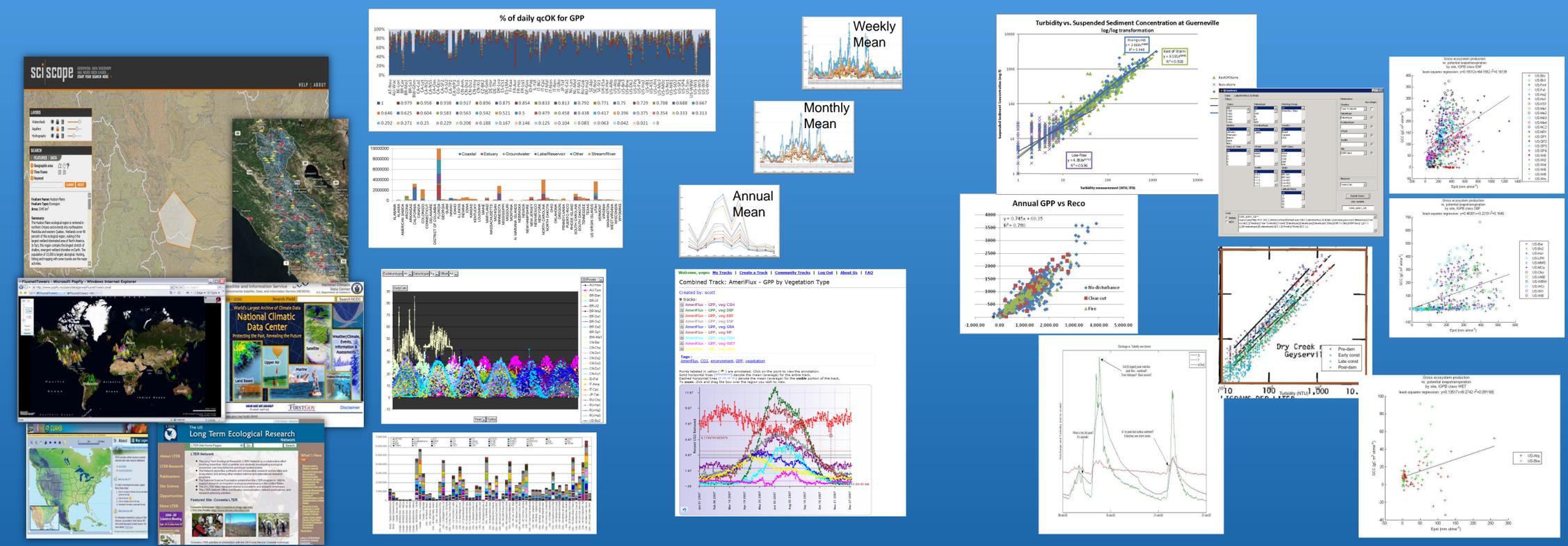


# EScience Technologies for the Science Data Pipeline

Catharine van Ingen<sup>c</sup>, Deborah Agarwal<sup>ab</sup>, James Hunt<sup>ad</sup>, Marty Humphrey<sup>e</sup>, Bora Beran<sup>c</sup>  
 Berkeley Water Center<sup>a</sup>, Lawrence Berkeley Laboratory<sup>b</sup>, Microsoft Research<sup>c</sup>, University of California, Berkeley<sup>d</sup>, University of Virginia<sup>e</sup>



## Data Gathering

“Raw” data includes sensor output, data downloaded from agency or collaboration web sites, papers (especially for ancillary data)

## Discovery and Browsing

“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 Exploration

“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 equation.

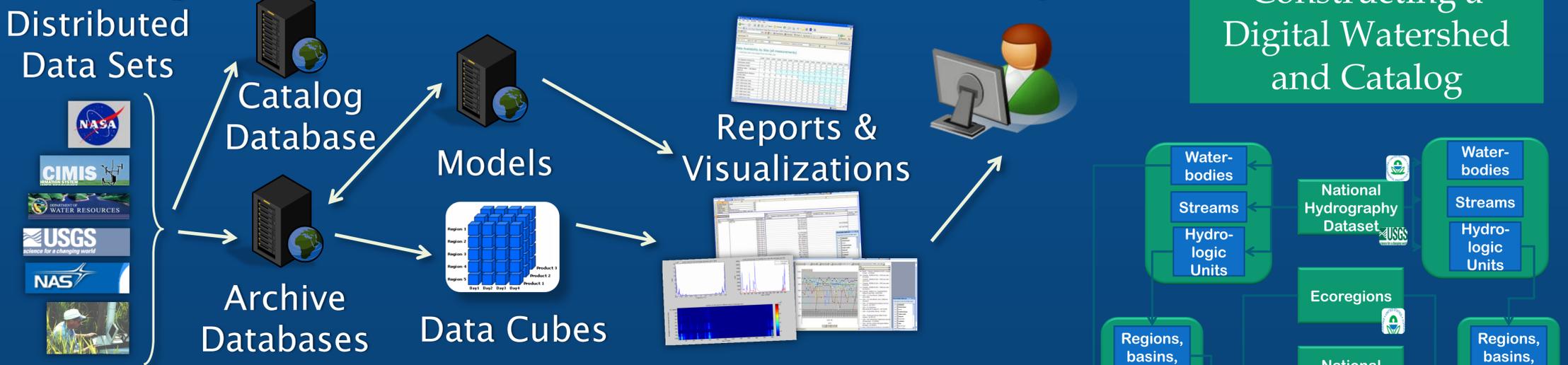
## Domain specific analyses

“Science variables” combined with models, other specialized code, or statistics for deep science understanding.

## Scientific Output

Scientific results via packages such as MatLab or R2. Special rendering package such as ArcGIS.  
 Paper preparation.

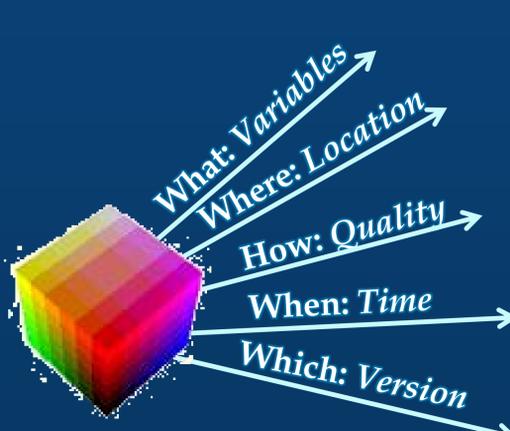
## Challenge is to Connect Data, Resources, and People!



## Behind the Scenes in the Cloud

- bwc.berkeley.edu Website and datacube access
- database archive data ingest and cube development
- DC for flux domain
- www.fluxdata.org Sharepoint and secured data file download
- heavy database queries and cube development
- Backup gateway
- wally.lbl.gov MatLab and ArcGIS for key scientists

## Common Data Cubes



www.noaa.gov  
 www.nationalatlas.gov  
 www.usgs.gov  
 www.epa.gov