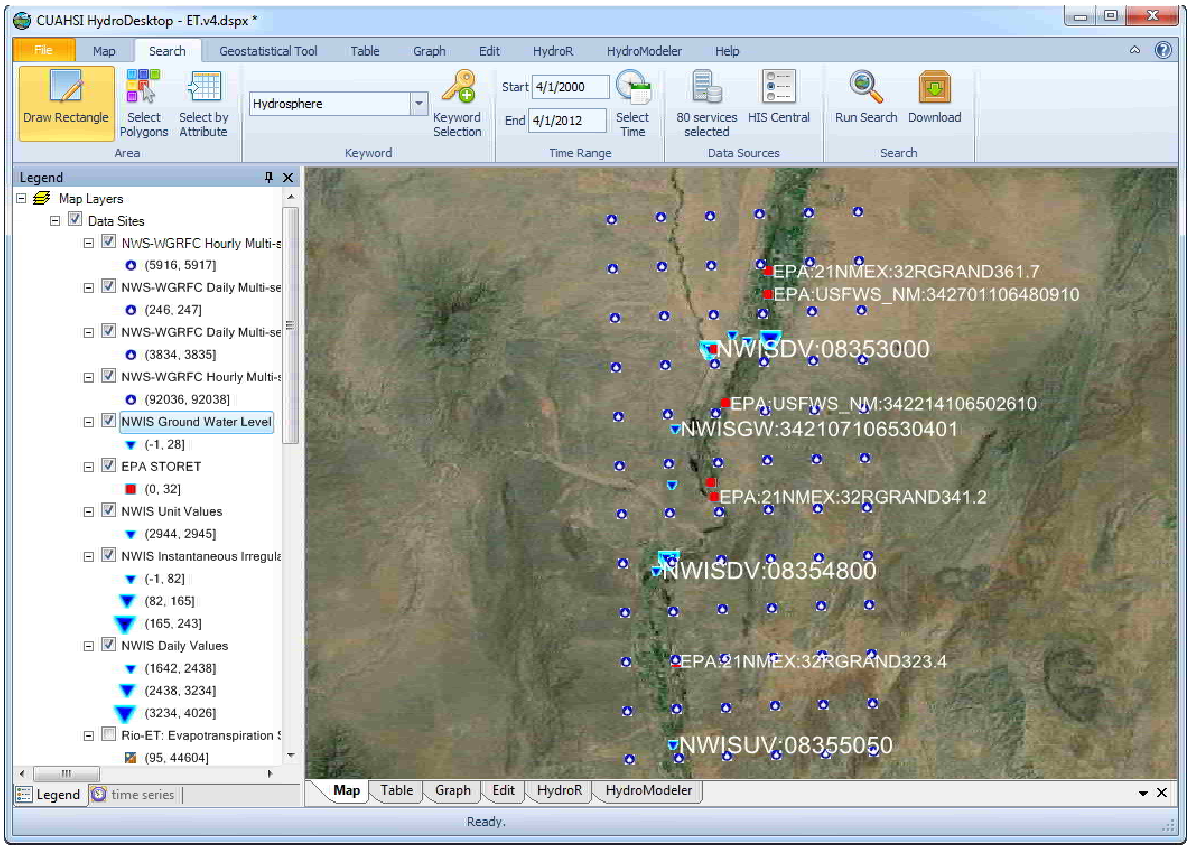


Figure 4: HydroDesktop: Query results for external data from HIS Central

2.2.3 Supplement with external data

By running a local HydroServer, datasets necessary for research but unavailable at HIS Central can be loaded into a separate CUAHSI database. Currently limited high resolution USGS stream discharge values are registered with HIS Central. Historic data are available in 15 minute increments via the USGS website. The query of HIS Central identified the gage near the region of interest where higher resolution data are needed. Downloaded data from the USGS website was ingested to the HydroServer for recurring visualization and analysis (Figure 5). This database can be added to as needed and shared by team members providing a robust local source of data for project analysis.

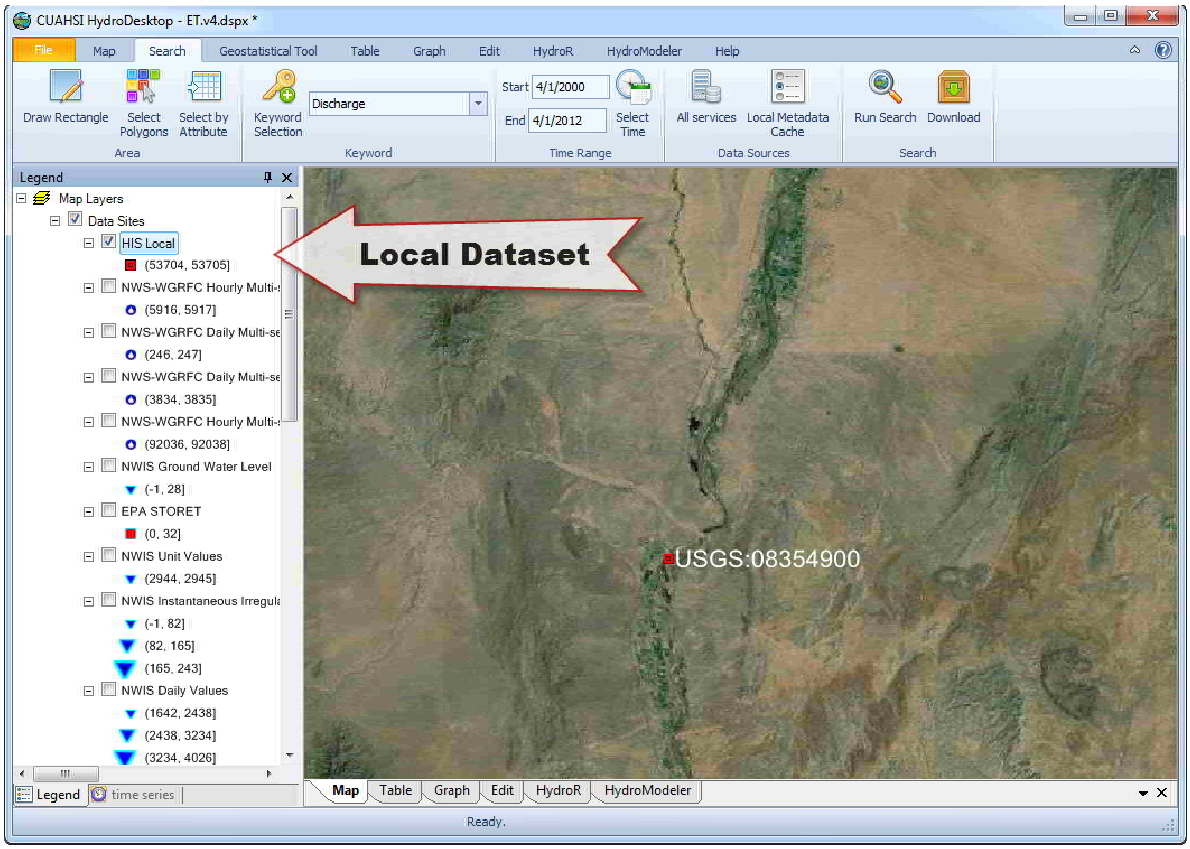


Figure 5: HydroDesktop: Adding external data from local database

2.3 Case Study: Gage Data Near Las Conchas Fire Boundary

In 2011, the Las Conchas fire burned 156,593 acres of land in northern New Mexico during the monsoon season (InciWeb, 2012). Flood risk increased due to vegetation and organic soil loss (Stoof, et al., 2011). Identification of current gage locations near the fire and obtaining historic climate and streamflow data are critical to building accurate models to predict potential flooding.

2.3.1 Load fire perimeter in HydroDesktop

A shapefile of the fire perimeter obtained from the US Army Corps of Engineers is loaded into HydroDesktop with the symbology changed appropriately. HydroDesktop is built upon the open source, MapWindow GIS software which allows efficient symbology of layers, including a feature ESRI's ArcGIS does not have . . . changing the opacity of an outline and fill separately (Figure 6a).

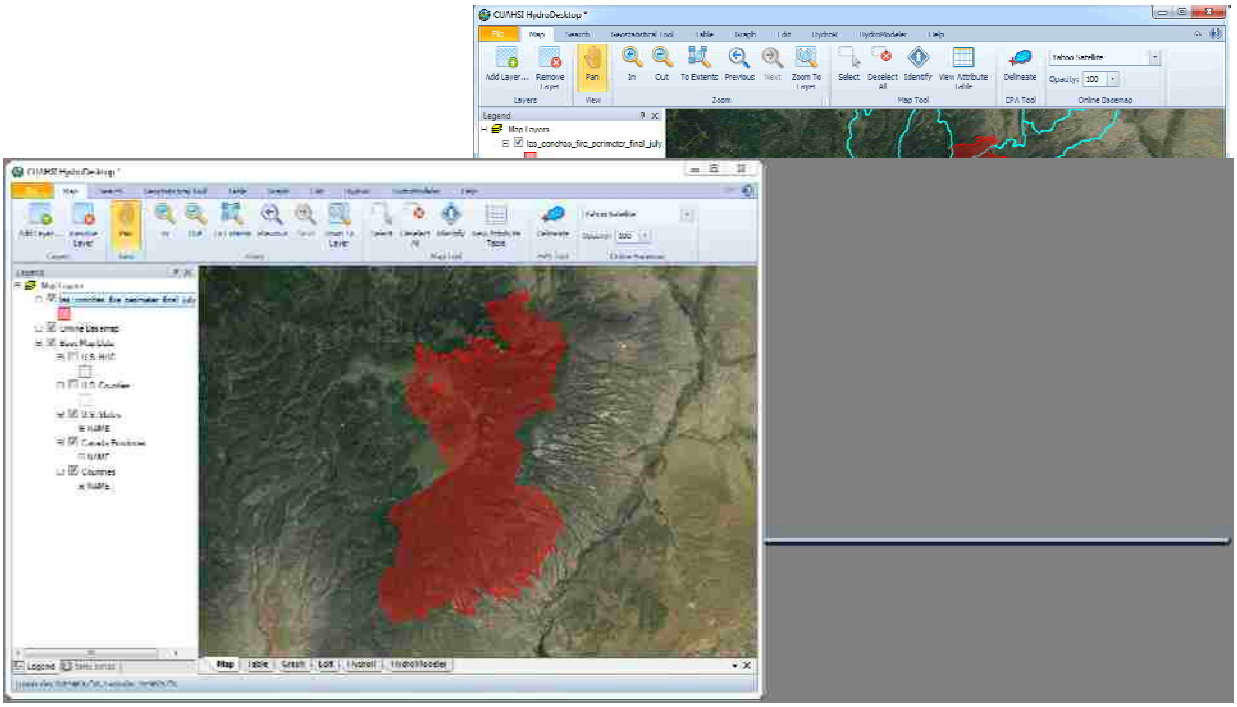


Figure 6ab: HydroDesktop: Adding shapefile as query extent

2.3.2 Search fire region for hydrologic data

There are several choices for searching the fire region. The shapefile of the fire boundary could be used to search for gages but that will not show measurement sites downstream of the fire. This would be useful to determine if any gages may have been damaged by the fire but not necessarily for downstream basin impacts.

The pour points of the basin affected by the fire need to be determined, ensuring our search includes a large enough extent. Selecting 'ESRI World Topo' from the Online Basemap options allows exploration of the hydrology in the region and drawing a search box large enough to encompass all watersheds. Alternatively, a shapefile of the National Hydrography Dataset (NHD) 12-digit Hydrologic Unit Codes (HUC) surrounding the fire perimeter was loaded (Figure 6b) and the region bounded by these polygons was queried. Multiple sites from three national agencies and two regional HydroServers were returned (Figure 7).

2.3.3 Export shapefile of results for future use

Right clicking on any of the Data Sites in the Legend allows export of the site locations as a shapefile with full metadata to maintain provenance of the downloaded data (Figure 8). This shapefile can be opened in any standard GIS application and joined with downloaded data values for visualization in a time aware environment.

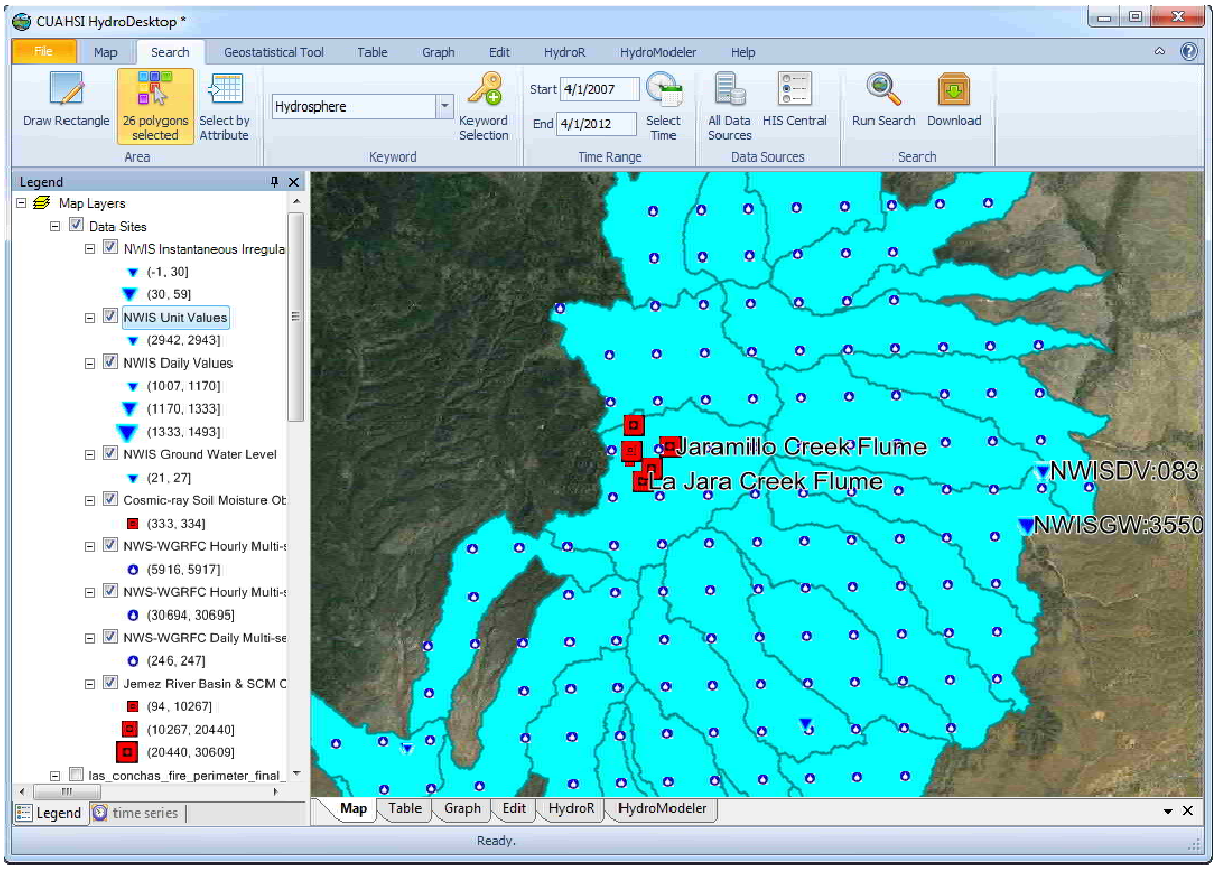


Figure 7: HydroDesktop: Search results using NHD HUC12 boundaries as query extent

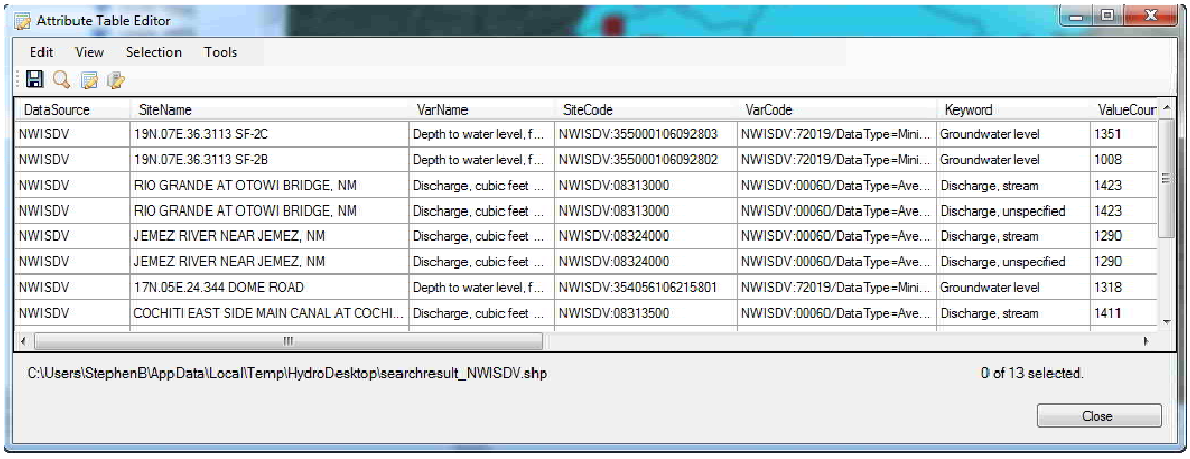


Figure 8: HydroDesktop: Attribute table of queried stations

CHAPTER 3: PROJECT AREAS

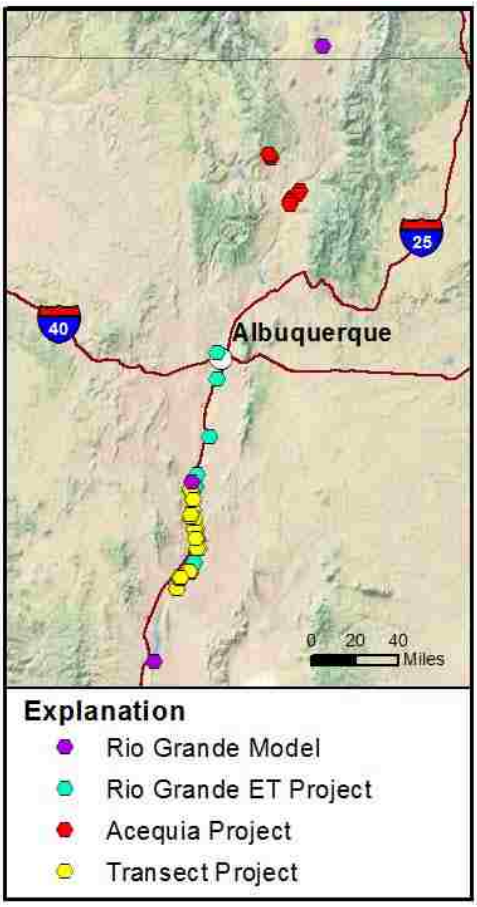


Figure 9: Project areas

The project areas were selected to provide a wide range of data types, collection methods, archive techniques, and research objectives (Figure 9). Each of the study area projects is managed by different principals, providing an opportunity to examine the complexity of working with different organizations.

The projects were at different stages of data acquisition, processing, and archiving, presenting special challenges when developing workflow methods for ingest to the Hydrologic Information System.

Two of the datasets selected for ingest presented special challenges when working with principals and attempting to obtain the raw data. The San Acacia Transects and the Acequia Project were both abandoned and are featured in the 'Migration Challenges' section.

3.1 Rio Grande Evapotranspiration (ET) Project

Eight ET tower locations and nine well locations (Figure 10) with up to five wells at each site, spreading from Albuquerque to the Bosque del Apache Wildlife Refuge near Socorro make up this project. Dozens of individual measurements per hour have been collected for up to ten years at each tower site. Detailed three dimensional wind speed and direction, air and soil

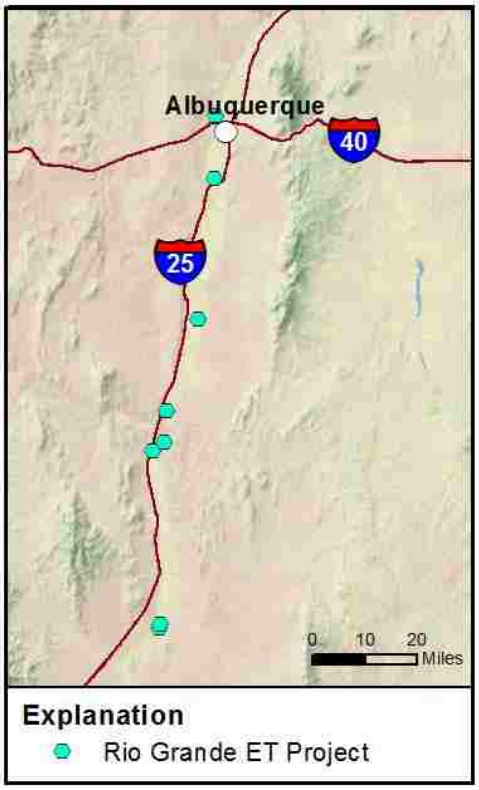


Figure 10: Rio Grande ET

temperature, precipitation, humidity, incoming and outgoing radiation have been collected at fifteen minute intervals. Ground water temperature and level have been recorded in thirty minute intervals.

This dataset provides a wealth of information to a wide variety of disciplines including, Biology, Ecology, Environmental Science, Hydrology, Civil and Environmental Engineering, and Climatology (Cleverly, et al., 2008).

3.1.1 Pre-HIS data access

The data from the ET towers have been stored on an Apple server in the UNM Biology Annex. Dr. James Cleverly, currently at University of Technology in Sydney, Australia, developed web based Perl scripts to deliver tables of data to researchers via an HTML interface. This method has been efficient but requires researchers to post-process the data to extract the desired time range and variables of interest. Collecting multiple stations, for multiple years, and multiple variables requires consecutive queries followed by post processing to merge the data for analysis.

The method of distribution for the groundwater data was via DVD or USB drive.

3.1.2 HIS Processing and Migration methods

During the migration process a complication arose involving the stability of the original server. A hardware failure had caused the server to become unresponsive resulting in data

and metadata being publicly unavailable via the Internet for several months. Using custom Perl scripts as the backend of the website interface for delivery of data to users have made transferring the data to a new server problematic. Much of the raw data is available as Excel tables but the daily measured evapotranspiration numbers are only served at the website.

The research scientist on the ET project, James Thibault, has been tasked with maintaining the server and incoming ET data after Dr. Cleverly's departure. Taking over server management and processing high frequency ET data using custom developed scripts by a programmer that is now living on a different continent is a challenge. Mr. Thibault was instrumental in obtaining the bulk of the raw data and getting the old Apple server to run long enough to obtain necessary data for migration.

The size and format of the ET data set required considerable pre-processing to prepare the data for ingest. Organizing the metadata was a complicated task. The CUAHSI-HIS data model contains a standardized structure in which to convert the measurements and metadata providing a well defined target. Knowing the required final format of the data allowed a path to standardization to be created. Many of the column names are cryptic, derived products that someone intimately familiar with ET tower data would understand but to a data manager unfamiliar with these products, they are completely foreign. Assistance from the principal investigator is essential when organizing the metadata and data for ingest.

HydroServer's Streaming Data Loader (SDL) allows a large table of dozens of variables and thousands of time stamps to be processed automatically after configuration. The decision was made to merge each station dataset of multiple years and multiple variables into one large comma delimited file for SDL ingest. Consolidating the data by station allowed for a

simplified, albeit time consuming workflow to gather the data, convert date formats, and merge data files. There were several files for each site for each year of record that needed to be merged, checked for errors, and ingested into the HIS. Date formats and alignment are particularly complex. Much of the data are returned from the Perl scripts, onscreen, in tabular form. These screens were copied and pasted into Excel. Some of the screens are incomplete datasets with months of time missing from the middle of the table. Aligning these missing chunks with columns of other time stamped data in Excel requires considerable attention to detail.

The ground water data have been well tended by James Thibault. Each year was sorted efficiently and the water levels with changes in datum due to cable changes or pressure transducer replacement were updated and referenced in master spreadsheets. The water levels were also converted from 'depth to ground water' to 'elevation' using the current North America Vertical Datum (NAVD88). Although the number of well data points exceeded two million, the processing was uncomplicated.

3.2 Modeled Rio Grande Climate Change Streamflow Data

Information entering the HIS does not need to be instrument derived. Researchers frequently produce valuable data from exhaustive model simulations. Using HEC-HMS software, estimates of future average streamflow of the Rio Grande for different climate change scenarios were modeled for the Rio Grande watershed above Elephant Butte Dam (Figure 11) through 2110. Having model data included in the HIS provides a foundation for comparison of various hydrologic scenarios.

Figure 11: Rio Grande Streamflow Model

3.2.1 Pre-HIS data access

This is a new dataset. Data has been stored in HEC-HMS and Excel tables for use by the researcher.

3.2.2 HIS Processing and Migration methods

This research was conducted by Chi Bui as the foundation of her Master's thesis in Civil Engineering at the University Of New Mexico. The dataset is well structured and with few variables making ingest trivial. Working closely with Mrs. Bui during the completion of her work ensured data were delivered in the proper format for compatibility with CUAHSI's Streaming Data Loader (SDL).

The most common and time consuming step in preparing data for ingest is formatting of dates in the standard HIS format. When the model scripts were being written, special consideration was taken for date formats. The output files were ready for ingest with minimal post processing. Configuration of the server, entering metadata, and ingest of the data took less than two hours.

3.3 Migration Challenges

3.3.1 Abandoned: Black Mesa and El Rito Acequia Projects

Test data for the Acequia project were being generated from a series of ground moisture sensors and a weather station in the El Rito region of New Mexico. Instrumentation was

installed in late 2010 and has gone through testing in 2011. The plan was to integrate HydroServer from the beginning of the project using telemetric data streaming, allowing semi-instantaneous visual access to the data. Several years of archive data were available from a currently running Acequia project on the Rio Grande in New Mexico to ingest after the test data (Fernald, et al., 2010).

Pre-HIS data access:

Multiple years of data are currently stored on a secure computer managed by the primary research team. No data are available to the public due to concern the stakeholders have regarding New Mexico water rights issues. New locations installed in 2011 did not have historic data.

HIS Processing and Migration Challenges:

In spring of 2011, a new deployment of instruments with 900MHz transmitters were installed in northern New Mexico. The initial plan was to stream data from these devices to the Principal Investigators (PI) office in Las Cruces then relayed to the HIS in Albuquerque. Once in the HIS, the Streaming Data Loader would automatically ingest the data into the database and make it available immediately. Data would be sent every half hour to the HIS for processing.

After the new installation was running, the archive datasets would be ingested manually. The project coordinator, a PhD candidate in Las Cruces has multiple years of back data on his computer. Much of the data is in a proprietary format requiring standardized export to streamline the flow into the HIS.

The PI of the Acequia project is very concerned about highly sensitive gage information being accidentally released to the public. Options have been presented to install the HIS on a system detached from all network connectivity, providing a secure yet uniform method of data management and analysis. After a year of negotiating and a site visit to help install equipment, the PI decided not to migrate any data to the HIS. The project has been abandoned.

3.3.2 Abandoned: San Acacia Transect Project

The project is located near San Acacia, NM both upstream and downstream of the Bosque del Apache Wildlife Refuge and consists of seven well transects, both pumping and passive. The data were collected several years ago for the New Mexico Interstate Stream Commission (NMISC) and have been stored on computers of the consulting firm S.S. Papadopulos and Associates, Inc. Distribution decisions for the dataset are handled by the New Mexico Office of the State Engineer (NMOSE). This data are a valuable addition to the Rio Grande ET dataset.

Pre-HIS data access:

All project data have been stored on the consultant's computer. Current dataset distribution method is unknown. The official final report is available on the NMOSE website but digital copies of the data values are unavailable for analysis without a FOIA request (New Mexico Office of the State Engineer, 2010).

HIS Processing and Migration Challenges:

When the NMISC was first contacted regarding the San Acacia transect data they were in the process of developing an internal policy and legal disclaimer for data distribution. After

waiting for many weeks for the state's lawyers, word finally came back from the state. They will release the data but not the database. In the time it took for the state to respond, other datasets from NMEPSCoR research projects became available for ingest and this project was abandoned.

CHAPTER 4: DATA MANAGEMENT PLAN VS DATA WORKFLOW PLAN

Creating the NSF required data management plan to ensure protection from loss and increase availability to other scientists may not provide a system to manage data during research. While contacting scientists to inquire about data management methods and possibilities of datasets ready for ingest to the CUAHSI-HIS, a recurring trend presented itself. Considerable amounts of time are being spent processing data. Before any analysis can be performed the raw data must be pre-processed, scrubbed for errors, have any gaps filled, and organized on disk. Data processing from instrumentation can be a tedious, repetitive process that may introduce errors if the technician is not diligent.

Scientists can be free (mostly) from monotonous data processing with the development and implementation of a Data Workflow Plan (DWP). When developing a DMP (Data Management Plan) the data repository can assist with creating a DWP containing the programming requirements for processing raw data and assist with deployment of tools to easily visualize and correct data inconsistencies. Scripting the initial data ingest and pre-processing ensures consistency from payload to payload. Any changes in the processing algorithms will be stored in the metadata by the programmers at the data archive. In the event the PI leaves the project, all the initial processing steps are recorded with the data repository.

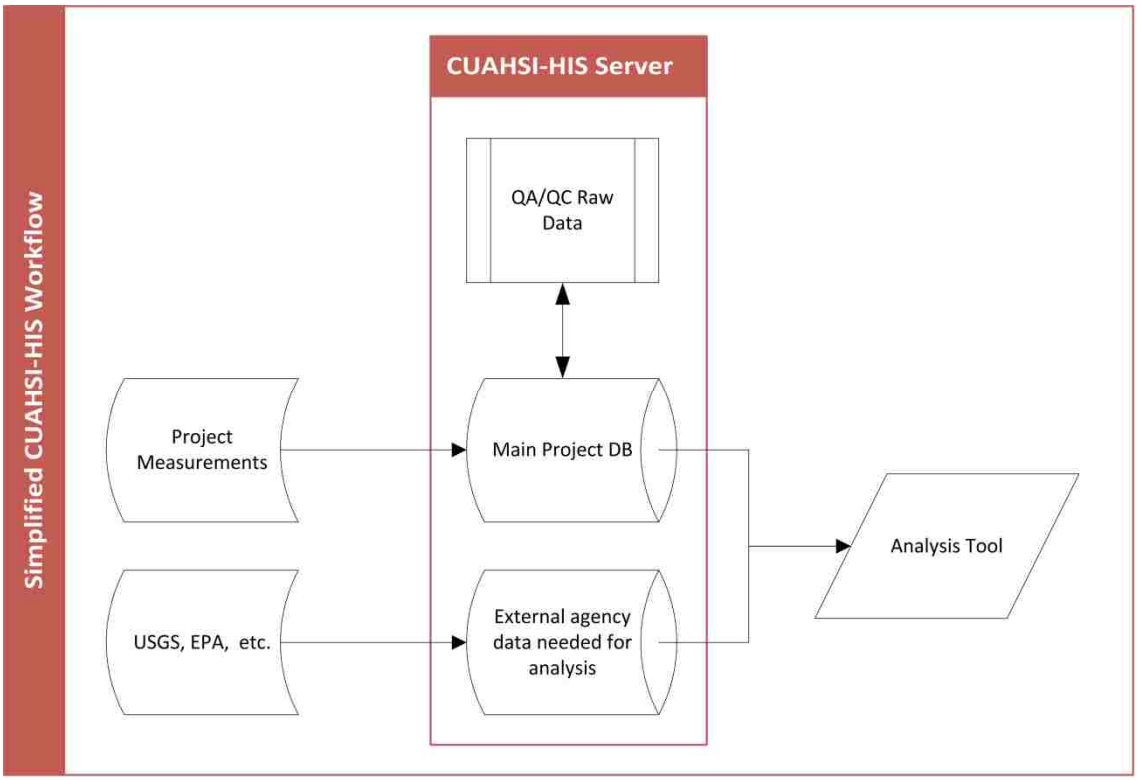


Figure 12: Simplified CUAHSI-HIS workflow

The CUASHI-HIS is ideally suited for implementing a DWP. HydroServer can run several discrete project databases allowing raw data to be ingested to one Observation Data Model (ODM), exported to standardized workflow, and returned to a new QA/QC ODM for visual review (Figure 12). The raw data are untouched and the QA/QC data can be served to the public. Workflow steps may include aggregation from high resolution data to hourly and daily averages, delivering a range of scales for modeling analysis. Using CUAHSI's ODM Tools application, a PI can visualize data in the database and make corrections easily. Datasets necessary for project data analysis from external agencies can be stored on the local HIS, integrated with the workflow and queried concurrently with project data for visualization.

Scientific workflows are popular in many disciplines but have yet to develop wide use in the hydrologic domain (Guru, et al., 2009). Development of Open Modelling Interface (OpenMI) arose out of the hydrologic domain but is generic in scope and has developed a strong user base. Twenty four hydrologic and hydraulic models are currently listed as compliant with OpenMI (OpenMI Association, 2012). CUAHSI's HydroDesktop has integrated an OpenMI plug-in for data analysis within the application.

The Kepler Project is another open source scientific workflow application. Kepler allows scientists to create, execute, and share models and analysis using a simple flow chart type interface (Kepler/CORE, 2012). Workflows may be developed for raw data streams that return the results to a database for final QA/QC. Kepler workflows can be reused, modified, and easily shared among researchers and data centers. As more organizations build workflows ingesting new instruments into the CUASHI-HIS, a shared repository can provide a foundation to streamline new deployments.

Figure 13 is presented to illustrate the complexity of processing data streams from modern instrumentation rather than provide a readable example. Evapotranspiration workflows are labor intensive to execute, requiring advanced scripting and great depth of understanding of the instrumentation, measurements, and processing requirements. Perl, R, and SAS Scripts for the Rio Grande ET processing were developed by Dr. James Cleverly when he was one of the PIs on the project. In 2009, Dr. Cleverly accepted a faculty position in Australia, more than 13000 km from New Mexico. The daily management of the ET data is now the responsibility of the senior research scientist, James Thibault. Tracking the workflow through a series of scripts in different languages on different operating systems is challenging no matter how well the code is documented. Using a workflow manager is essential for research

teams to develop, document, and catalog processing methods and changes through a project lifespan.

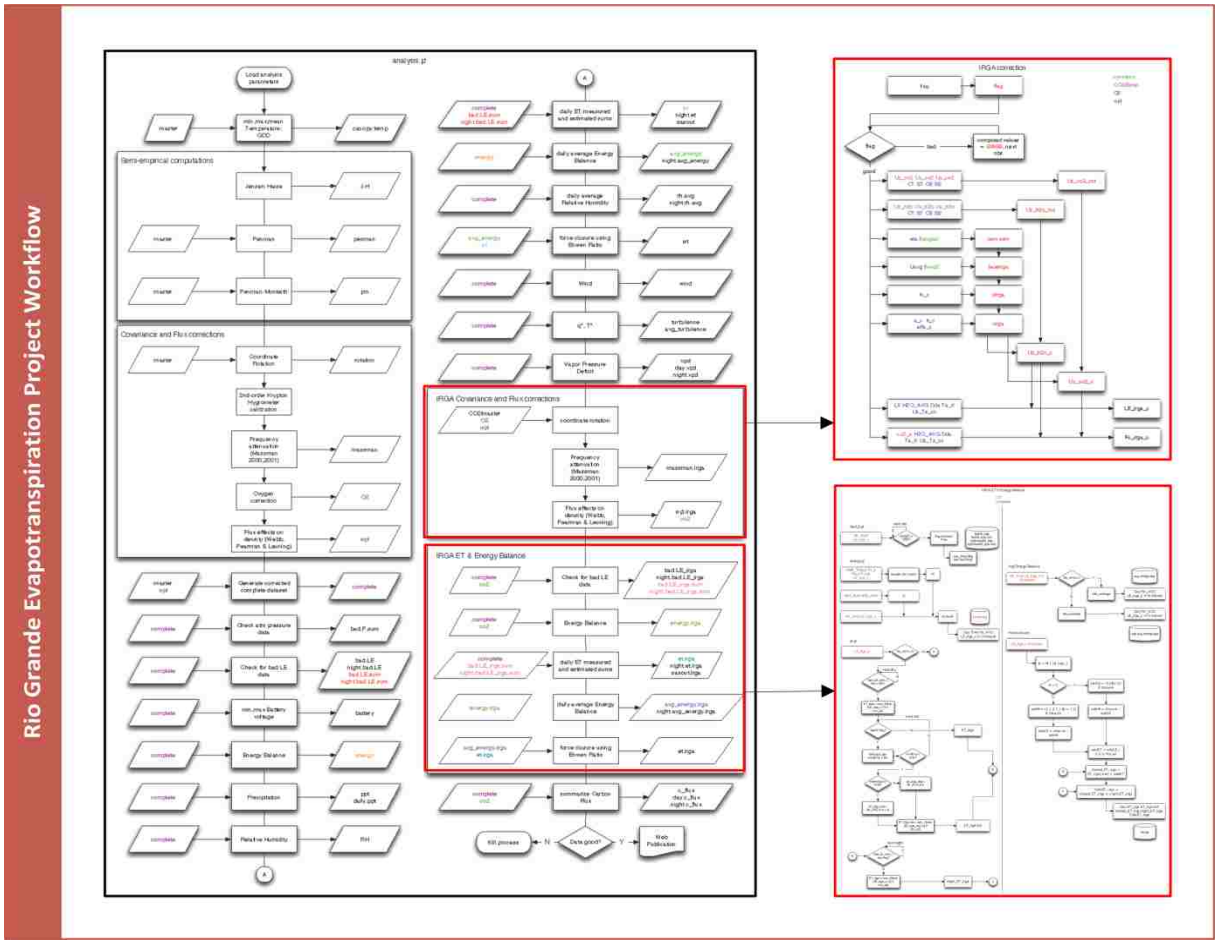


Figure 13: Rio Grande ET original workflow.

The Kepler Project workflow manager includes several key functionalities that assist with workflow development and increase transparency when handing long term research off to a new team.

Tagging: Workflows and workflow runs may be annotated with tags based on ontological concepts using Web Ontology Language (OWL). Tags applied a workflow are carried on to any runs of that workflow.

Provenance: Tracking of the lineage of workflows and data products. Changes to the processing algorithms are stored for future analysis of the QA/QC and derived products.

Reporting: GUI for generating reports from workflow runs customized for each specific workflow. Variables critical for analysis and reporting of the current workflow can be included in the report template.

Run Manager: A GUI for managing workflow runs histories stored by the Provenance module. Past workflows and reports can be browsed, tagged, exported, and uploaded to a remote repository.

The Kepler graphic interface uses 'Actors' to link components, building a workflow from many parts. Kepler ships with a large library of customizable actors to interact with the workflow including R, Matlab, Excel, command line, Web Service, and input/output. Workflows are re-usable and the platform supports grid and parallel processing technologies to maximize efficiency in server farm environments. Kepler Workflow System and CUAHSI-HIS are now part of a training program "Software Tools for Sensor Networks" (LTER, 2012) sponsored by the National Center for Ecological Analysis and Synthesis (NCEAS), Long Term Ecological Research Network (LTER), and DataONE showing increasing awareness of the need for standardized open source tools for data management and distribution.

CHAPTER 5: DISCUSSION

5.1 Project Discoveries

5.1.1 San Acacia Transect Project

The NMISC, responsible for the San Acacia Transect Project data, is in a position that many state and federal agencies find themselves in currently. Taxpayers have funded valuable hydrologic research and want access to the data but the agencies do not have standards of practice in place to digitally distribute the data. Managers and legal teams are making decisions about the formats of data availability without any understanding of modern hydrologic research workflows. The absence of accepted international standards for government agencies to follow when distributing time series hydrologic data adds to the problem.

The logical choice would be for one agency to provide a master framework with stable long term funding dedicated to preservation and distribution of taxpayer funded hydrologic data. The USGS would be one of the top choices as they already maintain a massive network of surface and ground water gages including historical data but they have suffered from reorganization and budget cuts that have weakened the data warehousing. FY2013 proposed federal budget includes 3.3 Million in cuts to *Hydrologic Networks and Analysis Information Management and Delivery (USGS Coalition, 2012).*

5.1.2 Acequia Project

HIS implementation for the Acequia team was largely hindered by a lack of understanding of how the system functioned. Primary concern for the project team was confidential data being

accidently released, compromising sensitive water rights. The original plan was to send the data from the instruments via cell phone network to the researcher computers at NMSU in Las Cruces and then on to the HydroServer at UNM in Albuquerque. Having the final data on a public server at UNM was unacceptable so the option to configure a local private HydroServer at NMSU was presented. The local private HydroServer was also dismissed by the project team as unsecure.

When discussing the CUAHSI-HIS with hydrologists during this research project, the most common response was *"I had no idea this existed"*. Describing the functionality brought people closer to understanding the system and benefit to active research but it took a live demonstration to see how the pieces fit together. The Acequia team is a prime example of this phenomenon. They have seen live demonstrations of HydroDesktop for data reconnaissance but HydroServer instruction is not readily available. Bridging the gap between HydroServer as an IT/programmer application and presenting the components as accessible tools for data management and discover is necessary.

5.1.3 Modeled Rio Grande Climate Change Streamflow

The Rio Grande Streamflow Model data is the only dataset with full principal investigator involvement in the HIS ingest process. Mrs. Bui was helpful and motivated to format her final output in a standardized method that streamlined migration to the HydroServer. All metadata was included and similarly formatted to aid migration to the HIS. Although the dataset contained just over one million measurements, the ingest process took less than two hours.

Several key elements contributed to the efficiency of this ingest process, the most important being the development of an Excel spreadsheet for metadata capture with preloaded CUAHSI Control Vocabularies. Collecting the metadata for the project was streamlined by dropdown menus in the Excel file containing the Control Vocabularies, facilitating rapid correlation of researcher data with CUAHSI standardized query terms. The spreadsheet is available in the LoboVault accompanying this manuscript. Providing tools to assist the researcher with organizing data and metadata in the proper format for ingest saves time and frustration for all involved parties.

5.1.4 Rio Grande Evapotranspiration Towers and Wells

Processing measurements from the ET towers was complicated by the volume of the data. Dozens of measurements and derivations contained in dozens of files with assorted header names for each station required significant attention to detail to manage. Metadata was available but cryptic for a data manager unfamiliar with eddy covariance towers. The PI was out of the country and not involved in the migration. Active participation by the researchers that deploy the instrumentation and process the data streams is imperative to ensure verification of ingest and proper assignment of metadata.

All the ET data for each tower were aggregated in to one large comma delimited file (CSV) with standard column arrangement to provide consistency when running the CUAHSI streaming data loader (SDL). The SDL queues multiple CSV files and the associated ingest instructions to automatically feed many stations into the database. Normally the SDL walks the researcher through dialog boxes to setup all associated metadata for each variable. This process works well for a handful of variables but loading 100+ became tedious and the chance

of data entry error increased. Using the metadata Excel spreadsheet created as part of this research, all the variable details were entered once, many times copied and pasted for groups of variables and loaded directly to SQL server with the Microsoft Import Wizard. Similar efficiency was realized by direct editing of the SDL Config.xml configuration file to setup the SDL for processing CSVs with multiple stations and variables.

Each ET tower CSV contained 100+ columns and up to 250,000 rows. These were the first files of this size processed by the SDL. Originally designed to automatically process small files dropped in a hot folder, the SDL handled the large archive datasets with few issues. The files processed and loaded into the database in a reasonable time but the SDL is configured to automatically update the Series Catalog, a relation of series attributes in the database, after each CSV, which increased processing time three or four fold. This 'feature' needs to be edited to run after all the CSVs have been loaded and only run once. The only other bug is when two rows have the same date/time stamp in a CSV. Due to differences in how some of the equipment was programmed for the ET data, a duplicate value for January 1 periodically appeared when aggregating the data. The CSV would fail to load without throwing a visible error. The error showed up in the log file, where the average researcher may not know to look. Some sort of easy to read pop-up summary report would be helpful for quick verification of ingest. Bug reports for both of these issues have been submitted to the project Codeplex site.

5.2 Principal Investigator Involvement

The projects that migrated smoothly to the HIS had active involvement from the scientist managing the data. Metadata were collected accurately and data files were formatted to

streamline ingest through the SDL. Processing datasets with millions of measurements became relatively trivial. The CUAHSI Controlled Vocabularies and standardized data structure provided a solid framework for organizing the data but knowledge of the dataset is critical. Ingesting archive datasets where the principal investigator is no longer available is a daunting task. In order to ingest any measurements, the data manager must become knowledgeable on the instrumentation, data types, methods, processes, derivations, units, etc. Gaining this knowledge does not rule out errors of interpretation when entering metadata or choosing methods accompany the data. The Excel metadata spreadsheet developed during this research study helped dramatically to organize that data and flag unknown variables.

5.3 Deployment Timing

Deploying a DMP, DWP, and configuring a HIS at the very beginning of a project achieves several goals. First, the instruments can be configured to output standardized data files. A standard of practice will be initiated to accompany each device on every deployment. All configuration settings will be decided before putting an instrument in the field. Second, the project overview and metadata will be clearly outlined. Processing steps and data provenance will be pre-determined allowing for rapid data analysis when measurements return from the field. Any necessary changes in the workflow after the first data payload returns from the field will be documented adding to the understanding of assumptions made before instrument deployment. Third, NSF reports will be easy to generate with standardized processing and HIS data storage in place.

Had the CUASHI-HIS been available at the start of the Rio Grande ET project, the Streaming Data Loader could have been configured to ingest the raw data streams directly from the field.

A solid foundation of metadata would be integrated into the server and adjustments could have been made directly within the server as instrumentation changed. The history of the equipment would be directly attached to the data itself and easily maintained. Any derivations, instrument drift, or equipment changes would be applied to a clearly labeled QA/QC data series removing the 'black box' from the analysis. All processing would be done once using consistent workflow management and the data would be ready for query and analysis.

5.4 Server Failure

Portability is a key component of the HIS. If a server failure occurs, a backup of the entire database can be copied to another instance of HydroServer and ready for access within an hour. One HydroServer can house dozens of separate projects, all running independently of each other. If a data set from a failed server needed to be available before a new physical server could be built, it is trivial to incorporate the dataset into a currently operating HydroServer or run a virtual server preconfigured with the base installation of the server package. Working with a data repository will ensure proper data archiving and quick restoration in the event of a system failure.

5.5 Data Management and Workflow

The preferred path to a viable data plan would be to work with a data manager like a geospatial repository or library already familiar with metadata and database architecture. Hardware, network, and programming experience are readily available from the data manager to save time and money. Many libraries are only storing spreadsheets or databases of

measurements from researchers that are not interactive. Coordination with the data manager to install an instance of HydroServer will allow spatial and temporal queries of data.

The raw data stream can go directly into HydroServer. QA/QC processing is done with a data workflow using CUAHSI's ODM Tools or Python/R scripting by the data management team. Final data are ported back into HydroServer with a full processing path recorded to have provenance for peer review and NSF reporting. Data backup is integrated with the repository management protocols and public data access can easily be activated when it is time to publish. While collaboration and analysis is underway, PIs in multiple states can securely view the data in real time (Figure 14).

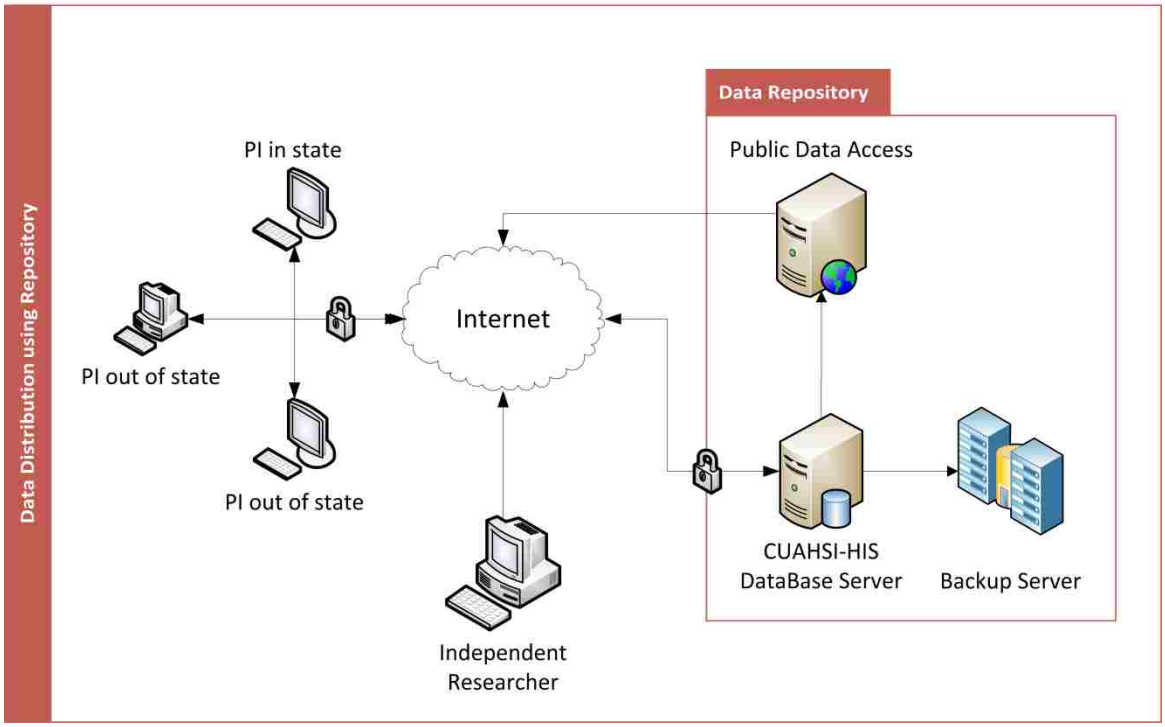


Figure 14: Data distribution using data repository

Long term data management is a critical concern for the future of earth sciences. Extended periods of record are required to conduct trend analysis that spans decadal oscillations. Data collected today needs to not only be available in fifty years but have full metadata allowing

new generations of researchers to examine the data integrity and collection methods (Robbins, 2012). Data centers should be the foundation of data storage, not a researcher's office computer. Data repositories need to be looked at as vital infrastructure with stable funding to allow for standards of practice regarding data sharing, schema evolution, and long term availability (Klump, 2011, Schofield, et al., 2010).

Data workflow plans will require patience to implement in the early stages. Planning any workflow is an iterative process, after an instrument returns data from the field the workflow steps need to be tested and adjusted. The iterative process will continue until stable results are generated. Stable workflow components can be used as foundations for similar devices and shared in a public repository. The modular structure of Kepler may provide direct interface with the instrument, allowing programming from the workflow before sending the device into the field. All input and output settings could be managed from one interface that maintains full provenance of data flow from device configuration to final publication in the HIS. The complete workflow may be shared with other researchers deploying similar equipment providing a 'plug and play' experience.

5.6 Budget Constraints

With tightened budgets, resources are limited to publish searchable hydrologic datasets using traditional database development (Robbins, 2012). Organizations may keep valuable datasets out of the public view until a Freedom Of Information Act (FOIA) request. CUAHSI's HIS is an affordable path to publishing these publicly funded datasets in either a Windows environment or by using the mySQL ODM, Linux/Unix. Commercial hydrologic database

packages like KISTERS WISKI and AQUARIUS Server can cost thousands of dollars per year per server license (KISTERS, 2012, Aquatic Informatics, 2012).

CUAHSI has been active in pursuing an international audience for the HIS. Many emerging countries have natural resource and hydrologic concerns with no organized system to store and distribute temporal data streams. A new HIS has been developed in the Czech Republic using CUAHSI's WaterML, providing the first free, complete coverage of hydrologic time series data in the country (Kadlec, 2010).

Sample scripts and methods used to process data for this study are available in Appendix B.

CHAPTER 6: CONCLUSION

Ingesting datasets into the HIS progressed smoothly when the principal data collector/creator was involved in the migration process. Date formats and column order can be standardized early in the data management process. Working with a researcher from the beginning of the project ensures consistent data quality and thorough metadata.

The two abandoned projects posed unique issues when attempting to acquire the data. State and Federal agencies that don't traditionally deliver on-demand digital data are complicated by legal issues and a lack of standards of practice regarding data distribution. These problems will likely continue to slow data access for independent researchers. Obtaining the data will be possible but it may not be in a format that is easy to ingest requiring extensive pre-processing. The Acequia project highlighted a lack of understanding of the functional operation and long term benefits of ingesting data into the HIS. Custom workflows for complex data management or security requirements can be attained with the assistance of a

data repository. The key is to be aware of what options are available and how to best implement them.

The two successfully ingested projects provided clarification of the best paths forward when processing data. PI involvement is critical when configuring the data and metadata for ingest. A data manager without in-depth knowledge of a discipline may overlook a description or miscategorize a unit causing invalid results for future researchers using the data. The NSF enforcement of DMPs will be a major asset to prevent this problem in the future. Overall the CUAHSI tools proved flexible when processing the forty million measurements from the Rio Grande ET project into the database. Additional programming is necessary to enhance the usability as is always the case for iterative software design. The more researchers contribute feedback to a product the more functional it becomes to a wide range of users. Working directly with the PI for the Rio Grande Streamflow Modeling project resulted in data and metadata delivered properly formatted and ready for ingest. Over one million measurements were ingested in less than two hours.

The project overviews are not intended to be critical of researcher's data management methods or standards of practice. As experts in our respective fields we use the tools at our disposal to produce the most thorough and accurate results possible. Most, if not all, scientific research conducted today requires broad depth of knowledge in the main field of interest (biology, engineering, geology, etc) as well as substantial understanding of computer processing methodology.

Team members of the four projects and technical support staff at EDAC come from a wide scope of backgrounds; hydrology, engineering, administrative, information technology,

biology, anthropology, range management, legal, geology, database management, and more. Working with these team members emphasized the need to make data management transparent to the research scientist. Achieving transparency can only be obtained through standardization and cooperation with a data repository to streamline the technical side of data management. The CUAHSI-HIS provides a ready platform for data storage and access needs while research is underway and satisfies key components of the NSF DMP requirements.

CHAPTER 7: FUTURE WORK

Incredible potential exists to satisfy NSF data management plan requirements and allow scientists to spend more time doing research in their fields. Working with dedicated teams in geospatial storehouses or library sciences, complete data workflows can be developed. From initial ingest of the raw data, visual or statistical QA/QC, derivations and calculation processing, and finally returning the processed data back into the database for distribution. The entire workflow would be automated and self-documenting, creating a built in provenance. More organization and time will be required in the beginning from the scientist and the archivist but soon standards of practice will be developed and a series of templates for specific data streams will be ready to drop into place.

Developing a Kepler full workflow, from device configuration to final QA/QC data entering the CUAHSI-HIS would be a valuable addition to hydrologic research. Error checking and statistical analysis can be accomplished in Python and R. When the workflow is complete a virtual server could be generated with all the necessary components installed, ready to spin up and connect a new measurement device. The result would be a ‘plug and play’ standardized hydrologic information system.

SECTION B: RESEARCH EFFICIENCIES

CHAPTER 1: INTRODUCTION AND BACKGROUND

1.1 Digital Watershed

Understanding complex basin scale hydrologic processes have long been critical to developing efficient land and water use practices. River and irrigation channel interaction with floodplain ground water is important to determine surface water use and aquifer recharge. Basin scale analysis of long river systems is complicated by large land areas with diverse ecosystems producing copious quantities of hydrologic measurements.

A *Digital Watershed* is an aggregation of spatial and temporal data combined with visualization and modeling tools that allow complex hydrologic systems to be analyzed for trends both graphically and numerically. Stream discharge, precipitation, eddy covariance, water quality, and any number of thousands of other hydrologic parameters can be recorded by instrumentation at high rates. Some of these devices take dozens of measurements per second. Processing and storing vast quantities of data have been a challenge due to computer and network speed and storage capacity.

Multi-agency, integrated hydrologic database systems provide the opportunity to examine basin scale river networks that have previously been too time consuming to explore. Computer workstations have the processing speed, storage capacity, and graphic capability to quickly visualize thousands of measurements from dozens of locations. Analysis that used to take days or weeks to obtain, enter, and standardize the data from national data sources followed by integration with locally collected measurement can now be conducted in minutes.

Rapid visual reconnaissance of a dozen years of observations can provide insight to the location of anomalies that require more detailed research. The rapid reconnaissance of anomalies will allow limited research funds to be spent wisely. New approaches of scientific exploration by novel data management, analysis, and visualization, referred to as 'data-driven discovery' is spreading from high-energy particle physics and astronomy into biology (Thessen and Patterson, 2011, Hey, et al., 2009) and other earth sciences.

1.2 Integrated Analysis

Hydrologic analysis today requires a variety of measurements from multiple agencies in addition to project data. The United States Geologic Survey (USGS) maintains a long history of surface and ground water. The National Weather Service (NWS) stores Next-Generation Radar (NEXRAD) precipitation data. The Environmental Protection Agency (EPA) has thousands of water quality sites throughout the US. Obtaining data from each of these agencies requires a different process and results in a different deliverable format. Once the data is downloaded, extensive time must be spent to aggregate and homogenize the data for analysis. Extending the analysis for an additional month requires another visit to each agency website, followed by aggregating and homogenizing the data all over again.

This cycle is broken with the Consortium of Universities for the Advancement of Hydrologic Science, Inc (CUAHSI) data reconnaissance and visualization tool, HydroDesktop. There are currently eighty searchable services registered with the CUAHSI central metadata repository, HIS Central, (Table 2) which are all queryable at one time by spatial and temporal extent in HydroDesktop. A local HydroServer can also be queried and the combined results shown on one map or graph. All data are formatted in WaterML dramatically reducing post processing

and simplifying analysis in the built-in R or OpenMI modules. A research team can register a HydroServer with HIS Central, making it queryable globally or keep the service private only allowing access by selected members of the research team. Having all project data in one HydroServer provides secure access to team members collaborating from different regions. New data, updates, and corrections are immediately available with appropriate data quality flags attached to the data values.

Table 2: HIS Central Data Services (April 2012)

EPA STORET	Niagara Peninsula Conservation Authority Water Quality Data Service
NWIS Daily Values	NWS-WGRFC Daily Multi-sensor Precipitation Estimates Recent Values
NWIS Ground Water Level	NWS-WGRFC Daily Multi-sensor Precipitation Estimates
NWIS Instantaneous Irregular Data	NWS-WGRFC Hourly Multi-sensor Precipitation Estimates
NWIS Unit Values	NWS-WGRFC Hourly Multi-sensor Precipitation Estimates Recent Values
USACE River Gages	Dry Creek Experimental Watershed, SW Idaho
Chesapeake Bay Information Management System	Panola Mountain Research Watershed, Georgia
NWS-ABRFC Hourly Multi Sensor Precipitation Estimates	Reynolds Creek Experimental Watershed, SW Idaho
Freeman Ranch Mesquite Juniper Flux Tower	Paradise Creek Watershed, Idaho
Baltimore Waters Test Bed Ground Water Level Data	Portneuf Watershed Observations, Idaho
Baltimore Precipitation	RIM Program
Benthic Data in Chesapeake Bay	Rio-ET: Evapotranspiration Studies in the Middle Rio Grande
Baltimore Ecosystem Study Stream Chemistry Data	Rio-ET Wells: Groundwater wells in the Middle Rio Grande
Baltimore Ecosystem Study Soils Data	Santa Fe Basin, Florida Daily Rain Tipping Bucket
Cosmic-ray Soil Moisture Observation System	San Diego River Park Foundation
Crown of the Continent Observatory Snow	Santa Fe Basin, Florida CTD Sondes
Coweeta Hydrologic Laboratory Public Data	Santa Fe - USGS Groundwater Data Florida

Czech Snow Cover	Santa Fe MICROWAVECITRA
Delaware Environmental Observing System	Santa Fe Basin, Florida SRWMD select river gages
EPA - East Fork Watershed in Ohio	Santa Fe, STORET
Glacial Ridge Project	Santa Fe, Southwest Florida Water Management District
Edwards Aquifer Groundwater Database	Storet Phosph and Nitr in Surf water
Hassberge catchment long-Term monitoring data	Santa Fe Ground Water Level SRWMD
Central European Climate Data	Snake River Basin, Modeled Streamflow
Hermine Flood	Susquehanna River Basin Hydrologic Observatory
HydroNEXRAD	TCEQ Surface Water Quality Monitoring (SWQM)
IIHR Tipping Bucket	Texas Instream Flow, Lower Sabine
IIHR Water Quality	Texas Instream Flow, Lower San Antonio
La Selva Hydrologic Data	Mountain Meadow Restoration with a Changing Climate
Grasslands Ecological Area of the San Joaquin Basin, California	TWDB_Sondes
Little Bear River Experimental Watershed, Northern Utah, USA	TWDB Wind
Logan River Observations, Northern Utah, USA	WRRC Acid Rain Monitoring Project
MAST	Weiherbach catchment long-Term monitoring data
McCall Outdoor Science School Observations	Jemez River Basin & SCM CZO
Multi-sensor Precipitation Estimates	Boulder Creek Critical Zone Observatory
Muddy River Water Quality Monitoring Project	JRB & Santa Catalina Mountains CZO
Mud Lake, Idaho, USA	Luquillo Critical Zone Observatory
National Atmospheric Deposition Program	Southern Sierra Critical Zone Observatory
NLDAS Hourly Mosaic Land Surface Model Output	Shale Hills Susquehanna CZO
Niagara Peninsula Conservation Authority Water Quantity Data Service	Christina River Basin Critical Zone Observatory

1.3 Visualization

Large temporal datasets have traditionally been challenging to visualize let alone analyze, this is where the real power and flexibility of storing data in a Hydrologic Information System comes into play. By loading project measurements in a relational database, quick reconnaissance can be performed by spatial and temporal extent.

CHAPTER 2: RIO GRANDE GROUND AND SURFACE WATER LEVELS

The Rio Grande Evapotranspiration Project Well data have been stored in dozens of Excel files by station and year in the project research scientist's computer. The files are very well organized and annual statistics have been complied but the data have not been convenient to access or analyze. Ingest of the data required merging the files and conversion of dates before using the CUAHSI Streaming Data Loader.

2.1 HIS Data Access: Regional

Using HydroDesktop to search the Rio Grande ET project area for 'Water depth' resulted in 30 groundwater sites. Using the HydroDesktop Expression Editor to create a simple query, the central wells are selected for each region (Figure 15). The San Acacia Alfalfa (ALF) site only has one well without a 'Central' label and is added manually. Ten series were downloaded to conduct visual trend analysis.

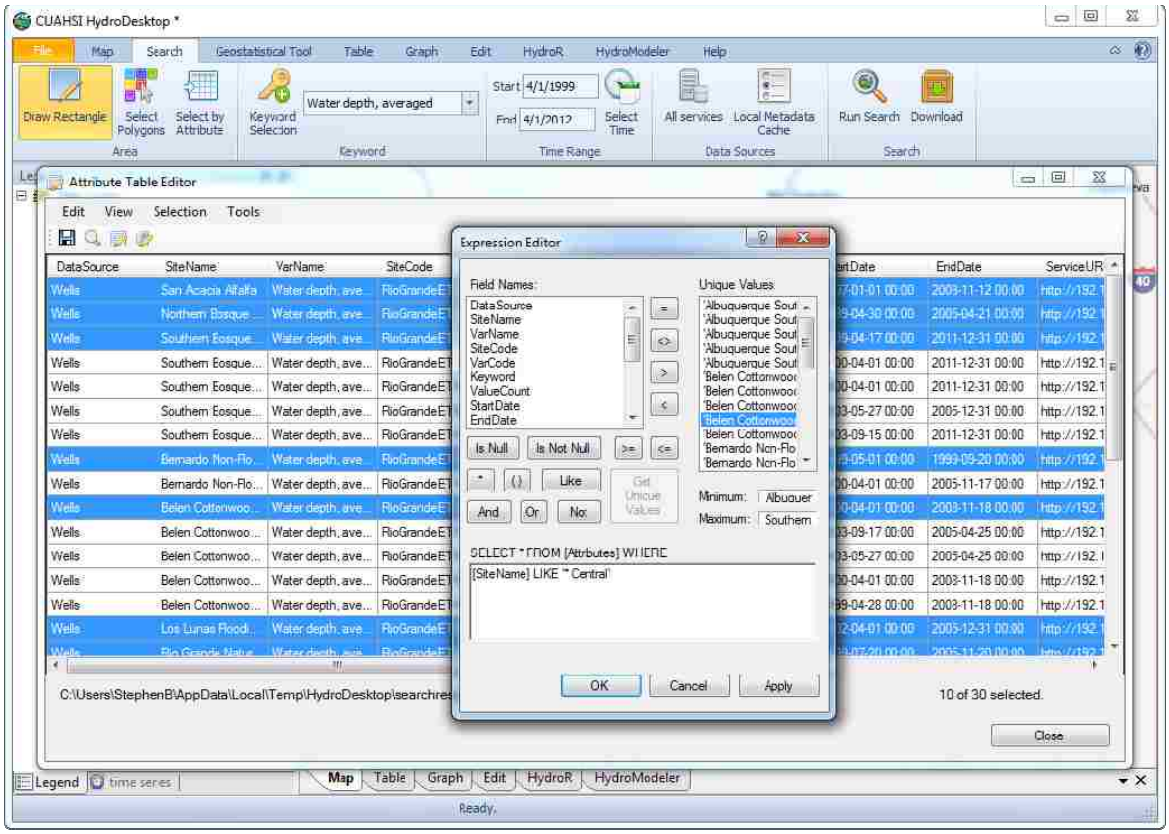


Figure 15: HydroDesktop: Select stations for data download

2.2 HIS Analysis: Tabular

First all ten stations were compared in parallel using the 'table' ribbon tab. Tables are often a quick method for determining critical events in a system. Figure 16 shows the ten stations displayed in parallel starting in 1999 and continuing through 2011. April 18 is highlighted as the first day of flooding in 2005, at La Joya State Game Refuge (LARO). Bosque del Apache (BDAS) did not flood until four days later. Observing multiple years of events for many sites is streamlined in HydroDesktop and provides a view into data previously unavailable without extensive data collection and preprocessing.

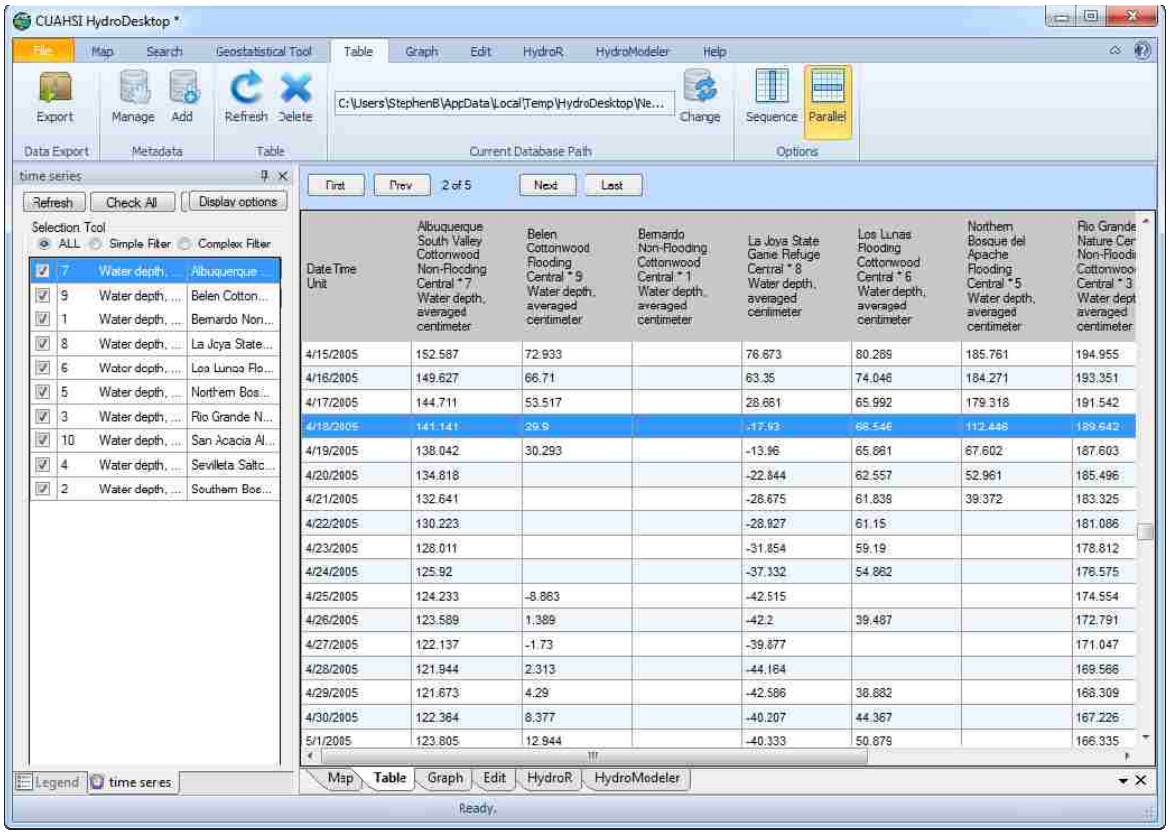


Figure 16: HydroDesktop: Comparing data in parallel

2.3 HIS Data Access: National

Comparing discharge data from the USGS NWIS database are trivial even for a long river reach. An extent from Albuquerque to Elephant Butte is chosen, followed by the variable, Discharge, the month of April, and NWIS Daily Values as the database to accelerate the search. Twenty one stations are returned for the search criteria. Three gages are of interest to the flooded locations, Rio Grande at Albuquerque (08330000) as a guide for what flow leaves Albuquerque, Rio Grande Floodway at San Acacia (08354900), and Rio Grande at San Marcial (08358400). Sorting the results attribute table by gage number organizes the list in order from upstream to downstream and simplifies choosing the site to download.

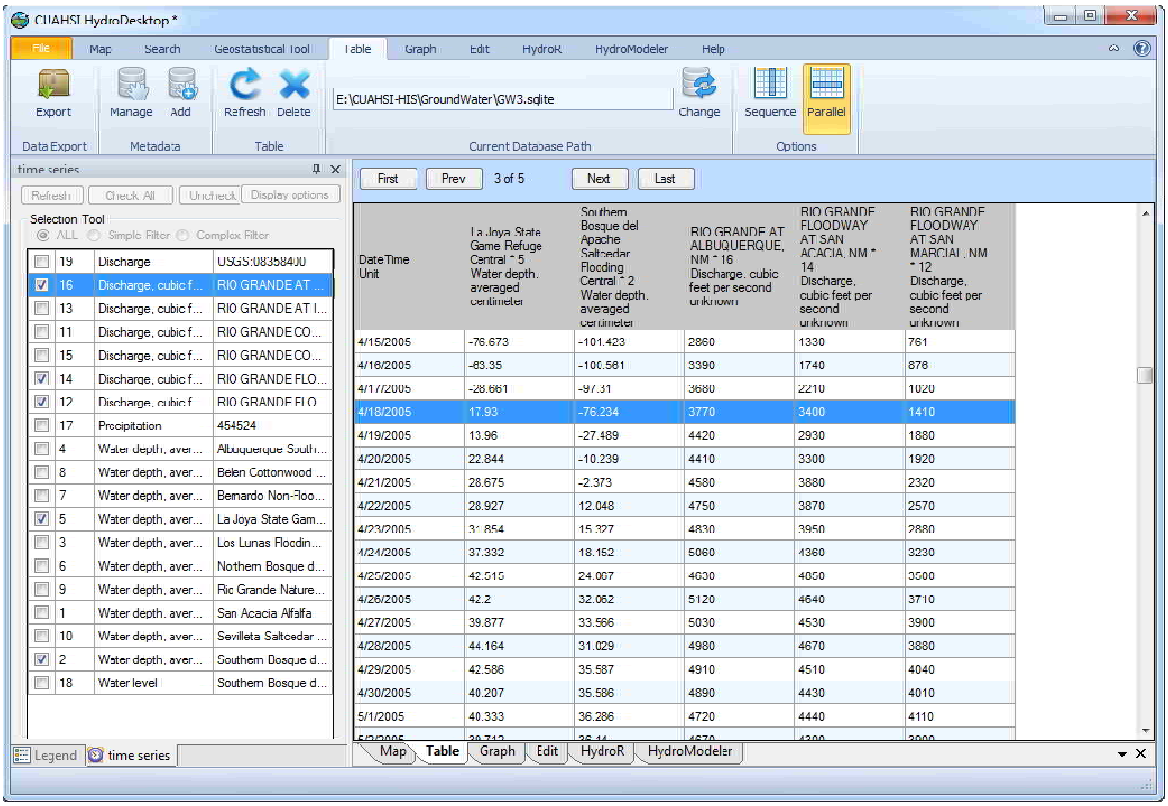


Figure 17: HydroDesktop: Identifying anomalies in tabular data

Loading the USGS gages and flooded well locations in the Table viewer (Figure 17) quickly illustrates the change in discharge in the main channel of the Rio Grande. On April 18, the flow in Albuquerque is 3770cfs, downstream of LARO is San Acacia at 3400cfs, and downstream of BDAS is San Marcial at 1410cfs. Bosque del Apache doesn't flood until the discharge reaches 2570cfs.

2.4 Identifying and exploring anomalies

Anomalies in the data can stand out when viewing in table form. For example, the San Acacia to San Marcial reach loses 1000cfs over the 90km while the Albuquerque to San Acacia reach loses only 370cfs. Exploring inconsistencies may be conducted in HydroDesktop using the high resolution base maps to look for large areas of crops or

diversion channels that would affect flow. Figure 18 shows a diversion near the San Acacia gage. Conducting another search of this reach for USGS discharge sites reveals Rio Grande Conveyance Channel at San Marcial.

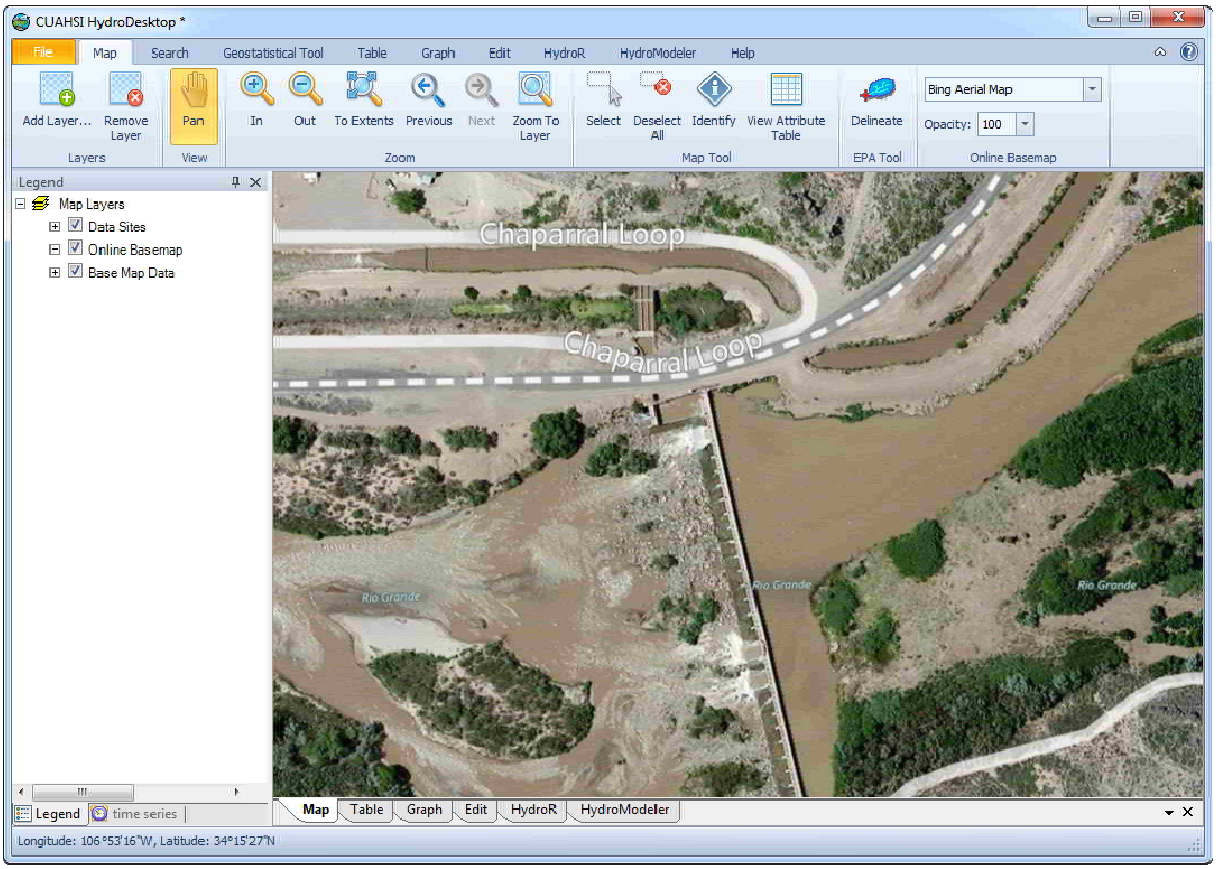


Figure 18: HydroDesktop: Base maps

The new data increases the total flow through San Marcial by 318cfs on April 18th bringing the total to 1728cfs, almost 1700cfs less than San Acacia. Looking back at the base maps for clues to the missing discharge shows considerable farming along the river with developed Acequia networks (Figure 19). Investigation outside of HydroDesktop resulted in an Interstate Stream Commission report indicating the Rio Grande experiences high seepage losses from Isleta to San Marcial (S.S. Papadopulos & Associates, 2002) that may account for the discharge shortage.

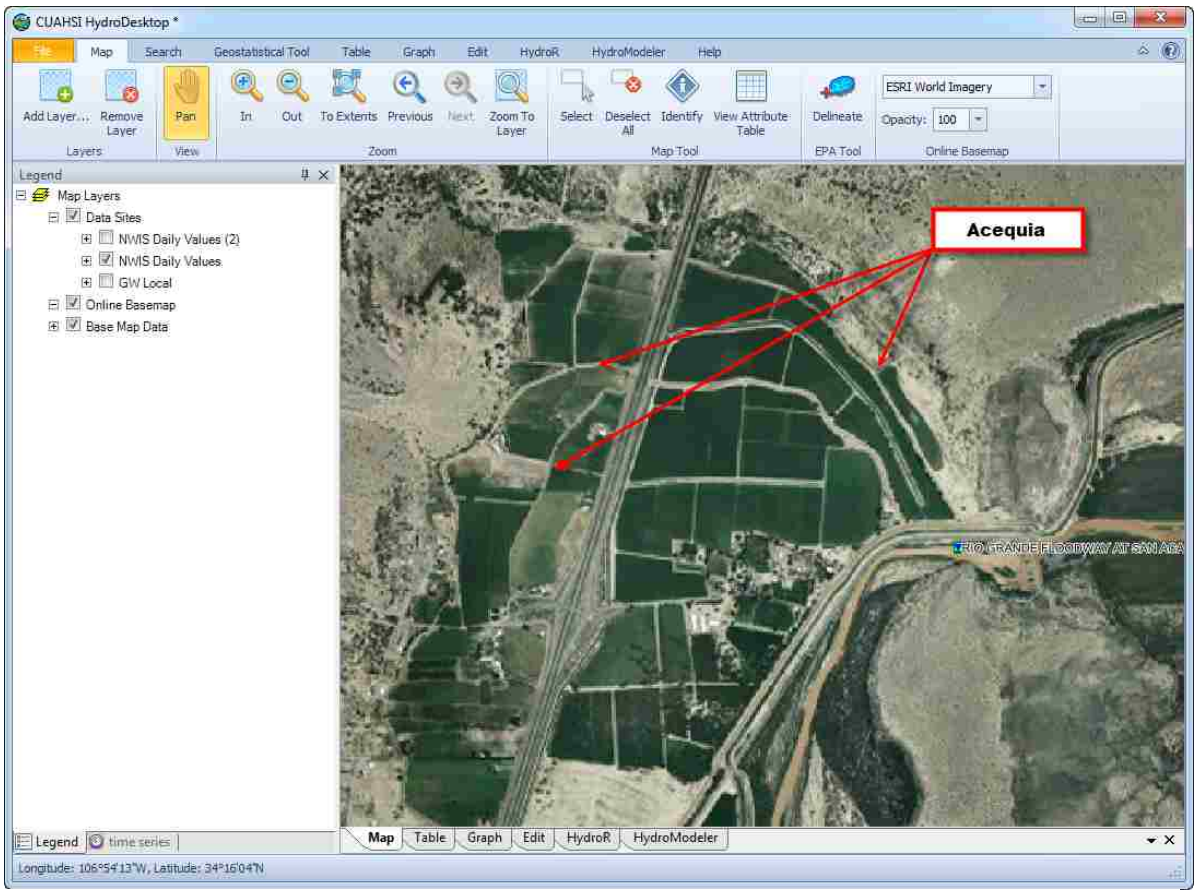


Figure 19: HydroDesktop: Identifying acequias

2.5 Graphing: Multiple stations and years

A common problem with visual trend analysis is getting enough data aggregated to conduct useful analysis. Few programs, if any, allow scrolling through time, zooming, easy add/remove of data series, and export to PDF as a vector for editing as an illustration. These graphing functions give HydroDesktop a unique position in data reconnaissance.

Creating a graph with twelve years of well data for ten locations is as simple as selecting checkboxes. Showing all the data at once makes for a crowded plot (Figure 20) but provides a starting point for analysis. The explanation/legend has been removed from this plot

providing a clear view of the data. The recurrence interval for flooding is obvious as well as the sites with the greatest swings in groundwater level throughout each season.

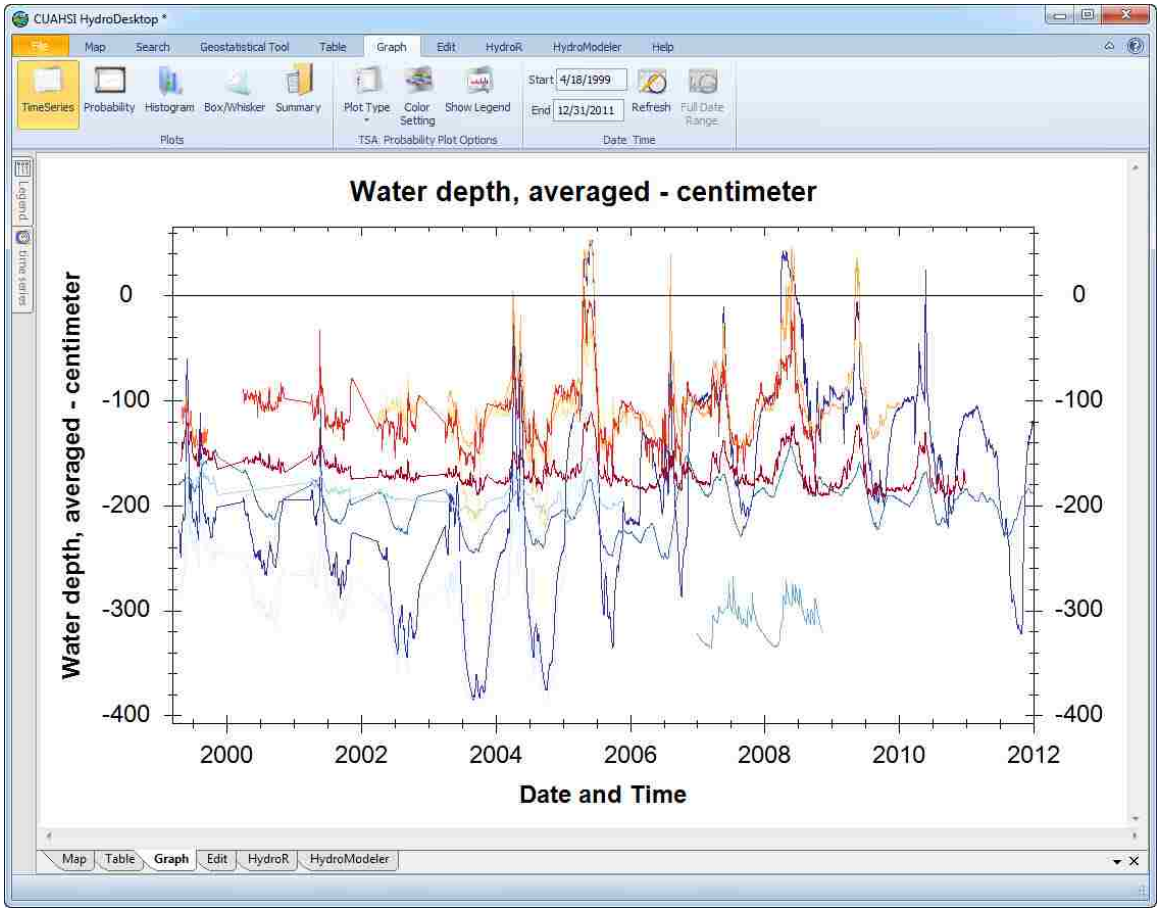


Figure 20: HydroDesktop: Eleven years of GW at ten stations

2.6 Graphing: Aggregation

Loading the Bosque del Apache groundwater level, San Marcial floodway and conveyance, and precipitation from a NWS virtual gage near the well, a picture can be painted of the influences of surface water on the ground water (Figure 21). Subsurface water levels are generally controlled by river stage with a 4-8 hour response time as stated by Martinet et. al. (2008) with the exception of small fluctuations of discharge similar to the event on April 12th

that are not reflected in the groundwater level (Figure 22). River stage was used in the Martinet et. al. analysis of surface and ground water hydrographs due to the impact of hydraulic head on ground water elevations (Martinet, et al., 2009). The use of discharge in the high resolution graph showing 30-minute data may account for the absence of surface water impact on the ground water level.

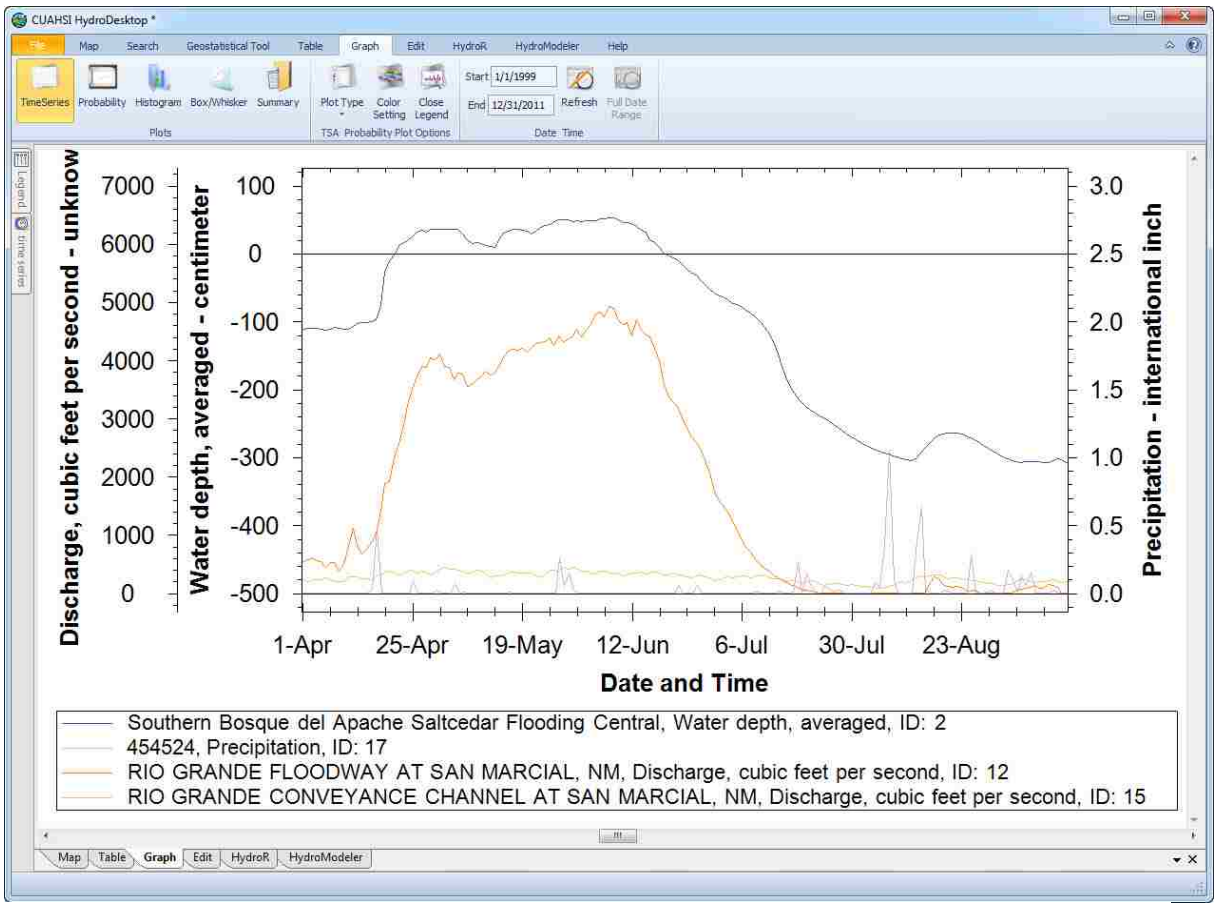


Figure 21: HydroDesktop: Graph aggregation

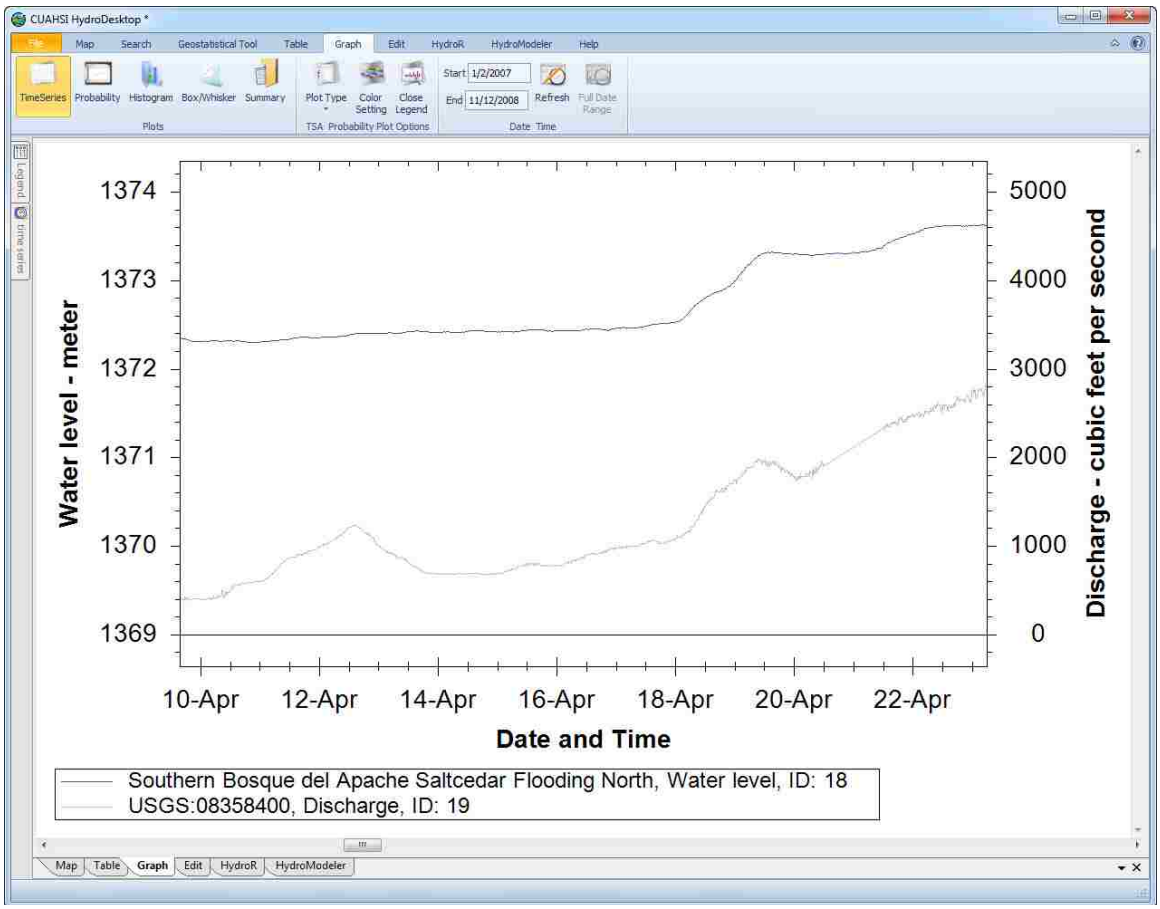


Figure 22: HydroDesktop: Zoom to detail

2.7 Graphing: Export for Publication

Telling a story with a graph in a publication requires precise control of the axes, explanation/legend, title, etc. HydroDesktop's graphing module falls short in these critical areas. The plots produced within the application will need to be exported to a vector illustration package to fine tune the layout or processed through HydroR. Exporting vector plot from HydroDesktop is accomplished by printing the graph as a PDF. Many researchers are familiar with Adobe Illustrator or Corel Draw allowing necessary customization and merging of plots. Explanations/Legends, axes and individual data series can be modified or deleted as shown in Figure 26 and Figure 27.

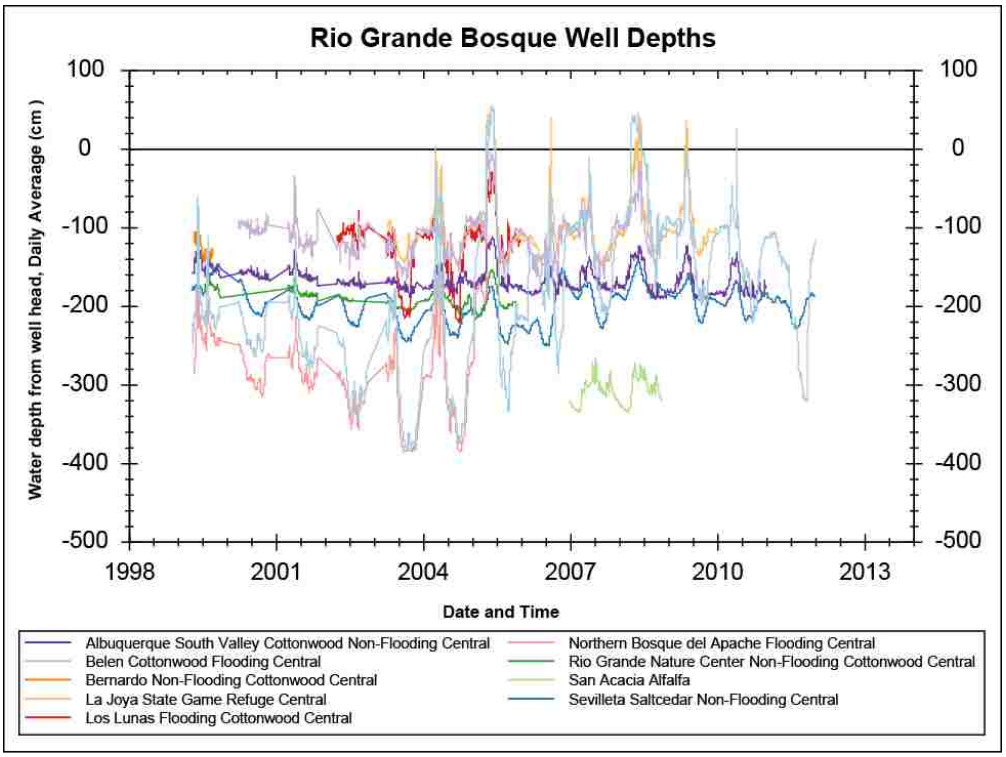


Figure 23: Adobe Illustrator editing

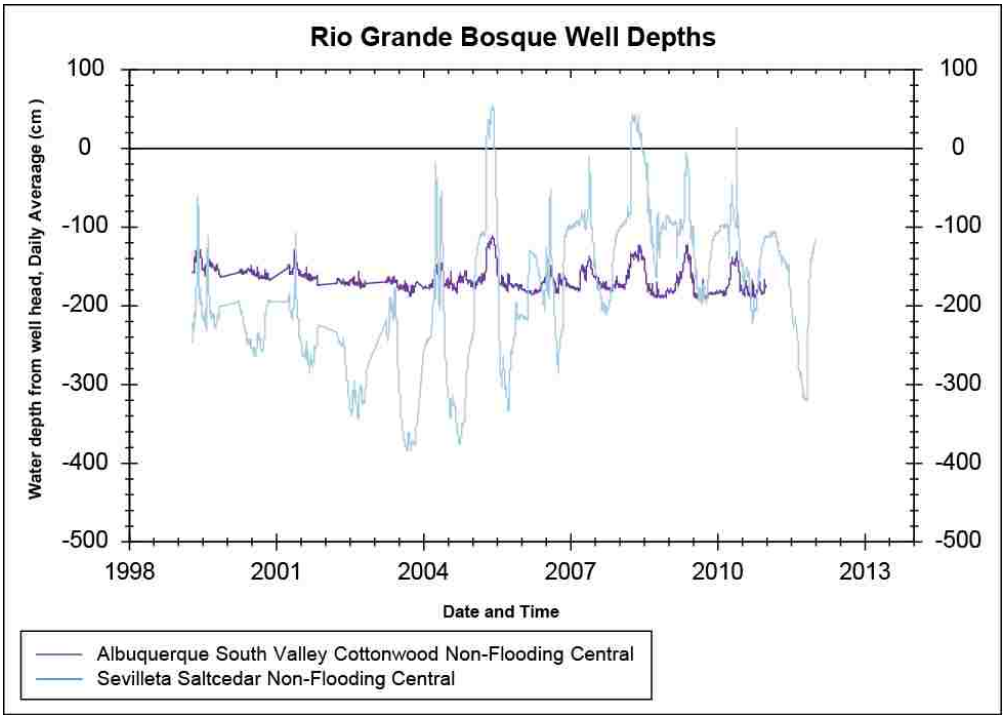


Figure 24: Adobe Illustrator editing

CHAPTER 3: RIO GRANDE EVAPOTRANSPIRATION

3.1 Data Access Methods: Original

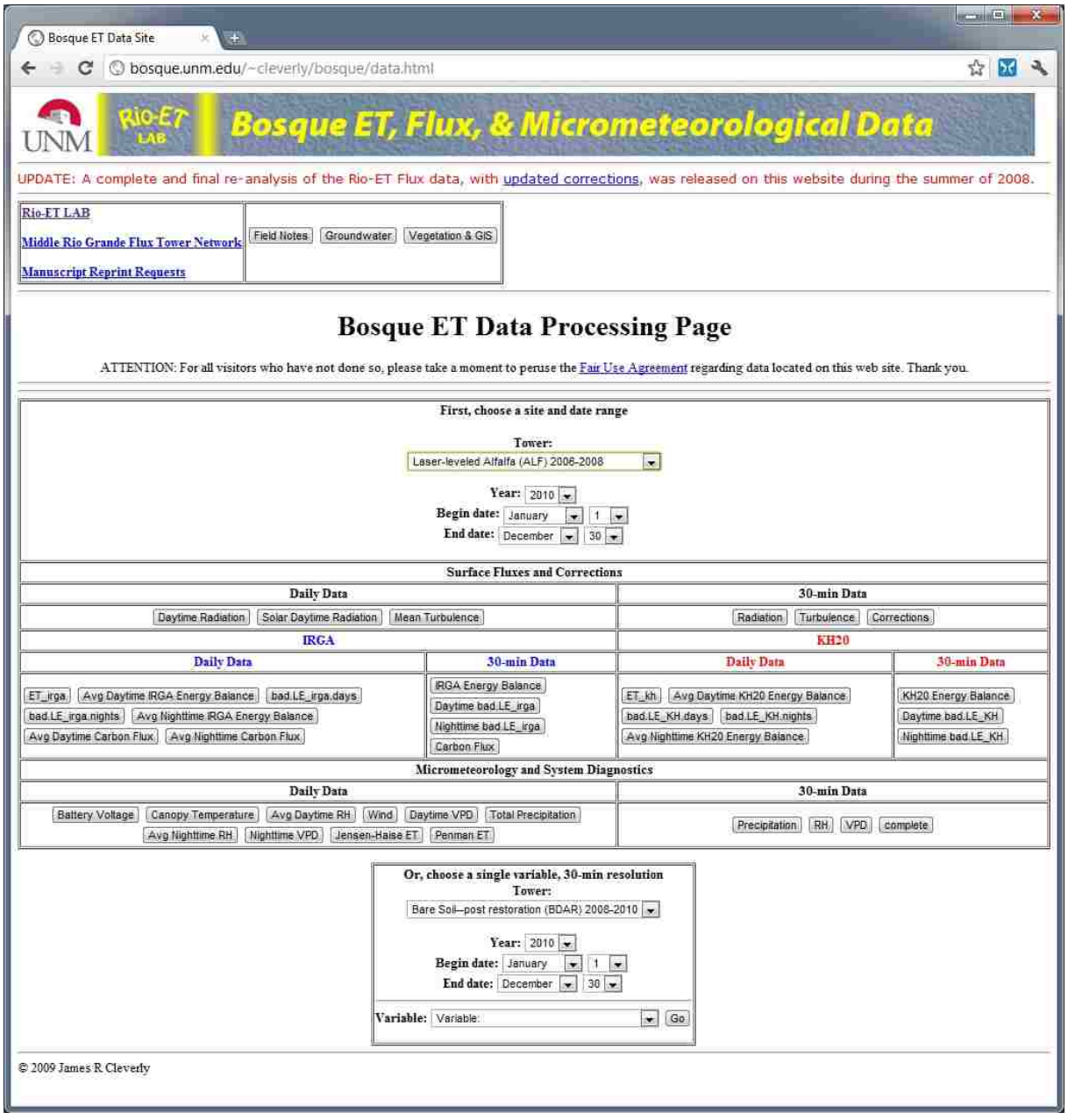
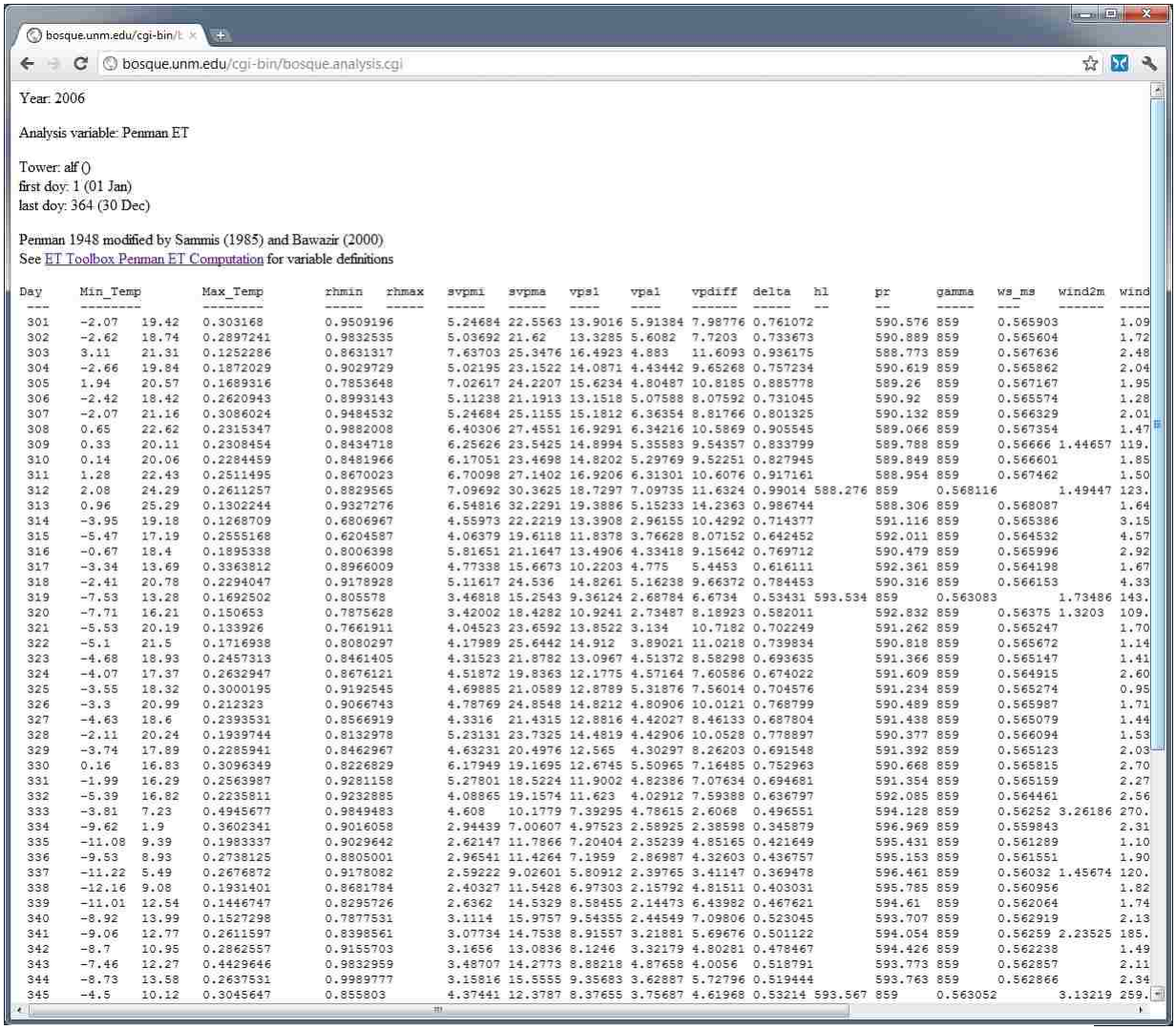


Figure 25: Original Rio Grande ET data access

The principal investigator developed a series of Perl scripts and an HTML portal to deliver the ET tower data (Figure 25). Users fill out the form on the project website for a specific tower,

year and set of variables, returning a tab delimited page in the web browser (Figure 26). The Perl scripts process the data fast and efficient, making data acquisition straight forward. Post processing requires extensive time if multi-year, multi-tower analysis is to be performed. The tab delimited pages must be saved or copied and pasted to a text editor or Excel and aggregated. Julian formatting of the dates adds further complication to multi-year compilation and analysis.



Day	Min_Temp		Max_Temp	rhmin	rhmax	svpmi	svpma	vps1	vpa1	vpdiff	delta	hl	pr	gamma	ws_ms	wind2m	wind
301	-2.07	19.42	0.303168	0.9509196		5.24684	22.5563	13.9016	5.91384	7.98776	0.761072		590.576	859	0.565903		1.09
302	-2.62	18.74	0.2897241	0.9832535		5.03692	21.62	13.3285	5.6082	7.7203	0.733673		590.889	859	0.565604		1.72
303	3.11	21.31	0.1252286	0.8631317		7.63703	25.3476	16.4923	4.883	11.6093	0.936175		588.773	859	0.567636		2.48
304	-2.66	19.84	0.1872029	0.9029729		5.02195	23.1522	14.0871	4.43442	9.65268	0.757234		590.619	859	0.565862		2.04
305	1.94	20.57	0.1689316	0.7853648		7.02617	24.2207	15.6234	4.80487	10.8185	0.885778		589.26	859	0.567167		1.95
306	-2.42	18.42	0.2620943	0.8993143		5.11238	21.1913	13.1518	5.07588	8.07592	0.731045		590.92	859	0.565574		1.28
307	-2.07	21.16	0.3086024	0.9484532		5.24684	25.1155	15.1812	6.36354	8.81766	0.801325		590.132	859	0.566329		2.01
308	0.65	22.62	0.2315347	0.9882008		6.40306	27.4551	16.9291	6.34216	10.5869	0.905545		589.066	859	0.567354		1.47
309	0.33	20.11	0.2308454	0.8434718		6.25626	23.5425	14.8994	5.35583	9.54357	0.833799		589.788	859	0.56666	1.44657	119.
310	0.14	20.06	0.2284459	0.8481966		6.17051	23.4698	14.8202	5.29769	9.52251	0.827945		589.849	859	0.566601		1.85
311	1.28	22.43	0.2511495	0.8670023		6.70098	27.1402	16.9206	6.31301	10.6076	0.917161		588.954	859	0.567462		1.50
312	2.08	24.29	0.2611257	0.8829565		7.09692	30.3625	18.7297	7.09735	11.6324	0.99014	588.276	859	0.568116	1.49447		123.
313	0.96	25.29	0.1302244	0.9327276		6.54816	32.2291	19.3886	5.15233	14.2363	0.986744		588.306	859	0.568087		1.64
314	-3.95	19.18	0.1268709	0.6806967		4.55973	22.2219	13.3908	2.96155	10.4292	0.714377		591.116	859	0.565386		3.15
315	-5.47	17.19	0.2555168	0.6204587		4.06379	19.6118	11.8378	3.76628	8.07152	0.642452		592.011	859	0.564532		4.57
316	-0.67	18.4	0.1895338	0.8006398		5.81651	21.1647	13.4906	4.33418	9.15642	0.769712		590.479	859	0.565996		2.92
317	-3.34	13.69	0.3363812	0.8966009		4.77338	15.6673	10.2203	4.775	5.4453	0.616111		592.361	859	0.564198		1.67
318	-2.41	20.78	0.2294047	0.9178928		5.11617	24.536	14.8261	5.16238	9.66372	0.784453		590.316	859	0.566153		4.33
319	-7.53	13.28	0.1692502	0.805578		3.46818	15.2543	9.36124	2.68784	6.6734	0.53431	593.534	859	0.563083	1.73486		143.
320	-7.71	16.21	0.150653	0.7875628		3.42002	18.4282	10.9241	2.73487	8.18923	0.582011		592.832	859	0.56375	1.3203	109.
321	-5.53	20.19	0.133926	0.7661911		4.04523	23.6592	13.8522	3.134	10.7182	0.702249		591.262	859	0.565247		1.70
322	-5.1	21.5	0.1716938	0.8080297		4.17989	25.6442	14.912	3.89021	11.0218	0.739834		590.818	859	0.565672		1.14
323	-4.68	18.93	0.2457313	0.8461405		4.31523	21.8782	13.0967	4.51372	8.58298	0.693635		591.366	859	0.565147		1.41
324	-4.07	17.37	0.2632947	0.8676121		4.51872	19.8363	12.1775	4.57164	7.60586	0.674022		591.609	859	0.564915		2.60
325	-3.55	18.32	0.3000195	0.9192545		4.69885	21.0589	12.8789	5.31876	7.56014	0.704576		591.234	859	0.565274		0.95
326	-3.3	20.99	0.212323	0.9066743		4.78769	24.8548	14.8212	4.80906	10.0121	0.768799		590.489	859	0.565987		1.71
327	-4.63	18.6	0.2393531	0.8566919		4.3316	21.4315	12.8816	4.42027	8.46133	0.687804		591.438	859	0.565079		1.44
328	-2.11	20.24	0.1939744	0.8132978		5.23131	23.7325	14.4819	4.42906	10.0528	0.778897		590.377	859	0.566094		1.53
329	-3.74	17.89	0.2285941	0.8462967		4.63231	20.4976	12.565	4.30297	8.26203	0.691548		591.392	859	0.565123		2.03
330	0.16	16.83	0.3096349	0.8226829		6.17949	19.1695	12.6745	5.50965	7.16485	0.752963		590.668	859	0.565815		2.70
331	-1.99	16.29	0.2563987	0.9281158		5.27801	18.5224	11.9002	4.82386	7.07634	0.694681		591.354	859	0.565159		2.27
332	-5.39	16.82	0.2235811	0.9232885		4.08865	19.1574	11.623	4.02912	7.59388	0.636797		592.085	859	0.564461		2.56
333	-3.81	7.23	0.4945677	0.9849483		4.608	10.1779	7.39295	4.78615	2.6068	0.496551		594.128	859	0.56252	3.26186	270.
334	-9.62	1.9	0.3602341	0.9016058		2.94439	7.00607	4.97523	2.58925	2.38598	0.345879		596.969	859	0.559843		2.31
335	-11.08	9.39	0.1983337	0.9029642		2.62147	11.7866	7.20404	2.35239	4.85165	0.421649		595.431	859	0.561289		1.10
336	-9.53	8.93	0.2738125	0.8805001		2.96541	11.4264	7.1959	2.86987	4.32603	0.436757		595.153	859	0.561551		1.90
337	-11.22	5.49	0.2676872	0.9178082		2.59222	9.02601	5.80912	2.39765	3.41147	0.369478		596.461	859	0.56032	1.45674	120.
338	-12.16	9.08	0.1931401	0.8681784		2.40327	11.5428	6.97303	2.15792	4.81511	0.403031		595.785	859	0.560956		1.82
339	-11.01	12.54	0.1446747	0.8295726		2.6362	14.5329	8.58455	2.14473	6.43982	0.467621		594.61	859	0.562064		1.74
340	-8.92	13.99	0.1527298	0.7877531		3.1114	15.9757	9.54355	2.44549	7.09806	0.523045		593.707	859	0.562919		2.13
341	-9.06	12.77	0.2611597	0.8398561		3.07734	14.7538	8.91557	3.21881	5.69676	0.501122		594.054	859	0.56259	2.23525	185.
342	-8.7	10.95	0.2862557	0.9155703		3.1656	13.0836	8.1246	3.32179	4.80281	0.478467		594.426	859	0.562238		1.49
343	-7.46	12.27	0.4429646	0.9832959		3.48707	14.2773	8.88218	4.87658	4.0056	0.518791		593.773	859	0.562857		2.11
344	-8.73	13.58	0.2637531	0.9989777		3.15816	15.5555	9.35683	3.62887	5.72796	0.519444		593.763	859	0.562866		2.34
345	-4.5	10.12	0.3045647	0.855803		4.37441	12.3787	8.37655	3.75687	4.61968	0.53214	593.567	859	0.563052	3.13219		259.

Figure 26: Original Rio Grande ET data access results

3.2 Data Access Methods: HydroServer

Project data from all towers, totaling over 35 million measurements, have been loaded in HydroServer with full metadata incorporated. Data are available from any REST aware application.

3.3 Data Portability: Original

Data for the ET towers are delivered via an Apple server in the research scientist's office. Using custom Perl scripts and an Apache web service seems like a stable, portable platform for delivering data but the server has experienced an unknown hardware problem resulting in frequent downtime. Without dedicated information technology staff to diagnose problems and return the machine to service it has been unavailable for extended periods of time. Upgrading the system is cost prohibitive and may require modification of the Perl scripts and web server configuration.

3.4 Data Portability: HydroServer

The server for the ET and Well data is a Windows Server 2008 virtual server running at the University of New Mexico in the Earth Data Analysis Center (EDAC) server farm. Initial configuration and data loading were conducted on a standalone system running VMware Workstation and moved to the data center when complete. Virtual environments and standardization of HydroServer allow streamlined migration of datasets.

One HydroServer can host multiple discrete projects, reducing cost to the research team. Each project is contained in its own database that is easily transferred to a new server if a

hardware failure occurs. Having portable databases allows for local development 'sandboxes' where researchers can conduct QA/QC and analysis on a local dataset without slowing the production server distributing data to the public.

3.5 Data Access: Download

Using HydroDesktop, a spatial search is conducted of the middle Rio Grande for Evapotranspiration over the time frame of the Rio-ET study, 1999-Present. 42 unique station/variable combinations are returned and queried to isolate TotalET for download resulting in 6 stations (Figure 27).

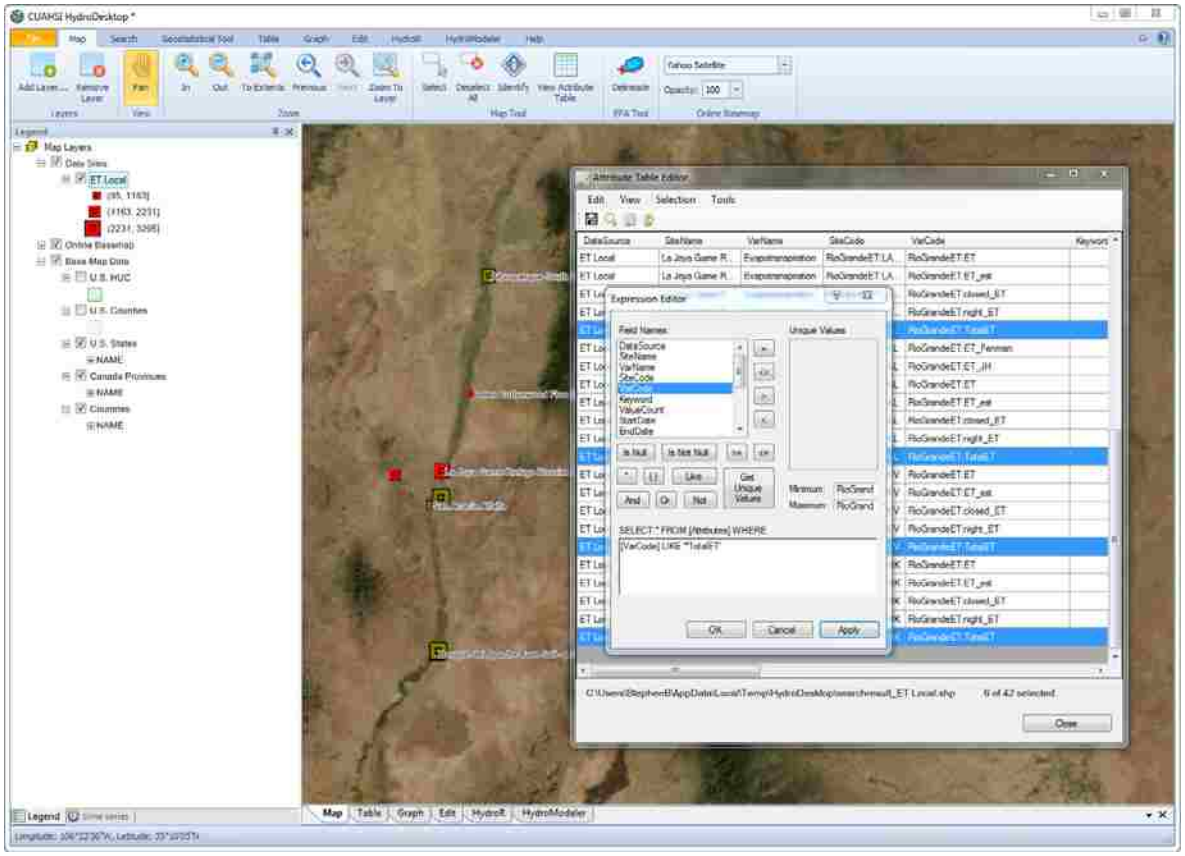


Figure 27: HydroDesktop: Rio Grande ET data access

3.6 Analysis: Tabular

Bringing the data into the Table viewer (Figure 28) allows rapid inspection of overlapping time series for system wide analysis and location of missing values that will need to be interpolated.

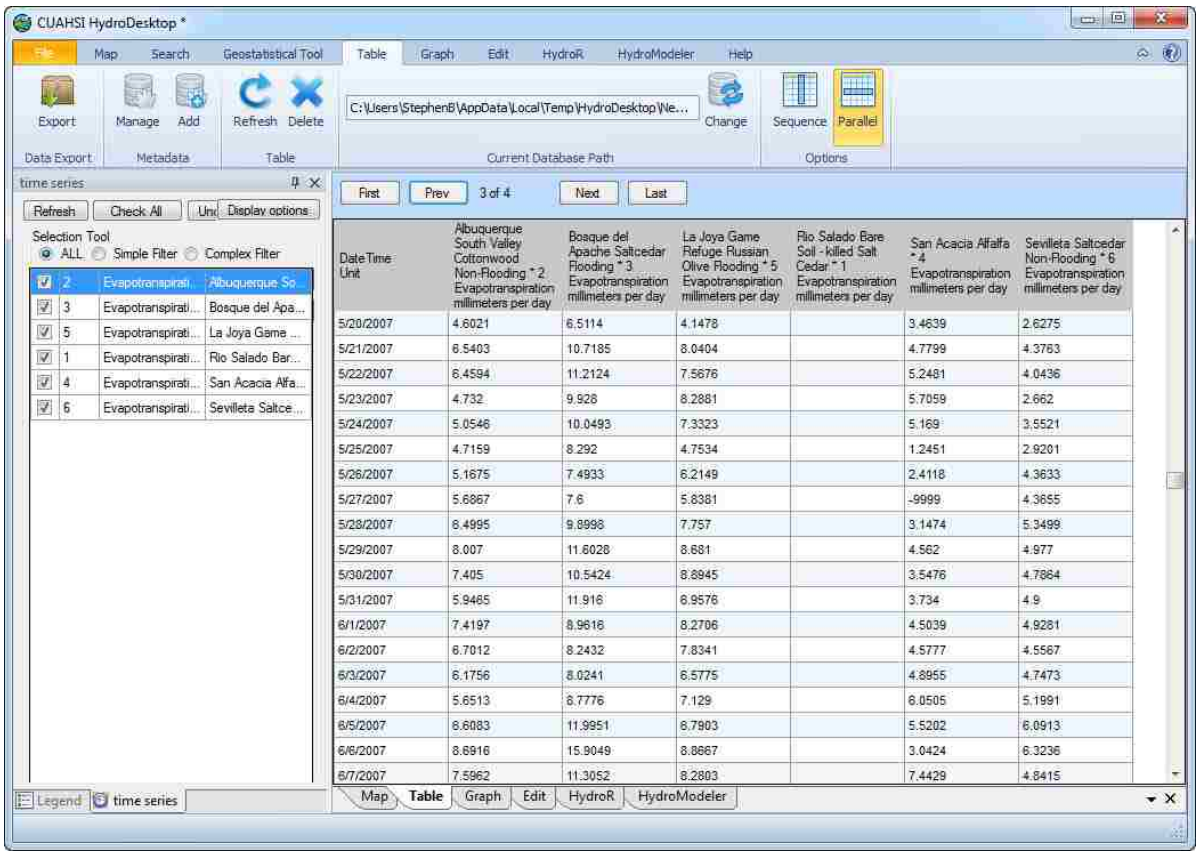


Figure 28: HydroDesktop: Tabular data review

3.7 Analysis: Graphing and Statistics

Switching to the Graph tab provides an overview of all years and stations. Entering a date range allows for quick isolation of a year (Figure 29). Here it becomes obvious the Bosque del Apache Saltcedar site experienced a large surge in ET during the first part of June.

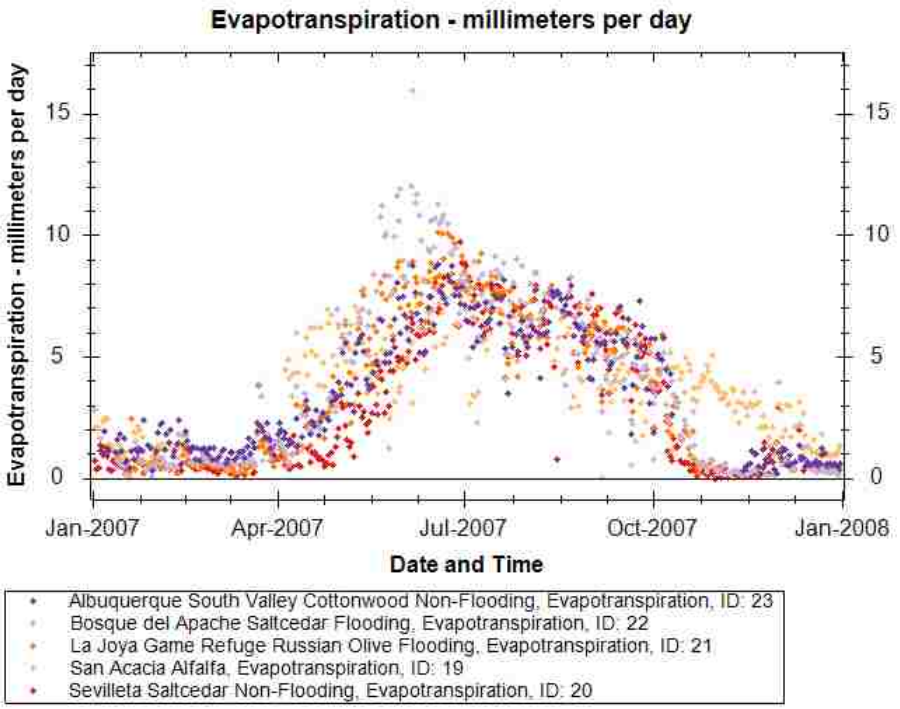


Figure 29: HydroDesktop graphing

La Joya Game Refuge, being the only other flooding site shows a slight elevation from the non-flooding sites. The other stations are relatively consistent with the exception of a few outliers. San Acacia Alfalfa experiences the largest dispersion, likely caused by irrigation. There is a stark contrast between the surge in ET at the flooding Bosque del Apache and the non-flooding saltcedar site, Sevilleta. Located roughly 100km apart in a losing reach of the Rio Grande, the riparian systems are similar (Martinet, et al., 2009).

Zooming in on May and June (Figure 30) shows ET steadily climbing at both flooding and non-flooding sites as expected in early season growth (Dahm, et al., 2002). When query results for local groundwater depth (green) and Rio Grande discharge (blue) are added, the impact of being connected to the river in a flooding system is evident. The San Acacia river discharge near the Sevilleta is 1000 cfs greater than at San Marcial near Bosque del Apache yet the groundwater level is flat for the Sevilleta.

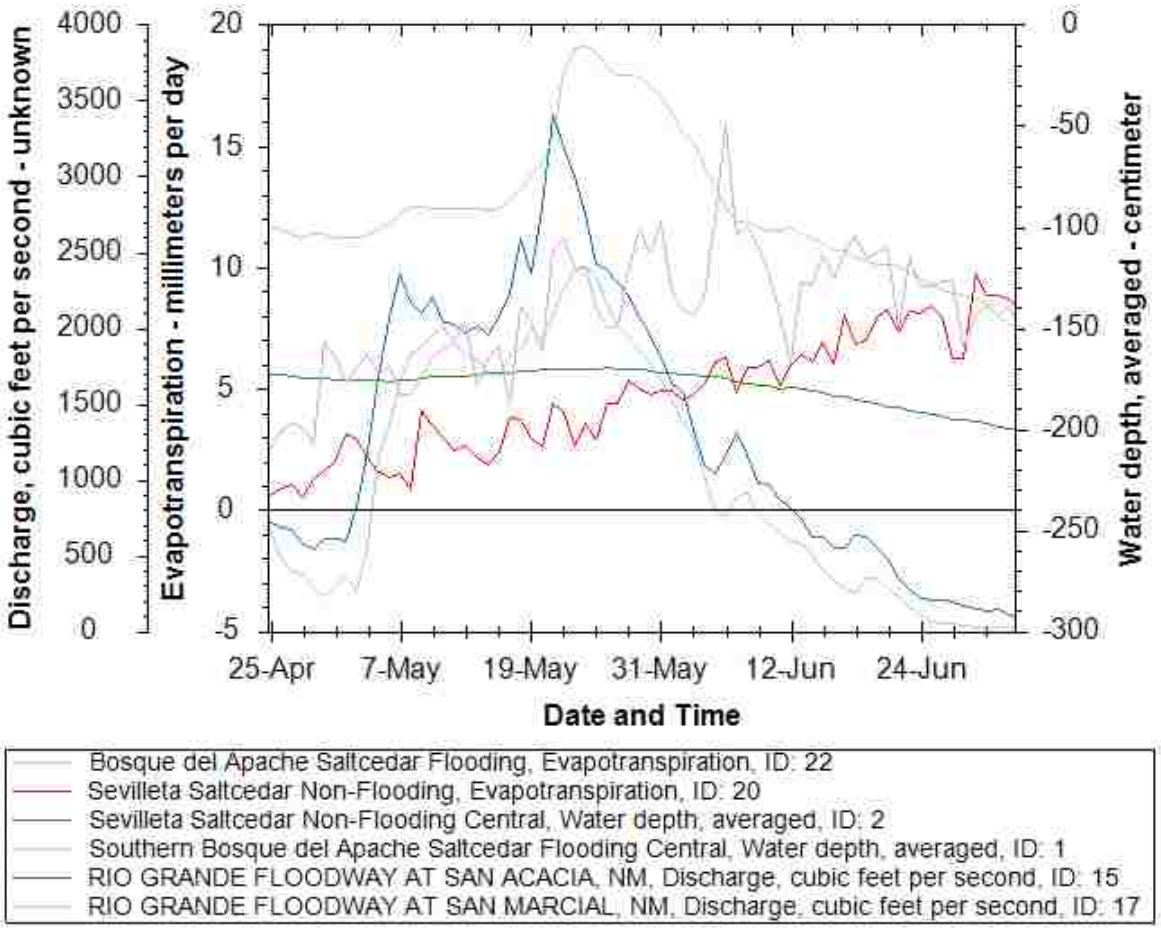


Figure 30: HydroDesktop Graphing Zoom

Several basic statistical functions are visible on the Graph ribbon. Probability, Histogram, and Box/Whisker return statistical plots for the Start and End Date range, Summary shows a list of general statistics for each site within the Start and End Date range (Table 3). Exploring the summary statistics provides insight to the quality of the dataset. Of the four sites in Table 3, the number of observations varies from 277 at San Acacia to 350 at Sevilleta. Bosque del Apache had the highest maximum ET, 15.9 mm/day, and San Acacia Alfalfa had the lowest maximum ET, 8.3 mm/day but the mean of the Alfalfa was higher, 3.87 versus 3.85. Inclusion of the Coefficient of Variation (CV, Standard Deviation/Arithmetic Mean) allows rapid assessment of the dispersion of the data (Wright, 2012). San Acacia Alfalfa has a CV

near 0.5, indicating a wide dispersion of data points that is evident in the scattered arrangement of values shown in the plot of the data (Figure 29).

Table 3: HydroDesktop: Rio Grande statistics

Rio Grande Evapotranspiration Statistics for 2007			
Albuquerque South Valley Cottonwood Non-Flooding, Evapotranspiration	ID 2	**San Acacia Alfalfa, Evapotranspiration**	ID 4
# Of Observations	314	**# Of Observations**	277
# Of Censored Obs.	0	**# Of Censored Obs.**	0
Arithmetic Mean	3.66390924	**Arithmetic Mean**	3.8735361
Geometric Mean	2.64619234	**Geometric Mean**	3.26454106
Maximum	8.7479	**Maximum**	8.2866
Minimum	0.1613	**Minimum**	0.0307
Standard Deviation	2.60079471	**Standard Deviation**	1.99050367
Coefficient of Variation	0.70984147	**Coefficient of Variation**	0.5138725
Percentiles 10%	0.6496	**Percentiles 10%**	1.1424
Percentiles 25%	1.1396	**Percentiles 25%**	2.1256
Percentiles 50%(median)	3.3708	**Percentiles 50%(median)**	3.8036
Percentiles 75%	6.0272	**Percentiles 75%**	5.3075
Percentiles 90%	7.3732	**Percentiles 90%**	6.6787
Bosque del Apache Saltcedar Flooding, Evapotranspiration	ID 3	**Sevilleta Saltcedar Non-Flooding, Evapotranspiration**	ID 6
# Of Observations	345	**# Of Observations**	350
# Of Censored Obs.	0	**# Of Censored Obs.**	0
Arithmetic Mean	3.8551171	**Arithmetic Mean**	2.90906429
Geometric Mean	2.54932926	**Geometric Mean**	2.04349897
Maximum	15.9049	**Maximum**	9.7005
Minimum	0.0306	**Minimum**	-0.073
Standard Deviation	3.47689682	**Standard Deviation**	2.81587007
Coefficient of Variation	0.90189136	**Coefficient of Variation**	0.9679642
Percentiles 10%	0.2964	**Percentiles 10%**	0.2514
Percentiles 25%	0.5891	**Percentiles 25%**	0.4636
Percentiles 50%(median)	2.7296	**Percentiles 50%(median)**	1.2904
Percentiles 75%	6.7759	**Percentiles 75%**	5.7634
Percentiles 90%	8.7119	**Percentiles 90%**	7.0651

3.8 Graphing: Publication Quality in HydroR

HydroR combined with R library ggplot2 (Wickham, 2009) provides a powerful tool for creating graphs in HydroDesktop. The R-Project website describes R as "a language and environment for statistical computing and graphics" and continues "One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed." (R Development Core Team, 2011) Producing plots in R comes at a small cost of learning curve and initial setup of the export script. R streamlines the graphing process by allowing user selection of time frames and station IDs within HydroDesktop or directly into the HydroServer ODM via SQL connection strings.

After several scripts are archived, they can be recycled and modified for future projects, saving considerable time. Scripts of all plots presented here are available in Appendix B.

Plotting for specific projects can be a redundant process requiring similarly formatted graphs with different temporal ranges or stations. For the Rio-ET Project Final Report annual ET plots for each station in the study were generated and nested for publication. Scripting in R streamlines these repetitive tasks and ensures consistency.

3.9 Graphing: Sample Plots from HydroR

Plots created in R are highly customizable. When working within the R interface, graphs can be stretched and enlarged dynamically without distorting the axis or title, and saved at the final size. Output formats include; PNG, JPG, TIFF, and PDF for editing as a vector image (Figure 31).

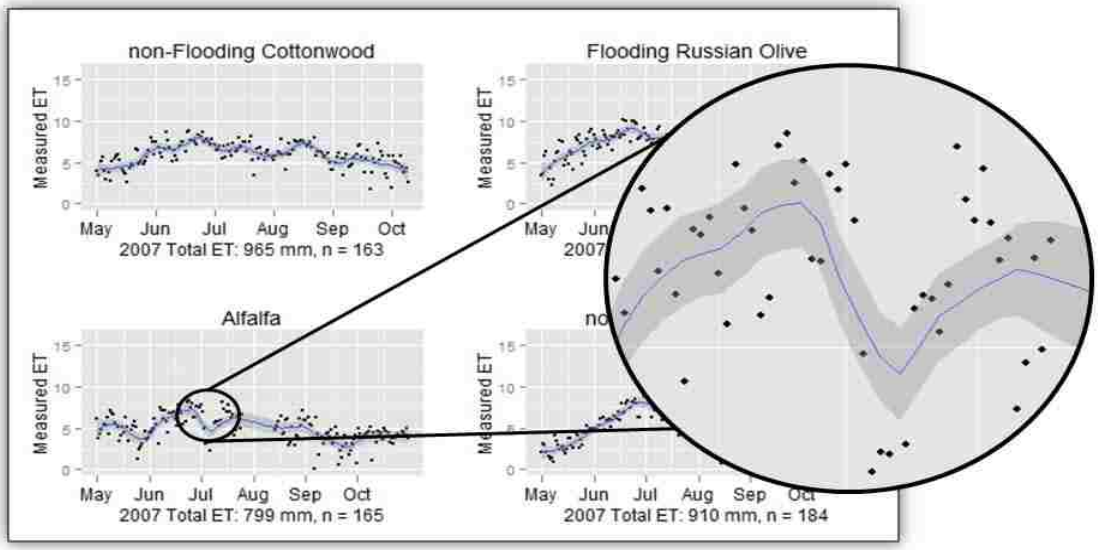


Figure 31: R plot as vector image

3.9.1 Multiple plots per page

Viewing multiple plots on one page with matching axes is trivial in R. Data calculations can also be included in the title or axes for additional user information. Figure 32 provides an overview of four stations for 2007, showing how the locations vary over the growing season and the total ET measured. The number of data points used to calculate total ET are included, allowing rapid identification of gaps in the datasets. The months of June and July appear to have the highest levels of ET, producing a plot to examine 60 days of detail takes seconds (Figure 32).

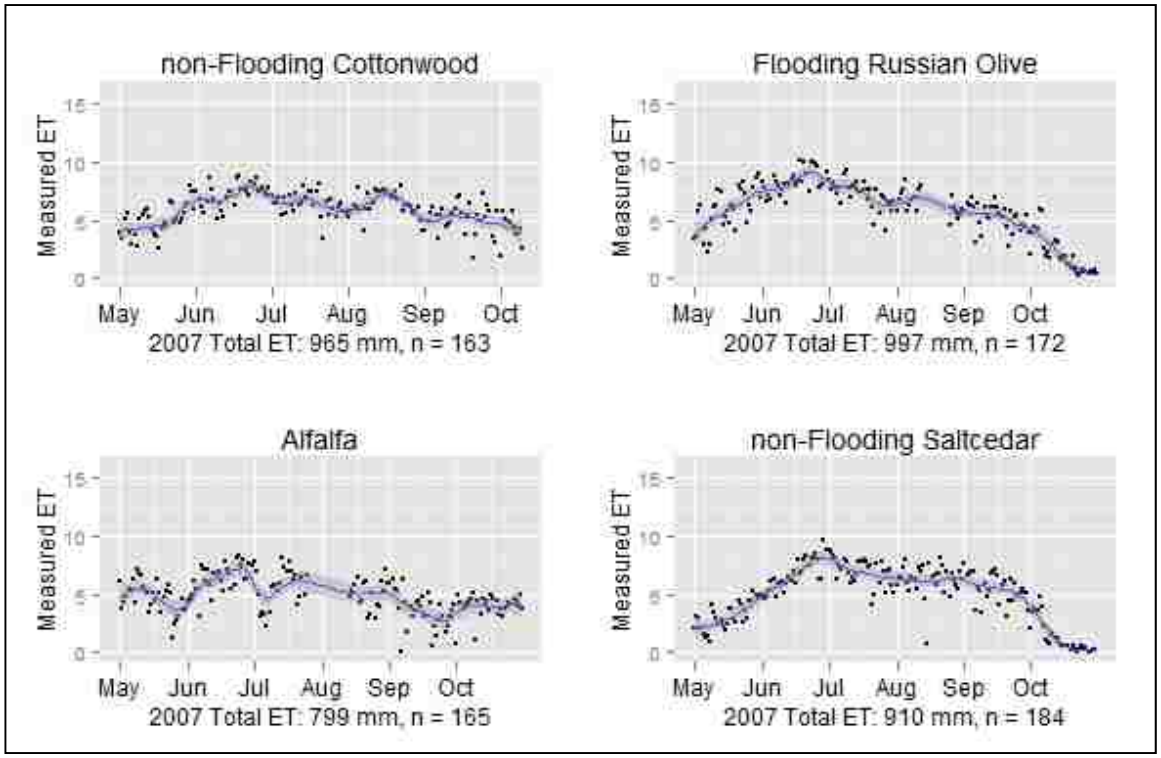


Figure 32: Rio Grande ET Stations :: 2007

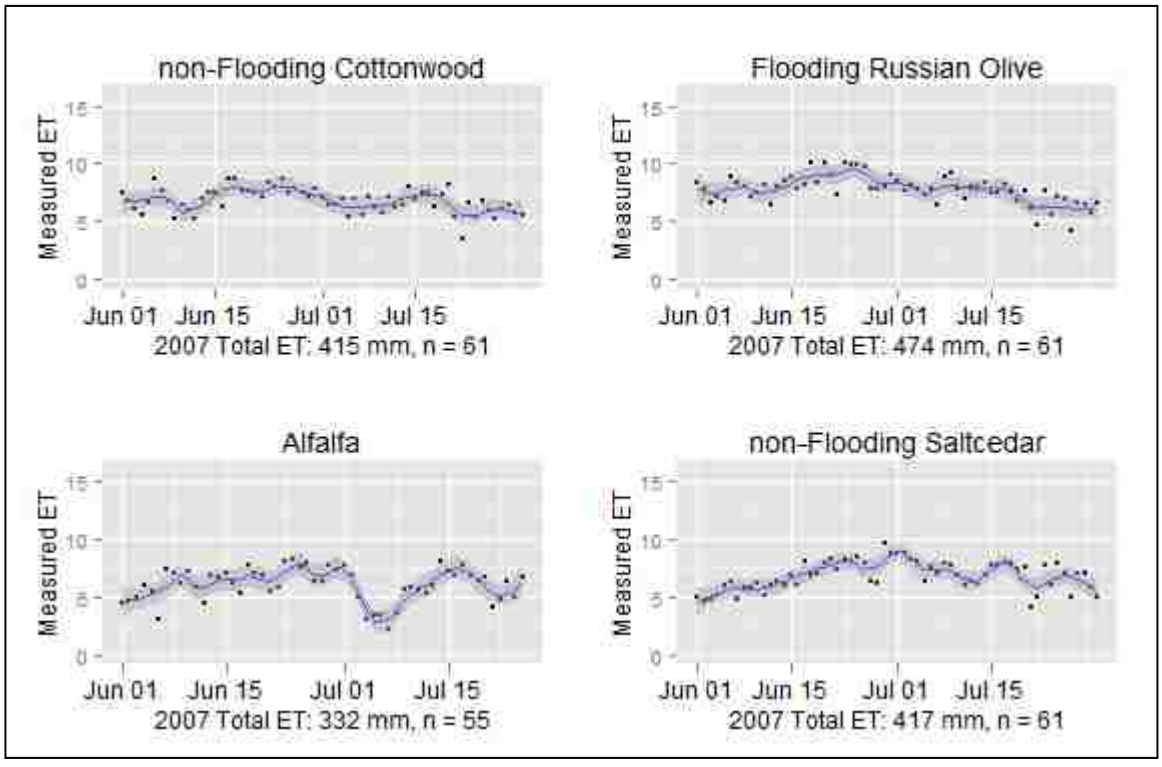


Figure 33: Rio Grande ET Stations :: June/July 2007

3.9.2 Changing Variables

HydroR scripts can be saved and loaded when different data needs analysis. Figure 34ab shows the full growing season for San Acacia Alfalfa and isolates the month of June.

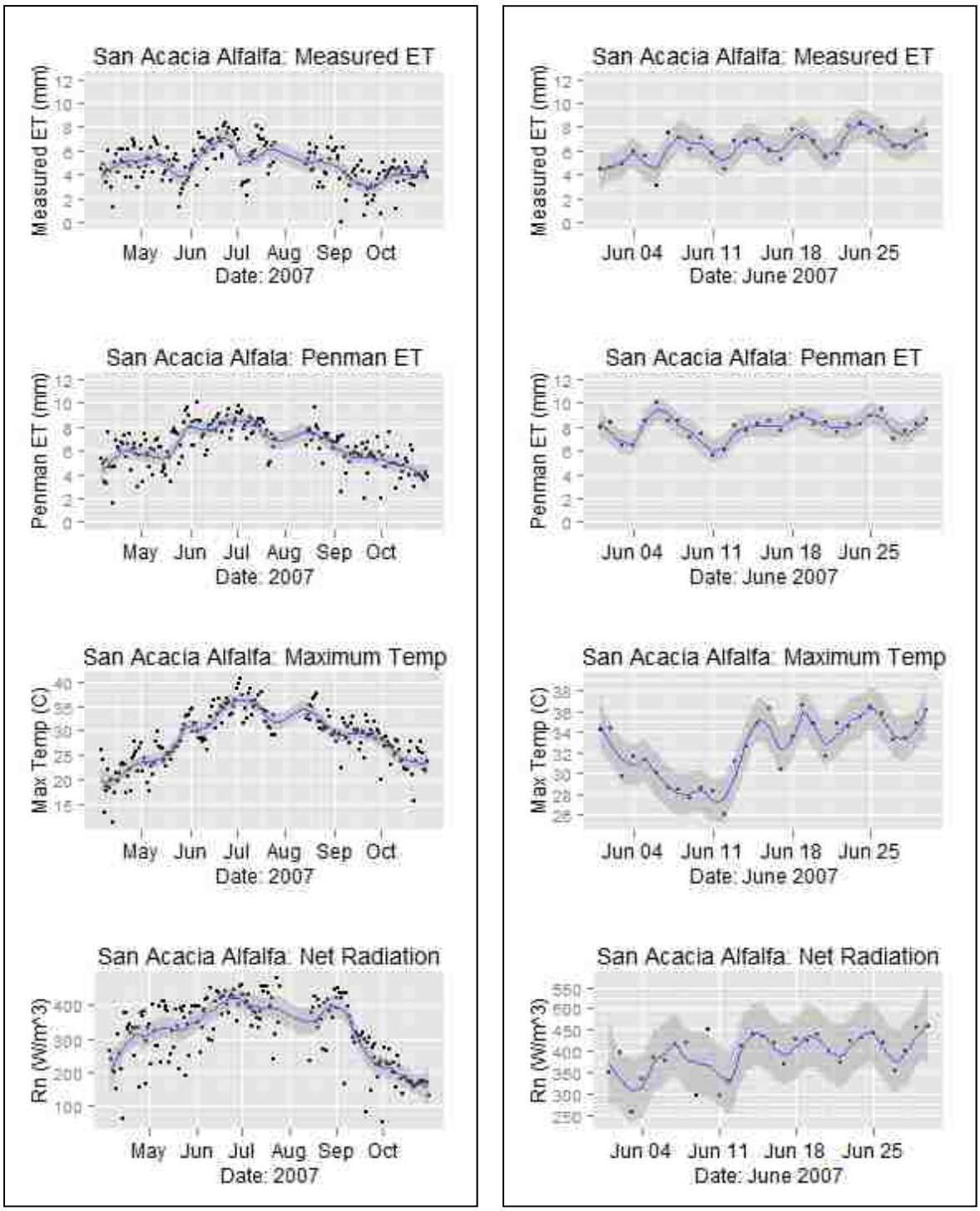


Figure 34ab: ALF :: Measured ET, Penman ET, Max Temp, Net Radiation

Looking at the entire growing season allows for rapid trend visualization, detailed analysis is available by zooming in to one month. Figure 34b shows the Penman ET approximation under-estimates ET for the month of June and does not capture the oscillations or increasing trend. When comparing the measured ET to maximum temperature and net radiation it is clear that net radiation is the driving force in ET for these variables. Bringing in other variables provides an opportunity to visually check for correlation. In Figure 35, we can see that river discharge is generally unrelated to measured ET.

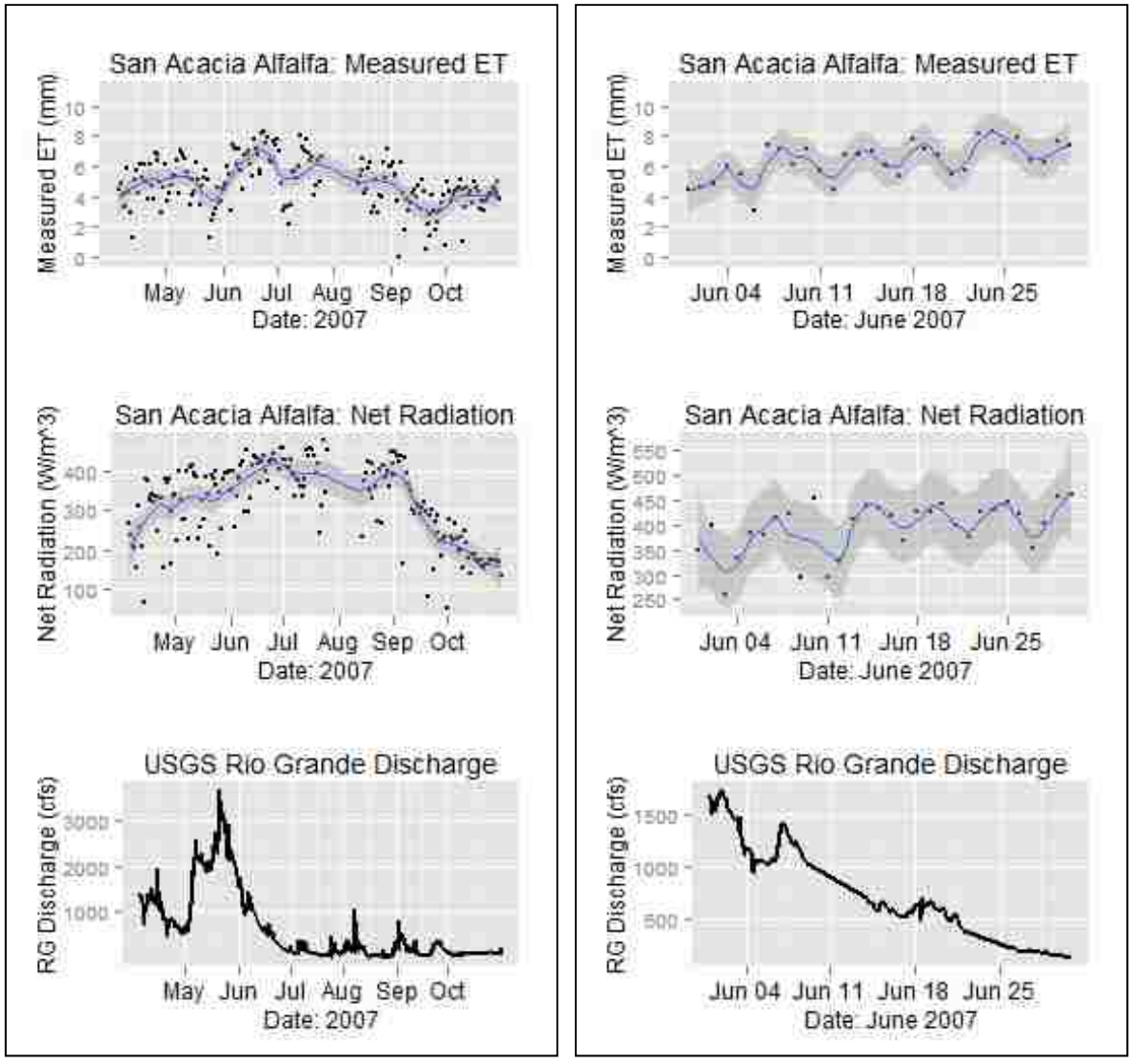


Figure 35ab: ALF :: Measured ET, Net Radiation, RG Discharge

3.9.3 Correlation

Correlation is a valuable analysis tool for comparing relationships between different sites or variables. Rapid reconnaissance of complex relationships is possible, highlighting trends that may need further investigation. Conducting correlation in natural systems is complicated by the fact that datasets are required with a matching number of values on the X and Y axis in the proper time stamp. Using the HIS to process and store data ensures consistency in time and variable formatting. HydroDesktop allows tabular review of multiple data streams in parallel to identify overlapping periods for correlation analysis.

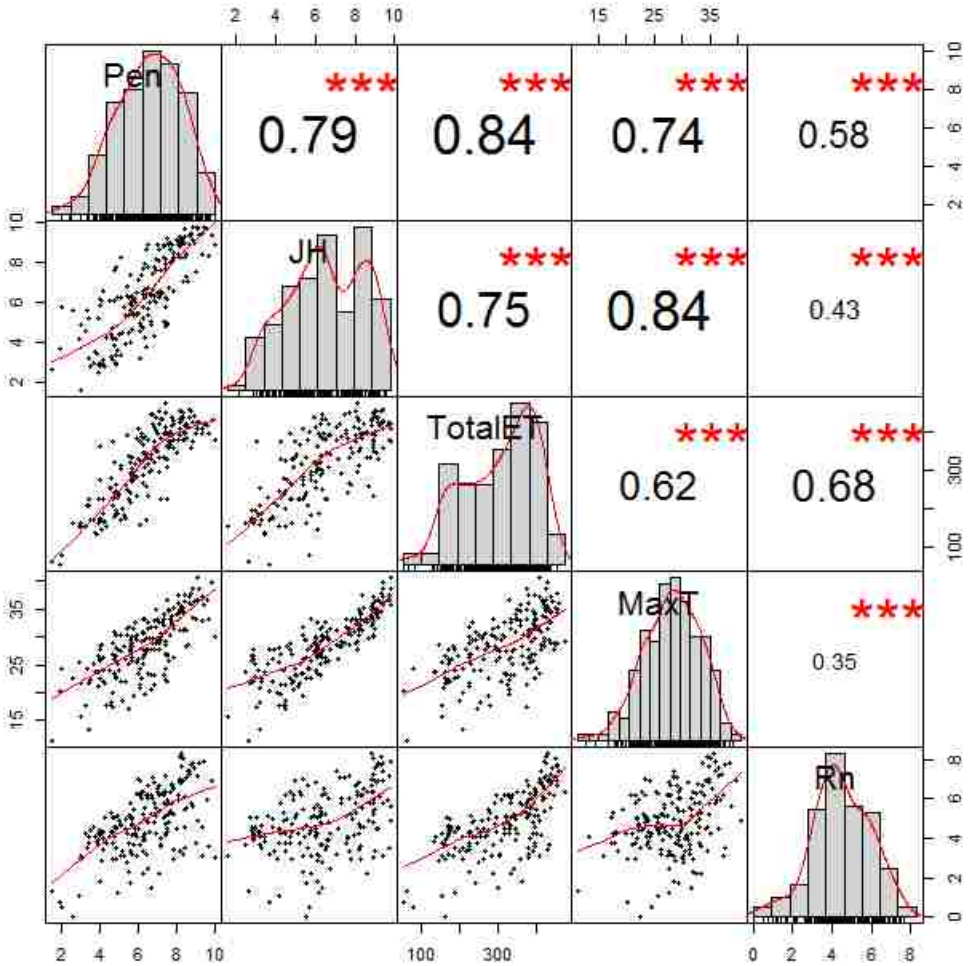


Figure 36: Alfalfa ET Growing Season :: 2007

Figure 36 and Figure 37 were generated with the open source Performance Analytics package in R (Carl, et al., 2012).

Using properly formatted data, this grid is produced from one line of code. Bottom panels show the scatter plot, top panels show the Pearson correlation value, and the middle shows the distribution histogram.

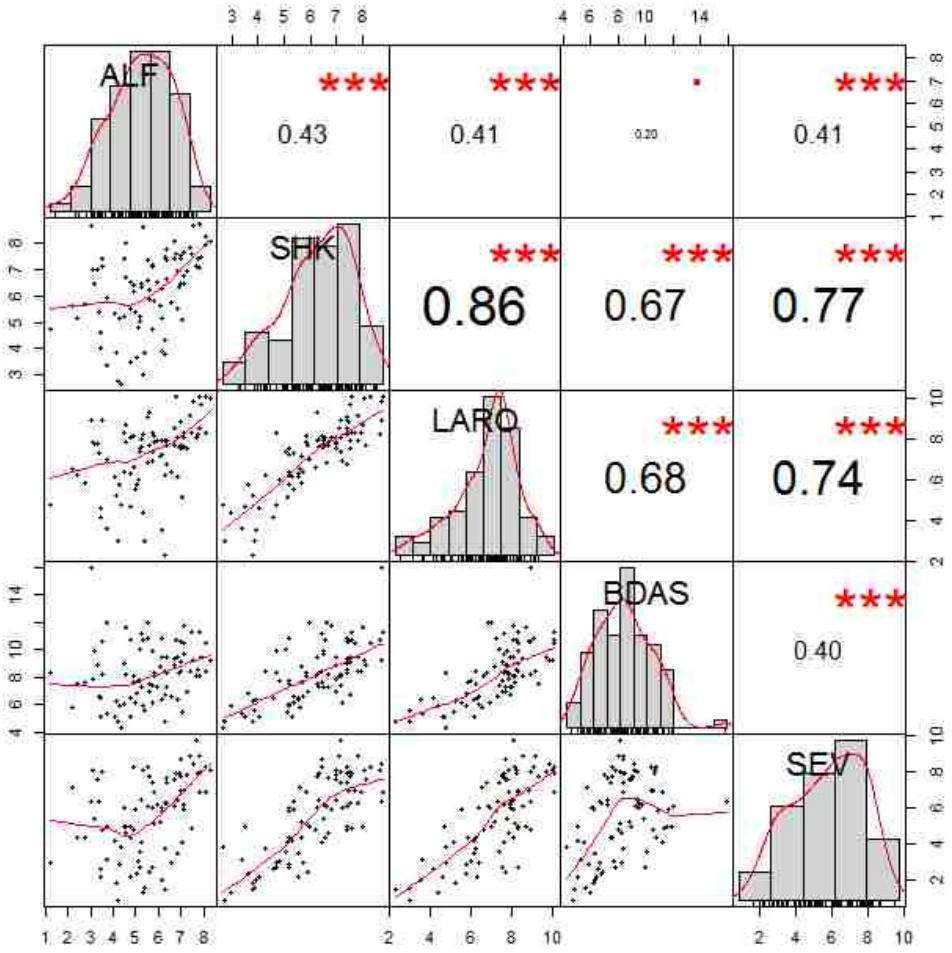


Figure 37: ET Correlation :: MJJ 2007

Five months of overlap were available for San Acacia Alfalfa after identifying and interpolating several missing values (Figure 36). Five values are correlated; Penman ET,

Jensen-Haise ET, Measured ET, Maximum Temperature, and Net Radiation. When comparing the entire growing season, a high correlation between Penmen (Pen) and Measured ET (TotalET) is evident. Correlations of five ET towers for the months of May, June, and July are shown in Figure 37. Although a limited sample, it is clear that ET fluctuates in space, time, and by species. Native Cottonwood (SHK) shows a high ET correlation with invasive species, Russian Olive (LARO) and Saltcedar (BDAS, SEV) but not Alfalfa (ALF).

At the present time there are gaps in the ET dataset preventing long-term correlation of variables and stations. Several algorithms are available for filling gaps in eddy covariance tower data (Dafeng, et al., 2004, Hui, et al., 2004, Andrew and David, 2007, Antje, et al., 2007) similar to the Rio Grande ET stations. Adjustment could be integrated into a workflow plan that would allow long term correlations.

3.9.4 New views

Correlation provides a deeper insight to complex systems but gaps in the correlation analysis may result in misleading Pearson coefficients. Traditional correlation methods must be augmented with emerging technologies for conducting trend visualization and analysis. Figure 38 shows Measured ET with the corresponding Maximum Temperature and Net Radiation as varying marker sizes. Showing a third variable as a marker size allows visual correlation of data. The larger marker sizes in Figure 38b showing net radiation at high measured ET values support the earlier findings of radiation consistently playing a significant role in ET.

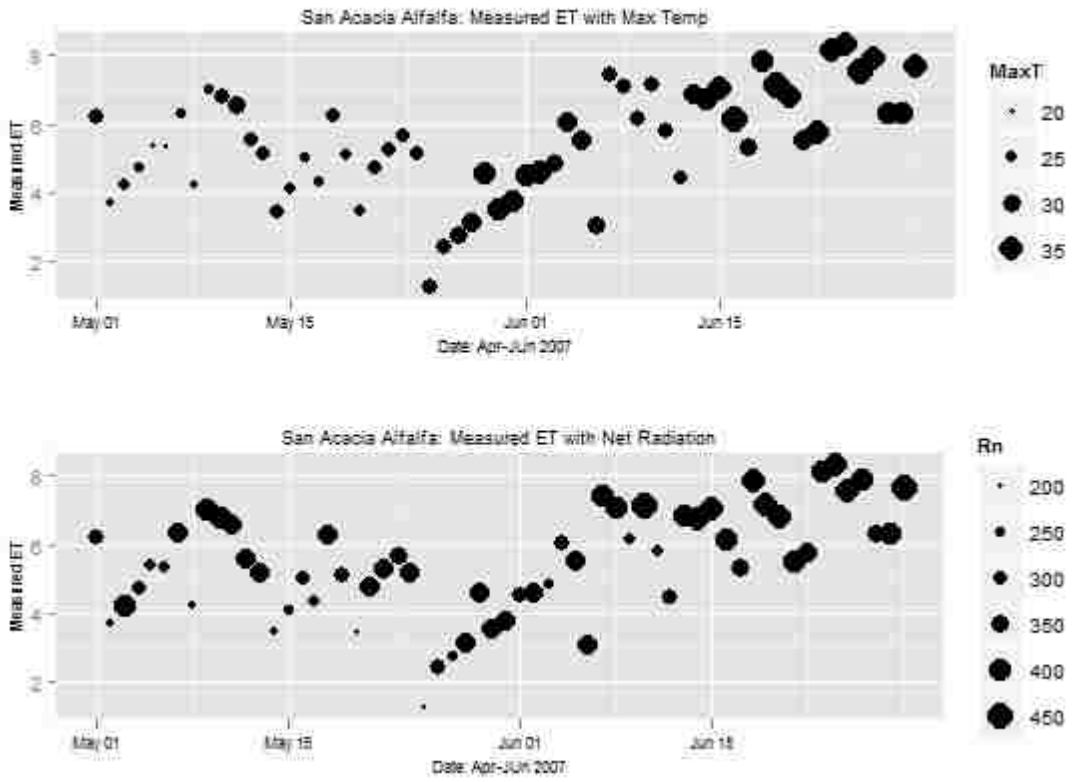


Figure 38ab: ALF :: ET varying by Max Temperature and Net

3.9.5 Automated workflows

When incorporated into a data workflow plan, plots from R can be automatically emailed to a research team or added to a project website at defined intervals. Web based queries may produce plots on demand directly from the HydroServer database.

CHAPTER 4: DISCUSSION

4.1 HydroServer

HydroServer provides a solid foundation for standardization and storage of hydrologic data. Ingest of raw data followed by versioned QA/QC data allows for full provenance to be

maintained for future scientists to make full use of field measurements. Developing a data ingest workflow at the initiation of a research project allows the opportunity to store raw data for future discovery and verification of processing methods. Server based tools allow for visual and statistical quality assurance and quality control.

Much of the basic functionality of HydroServer may be unleashed with a small learning curve from a researcher. Initial Windows Server setup, configuration of SQL Server, and management of the Internet Information Server are well documented but require time and patience to implement properly. Windows operating system experience is beneficial. Loading data into the database is likewise well documented although care must be taken to provide useful project metadata and ensure files are properly formatted for ingest.

The learning curve to fully exploit the workflow possibilities would best be coordinated with a local data repository team. The real power of HydroServer is revealed when viewing and analyzing data. Developing a data capture and processing workflow at the beginning of the project may increase the early work but dividends will be paid for the lifetime of the project with workflow and analysis efficiencies. Storing project data in a standardized format provides the opportunity to create application plug-ins to query the database. Several tools are currently available to download data in WaterML format including Excel, Matlab, HEC-DSSVue, and HydroDesktop.

One HydroServer may host multiple databases reducing cost for the organization collecting data. Temporal datasets from external sources can easily be loaded into a new database and served to a project team. The project team can begin collecting a wide assortment of data from many sources and integrate the measurements into current research. Multiple principals

can access the data from any location, critically important, as collaboration frequently crosses continents. Everyone on the research will be working with the same dataset, ensuring project consistency.

Database portability between HydroServers ensures data availability in the event of a server hardware or software failure. A backup database can be migrated to a new physical or virtual server and returned to service in a matter of minutes. Storing project data on one desktop system in a research office presents a chance of losing the data permanently in the event of a fire or flood. Keeping measurements in an organized data center with dedicated information technology professionals managing the servers ensures proper steps are taken to preserve data integrity.

4.2 HydroDesktop:

HydroDesktop is a valuable tool for data discovery and visualization. Integration of spatial searching with online basemap and local shapefiles provides powerful methods for identifying national and regional hydrologic resources. Exploring a large spatial and temporal extent of national and regional datasets is trivial. The initial search provides detailed metadata on the discovered stations allowing additional SQL style queries to fine tune the data download.

Graphing system needs improvement to produce publication quality graphs. As of this writing, the HydroDesktop working group is updating the graphing plug-in to address some of the limitations. A vector illustration package is necessary to properly customize axes and legends. Highly customizable graphical output is possible using HydroR. For those unfamiliar with R, getting HydroR running may be frustrating. Several dependencies, or extra packages, are required by HydroR. From the HydroR console the necessary packages can

easily be installed. Required packages are *DBI*, *RSQLite*, and *tcltk*. *ggplot2* was used for creating more complicated plots in addition to a custom function, *multiplot*, from the Cookbook for R (Chang, 2012).

Once scripts are developed in R it is trivial to recycle them for other visualization. Further programming in R would allow detailed statistical analysis. Data may also be exported to Matlab or Excel to reduce researcher learning curve. When developing a data workflow plan for final deliverables, HydroDesktop can assist with data discovery and fine tuning the R scripts to add to the automated workflow.

The CUAHSI team has written conduits in R to connect to local HydroDesktop databases and remote ODM SQL databases. There are times when discovering data in HydroDesktop is necessary, resulting in a dynamic exploration and analysis workflow. Linking directly to the HydroServer ODM in an automated workflow, delivering daily, weekly, or monthly reports and plots may provide a streamlined method to follow trends in a streaming data feed when HydroDesktop is unnecessary.

CHAPTER 5: CONCLUSION

Understanding complex hydrologic systems is easier when storing data in an organized, standardized database system. Standardization increases the possibility that external visualization and modeling software will accept the data without manipulation. These standardized datasets will be available for visualization and analysis tools yet to be discovered. Data discovery is streamlined by use of spatial and temporal searches of local,

regional, and national datasets. The temporal range can be easily adjusted to identify periods with anomalies requiring additional inspection or introduction of new variables.

Moving the Rio Grande ET dataset to the HIS provided the opportunity to conduct data reconnaissance in time and space that has not been possible. Ten years of ET tower and ground water measurements are now queryable across all stations and variables within the project. Loading ground water sites into HydroDesktop provided rapid visual analysis of flood recurrence intervals and seasonal depth fluctuation for multiple ecosystems. Using the CUAHSI-HIS provided a standardized platform where ground water data are now easily visualized with ET tower variables for any time range. Data standardization in WaterML also allowed efficient integration with national datasets like USGS Rio Grande discharge measurements.

One powerful tool made available by standardization in the CUAHSI-HIS is integration with R through HydroR. R allows consistently reproducible graphs with varying time ranges for visual comparison. Using the table viewer in HydroDesktop to locate overlapping data sequences, advanced correlation analyses are simplified. Multiple variables, across many stations, can now be expressed with a quantitative correlation factor amplifying previously hidden relationships.

Working with a data repository to develop a data management and workflow plan simplifies satisfying the NSF data management plan requirements and enables automated processing and reporting on incoming data streams. Using the CUAHSI-HIS in the data repository, provides a stable environment for housing data that are managed by IT professionals, ensuring reliable data backup and management resulting in rapid restoration in the event of a server failure.

Raw data can be streamed directly into the HIS for providing the first step in long term archiving followed by QA/QC scripts processing a new stream back into the HIS for verification with ODM Tools by the researcher. Standard statistical analysis and plots can be generated at the repository and placed into dynamic HTML reports. Project teams spread over large geographic regions are able to use raw and QA/QC data as soon as they are ingested into HydroServer.

Deployment of a CUAHSI-HIS provides a stable, standardized platform for storage and distribution of hydrologic information. Data, ingest, QA/QC, and visualization of local project data is streamlined with tools specially developed by CUAHSI. New research discoveries unavailable with standard data management methods are now attainable with the open source CUAHSI-HIS platform.

APPENDIX A: DETAILED SERVER CONFIGURATION

The virtual server is equivalent to an Intel XEON 2.53GHz quad core processor, 4GB of RAM, and 60GB of hard drive storage. These specifications exceed the minimum recommendations in the HydroServer System Specifications with the exception of the hard drive capacity (Valentine, 2012). A production HydroServer is recommended to contain at least 500GB of storage space. As a test bed/sandbox system this configuration was adequate.

Required commercial software includes Microsoft Windows Server 2008 R2 and Microsoft SQL Server 2008 R2 Standard Edition. These licenses were obtained through the UNM's ULA with Microsoft. ArcGIS Server is installed on the system to facilitate dynamic map integration with the HIS but this service was not implemented. Microsoft Visual Studio was installed per setup instructions.

All CUAHSI HydroServer products are open source and free of charge. The latest versions have been obtained from the HydroServer Codeplex site, http://hydroserver.codeplex.com.

Installed HydroServer Components (CUAHSI, 2012):

> ***Observations Data Model*** *- A relational schema for storing point hydrologic observations in a relational database managmenet system.*
>
> ***ODM Tools*** *- A software application for querying, visualizing, and editing data stored in an ODM database.*
>
> ***ODM Data Loader*** *- A software application for loading data from CSV or Excel files into an ODM database.*

ODM Streaming Data Loader *- A software application for automating the loading of streaming sensor data into an ODM database.*

WaterOneFlow Web Services *- A web application for publishing the contents of an ODM database on the Internet in WaterML format.*

HydroServer Capabilities *- A database, configuration tool, and web service for publishing the capabilities of a HydroServer on the Internet in a machine readable format.*

HydroServer Website *- A public website for publishing the capabilities of a HydroServer.*

Time Series Analyst *- A web application that provides data visualization, summary, and download for observational data stored in ODM databases on a HydroServer.*

HydroServer Map *- A dynamic web map application for presenting both spatial (GIS) datasets and observational data for a reasearch watershed or region for which data have been published.*

Installed Utilities:

Notepad ++ - Excellent notepad viewer for reviewing SDL logs and editing configuration files (Ho, 2011).

RStudio – Open source IDE for writing R scripts (RStudio, 2012).

APPENDIX B: METHODS, TIPS, AND R SCRIPTS

Provided I had more programming experience, much of the data processing methods presented here would be executed in Python. As I am not a programmer, the familiar hammer with which most problems became nails, was Excel. Looking back, I recommend taking a couple weeks and learning Python now if you are not already familiar with the program. Pythonxy (http://code.google.com/p/pythonxy/) is a well packaged scientific distribution for WinTel systems. Spyder (http://code.google.com/p/spyderlib/) is a valuable cross platform Python IDE.

EXCEL TIPS:

These tips came from many sources and some are aggregations or customization of discussions board topics. http://www.excelforum.com was a regular source of valuable information.

Editing multiple sheets at one time:

When multiple sheets in an Excel workbook all need the same change (Ex. adding a row or inserting a function into a column) shift-click on all the sheet tabs to select them and perform the change on one sheet. All the sheets will reflect the change. Applying the changes to the longest sheet will ensure any column copies will encompass the shorter sheets too.

Changing Julian/Ordinal time to CUAHSI time:

All in one date time fix for Raw Data, starting with three columns, Julian day, year, second. Make sure to change to column format to mm/dd/yyyy hh:mm:ss and double check that leap years transferred properly:

=DATE(INT(VALUE(TEXT(B2,"0000")&TEXT(A2,"000"))/1000),1,MOD(VALUE(TEXT(B2,"0000")&TEXT(A2,"000")),1000)) + (MID(TEXT(C2,"0000"),1,2) & ":" & MID(TEXT(C2, "0000"),3,2))

Day	YR	Time	LocalDateTime
1	2006	30	01/01/2006 00:30:00
1	2006	100	01/01/2006 01:00:00

All in one date time fix for ET data, starting with two columns, Julian day, year. Make sure to change to column format to mm/dd/yyyy hh:mm:ss and double check that leap years transferred properly:

=DATE(INT(VALUE(TEXT(B2,"0000")&TEXT(A2,"000"))/1000),1,MOD(VALUE(TEXT(B2,"0000")&TEXT(A2,"000")),1000))

Day	YR	LocalDateTime
60	2006	03/01/2006 00:00:00
61	2006	03/02/2006 00:00:00

Multiply multiple selected cells by a value:

Enter the multiplier in a cell

Copy that cell to the clipboard

Select the range you want to multiply by the multiplier

(Excel 2003 or earlier) Choose Edit | Paste Special | Multiply

(Excel 2007 or later) Click on the Paste down arrow | Paste Special | Multiply

---

Export multiple Excel sheets at CSV files:

Option Explicit

Sub ExportAllSheetsAsCSV()

Dim newWks As Worksheet

```
Dim wks As Worksheet

For Each wks In ActiveWorkbook.Worksheets

wks.Copy 'to a new workbook

Set newWks = ActiveSheet

With newWks

.SaveAs Filename:="C:\temp\" & "2006." & wks.Name & ".csv", FileFormat:=xlCSV

.Parent.Close savechanges:=False

End With

Next wks

MsgBox "Done with: " & ActiveWorkbook.Name

End Sub
```

R SCRIPTS:

These are general scripts that need to be customized for each SQL database of interest.

Correlate five ET tower locations:

```
# ET Graphing Correlation.

library(HydroR)
library(PerformanceAnalytics)

# Change date range

inputStartDate <- "2007-05-01"
inputEndDate <- "2007-07-25"

# Data connections

data1 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ET.v4.sqlite",
                       seriesID=1,
                       SQLite=TRUE,
                       startDate= inputStartDate,
```

```
                        endDate= inputEndDate)
data2 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ET.v4.sqlite",
                        seriesID=2,
                        SQLite=TRUE,
                        startDate= inputStartDate,
                        endDate= inputEndDate)
data3 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ET.v4.sqlite",
                        seriesID=3,
                        SQLite=TRUE,
                        startDate= inputStartDate,
                        endDate= inputEndDate)
data5 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ET.v4.sqlite",
                        seriesID=5,
                        SQLite=TRUE,
                        startDate= inputStartDate,
                        endDate= inputEndDate)
data6 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ET.v4.sqlite",
                        seriesID=6,
                        SQLite=TRUE,
                        startDate= inputStartDate,
                        endDate= inputEndDate)

# count each series to verify same length, does not verify if dates match

d0 <- nrow(data1$DataValues)
d1 <- nrow(data2$DataValues)
d2 <- nrow(data3$DataValues)
d3 <- nrow(data5$DataValues)
d4 <- nrow(data6$DataValues)

if ((d1+d2+d3+d5+d6)/5 != d1)  stop("Stream lengths don't match, cannot run
correlation")

# 1 = ALF, 2 = SHK,  3 = BDAS,  5 = SEV,  6 = LARO

# define variables

ALF <- data1$DataValues$DataValue
SHK <- data2$DataValues$DataValue
BDAS <- data3$DataValues$DataValue
SEV <- data5$DataValues$DataValue
LARO <- data6$DataValues$DataValue

# Create time series

ALFDateTime <- data1$DataValues$LocalDateTime
SHKDateTime <- data2$DataValues$LocalDateTime
BDASDateTime <- data3$DataValues$LocalDateTime
SEVDateTime <- data5$DataValues$LocalDateTime
LARODateTime <- data6$DataValues$LocalDateTime

# Create matrix of values

ET <- cbind(ALF, SHK, LARO, BDAS, SEV)

# Use PerformanceAnalytics library to plot correlation

chart.Correlation(ET[,1:5], histogram=TRUE, pch=20)
```

Graphing five variables horizontally:

```
# ET Graphing

library(HydroR)
library(ggplot2)

# input dage range and axis labels

inputStartDate <- "2007-05-01"
inputEndDate <- "2007-11-01"
strXaxisLabel <- "2007"
strYaxisLabel <- "Measured ET"

# data connection

data0 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ET.v4.sqlite",
       seriesID=2,
       SQLite=TRUE,
       startDate= inputStartDate,
       endDate= inputEndDate)
data1 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ET.v4.sqlite",
       seriesID=3,
       SQLite=TRUE,
       startDate= inputStartDate,
       endDate= inputEndDate)
data2 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ET.v4.sqlite",
       seriesID=6,
       SQLite=TRUE,
       startDate= inputStartDate,
       endDate= inputEndDate)
data3 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ET.v4.sqlite",
       seriesID=1,
       SQLite=TRUE,
       startDate= inputStartDate,
       endDate= inputEndDate)
data4 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ET.v4.sqlite",
       seriesID=5,
       SQLite=TRUE,
       startDate= inputStartDate,
       endDate= inputEndDate)


# define variables

SHKValues <- data0$DataValues$DataValue
BDASValues <- data1$DataValues$DataValue
LAROValues <- data2$DataValues$DataValue
ALFValues <- data3$DataValues$DataValue
SEVValues <- data4$DataValues$DataValue

# Create time series

SHKDateTime <- data0$DataValues$LocalDateTime
BDASDateTime <- data1$DataValues$LocalDateTime
LARODateTime <- data2$DataValues$LocalDateTime
ALFDateTime <- data3$DataValues$LocalDateTime
SEVDateTime <- data4$DataValues$LocalDateTime

# Create data frame of date and values
```

```
SHK <- data.frame(SHKDateTime, SHKValues)
BDAS <- data.frame(BDASDateTime, BDASValues)
LARO <- data.frame(LARODateTime, LAROValues)
ALF <- data.frame(ALFDateTime, ALFValues)
SEV <- data.frame(SEVDateTime, SEVValues)

# Remove -9999 value to plot data

SHK.clean <- SHK[SHK$SHKValues !=(-9999.000),]
BDAS.clean <- BDAS[BDAS$BDASValues !=(-9999.000),]
LARO.clean <- LARO[LARO$LAROValues !=(-9999.000),]
ALF.clean <- ALF[ALF$ALFValues !=(-9999.000),]
SEV.clean <- SEV[SEV$SEVValues !=(-9999.000),]

# Plot Values

# setup grid for ggplot from:
http://wiki.stdout.org/rcookbook/Graphs/Multiple%20graphs%20on%20one%20page%20%28gg
plot2%29/

multiplot <- function(..., plotlist=NULL, cols) {
       require(grid)

       # Make a list from the ... arguments and plotlist

       plots <- c(list(...), plotlist)
       numPlots = length(plots)

       # Make the panel
       plotCols = cols           # Number of columns of plots
       plotRows = ceiling(numPlots/plotCols) #Number of rows needed, calculated from
# of cols

       # Set up the page
       grid.newpage()
       pushViewport(viewport(layout = grid.layout(plotRows,plotCols)))
       vplayout <- function(x, y)
              viewport(layout.pos.row = x, layout.pos.col = y)

       # Make each plot, in the correct location
       for (i in 1:numPlots) {
              curRow = ceiling(i/plotCols)
              curCol = (i-1) %% plotCols + 1
              print(plots[[i]], vp = vplayout(curRow, curCol ))
       }
}

# Customize plots

SHK.2 <- ggplot(SHK.clean, aes(SHKDateTime, SHKValues)) + xlab(paste(strXaxisLabel,
":: Total ET:",round(sum(BDAS.clean$BDASValues), 0), "mm")) + ylab(strYaxisLabel) +
geom_point(size = 1) + stat_smooth(span = 0.2) + ylim(0,16) +
opts(axis.text.x=theme_text(size=9)) + opts(axis.title.x=theme_text(size=9)) +
opts(axis.title.y=theme_text(size=9, angle = 90))
+opts(plot.title=theme_text(size=10)) + opts(title="non-Flooding Cottonwood")

BDAS.2 <- ggplot(BDAS.clean, aes(BDASDateTime, BDASValues)) +
xlab(paste(strXaxisLabel, ":: Total ET:",round(sum(BDAS.clean$BDASValues), 0),
"mm")) + ylab(strYaxisLabel) + geom_point(size = 1) + stat_smooth(span = 0.2) +
ylim(0,16) + opts(axis.text.x=theme_text(size=9)) +
opts(axis.title.x=theme_text(size=9)) + opts(axis.title.y=theme_text(size=9, angle
= 90)) +opts(plot.title=theme_text(size=10)) + opts(title="Flooding Saltcedar")
```

```
LARO.2 <- ggplot(LARO.clean, aes(LARODateTime, LAROValues)) +
xlab(paste(strXaxisLabel, ":: Total ET:",round(sum(LARO.clean$LAROValues), 0),
"mm")) + ylab(strYaxisLabel)     + geom_point(size = 1) + stat_smooth(span = 0.2) +
ylim(0,16) + opts(axis.text.x=theme_text(size=9)) +
opts(axis.title.x=theme_text(size=9)) + opts(axis.title.y=theme_text(size=9, angle
= 90)) +opts(plot.title=theme_text(size=10)) + opts(title="Flooding Russian Olive")

ALF.2 <- ggplot(ALF.clean, aes(ALFDateTime, ALFValues)) + xlab(paste(strXaxisLabel,
":: Total ET:",round(sum(ALF.clean$ALFValues), 0), "mm")) + ylab(strYaxisLabel) +
geom_point(size = 1) + stat_smooth(span = 0.2) + ylim(0,16) +
opts(axis.text.x=theme_text(size=9)) + opts(axis.title.x=theme_text(size=9)) +
opts(axis.title.y=theme_text(size=9, angle = 90))
+opts(plot.title=theme_text(size=10)) + opts(title="Alfalfa")

SEV.2 <- ggplot(SEV.clean, aes(SEVDateTime, SEVValues)) + xlab(paste(strXaxisLabel,
":: Total ET:",round(sum(SEV.clean$SEVValues), 0), "mm")) + ylab(strYaxisLabel) +
geom_point(size = 1) + stat_smooth(span = 0.2) + ylim(0,16) +
opts(axis.text.x=theme_text(size=9)) + opts(axis.title.x=theme_text(size=9)) +
opts(axis.title.y=theme_text(size=9, angle = 90))
+opts(plot.title=theme_text(size=10)) + opts(title="non-Flooding Saltcedar")

# output plots
multiplot(SHK.2,LARO.2,ALF.2,BDAS.2,SEV.2, cols=3)
```

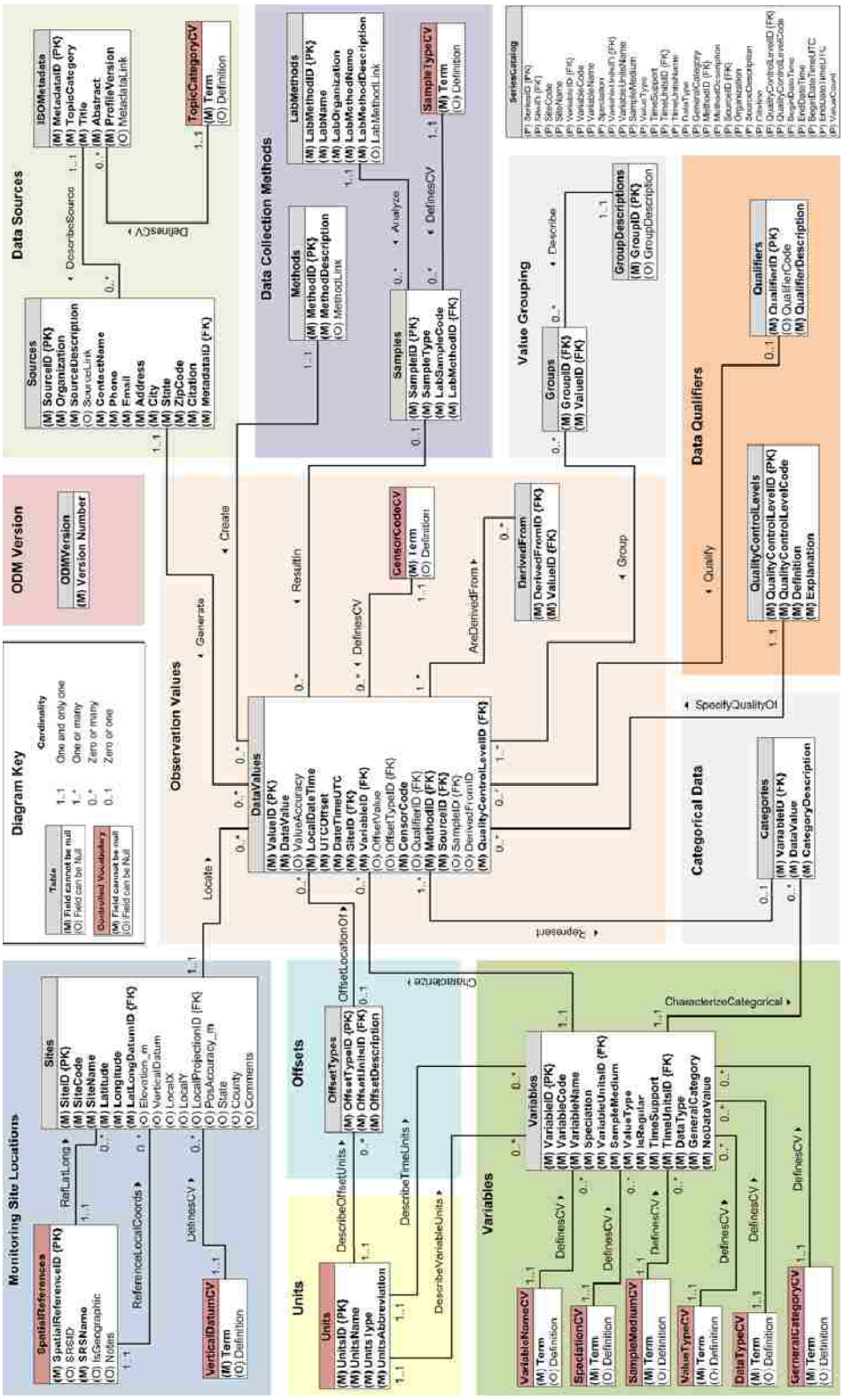
Create plots with third variable mapped as dot size:

```
# ET Graphing

library(HydroR)
library(ggplot2)

# input date range and axes labels

inputStartDate <- "2007-05-01"
inputEndDate <- "2007-06-30"
strXaxisLabel <- "Date: Apr-JUn 2007"
strYaxisLabel <- "Measured ET"

# connection strings

data0 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ALF.sqlite",
       seriesID=1,
       SQLite=TRUE,
       startDate= inputStartDate,
       endDate= inputEndDate)
data1 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ALF.sqlite",
       seriesID=2,
       SQLite=TRUE,
       startDate= inputStartDate,
       endDate= inputEndDate)
data2 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ALF.sqlite",
       seriesID=4,
       SQLite=TRUE,
       startDate= inputStartDate,
       endDate= inputEndDate)
data3 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ALF.sqlite",
       seriesID=3,
```

```
       SQLite=TRUE,
       startDate= inputStartDate,
       endDate= inputEndDate)
data4 <- getDataSeries(connectionString="E:/CUAHSI-HIS/ET/ALF.sqlite",
       seriesID=5,
       SQLite=TRUE,
       startDate= inputStartDate,
       endDate= inputEndDate)

# 1 = Pen, 2 = JH, 4 = Total,  5 = Max Temp,  3 = Net Rad

# define variables

Pen <- data0$DataValues$DataValue
JH <- data1$DataValues$DataValue
TotalET <- data2$DataValues$DataValue
Rn <- data3$DataValues$DataValue
MaxT <- data4$DataValues$DataValue

# Create time series

DateTime <- data0$DataValues$LocalDateTime

# Create data frame of date and values

ALF <- data.frame(DateTime, Pen, JH, TotalET, Rn, MaxT)

# Plot Values

# setup grid for ggplot from:
http://wiki.stdout.org/rcookbook/Graphs/Multiple%20graphs%20on%20one%20page%20%28gg
plot2%29/

multiplot <- function(..., plotlist=NULL, cols) {
       require(grid)

       # Make a list from the ... arguments and plotlist

       plots <- c(list(...), plotlist)
       numPlots = length(plots)

       # Make the panel
       plotCols = cols            # Number of columns of plots
       plotRows = ceiling(numPlots/plotCols) #Number of rows needed, calculated from
# of cols

       # Set up the page
       grid.newpage()
       pushViewport(viewport(layout = grid.layout(plotRows,plotCols)))
       vplayout <- function(x, y)
              viewport(layout.pos.row = x, layout.pos.col = y)

       # Make each plot, in the correct location
       for (i in 1:numPlots) {
              curRow = ceiling(i/plotCols)
              curCol = (i-1) %% plotCols + 1
              print(plots[[i]], vp = vplayout(curRow, curCol ))
       }
}

# Customize plots
```

```
ALF.1 <- ggplot(ALF, aes(DateTime, TotalET)) + xlab(strXaxisLabel) + ylab("Measured
ET") + geom_point(size = 2) + stat_smooth(span = 0.2) +
opts(axis.text.x=theme_text(size=8)) + opts(axis.title.x=theme_text(size=8)) +
opts(axis.title.y=theme_text(size=8, angle = 90))
+opts(plot.title=theme_text(size=9)) + opts(title="San Acacia Alfalfa: Measured
ET")

ALF.2 <- ggplot(ALF, aes(DateTime, TotalET)) + xlab(strXaxisLabel) + ylab("Measured
ET") + geom_point(aes(size = MaxT)) + opts(axis.text.x=theme_text(size=8)) +
opts(axis.title.x=theme_text(size=8)) + opts(axis.title.y=theme_text(size=8, angle
= 90)) +opts(plot.title=theme_text(size=9)) + opts(title="San Acacia Alfalfa:
Measured ET with Max Temp")

ALF.3 <- ggplot(ALF, aes(DateTime, TotalET)) + xlab(strXaxisLabel) + ylab("Measured
ET") + geom_point(aes(size = Rn)) + opts(axis.text.x=theme_text(size=8)) +
opts(axis.title.x=theme_text(size=8)) + opts(axis.title.y=theme_text(size=8, angle
= 90)) +opts(plot.title=theme_text(size=9)) + opts(title="San Acacia Alfalfa:
Measured ET with Net Radiation")

# generate plots

multiplot(ALF.2, ALF.3, cols=1)
```

APPENDIX C: CUAHSI-HIS DATABASE SCHEMA



REFERENCES CITED

[1] IMT Strategies, I. (1999). *The Sales and Marketing Imperative: The Impact of Technology on Business Strategy*, Birkhauser.
[2] O'Brien, T. V. (1971). "Tracking Consumer Decision Making." *Journal of Marketing*, 35(1), 34-40.
[3] Baxendale, P., and Codd, E. F. (1970). "A Relational Model of Data for Large Shared Data Banks." *Communications of the ACM*, 13(6), 377-387.
[4] Abbott, M. B. (1991). *Hydroinformatics: information technology and the aquatic environment*, Avebury Technical.
[5] CSDMS (2012). "NSF data management plan - csdms." <http://csdms.colorado.edu/wiki/NSF_data_management_plan>. (03/17/2012, 2012).
[6] DataONE (2012). "Data Management Planning | DataONE." <http://www.dataone.org/data-management-planning>. (03/17/2012, 2012).
[7] Brunt, J. (2012). "How to Write a Data Management Plan for a National Science Foundation (NSF) Proposal | LTER Network Office." LTER Network Office, http://lno.lternet.edu/.
[8] Olendorf, R., Townsend, L., van Reenen, J., Barkley, D., Benedict, K., and Quinn, T. (2012). "Data Management Plans - Digital Data Management, Curation and Archiving - Research Guides at University of New Mexico." <http://libguides.unm.edu/content.php?pid=137795&sid=2543372>.
[9] NSF (2010). "ENG_DMP_Policy.pdf (application/pdf Object)." *Grant Proposal Guide*, National Science Foundation.
[10] NSF (2012). "Special Information and Supplementary Documentation." *Grant Proposal Guide Chapter II*, <http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp>. (20FEB2012, 2012).
[11] UC Curation Center (2012). "Data Management Plan Tool." <https://dmp.cdlib.org/>. (03/17/2012, 2012).
[12] Randelshofer, W. (2012). "Tree Visualization." <http://www.randelshofer.ch/treeviz/>. (4MAR2012, 2012).
[13] Beran, B., and Piasecki, M. (2008). "Availability and coverage of hydrologic data in the US geological survey National Water Information System (NWIS) and US Environmental Protection Agency Storage and Retrieval System (STORET)." *EARTH SCIENCE INFORMATICS*, 1(3-4), 119-129.
[14] Office Watch (2012). "Excel – a history of rows and columns - Office Watch." <http://office-watch.com/t/n.aspx?articleid=1408&zoneid=29>. (21FEB2012, 2012).
[15] National Research Council (1991). *Opportunities in the hydrologic sciences / Committee on Opportunities in the Hydrologic Sciences, Water Science and Technology Board, Commission on Geosciences, Environment, and Resources, National Research Council*, Washington, D.C. : National Academy Press, 1991.
[16] Dozier, J. (1992). "OPPORTUNITIES TO IMPROVE HYDROLOGIC DATA." *REVIEWS OF GEOPHYSICS*, 30(4), 315-331.
[17] Miller, F. P., Vandome, A. F., and McBrewster, J. (2009). *Moore's Law*, VDM Publishing House Ltd.
[18] CUAHSI, Inc (2012). "CUAHSI Hydrologic Information System." <http://his.cuahsi.org/system.html>. (22FEB2012, 2012).
[19] Open Geospatial Consortium, Inc. (2008). "WaterML Web Services." *OGC® Discussion Paper*, Open Geospatial Consortium, Inc.
[20] Open Geospatial Consortium (2012). "WaterML 2.0 SWG | OGC(R)." *WaterML 2.0 Standards Working Group*, <http://www.opengeospatial.org/projects/groups/waterml2.0swg>. (22FEB2012, 2012).

[21] CUAHSI (2012). "CUAHSI Hydrologic Information System - Master Controlled Vocabulary Registry." *Master Controlled Vocabulary Registry*, <http://his.cuahsi.org/mastercvreg/cv11.aspx>. (04/11/2012, 2012).
[22] Valentine, D. (2012). "HydroServer - CUAHSI Hydrologic Information System Server." *HydroServer Documentation*, <http://hydroserver.codeplex.com/documentation>. (22FEB2012, 2012).
[23] Horsburgh, J. S., Tarboton, D. G., Maidment, D. R., and Zaslavsky, I. (2008). "A relational model for environmental and water resources data."
[24] Vinoski, S. (2007). "REST Eye for the SOA Guy." *IEEE Internet Computing*, 11(1), 82-84.
[25] (2002). *NWISWeb, new site for the Nation's water data*, [Reston, Va.] : U.S. Dept. of the Interior, U.S. Geological Survey, [2002].
[26] InciWeb (2012). "Las Conchas Fire Info." *Incident Information System*, <http://www.inciweb.org/incident/2385/>. (11/01/2011, 2011).
[27] Stoof, C. R., Vervoort, R. W., Iwema, J., Den Elsen, E. v., Ferreira, A. J. D., and Ritsema, C. J. (2011). "Hydrological response of a small catchment burned by experimental fire." *Hydrology & Earth System Sciences Discussions*, 8(2), 4053-4098.
[28] Cleverly, J. R., Dahm, C. N., Coonrod, J. E., Vanderbilt, K., and Thibault, J. R. (2008). "Middle Rio Grande Flux Network." *Rio-ET lab*.
[29] Fernald, A. G., Cevik, S. Y., Ochoa, C. G., Tidwell, V. C., King, J. P., and Guldan, S. J. (2010). "River Hydrograph Retransmission Functions of Irrigated Valley Surface Water-Groundwater Interactions." *Journal of Irrigation & Drainage Engineering*, 136(12), 823-835.
[30] New Mexico Office of the State Engineer (2010). "Rio Grande Watershed Study, San Acacia Surface Water - Groundwater Investigation Data Summary." <http://www.ose.state.nm.us/newtstweb/isc_rio_grande_tech_WatershedStudy.html>. (20FEB2012, 2012).
[31] Guru, S. M., Kearney, M., Fitch, P., and Peters, C. (2009). "Challenges in using scientific workflow tools in the

hydrology domain." *18th World IMACS / MODSIM Congress*Cairns, Australia.
[32] OpenMI Association (2012). "OpenMI Compliant Software." <http://www.openmi.org/reloaded/users/compliant-software.php>. (03/17/2012, 2012).
[33] Kepler/CORE (2012). "The Kepler Project — Kepler." <https://kepler-project.org/>. (03/17/2012, 2012).
[34] LTER (2012). "Software Tools for Sensor Networks." <http://sensor-workshop.ecoinformatics.org/>. (2012).
[35] USGS Coalition (2012). "HHRG-112-AP06-WTestimony-RGropp-20120321.pdf (application/pdf Object)." US House of Representatives, http://appropriations.house.gov/.
[36] Robbins, R. (2012). "Data Management for LTER: 1980-2010." <http://www.nsf.gov/publications/pub_summ.jsp?ods_key=bio12002>. (03/17/2012, 2012).
[37] Klump, J. (2011). "Criteria for the Trustworthiness of Data Centres." *D-Lib Magazine*, 17(1/2), 8-8.
[38] Schofield, P. N., Eppig, J., Huala, E., de Angelis, M. H., Harvey, M., Davidson, D., Weaver, T., Brown, S., Smedley, D., Rosenthal, N., Schughart, K., Aidinis, V., Tocchini-Valentini, G., and Hancock, J. M. (2010). "Sustaining the Data and Bioresource Commons." *SCIENCE*, 330(6004), 592-593.
[39] KISTERS (2012). "WISKI | Water Management." <http://www.kisters.net/wiski.html>. (2012).
[40] Aquatic Informatics (2012). "AQUARIUS Server | Aquatic Informatics." <http://aquaticinformatics.com/aquarius-server>. (2012).

[41] Kadlec, J. (2010). "Kadlec_abs_17.pdf (application/pdf Object)." *Proc., AWRA 2010 SPRING SPECIALTY CONFERENCE*, AWRA.
[42] Thessen, A. E., and Patterson, D. J. (2011). "Data issues in the life sciences." *ZooKeys*, 150, 15-51.
[43] Hey, T., Tansley, S., and Tolle, K. (2009). *The Fourth Paradigm*, Microsoft Research.
[44] S.S. Papadopulos & Associates, I. (2002). *Assessment of flow conditions and seepage on the Rio Grande and adjacent channels, Isleta to San Marcial [electronic resource]*, [Santa Fe, N.M. : New Mexico Interstate Stream Commission], [2002].
[45] Martinet, M. C., Vivoni, E. R., Cleverly, J. R., Thibault, J. R., Schuetz, J. F., and Dahm, C. N. (2009). "On groundwater fluctuations, evapotranspiration, and understory removal in riparian corridors." *Water Resour. Res.*, 45(5), W05425.
[46] Dahm, C. N., Cleverly, J. R., Allred Coonrod, J. E., Thibault, J. R., McDonnell, D. E., and Gilroy, D. J. (2002). "Evapotranspiration at the land/water interface in a semi-arid drainage basin." *Freshwater Biology*, 47(4), 831-843.
[47] Wright, J. (2012). "Measures of Dispersion: Coefficient of Variation." <http://www.jimwright.org/WebEd/u02/we020304.htm>. (04/02/2012, 2012).
[48] Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*, Springer New York.
[49] R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing.
[50] Carl, P., Peterson, B., Boudt, K., and Zivot, E. (2012). "Econometric tools for performance and risk analysis." *Package 'Performance Analytics'*, r-project.org, http://cran.r-project.org/web/packages/PerformanceAnalytics/PerformanceAnalytics.pdf, 141.
[51] Dafeng, H., Shiqiang, W., Bo, S., Gabriel, K., Russell, M., and Yiqi, L. (2004). "Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations." *Agricultural and Forest Meteorology*, 121, 93-111.
[52] Hui, D., Wan, S., Su, B., Katul, G., Monson, R., and Luo, Y. (2004). "Gap-filling missing data in eddy covariance measurements using multiple imputation (MI) for annual estimations." *Agricultural & Forest Meteorology*, 121(1/2), 93.
[53] Andrew, D. R., and David, Y. H. (2007). "A method to estimate the additional uncertainty in gap-filled NEE resulting from long gaps in the CO2 flux record." *Agricultural and Forest Meteorology*, 147, 199-208.
[54] Antje, M. M., Dario, P., Markus, R., David, Y. H., Andrew, D. R., Alan, G. B., Clemens, B., Bobby, H. B., Galina, C., Ankur, R. D., Eva, F., Jeffrey, H. G., Martin, H., Dafeng, H., Andrew, J. J., Jens, K., Asko, N., and Vanessa, J. S. (2007). "Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes." *Agricultural and Forest Meteorology*, 147, 209-232.
[55] Chang, W. (2012). "Cookbook for R » Multiple graphs on one page (ggplot2)." <http://wiki.stdout.org/rcookbook/Graphs/Multiple%20graphs%20on%20one%20page%20(ggplot2)/>. (03/17/2012, 2012).
[56] CUAHSI, I. (2012). "HydroServer - CUAHSI Hydrologic Information System Server." <http://hydroserver.codeplex.com/documentation>. (03/17/2012, 2012).
[57] Ho, D. (2011). "Notepad++ Home." <http://notepad-plus-plus.org/>. (03/17/2012).
[58] RStudio, I., , (2012). "RStudio." <http://rstudio.org/>. (03/23/2012, 2012).