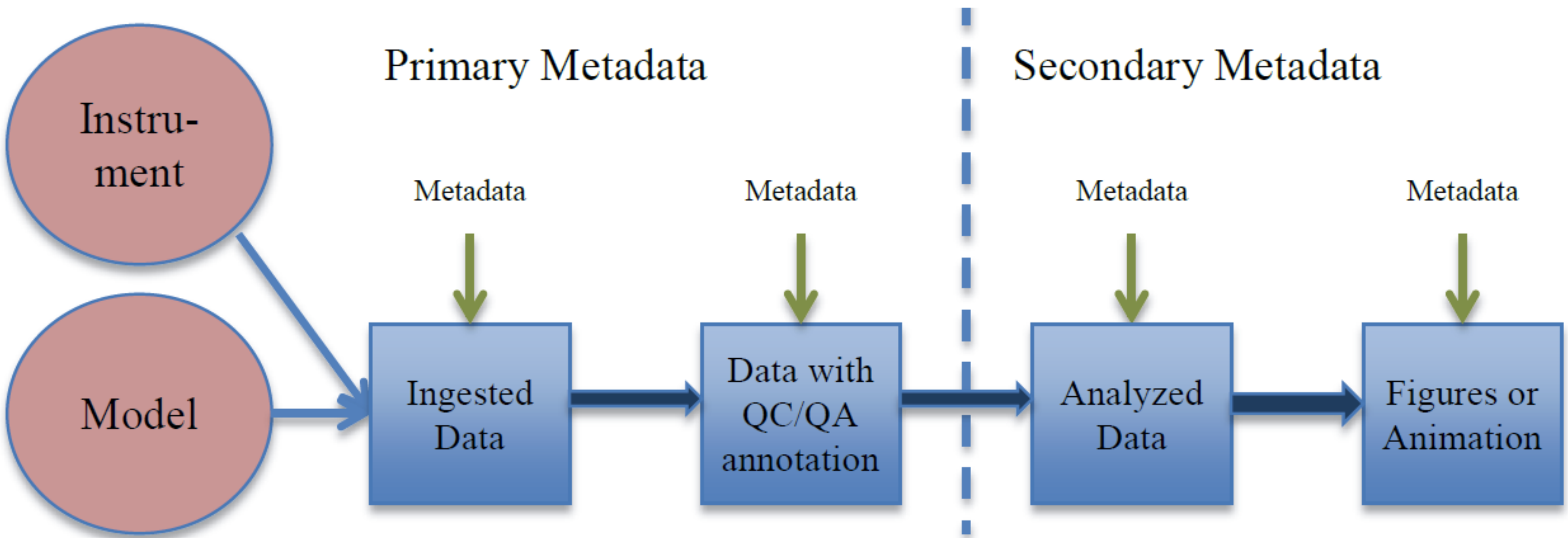
EarthCube Metadata:

Helping Modelers Figure Out What to Assimilate

John Horel
Department of Atmospheric Sciences
University of Utah
John.horel@utah.edu

Definitions of Metadata

- data about data
  - data containers: structures
  - data content: descriptive



EarthCube metadata research: addressing metadata gaps and integration opportunities in the geosciences

Definitions of Metadata

- *Context: data that allow users to evaluate and apply information for their specific applications: characteristics, value, impact, feedback from data assimilation, etc.*
- Search engines: building context from content linkages and users
- ontologies vs. folksonomies: classification designed by experts *vs.* derived from users

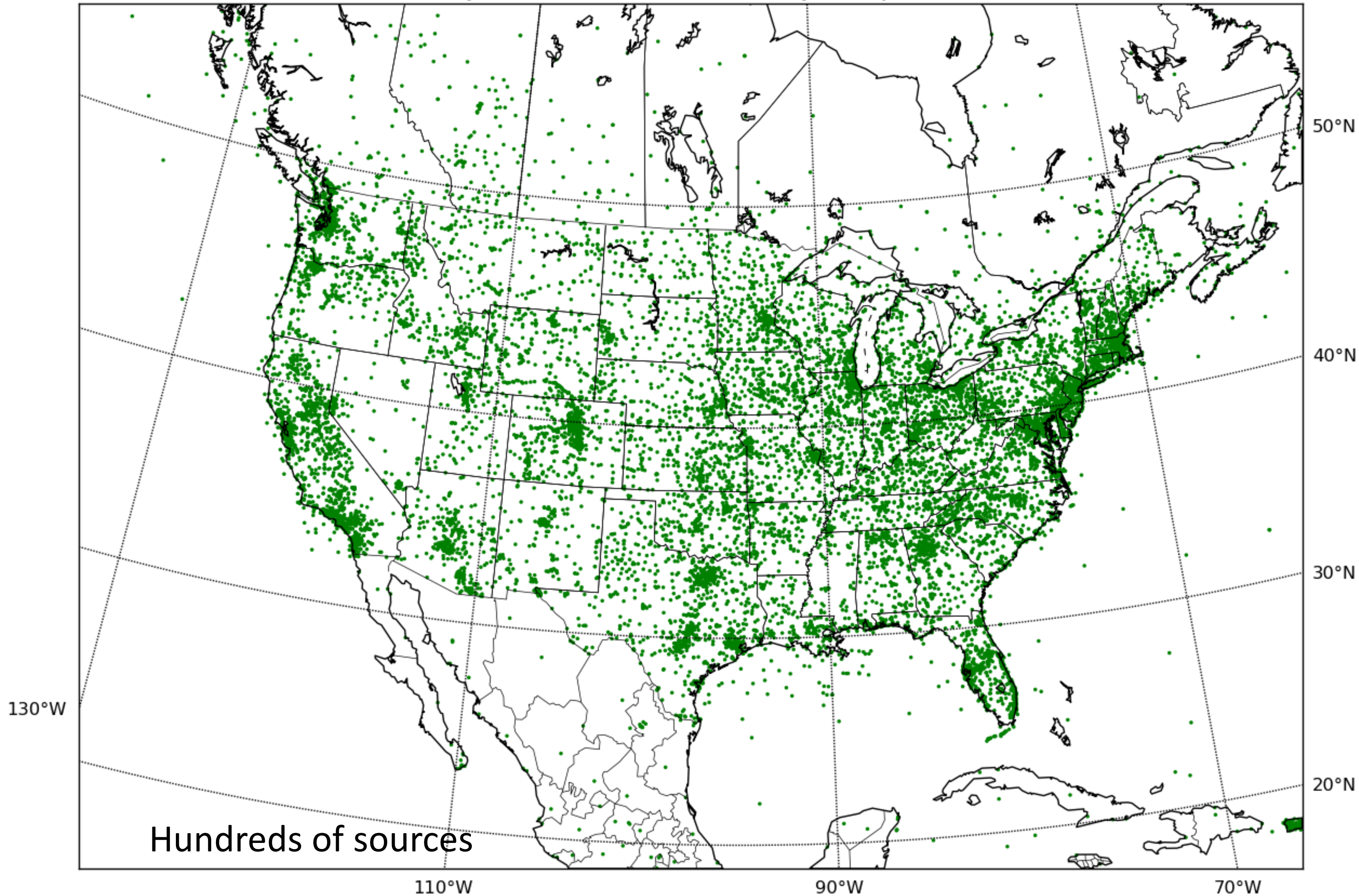
Intersecting Interests: Relevance for Computing Science via EarthCube Data Assimilation Efforts

- Computing Research for Sustainability (NAS 2012)
  - "There should be strong incentives at all stages of research for focusing on solving real problems whose solution can make a substantial contribution to sustainability challenges"
  - "Encourage research at and across disciplinary boundaries, well informed by specifics and well structured to handle scale, data, integration, architecture, simulation, optimization, iteration, and human and systems aspects.
  - CS research in sustainability should be an interdisciplinary effort, with experts in the various fields of sustainability being equal partners in the research."
  - "Undergraduate and graduate education in computer science should provide experience in working across disciplinary boundaries."

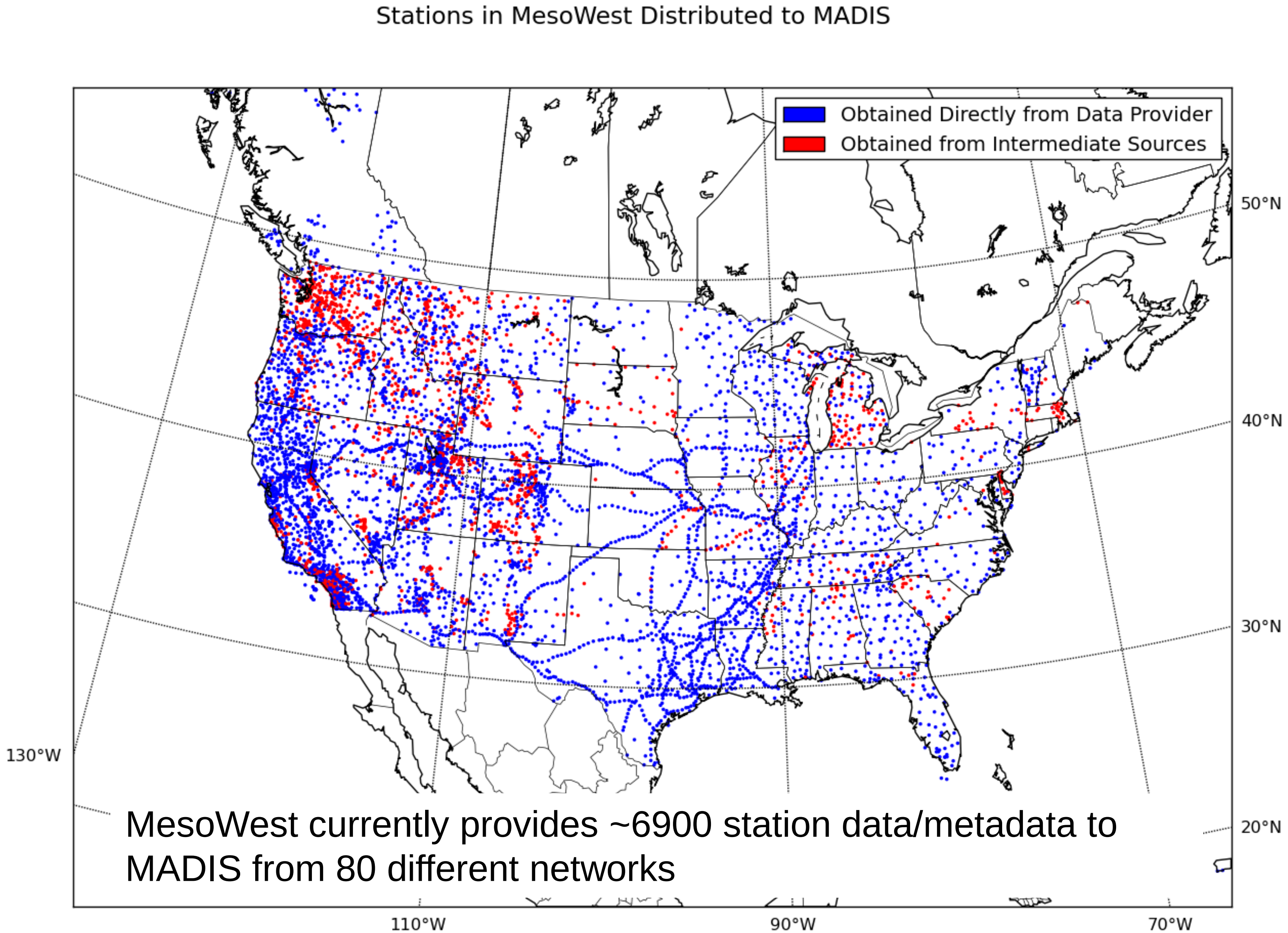
Existing Publicly-Accessible Surface Mesonet Observations

Stations in MesoWest Obtained from MADIS

Meteorological Assimilation Data Ingest System

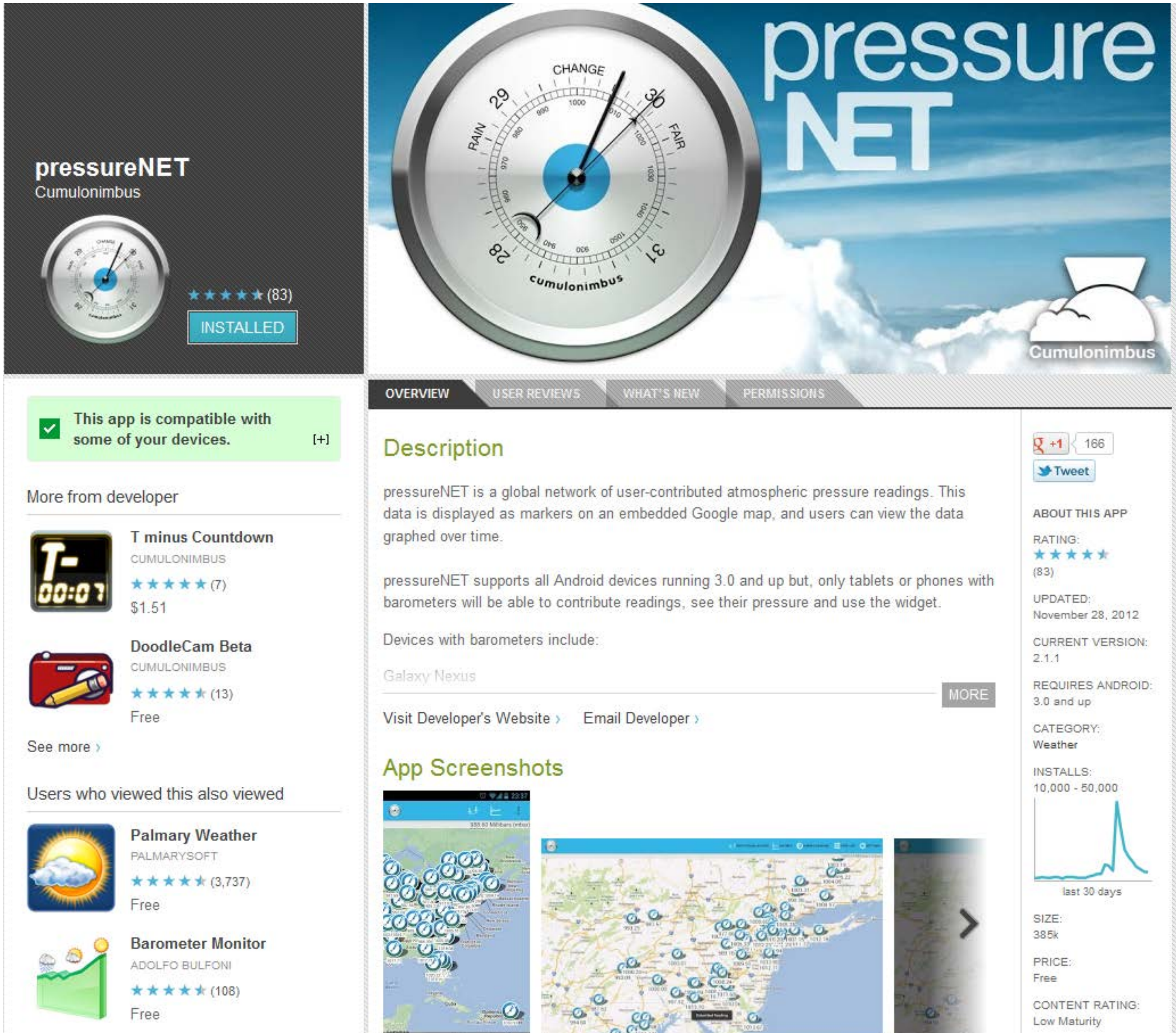


Existing Surface Mesonet Observations

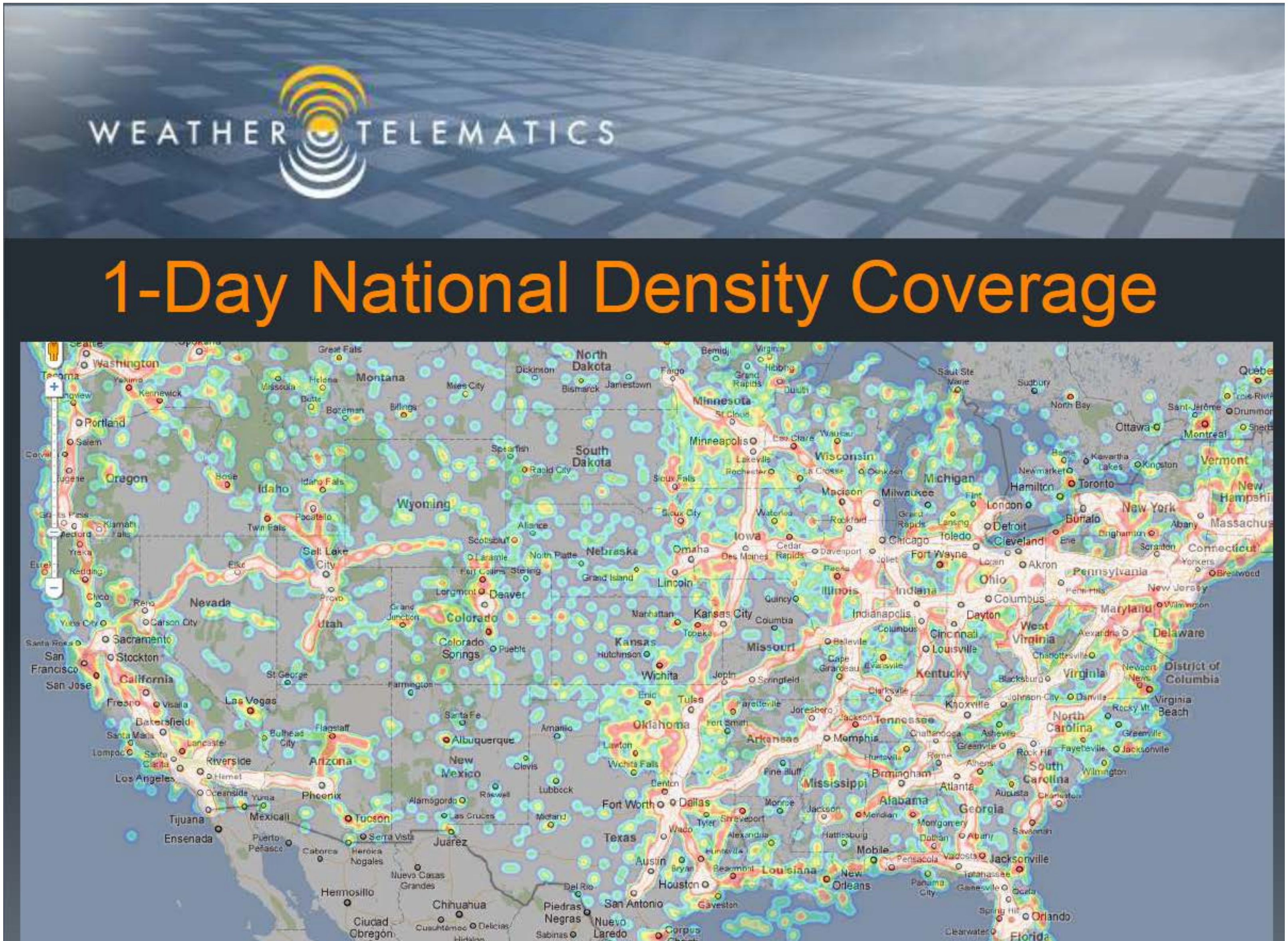


MesoWest currently provides ~6900 station data/metadata to MADIS from 80 different networks

Viral Deployments: Phone Sensors



Mobile Fleet Observations



New Research Deployment Opportunities with Inexpensive Computers, Communication Devices and Sensors

Raspberry Pi® | Quick Start | Downloads | Store | Foru

Image courtesy of Switched On Tech Design (www.sotechdesign.com.au) – thanks guys!

INTRODUCTION

What's a Raspberry Pi?

The Raspberry Pi is a credit-card sized computer that plugs into your TV and a keyboard. It's a capable little PC which can be used for many of the things that your desktop PC does, like spreadsheets, word-processing and games. It also plays high-definition video. We want to see it being used by kids all over the world to learn programming.

ARDUINO

search

Buy | Download | Getting Started | Learning | Reference | Products | FAQ | Contact Us

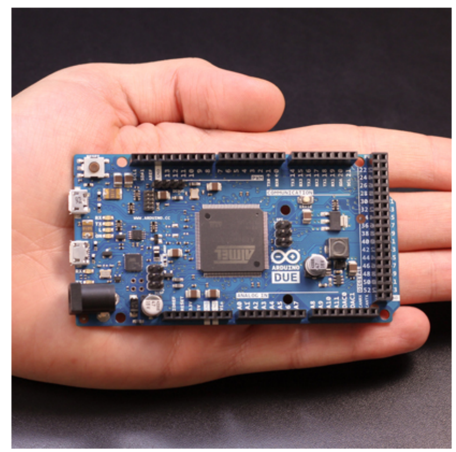


Photo by the Arduino Team

Arduino is an open-source electronics prototyping platform based on flexible, easy-to-use hardware and software. It's intended for artists, designers, hobbyists, and anyone interested in creating interactive objects or environments.

Arduino can sense the environment by receiving input from a variety of sensors and can affect its surroundings by controlling lights, motors, and other actuators. The microcontroller on the board is programmed using the Arduino programming language (based on Wiring) and the Arduino development environment (based on Processing). Arduino projects can be stand-alone or they can communicate with software running on a computer (e.g. Flash, Processing, MaxMSP).

The boards can be built by hand or purchased preassembled; the software can be downloaded for free. The hardware reference designs (CAD files) are available under an open-source license, you are free to adapt them to your needs.

Arduino received an Honorary Mention in the Digital Communities section of the 2006 Ars Electronica Prix. The Arduino team is: Massimo Banzi, David Cuartielles, Tom Igoe, Gianluca

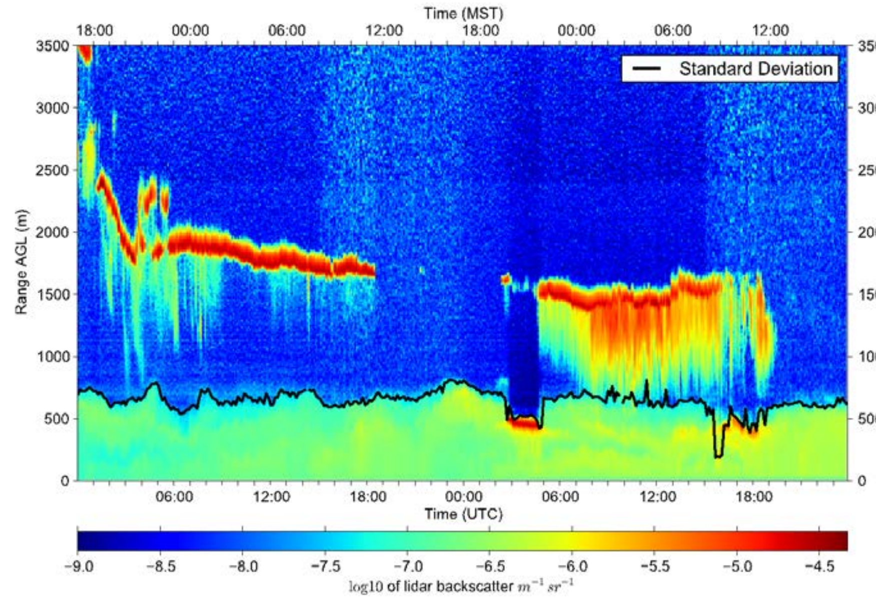
Arduino Blog (more)

66 days ago

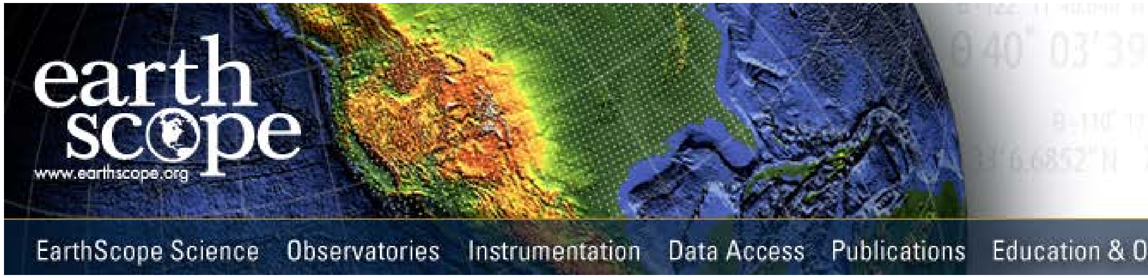
Arduino-controlled blinds: a tutorial http://t.co/KWvTGw3g

68 days ago

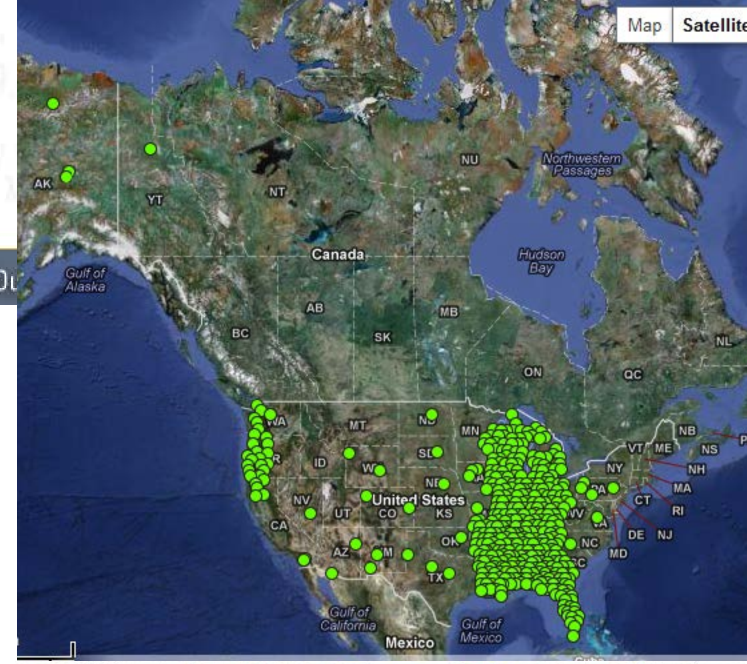
Improved Boundary Layer Assets: CASA Radars, Radiometers, Lidars, Sodars, Ceilometers, Mobile Sondes

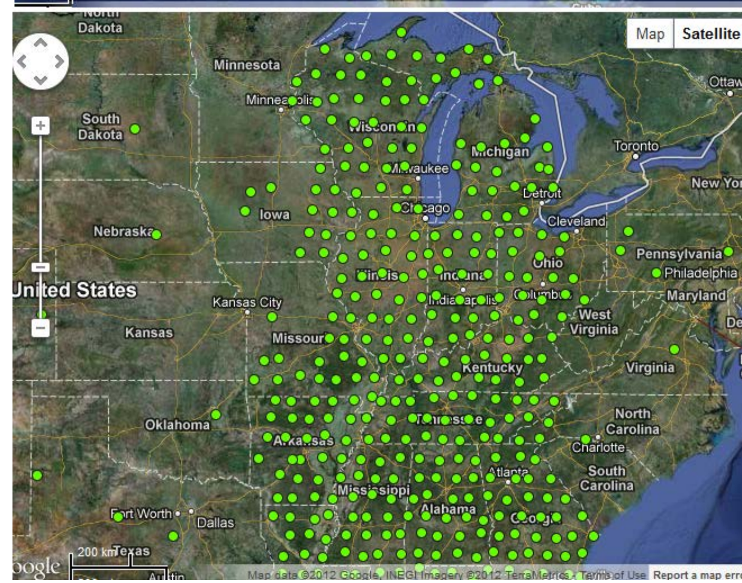


- US Array pressure data: Example of multidisciplinary earth observing system
- EarthScope: NSF sponsored geological observing network utilizing freely available data from instruments that measure motions of the Earth's surface, record seismic waves, and recover rock samples from depths at which earthquakes originate

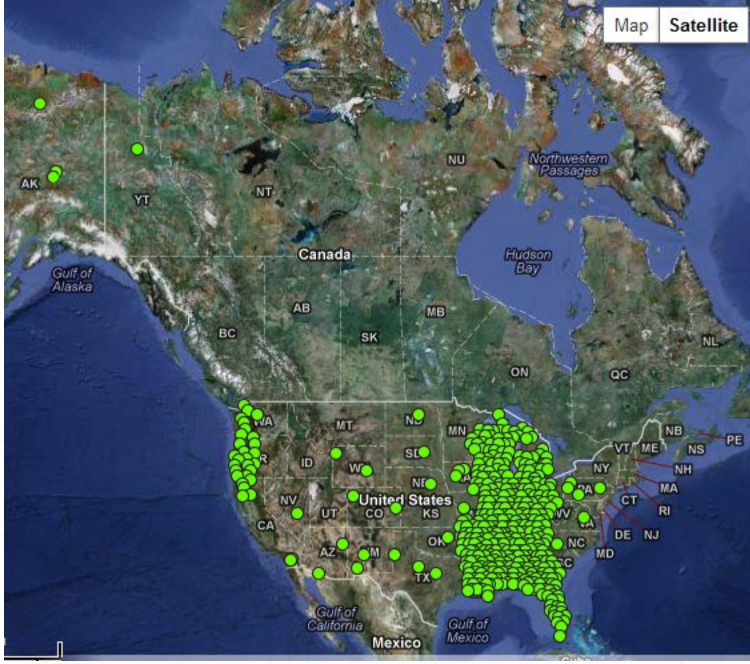


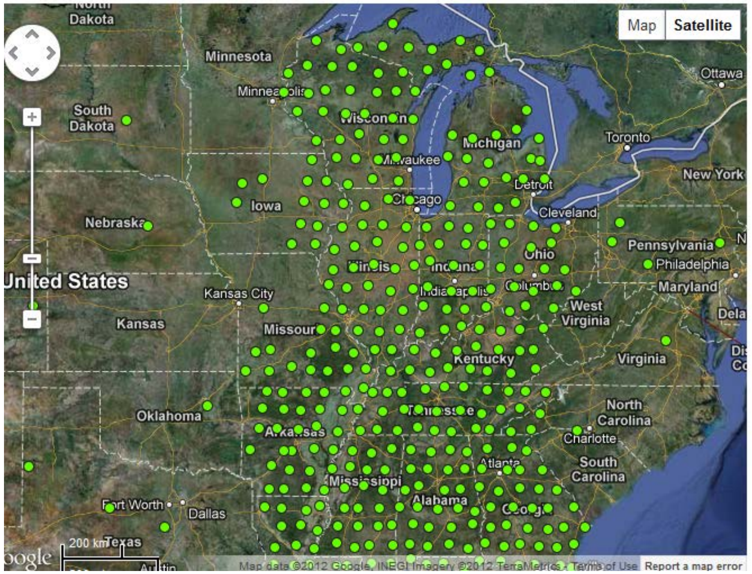
- Transportable Array: network of 400 seismographs placed at temporary sites across the United States from west to east in a regular grid pattern
- NSF ATMOS supported UCSD to deploy atmospheric pressure sensors as part of US Array beginning in mid-2010
- 1 Hz pressure data from each station transmitted to UCSD and EarthScope IRIS Data Management Center
- Transportable Array supplements Reference Network





- US Array pressure data processed at U/Utah since early 2012
- 5 minute averages sent from U/Utah to MADIS-NCEP for use in operational models
- Data and metadata stored in MesoWest relational databases: http://mesowest.utah.edu/cgi-bin/droman/stn_mnet.cgi?mnet=170
- *U/Utah goal: increase utilization of US Array pressure data in the atmospheric sciences*
  - *leverage NSF's investment*
  - *create opportunities for collaboration between atmospheric scientists and geophysicists*
  - *examine pressure fluctuations on appropriate spatials and temporal scales*





Annoyances…

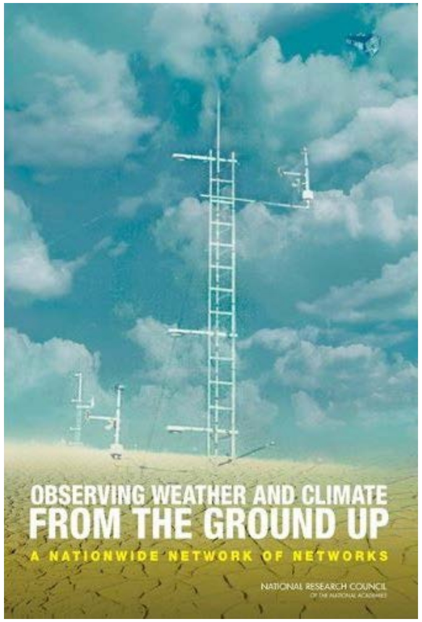
- Obtaining basic metadata can be more complicated than obtaining the associated data
- Heterogeneous nature of metadata and reporting standards
- Overcoming outdated traditions and reporting practices of data and metadata imposed by specific user communities, e.g., METAR, SHEF, BUFR, NWSLI metadata
- Unnecessary burden placed on data providers to provide metadata and data in common formats

EarthCube Issues

- How to organize and access ever increasing amounts of data for data assimilation and other applications?
- Improvements needed for data assimilation systems to effectively use surface and boundary layer observations
- What are proper roles for automated/semi-automated/manually entered data catalogs that often become rapidly out of date?
- Restrictive data usage agreements in some sectors and roles of research, government, and private sectors
- What can be done to direct the national efforts to improve boundary layer data and its metadata for widespread applications?

NWS: National Mesonet Program

- National Academy of Science (2009) recommendation to build national network of networks from existing and future mesonets
- National Mesonet Program collaborative effort to develop:
  - A national strategy to integrate disparate observing systems;
  - Increase coordination among existing surface networks, including improved quality checking, more complete metadata, increased access to observations, and broader usage of data serving multiple locally driven needs
- Legislatively directed funding to NWS to support commercial firms and educational institutions to address some of these issues
- Adoption of StarFL as alternative to Sensor Markup Language for metadata http://www.opengeospatial.org/standards/dp

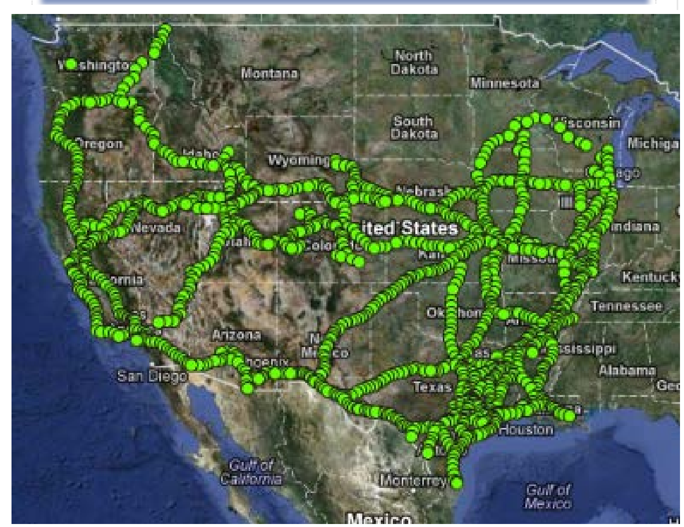


Approaches For Obtaining Metadata

- Pay data providers for data & metadata?
  - "gold standard" equipment with "gold standard" metadata paid more?
  - Identify sensors that provide more "value" based on location, weather, economic impact, usage?
- Appeal to data providers' self interest?
  - Provide information to data providers that shows the public, research, and commercial use and benefits of installed equipment
  - Enhance utility of data through quality control and dissemination to broader communities that increases utilization of equipment

Environets: An open resource for environmental monitoring networks

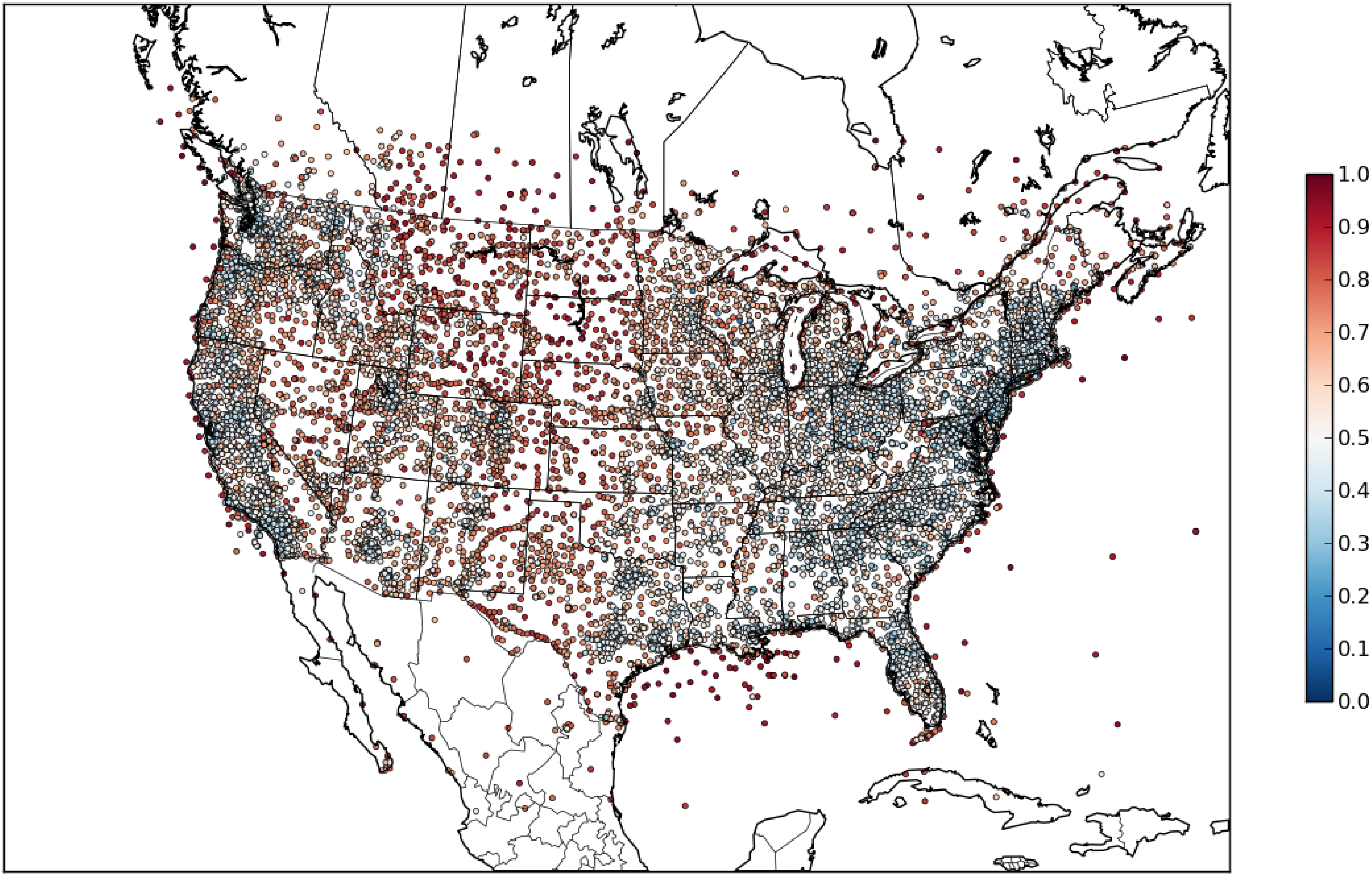
- *In development- soft rollout 12/12/12*
- Provide alternative to traditional one-way flow of metadata from station owner to users
  - Use social media to engage data users and data providers
  - What/Why data providers installed the equipment
  - Who/how/why people are using it
- Use data mining techniques to provide metrics on value of environmental information as a function of location and weather
  - Which stations are users accessing?
  - Relative bias and impact of stations and networks on high-resolution surface analyses (Tyndall and Horel, WAF 2012)



Summary

- Realistic and flexible standards, data and metadata exchange protocols, etc. that take into consideration diverse needs and communities
- Encourage broader interpretation of metadata focused on user needs to make good decisions about data beyond static nuts and bolts of sensors and equipment deployment
  - impact of observing networks via OSSE and OSE from research and operational models
  - data mining for discovery of dynamic metadata: quality control, realistic bounds and temporal/spatial changes
- New tools through CS and sensor design collaborations:
  - relieve burden for data providers to manually update metadata through metadata discovery of smart sensors
  - user data discovery akin to Google search engine rather than direct user access of data catalogs
  - middleware layers to provide access to data without need to have centralized data repositories

Median Impact of Wind Speed Observations During November 2012



Tyndall and Horel (WAF 2012)

US Array Pressure Data

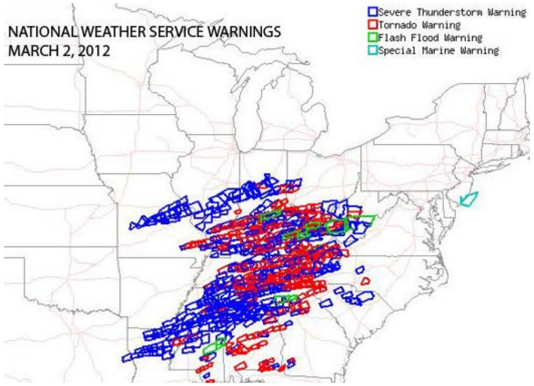


Fig. 3. Severe thunderstorm (blue) and tornado warnings (red) on 2 March 2012.

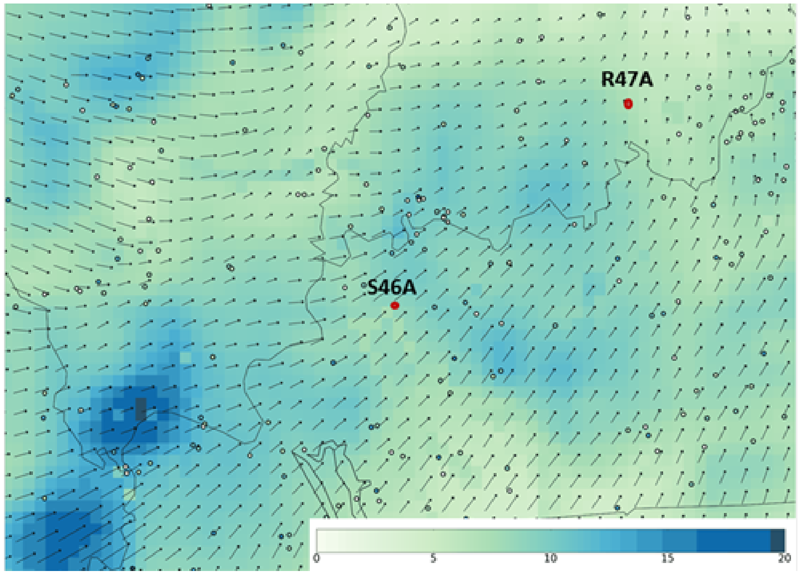


Fig. 7. Analysis of 10-m surface wind speed according to scale at the bottom (shading m s^{-1}) and wind vectors for 20 UTC 2 March 2012 from the University of Utah UU2DVAR analysis system (Tyndall and Horel 2012). Strong southwesterly winds (> 10 m s^{-1}) associated with the convective line evident in Fig. 5 lie in between the S46A and R47A US Array stations at this time. Stations archived in the MesoWest database are indicated by circles.

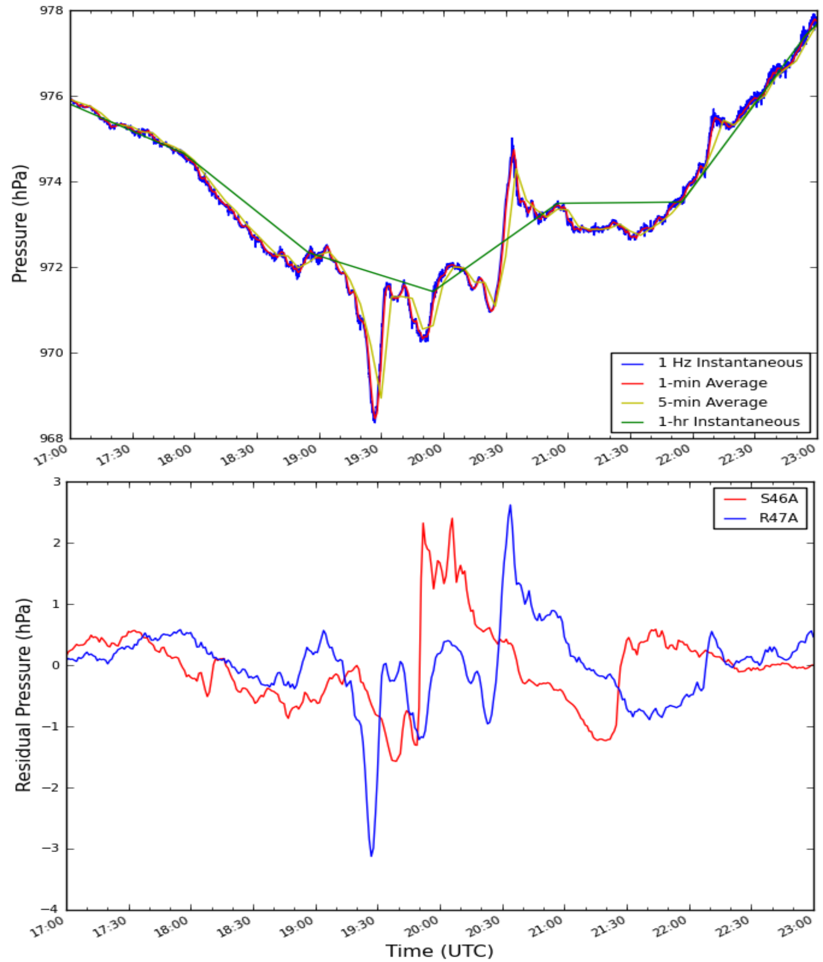
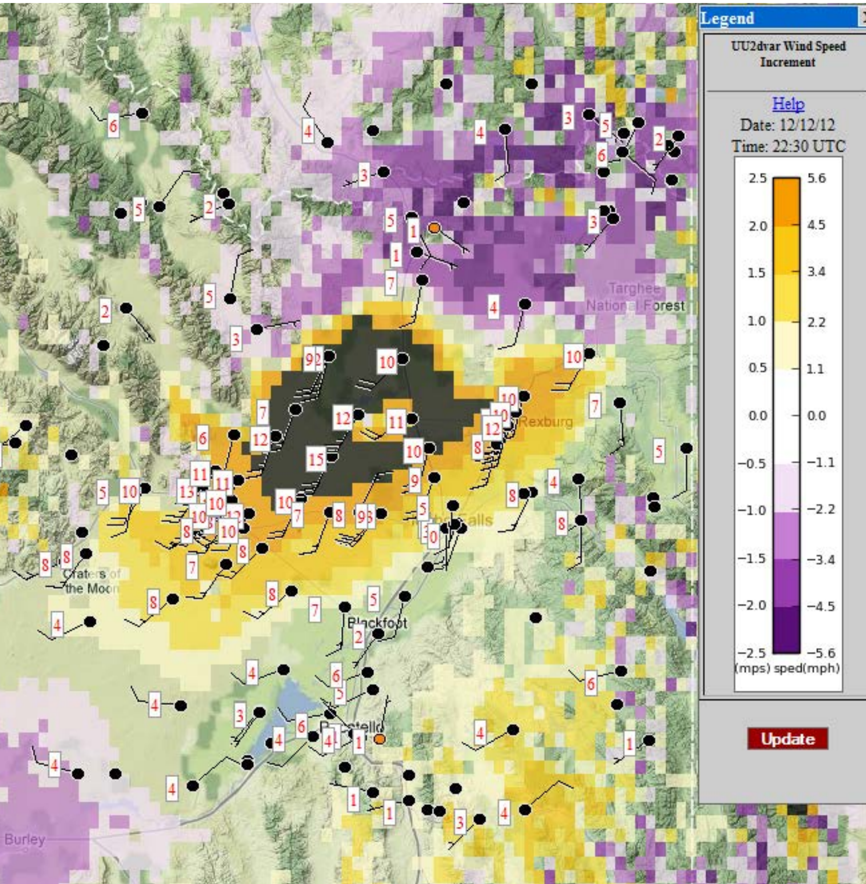
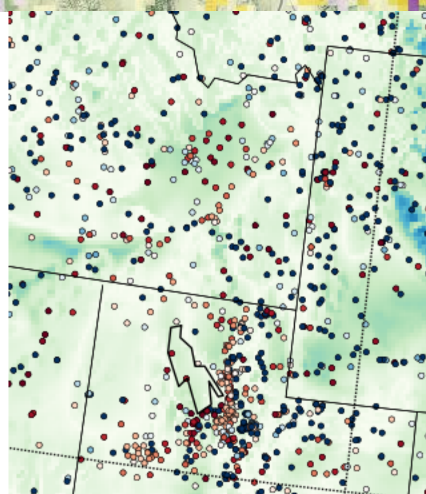
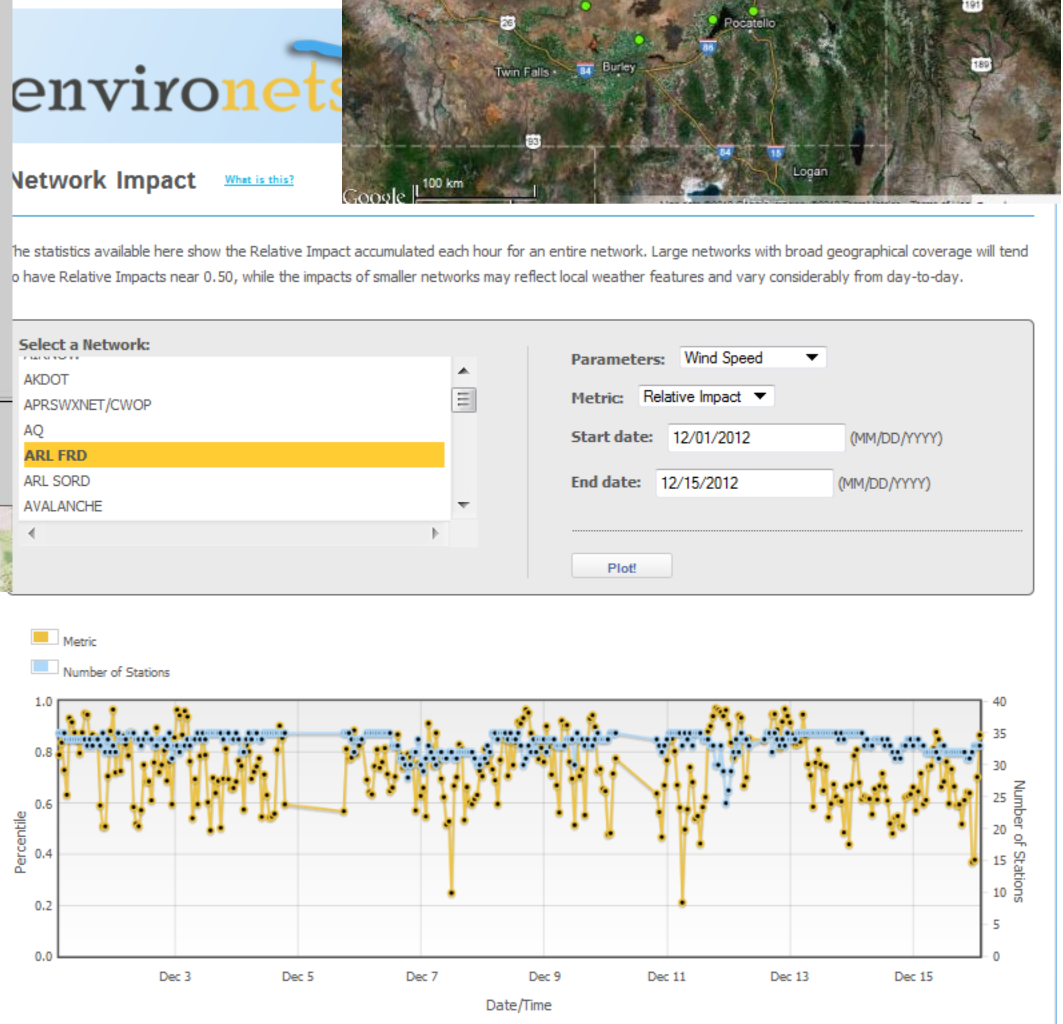


Fig. 5. Upper panel. Pressure (hPa) at R47A sampled at 1 Hz and 1 h (blue and green lines), and averaged over 1 and 5 min (red and yellow lines) during 17-23 UTC 2 March 2012. Lower panel. 1 min pressure after filtering of periods > 2 h at S46A (red) and R47A (blue).


Legend
UU2dvar Wind Speed Increment
Help
Date: 12/12/12
Time: 22:30 UTC
2.5 5.6
2.0 4.5
1.5 3.4
1.0 2.2
0.5 1.1
0.0 0.0
−0.5 −1.1
−1.0 −2.2
−1.5 −3.4
−2.0 −4.5
−2.5 −5.6
(mps) sped(mph)
Update
Targhee National Forest
Rexburg
Craters of the Moon
Blackfoot
Pocatello
Burley

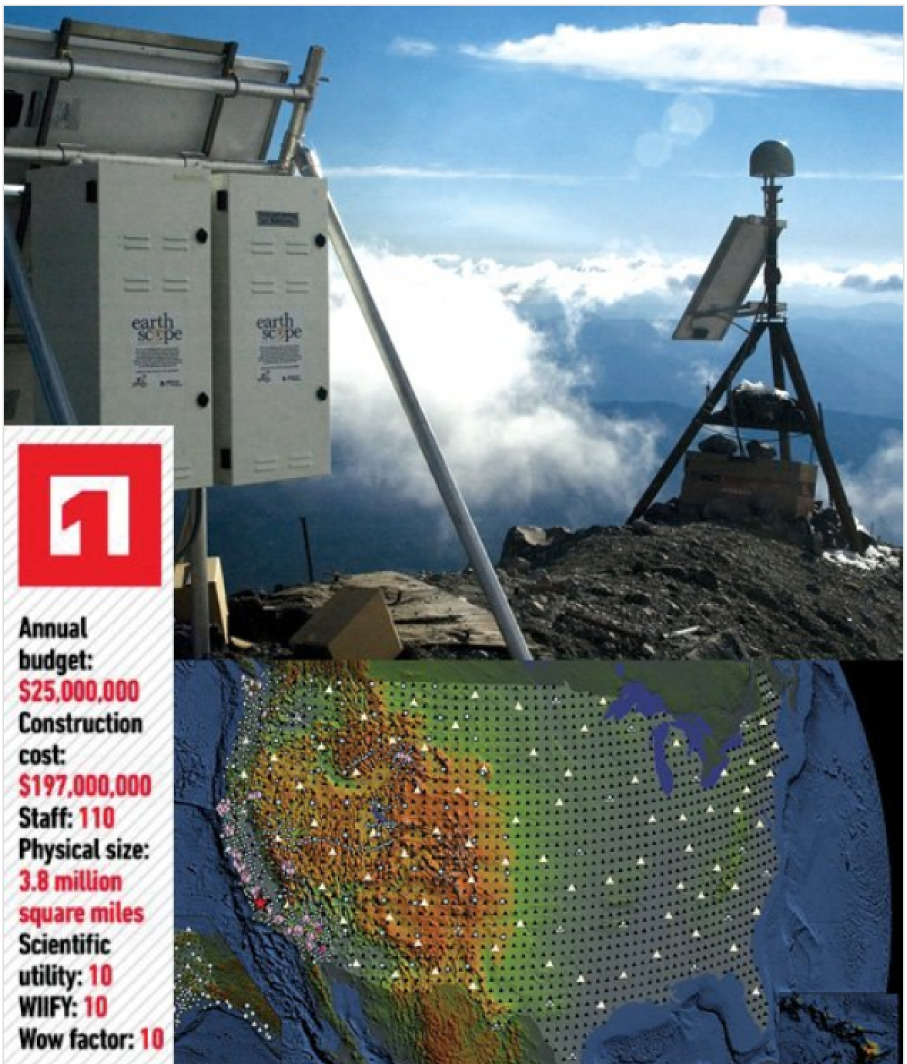



Map
Satellite
Idaho
Idaho Falls
Blackfoot
Pocatello
Mountain Home
Craters of the Moon
Twin Falls
Burley
Logan
Grand Teton National Park
100 km
Google


environets
Network Impact
What is this?
The statistics available here show the Relative Impact accumulated each hour for an entire network. Large networks with broad geographical coverage will tend to have Relative Impacts near 0.50, while the impacts of smaller networks may reflect local weather features and vary considerably from day-to-day.
Select a Network:
AKDOT
APRSWXNET/CWOP
AQ
ARL FRD
ARL SORD
AVALANCHE
Parameters: Wind Speed
Metric: Relative Impact
Start date: 12/01/2012 (MM/DD/YYYY)
End date: 12/15/2012 (MM/DD/YYYY)
Plot!
Metric
Number of Stations
Percentile
Number of Stations
Dec 3
Dec 5
Dec 7
Dec 9
Dec 11
Dec 13
Dec 15
Date/Time

Big Science: The Universe's Ten Most Epic Projects

By Gregory Mone, Brooke Borel, Katherine Bagley and Jennifer Abbasi Posted 7.16.11 at 8:06 pm 12 Comments



1: The Earthscope

EarthScope

A telescope to peer deep into the heart of our planet

Designed to track North America's geological evolution, EarthScope is the largest science project on the planet. This earth-sciences observatory records data over 3.8 million square miles. Since 2003, its more than 4,000 instruments have amassed 67 terabytes of data—that's equivalent to more than a quarter of the data in the Library of Congress—and add another terabyte every six to eight weeks

Scientific Utility

Researchers are using EarthScope, which consists of many kinds of experiments, to examine all facets of North America's geological composition. Across the continental U.S. and Puerto Rico, 1,100 permanent GPS units track deformations in the land's surface caused by tectonic shifts below. Seismic sensors next to the active San Andreas Fault in California record its tiniest slips, while rock samples pulled from a drill site that extends two miles into the fault reveal the grinding and strain on the rocks that occur when the two sides of the fault slide past each other during an earthquake. And over the course of 10 years, small crews have hauled a moveable array of 400 seismographs across the country using backhoes and sweat. By the time the stations reach the East Coast next year, they will have collected data from almost 2,000 locations.

What's In It For You

Collectively, EarthScope's measurements could help explain the forces behind geological events such as earthquakes and volcanic eruptions, leading to better detection. So far, data from the project

Tremendous Opportunities…

- Computers are cheap
- Networking is great
- Huge volumes of in situ and remote sensing data at finger tips
- New data sources becoming available
- Flexible open source software is available to develop quality control, data analysis, and data mining software
- Expanded utilization of ensemble data assimilation methods
- Enhanced interest by computing science faculty to collaborate due to NSF funding opportunities in “big data and data mining”, “sustainable computing”, and “extreme-scale computing”