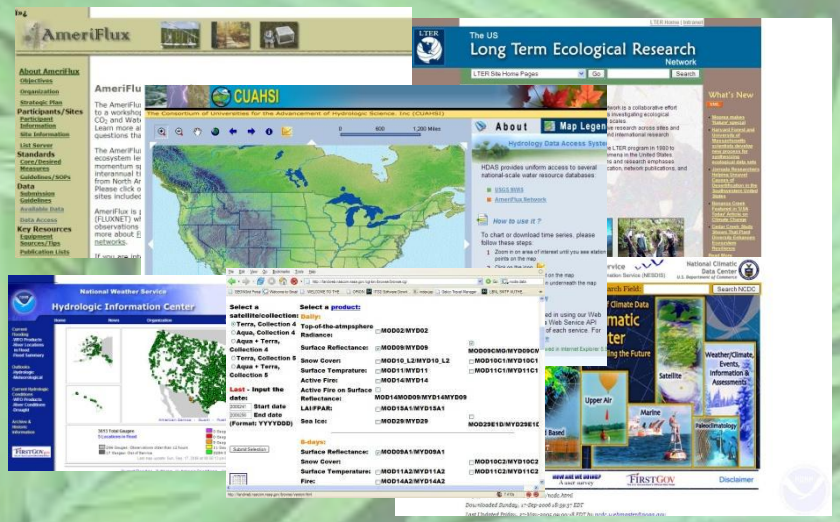
ENVIRONMENTAL SCIENCE AS A DATA INTENSIVE SCIENCE

Catharine van Ingen
Microsoft Research eScience Group

Introduction

- Initial research: Can you use traditional database technologies to accelerate this science?
  - While small by computer science standards, the data management is an impediment to the science.
  - "All your eggs in one basket" helps a lot
  - Browsing for science is a precursor for analysis
  - Science data is different from commercial data
- Current research: What it means to do science in the cloud ?
  - Heavy technology managed in the cloud
  - Leveraging technology to break down social barriers

The Perfect Data Storm

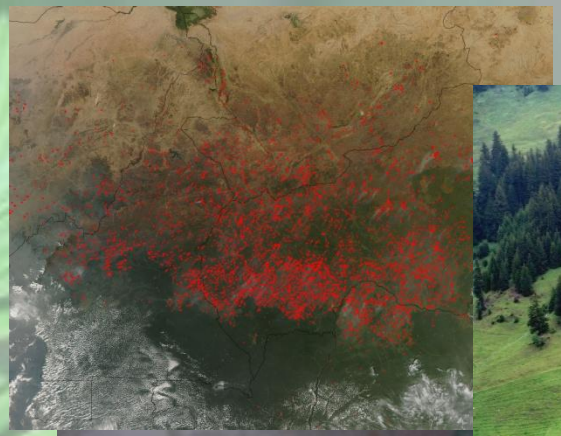


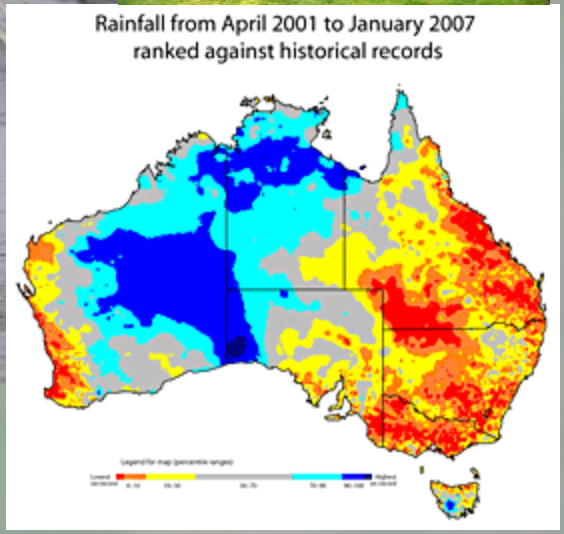
- The era of remote sensing, cheap ground-based sensors and web service access to agency repositories is here

- Extracting and deriving the data needed for the science remains problematic
  - Specialized knowledge
  - Finding the right needle in the haystack

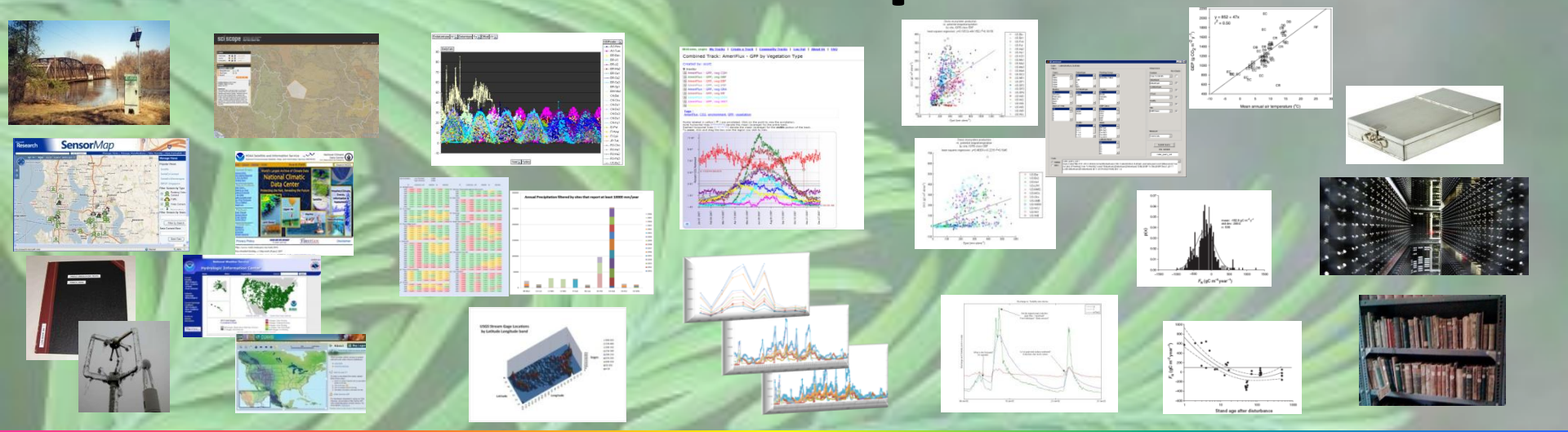
Environmental Science Synthesis

- Everything is connected
  - Bridging from global to regional to local spatial scales is more and more important, yet remains problematic
  - Bridging across disciplines (eg climatology, hydrology, forestry) and across methodologies (eg modeling and field observations) always challenging
- Very heterogeneous data
  - Satellite imagery, older sensor data, and new sensor field deployments all provide pieces of the puzzle
- Politics plays a part
  - Water resource management, electric generation, disaster response, etc. etc.
  - Critical information sometimes restricted





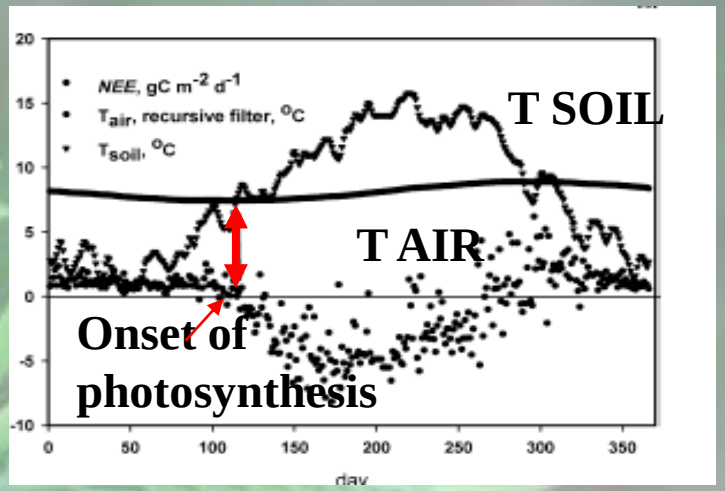
The Data Pipeline

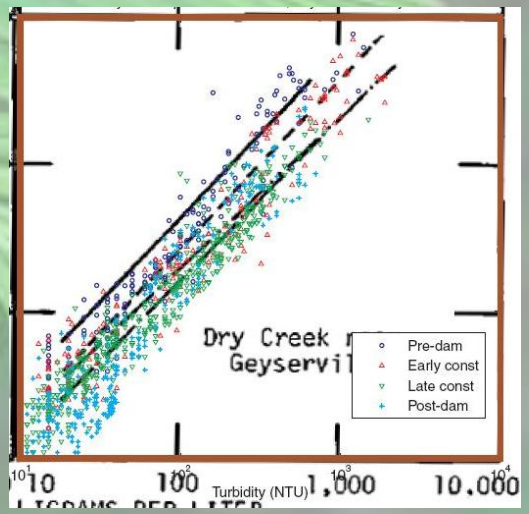


Data Acquisition	Data Assembly	Discovery and Browsing	Science Exploration	Domain specific analyses	Scientific Output	Archive
Field sensor deployments and operations; field campaigns measuring site properties.	"Raw" data includes sensor output, data downloaded from agency or collaboration web sites, papers (especially for ancillary data.	"Raw" data browsing for discovery (do I have enough data in the right places?), cleaning (does the data look obviously wrong?), and light weight science via browsing	"Science variables" and data summaries for hypothesis testing and early exploration. Like discovery and browsing, but variables are computed via gap filling, units conversions, or simple equations.	"Science variables" combined with models, other specialized code, or statistics for deep science understanding.	Scientific results via packages such as MatLab or R2. Special rendering package such as ArcGIS. Paper preparation.	Data and analysis methodology stored for data reuse, or repeating an analysis .

Environmental Science Data

- Time series data
  - Over some period of time at some time frequency at some spatial location.
  - May be actual measurement (L0) or derived quantities (L1+)
- (Re)calibrations are a way of life.
  - Various quality assessment algorithms used to mark and/or correct spikes, drifts, etc.
- Gaps and errors are a way of life.
  - Birds poop, batteries die, sensors fail.
  - Gap filling algorithms becoming more and more common because regularly spaced time series are much simpler to analyze
- Space and time are fundamental drivers
- Versioning is essential



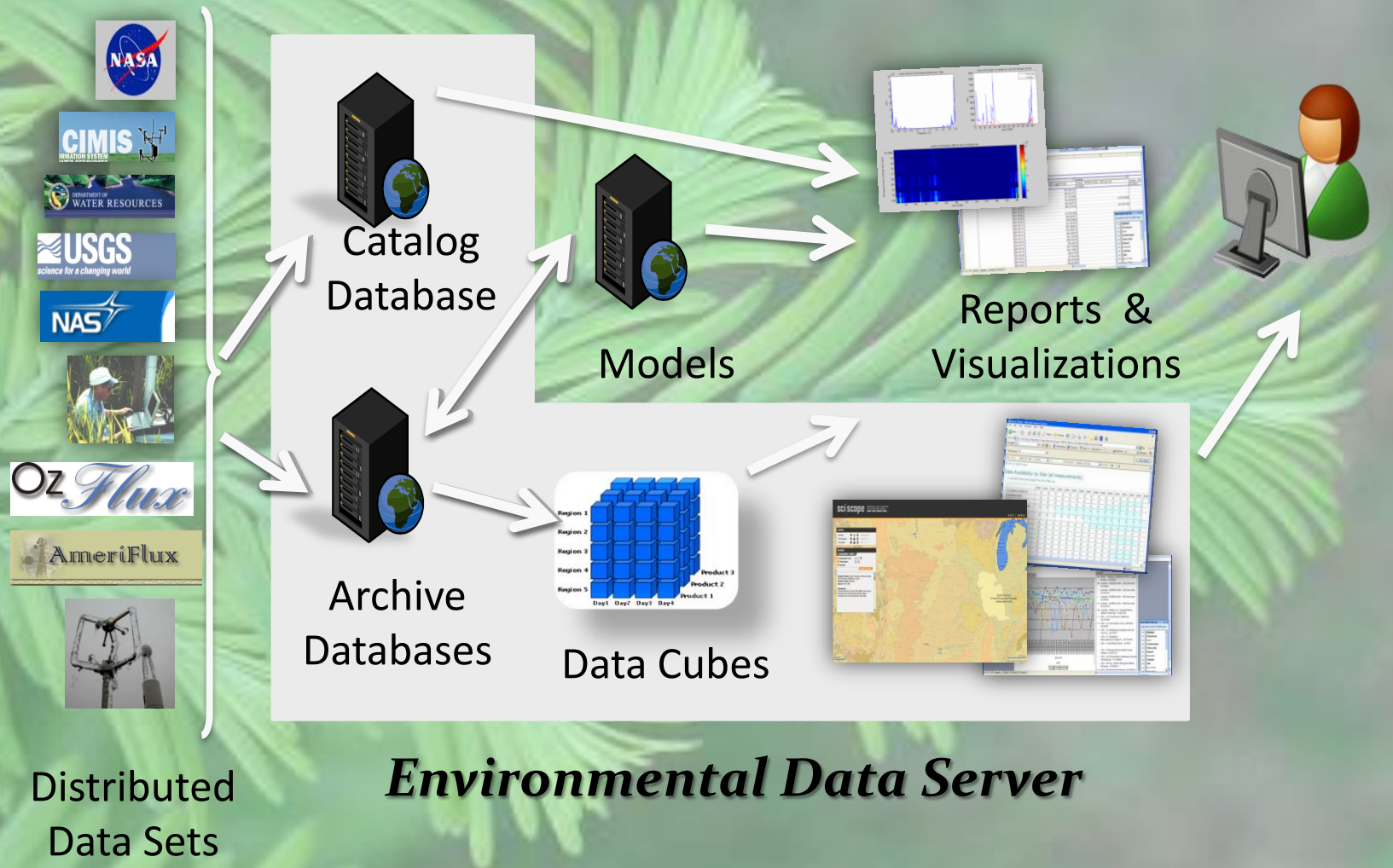


"Time is not just another axis"

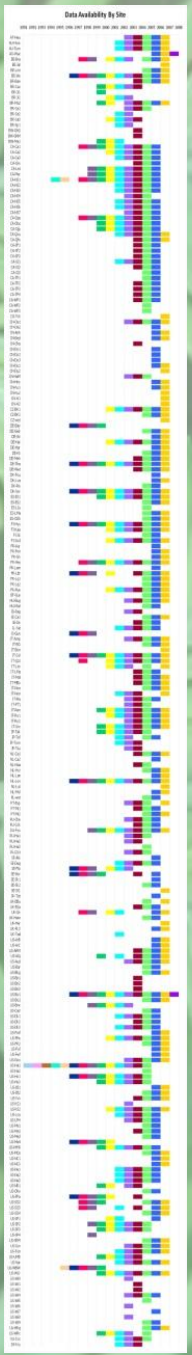
Environmental Ancillary Data

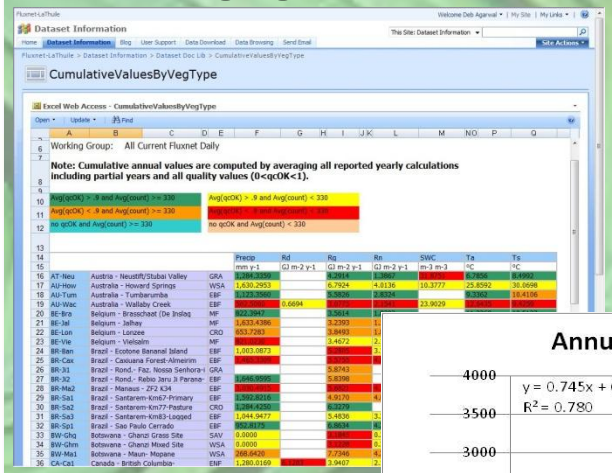
- Non-time series data
  - May be ‘constant’ such as latitude or longitude
  - May be measured intermittently such as LAI (leaf cross-sectional area) or sediment grain size distribution
  - May be a range and estimated time
  - May be a disturbance such as a fire, harvest, landslide
  - May be derived from the data such as flood
  - Not metadata such as instrument type, derivation algorithm, etc.
- Usage pattern is key
  - Constant location attributes or aliases
  - Time series data (by interpolating or “gap filling” irregular samples
  - Time filters (short periods before or after an event or sampled variable)
  - Time benders (“since <event>” including the deconvolution of closely spaced events a fire

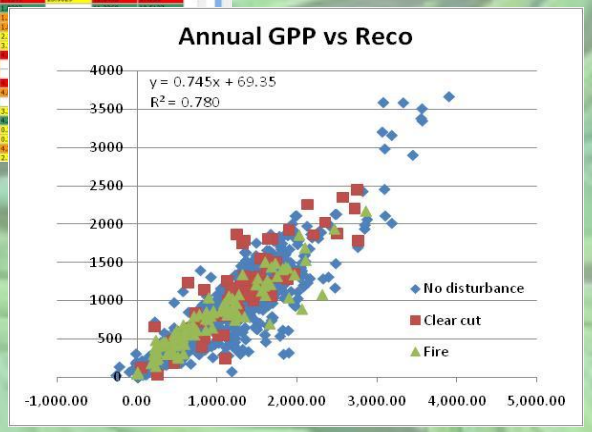
Connecting Data, Resources, and People

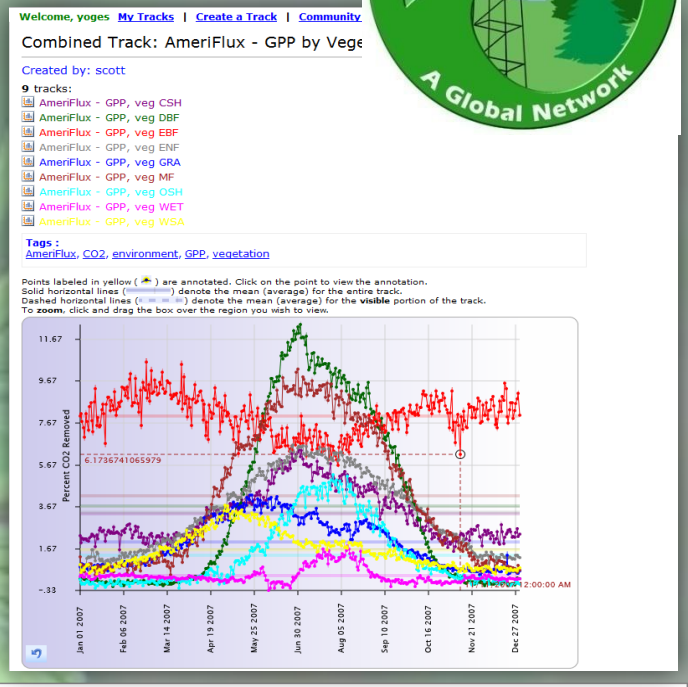


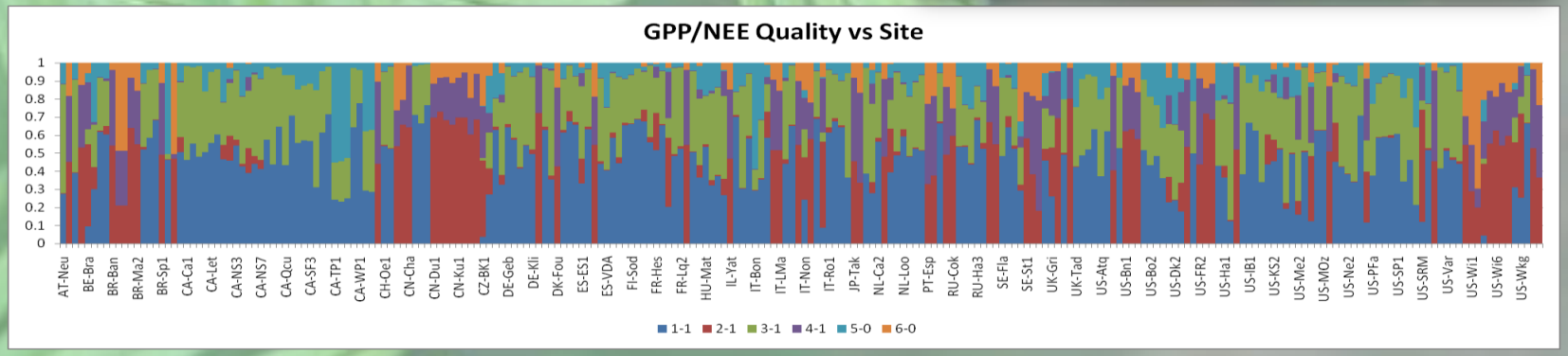
Browsing For Data Availability, Applicability, Early Science



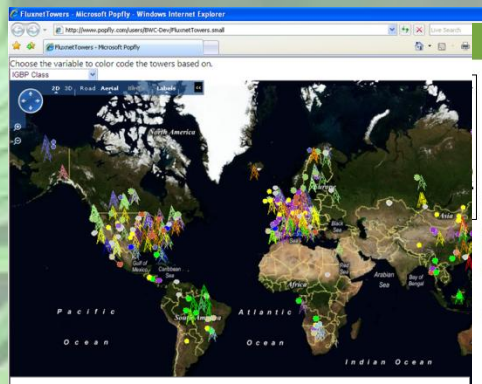
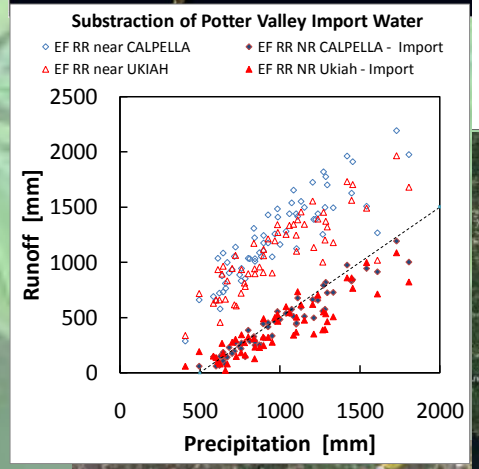
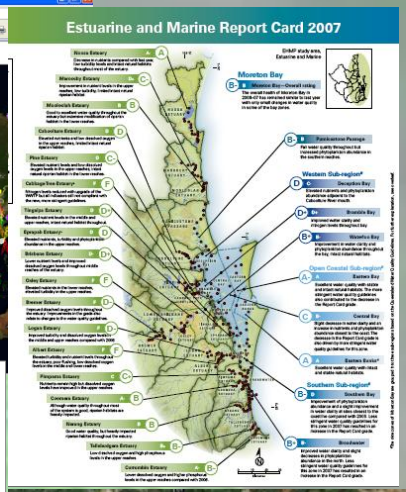








Related projects leveraging same approach and technology base

- Digital watersheds (BWC)
  - Russian River digital watershed built from existing data to look at changes over time affecting salmon
- RENCI and CUASHI
  - 100 universities collaborating on hydrology
  - Working with SDSC team to redesign schema and tools to enable simple data download/discovery
  - RENCI team will leverage that redesign to link with models
- Queensland Health E-Water
  - Facing extreme water shortage and endangered wetlands due to rapid population growth and
  - Working with Jane Hunter and Eva Abal to archive existing data and produce new data summaries in a timely fashion
- Fluxdata.org (BWC)

- Pajaro River (BWC)
  - Digital watershed to be built from existing data to look at natural channel flood control alternatives (listed in top 10 "endangered river" in 2006)
  - SciScope (UCB/SDSC)
  - Virtual Earth mashup discovery engine for CUASHI