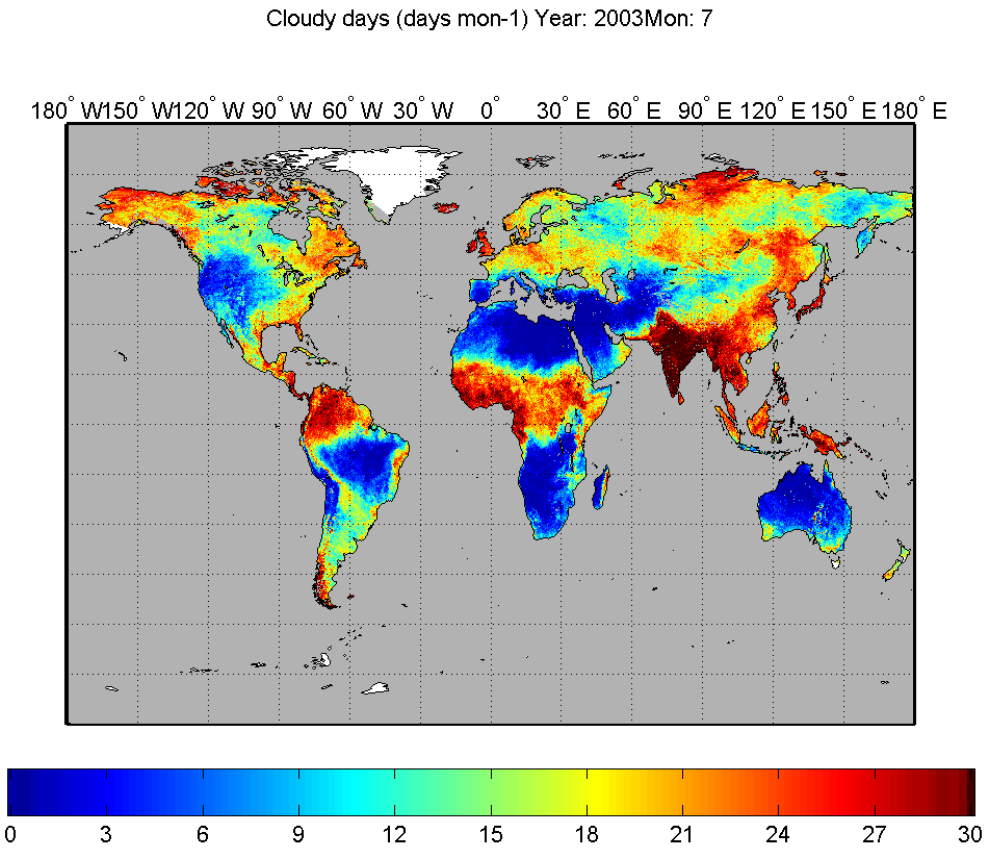
Scaling Science: HPC in a Big Data World

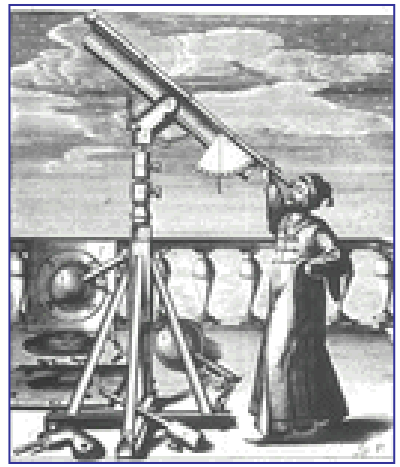


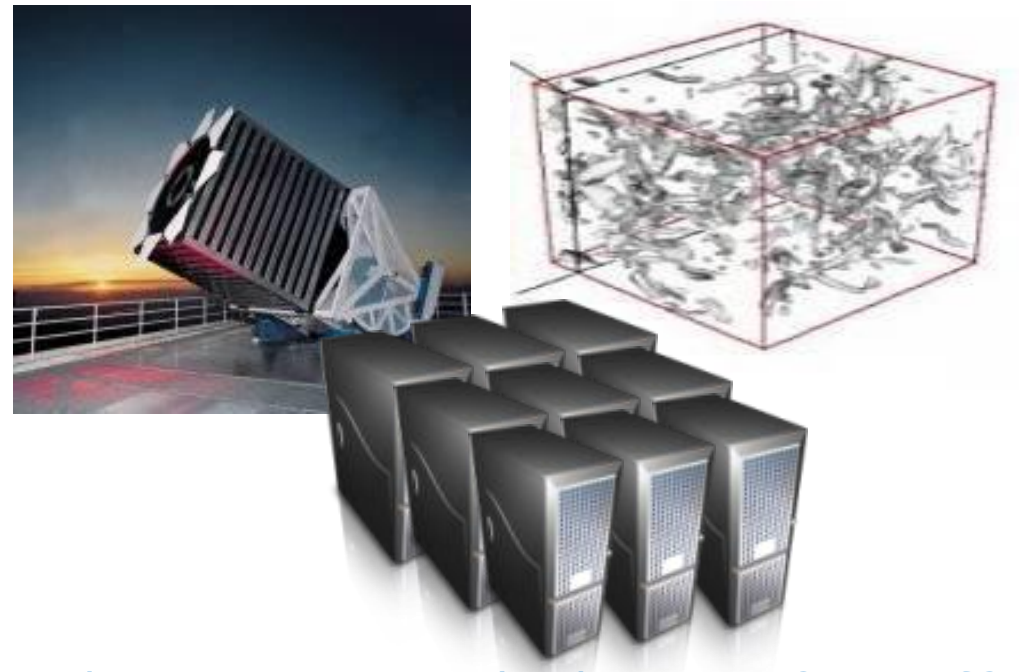
Catharine van Ingen
Microsoft Research

Emergence of a Fourth Paradigm

- Thousand years ago – **Experimental Science**
  - Description of natural phenomena
- Last few hundred years – **Theoretical Science**
  - Newton's Laws, Maxwell's Equations…
- Last few decades – **Computational Science**
  - Simulation of complex phenomena
- Today – **Data-Intensive Science**
  - Scientists overwhelmed with data sets from many different sources
    - Data captured by instruments
    - Data generated by simulations
    - Data generated by sensor networks
- eScience is the set of tools and technologies to support data federation and collaboration
  - For analysis and data mining
  - For data visualization and exploration
  - For scholarly communication and dissemination

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



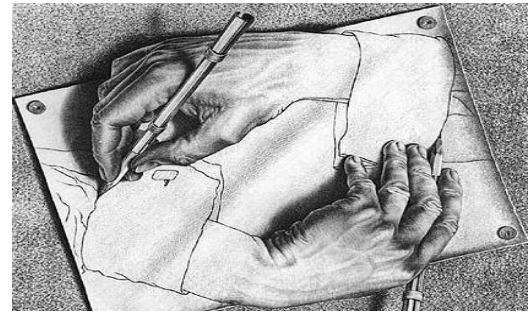


http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Thanks to Jim Gray (MSR) 2007

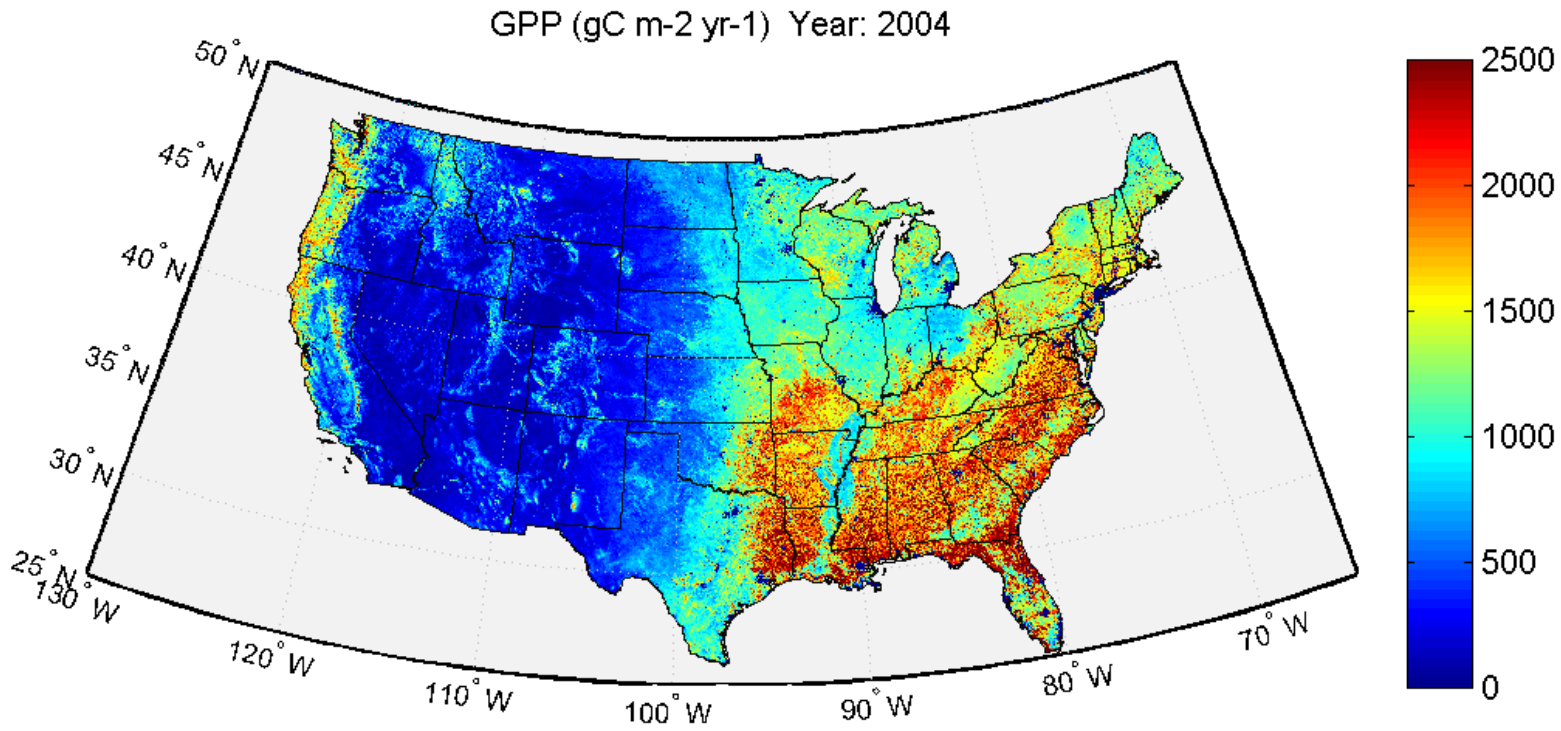
Transformation and Synergy

- We are now in the second phase of the IT revolution: the rise of the ***information/data driven computing***
  - In addition to the traditional numerically-intensive science
  - IT as a primary publishing and communication technology
- ***All science*** in the 21st century is becoming cyber-science (aka e-Science) - and with this change comes the need for ***a new scientific methodology***
- The challenges we are tackling:
  - Management of large, complex, distributed data sets
  - Effective exploration of such data ⮕ new knowledge
  - **These challenges are universal**
- A great synergy of the computationally enabled science, and the science-driven IT



Thanks to George Djorgovski (CalTech) 2010

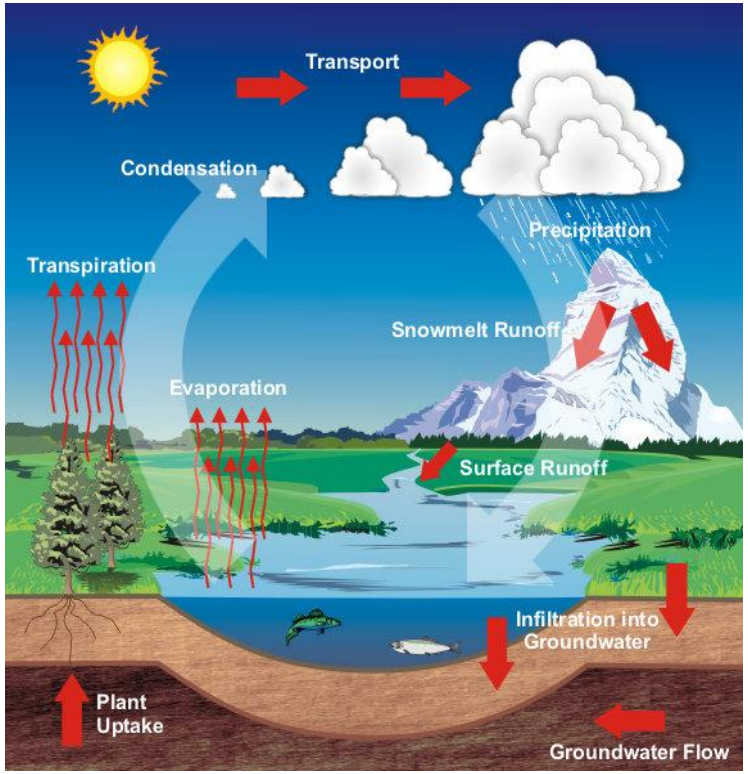
Data-Driven Science, Public Policy, and Economics



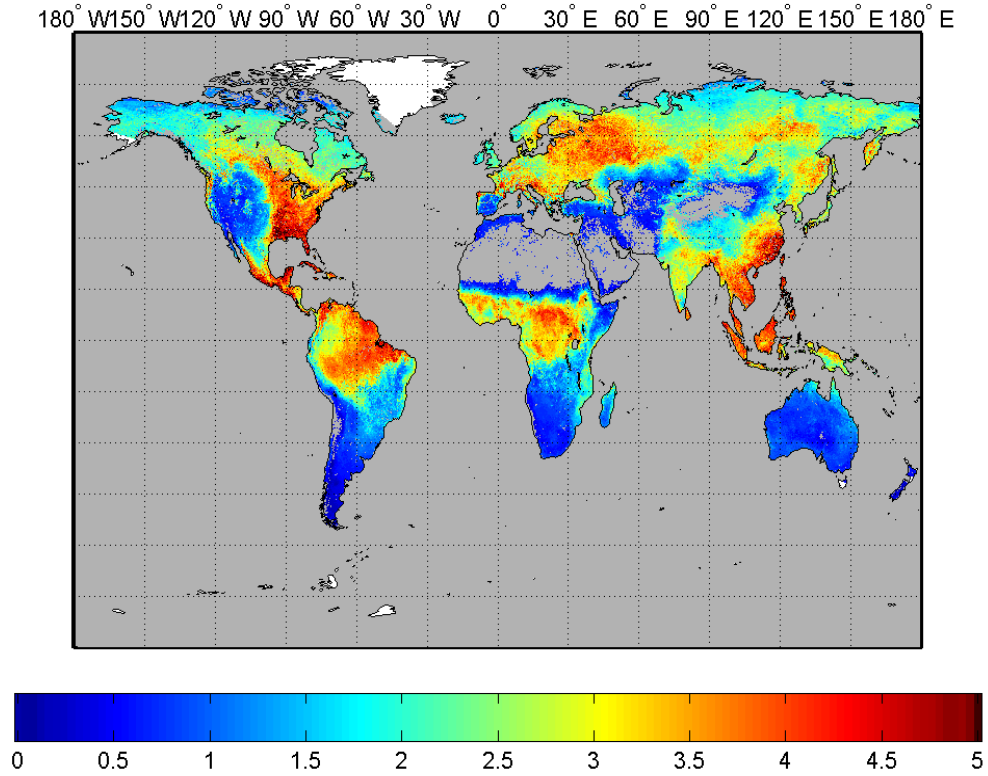
- GPP or Gross Photosynthetic Production is a measure of carbon fixation and related to biosphere water balance
- High GPP indicates high potential for biofuel production
- Computation suggests that the Southern temperate forests have higher GPP than the mid-west Corn Belt

Thanks to Dennis Baldocchi (UC Berkeley) 2010

Computing Water Balance at Global Scale with Local Fidelity



- Evapotranspiration is loss of water to atmosphere by biosphere (transpiration) and standing water bodies (evaporation)
- Computation is necessary for estimates of carbon fixation



- Computed monthly average evapotranspiration (ET) for July 2003
- Red indicates high transpiration (hence likely high carbon fixation)

Thanks to (left) NOAA and (right) Youngryel Ryu (Seoul National University)

Computing Water Balance: First Principles

$$ET = \frac{\Delta Rn + \rho_a c_p (\delta q)/r_a}{(\Delta + \gamma\left(1 + \frac{r_s}{r_a}\right))\lambda_v}$$

Penman-Monteith (1964)

ET = Water volume evapotranspired ($m^3\ s^{-1}\ m^{-2}$)

Δ = Rate of change of saturation specific humidity with air temp.($Pa\ K^{-1}$)

λ_v = Latent heat of vaporization (J/g)

R_n = Net radiation ($W\ m^{-2}$)

c_p = Specific heat capacity of air ($J\ kg^{-1}\ K^{-1}$)

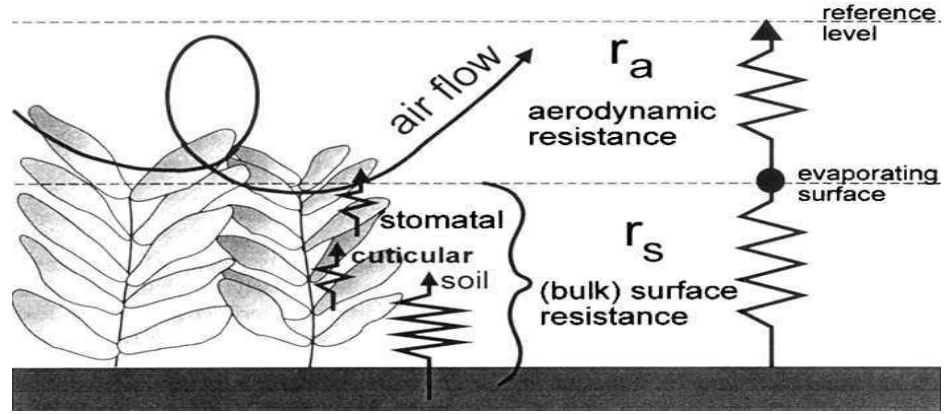
ρ_a = dry air density ($kg\ m^{-3}$)

δq = vapor pressure deficit (Pa)

r_a = Resistance of air ($m\ s^{-1}$)

r_s = Resistance of plant stoma, air ($m\ s^{-1}$)

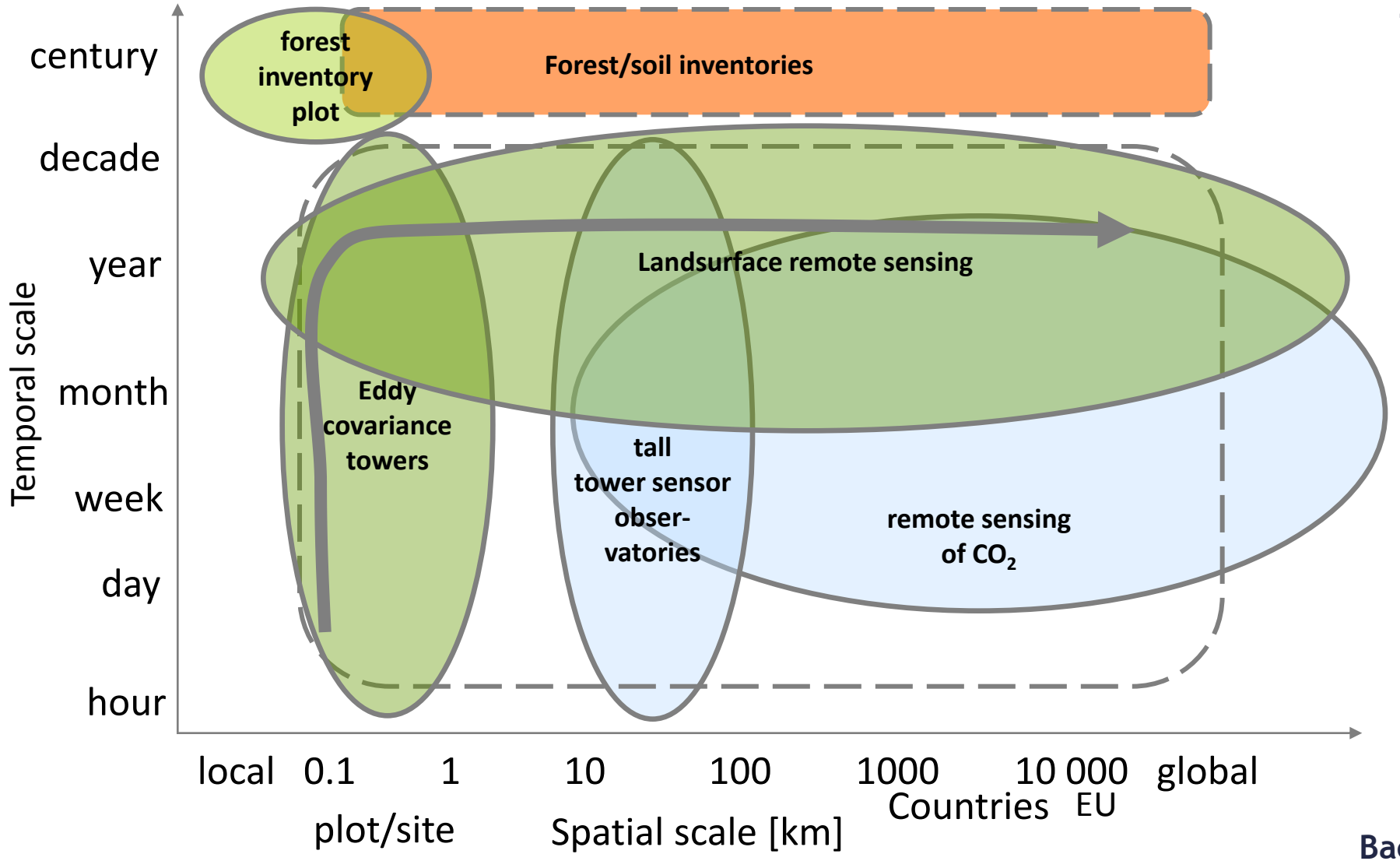
γ = Psychrometric constant ($\gamma \approx 66\ Pa\ K^{-1}$)



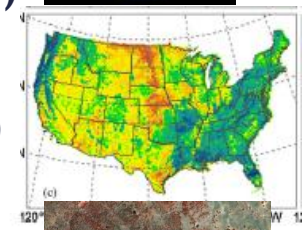
Estimating resistance/conductivity across a catchment can be tricky

- Big reduction: many inputs
- Not a matrix computation: some inputs have categorical dependencies

Massive Science Data Mashup



Globe
10,000 km (10^7 m)



Continent
1000 km (10^6 m)

Landscape
1-100 km

Canopy
100-1000 m

Plant
1-10 m

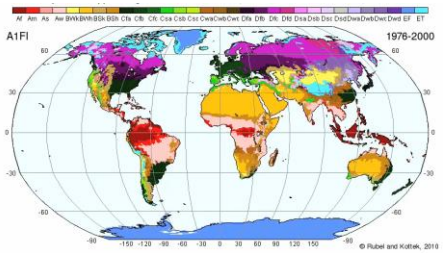
Leaf
0.01-0.1 m

Stomata
10^{-5} m

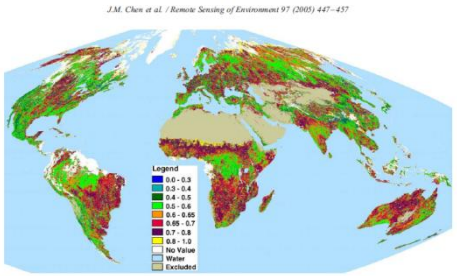
Bacteria/Chloroplast
10^{-6} m

Thanks to Markus Reichstein (Max Planck) and Dennis Baldocchi (UC Berkeley) 2010

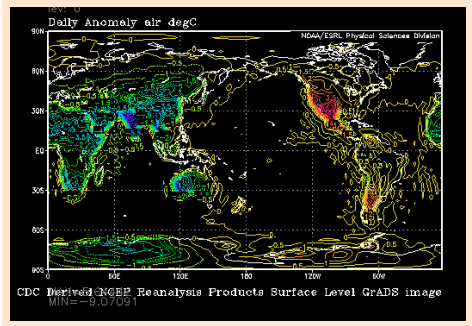
Synthesizing Imagery, Sensors, Models and Field Data



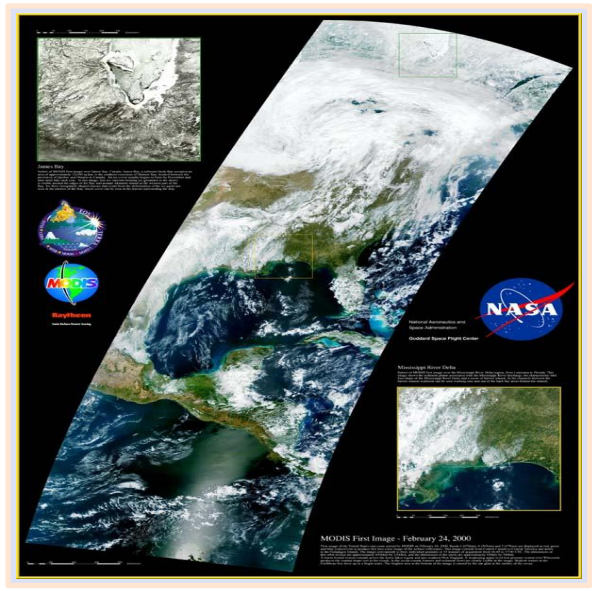
Climate classification
~1MB (1file)



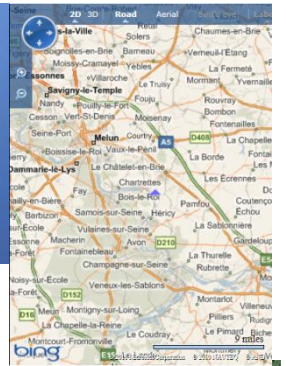
Vegetative clumping
~5MB (1file)



NCEP/NCAR ~100MB
(4K files)



NASA MODIS imagery archives
5 TB (600K files)



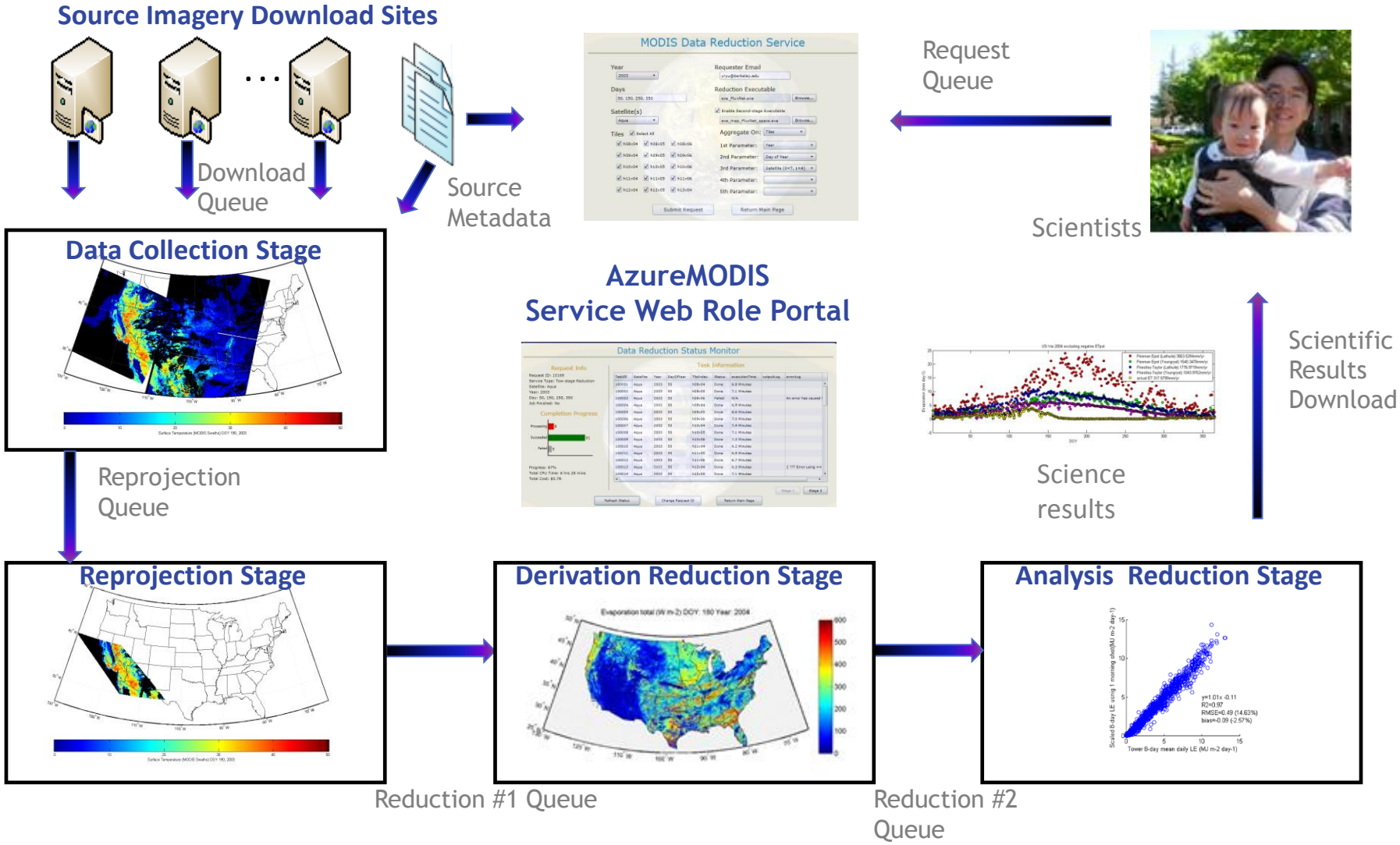
FLUXNET
Curated
sensor
dataset
30GB
(960 files)

FLUXNET
curated field
dataset
2 KB (1 file)

Sizes given are 1 US year
20 US years ~ 1 global land surface year

Four Stage Image Processing Pipeline

- Data collection stage
  - Downloads requested input tiles from NASA ftp sites
  - Includes geospatial lookup for non-sinusoidal tiles that will contribute to a reprojected sinusoidal tile
- Reprojection stage
  - Converts source tile(s) to intermediate result sinusoidal tiles
  - Simple nearest neighbor or spline algorithms
- Derivation reduction stage
  - First stage visible to scientist
  - Computes ET in our initial use
- Analysis reduction stage
  - Optional second stage visible to scientist
  - Enables production of science analysis artifacts such as maps, tables, virtual sensors

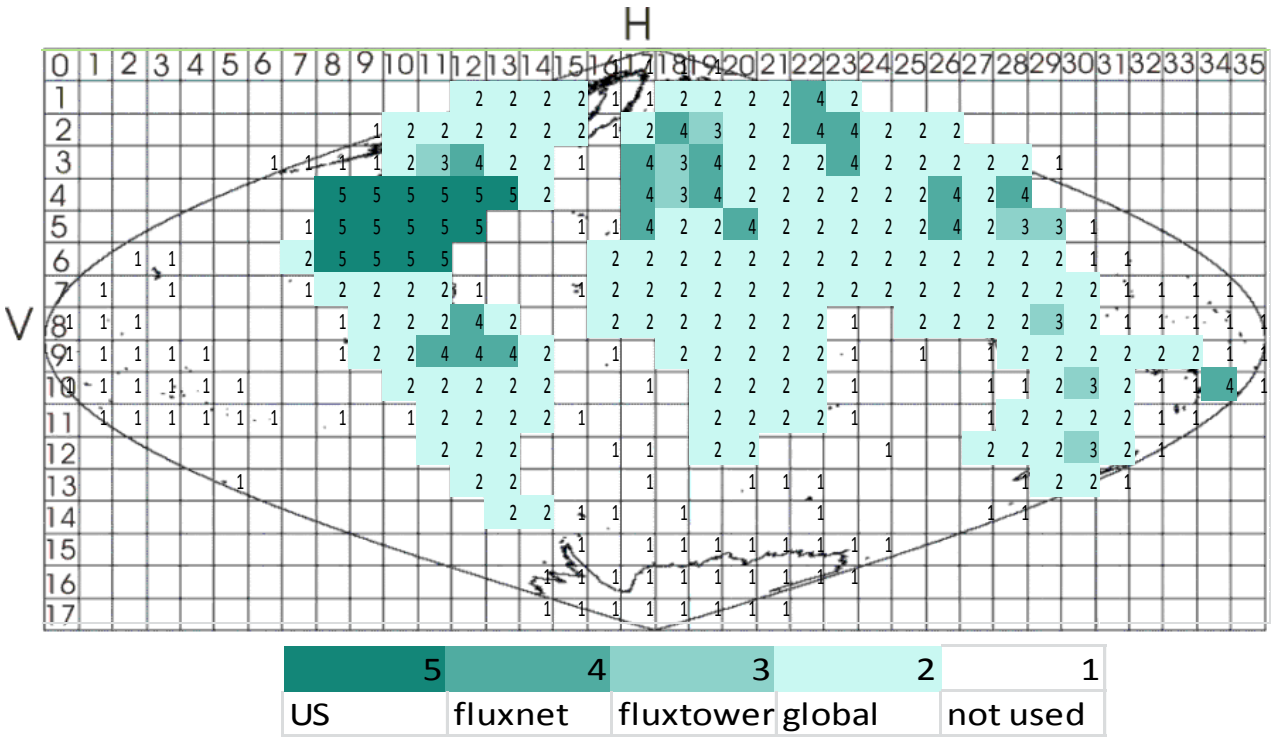


http://research.microsoft.com/en-us/projects/azure/azuremodis.aspx

Sizing a One Year Global Computation

- 194 sinusoidal cells, each cell covers 1.2x1.2 KM or a total of 11M 5 KM pixels
- 350M reprojected tiles and 14K source sinusoidal tiles
- 2.7 TB (>3.3 M files) downloaded from NASA ftp

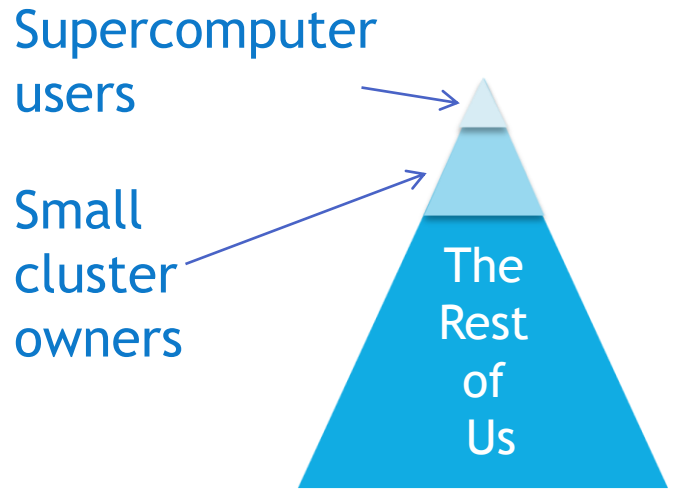
- 19 NASA download days
- 50K reprojection compute hours
- 315 TB moved across Azure fabric
- 3 download result days via IN2
- 70K derivation reduction tasks



5 seconds on the Cray Jaguar (1.75 PFLOPs), but only after the data pass through !

Implications for Traditional HPC

- Massive data, yet many small files and small IOs
  - Creates metadata pressure, very hot file contention, and exceeds may user storage quotas
- "Delightfully parallel" "Bag-of-Tasks" scheduling
  - Loosely connected data-centric dependencies are unlike parameter sweeps for PDE simulation, Monte Carlo estimation, or Map-Reduce
- Scaling computations means scaling science knowledge
  - On-demand, pay-as-you-go resource allocation a good match for algorithm development and sensitivity analysis validation
  - Enabled by seamless bridge from desktop, local cluster, and beyond



Where do you want your data?

Clouds optimized for commercial applications and textural search may be a better match?