Data Cubes for Eco-Science Can One Size Fit All?

Deborah Agarwal (UCB/LBL)
Catharine van Ingen (MSR)
Berkeley Water Center
22 October 2007

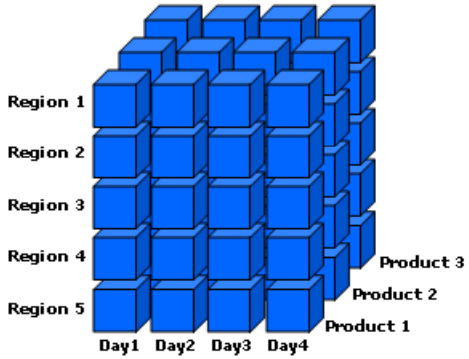
Introduction

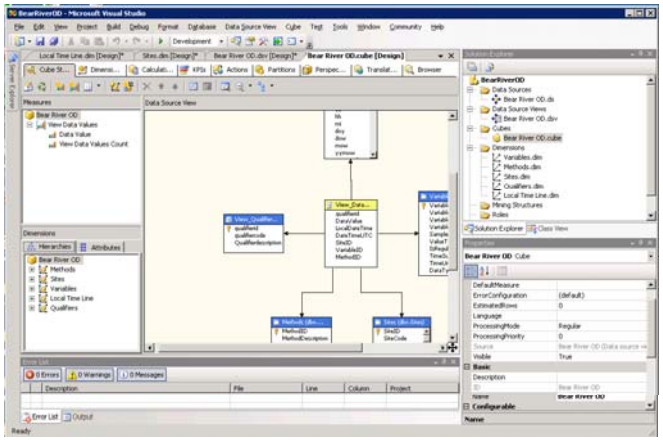
- Over the past year, we've been experimenting using data cubes to support carbon-climate, hydrology, and other eco-scientists
  - While the science differs, the data sets have much in common
  - The cube is a useful tool in the data analysis pipeline
- Along the way, we've wondered how to build toward a "My Cube" service
  - Empower the scientist to build a custom cube for a specific analysis

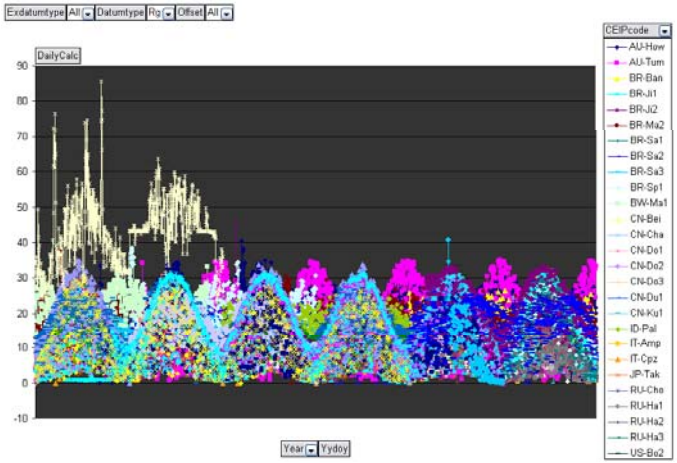
http://bwc.berkeley.edu/
http://www.fluxdata.org/

Data Cube Basics

- A data cube is a database specifically for data mining (OLAP)
  - Initially developed for commercial needs like tracking sales of Oreos and milk
  - Simple *aggregations* (sum, min, or max) can be pre-computed for speed
  - *Hierarchies* for simple filtering with drilldown capability
  - Additional *calculations* (median) can be computed dynamically or pre-computed
  - All operate along *dimensions* such as time, site, or datumtype
  - Constructed from a relational database
  - A specialized query language (MDX) is used
- Client tool integration is evolving
  - Excel PivotTables allow simple data viewing
  - More powerful analysis and plotting using Matlab and statistics software







Daily Rg 2000-2005 72 sites, 276 site-years

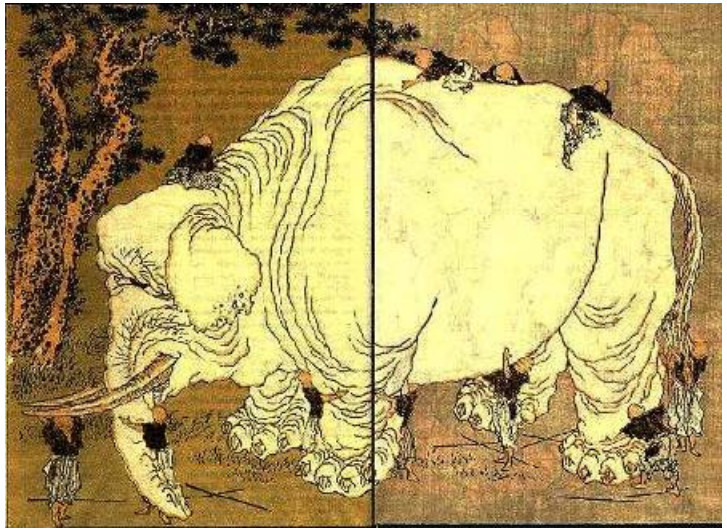
Eco-Science Data

- What we start with

Ecological Data Avalanche/Landslide/Tsunami

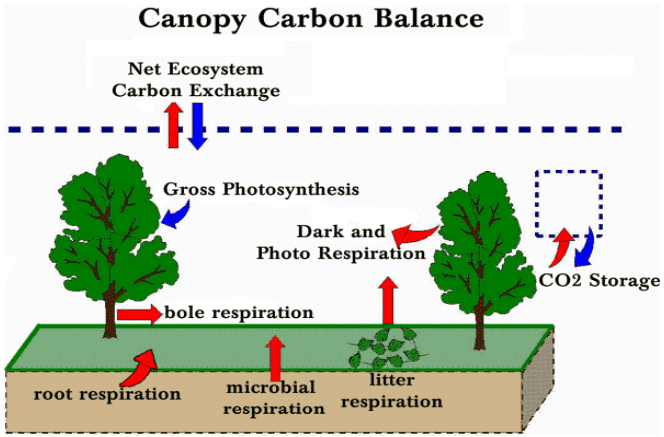
The era of remote sensing, cheap ground-based sensors and web service access to agency repositories is here

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



Carbon-Climate Research Overview

- What is the role of photosynthesis in global warming?
  - Measurements of Co2 in the atmosphere show 16-20% less than emissions estimates predict
  - Do plants absorb more than we expect?
- Communal field science - each principle investigator acts independently to prepare and publish data.
- 496 sites world wide organized into 13 networks plus some unaffiliated sites
  - AmeriFlux: 149 sites across the Americas
  - CarboEuropeIP: 129 sites across Europe
- Data sharing across investigators just beginning
  - Level 2 data published to and archived at network repository
  - Level 3 & 4 data now being produced in cooperation with CarboEuropeIP and served by BWC TCI
- Total fluxnet data accumulated to date ~800M individual measurements

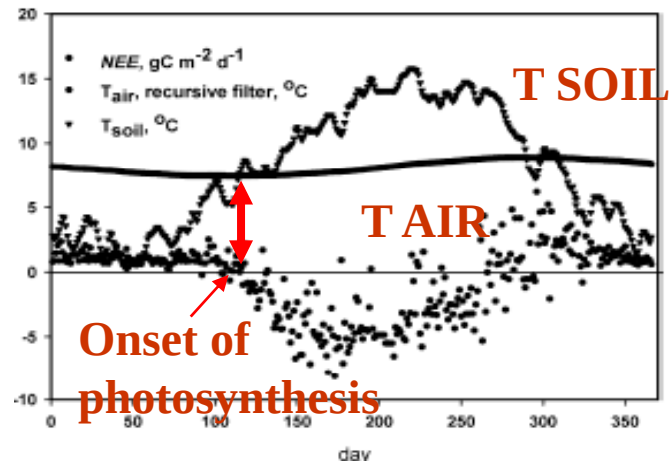


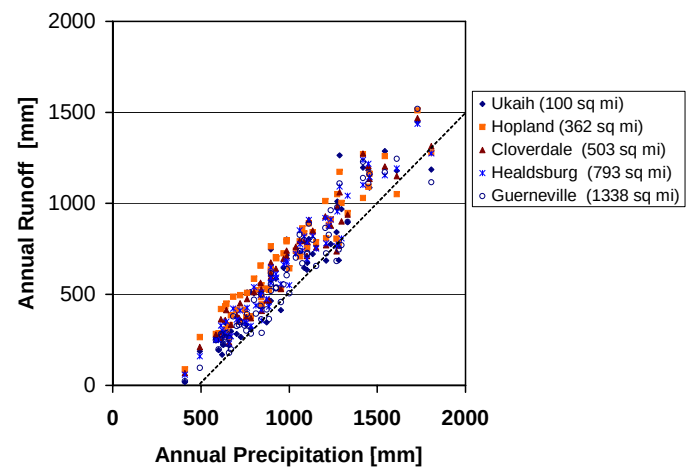
http://www.fluxdata.org

http://gaia.agraria.unitus.it/cpz/index3.asp

Eco-Science Data

- When we say data we mean predominantly 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, and sensors fail.
  - Gap filling algorithms becoming more and more common because a regularly spaced time series is much simpler to analyze
- Space and time are fundamental drivers
- Versioning is essential





Eco-Science Ancillary Data

- When we say ancillary data, we mean 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, or flood
  - 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

Eco-Science Analyses

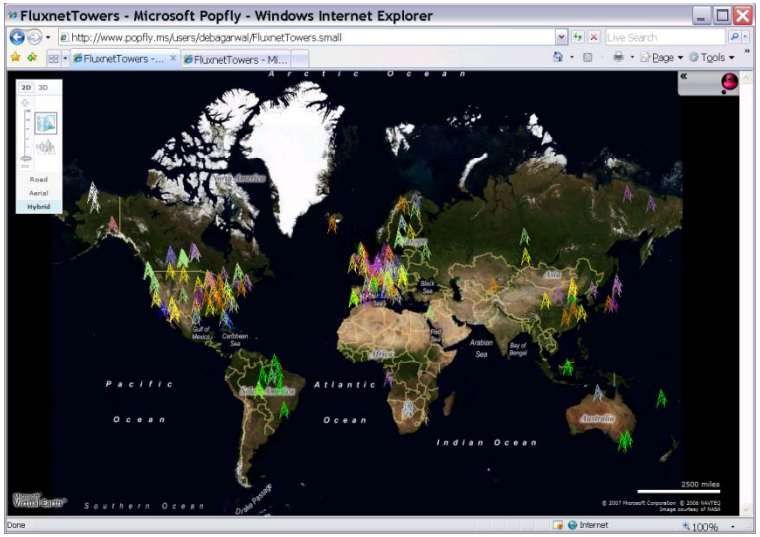
- Why use a datacube?

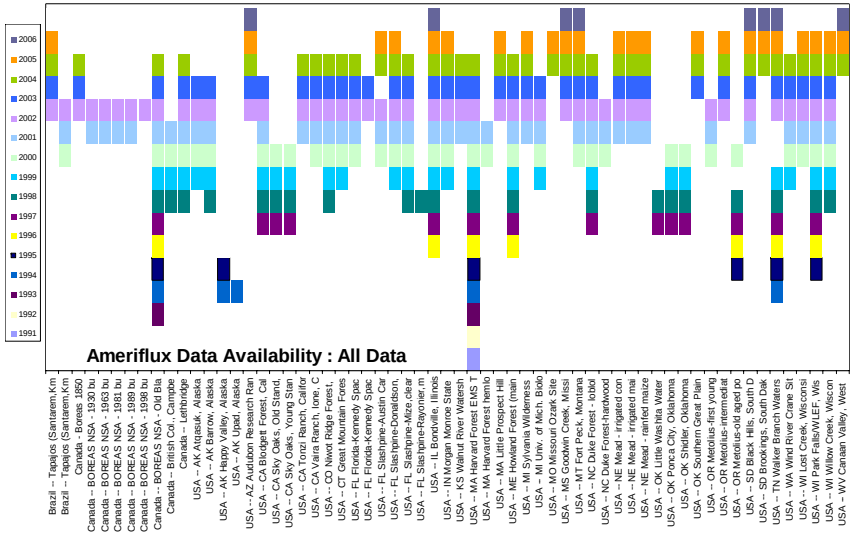
The Data Pipeline

Data Gathering	Discovery and Browsing	Science Exploration	Domain specific analyses	Scientific Output
"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 early science exploration and hypothesis testing. Similar to discovery and browsing, but with science variables computed via gap filling, units conversions, or simple equation.	"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.

Browsing for Data Availability

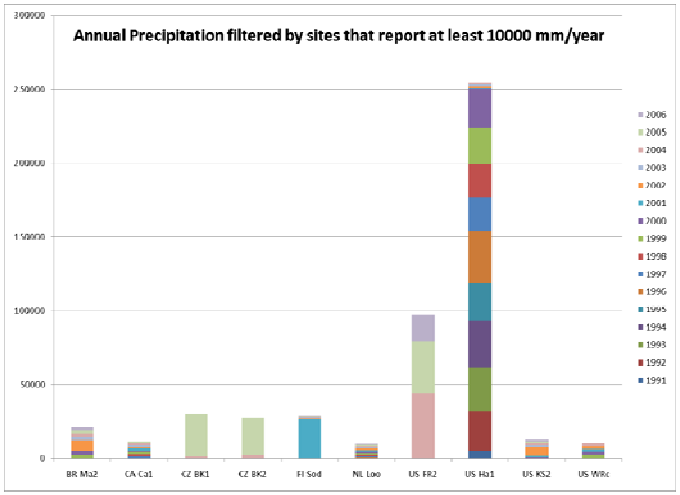
- Summary data products (yearly min/max/avg) almost trivially
- Simple mashups and data cubes aid discovery of available data
- Simple Excel graphics show cross-site comparisons and availability filtered by one variable or another

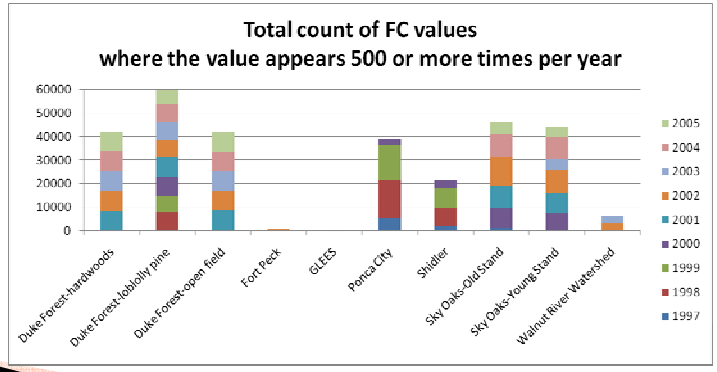


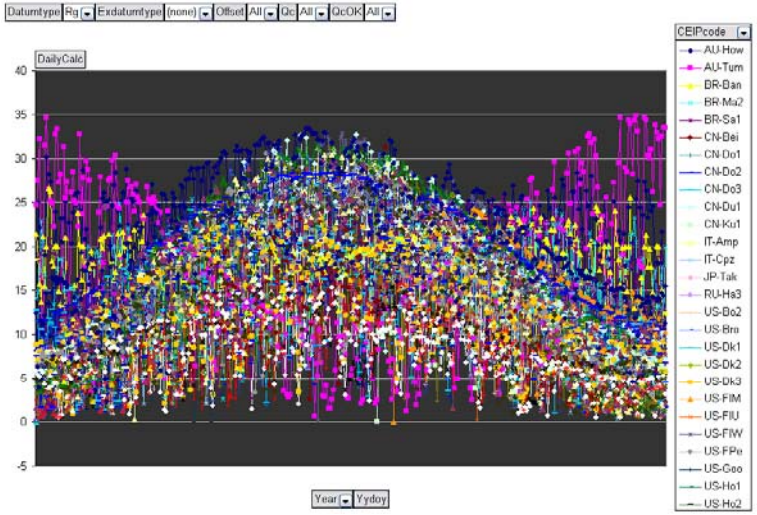


Browsing for Data Quality

- Data cleaning never ends
  - Existing practice of running scripts on specific site years often misses the big picture
  - Corrections to calibration





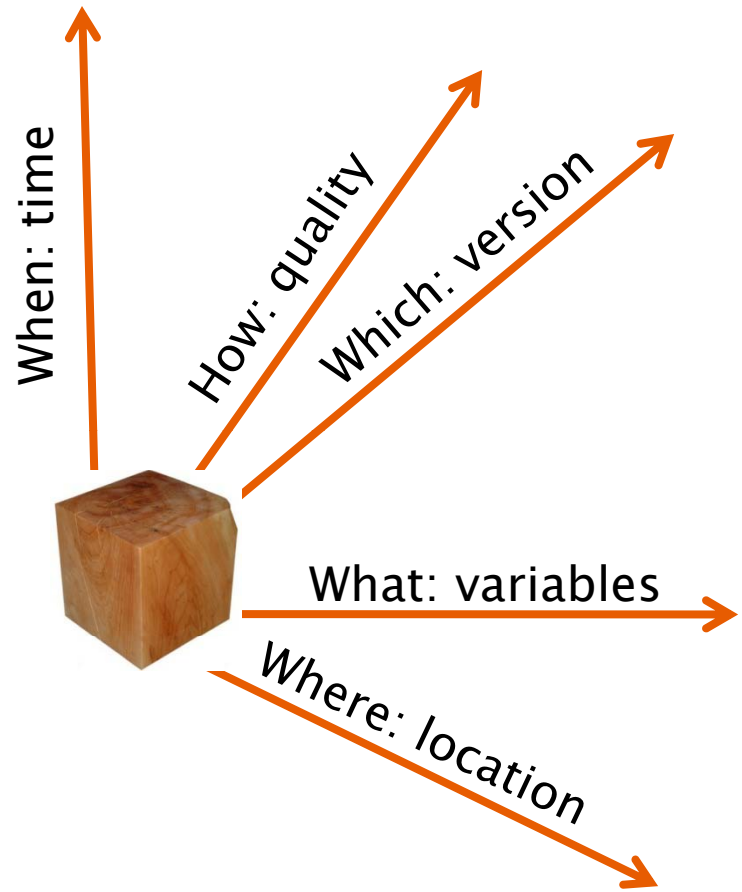


Eco-Science Datacubes

- Building our datacube family

Cube Dimensions

- We've been building cubes with 5 dimensions
  - What: variables
  - When: time, time, time
  - Where: (x, y, z) location or attribute where (x,y) is the site location and (z) is the vertical elevation at the site.
  - Which: versioning and other collections
  - How: gap filling and other data quality assessments



Driven by the nature of the data – space and time are fundamental drivers for all earth sciences.

Cube Measures

- We've been including a few computed members in addition to the usual count, sum, minimum and maximum
  - hasDataRatio: fraction of data actually present across time and/or variables
  - DailyCalc: average, sum or maximum depending on variable and includes units conversion
  - YearlyCalc: similar to DailyCalc
  - RMS or sigma: standard deviation or variance for fast error or spread viewing

Driven by the nature of the analyses – gaps, errors, conversions, and scientific variable derivations are facts of life for earth science data.

What: variables and measures

Core variables or datumtypes	Non-core or extended datumtypes	Ancillary data treated as time-series data or filters. Note that the gap filling or interpolation algorithm is likely domain-specific.
Daily and yearly value calculation, data counts, min/max values	Hierarchies to solve very large namespace navigation. Note that most science cannot leverage cube aggregations here.	Special calculation such as potential evapotranspiration or bedload sediment transport.

When: handling time

Core time hierarchies. Includes simple calendar, water year, MODIS week.

"tunable" time filters such as morning, afternoon, night, winter; each defined by a start/stop at a hierarchical level.

Time period definition determined by time-series variable (eg. PAR-day determined by photosynthetic activity)

Selectable hierarchy top and bottom: decade, year, month, day

Time folding based on data value (eg. after a rain) or ancillary data value (eg. after a fire).

"time is not just another axis"

Where: handling location

Site (location) presented by friendly name and selectable constant site ancillary data such as latitude band or vegetationtype.

Selectable site hierarchies for simple navigation (ala variable) or aggregation (eg. state or HUC).

Site selection determined by time-series variable (eg. minimum temperature) or non-constant ancillary variable (eg. above a soil nitrogen threshold).

No offset. Either one vertical location or aggregated over the vertical.

Geo-spatial calculations

Vertical profiles

Which: versions and subsets

No version (dataset chosen when cube is built).

Selectable site hierarchies for simple navigation (ala variable) or aggregation (eg state or HUC).

Datasets used to include/exclude location subsets, datumtype subsets, or the same data at different processing levels or measurement granularity (eg USGS daily vs 15 minute stream flow).

How: quality

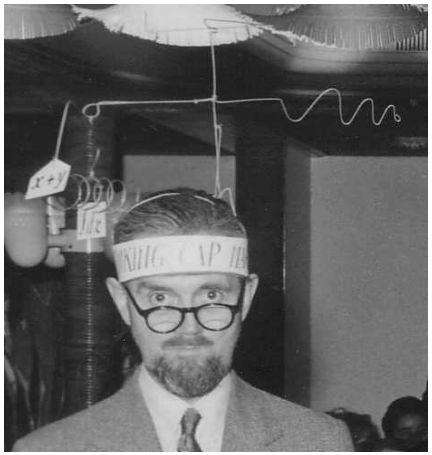
Ignore any quality metric

Simple statistics, gap-filling, spike detection, level delta and drift checks.

Quality dimension to allow visualization and filtering

??????

This is clearly where we'll be spending more time!



How Does This Fit in the Scientific Process?

Data Gathering	Discover y and Browsing	Science Exploration	Domain specific analyses	Outputs
Automation of data ingest.	Data cube.	Data cube calculated dimensions/ aggregations.	My calculation/ My cube/My database.	Interfaces to commercial products and shareware.
As standards emerge, sensor data acquisition and ingest will become easier.	Should be reasonably straight forward to generalize. (Data mining could be interesting here)	Some conversions are simple. Some exploration is just browsing with different variables.	"Special purpose – may be hard to do given the base database/ datacube technologies. (Workflow technologies might help here)	Should be reasonably straight forward to generalize. (Web services and Collaborative tools help here)

Acknowledgements

Berkeley Water Center, University of California, Berkeley, Lawrence Berkeley Laboratory

- Jim Hunt
- Deb Agarwal
- Robin Weber
- Monte Good
- Rebecca Leonardson (student)
- Matt Rodriguez
- Carolyn Remick
- Susan Hubbard

University of Virginia

- Marty Humphrey
- Norm Beekwilder

Microsoft

- Catharine van Ingen
- Jayant Gupchup (student)
- Savas Parastidis
- Andy Sterland
- Nolan Li (student)
- Tony Hey
- Dan Fay
- Jing De Jong-Chen
- Stuart Ozer
- SQL product team
- Jim Gray

Ameriflux Collaboration

- Beverly Law
- Youngryel Ryu (postdoc)
- Tara Stiefl (student)
- Gretchen Miller (student)
- Mattias Falk
- Tom Boden
- Bruce Wilson

Fluxnet Collaboration

- Dennis Baldocchi
- Rodrigo Vargas (postdoc)
- Dario Papale
- Markus Reichstein
- Bob Cook
- Susan Holladay
- Dorothea Frank

The saga continues at

http://bwc.berkeley.edu and

http://www.fluxdata.org