Collaborative Data Sharing in Climate Science:

Acknowledgement, Transparency, & Access

Scott Denning

Atmospheric Science, CSU

Email Scott.Denning@ColoState.edu for a copy of this presentation

Outline

1. Carbon & climate
2. Eddy Covariance
   - Data collection, analysis, reduction
   - Network is bigger than collection of sites
   - Credit for site data
3. Multiscale Global Modeling
   - Computing
   - Source, docs, reproducibility
   - Archival & Serving Model output

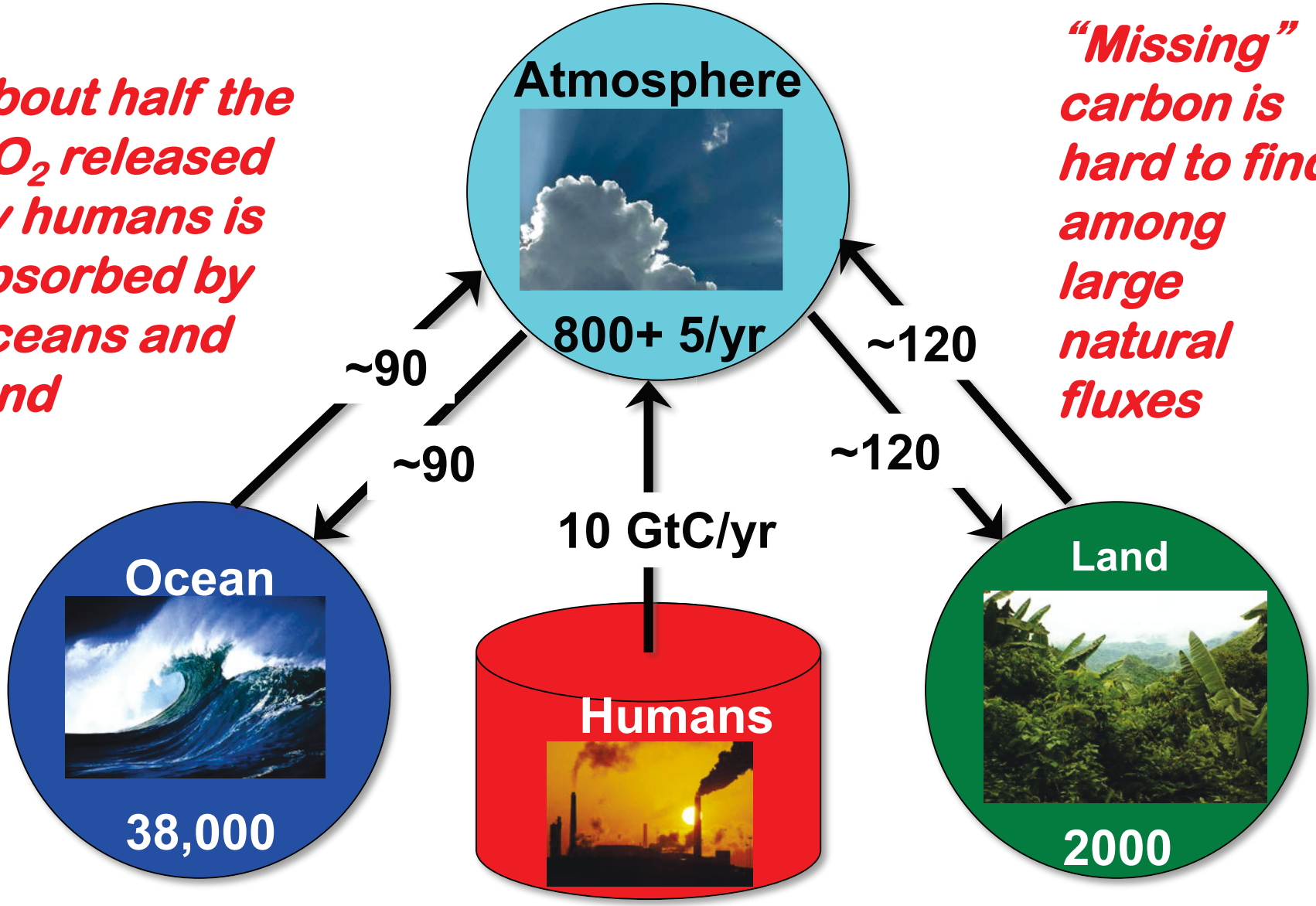
Outline

1. Carbon & climate
2. Eddy Covariance
   - Data collection, analysis, reduction
   - Network is bigger than collection of sites
   - Credit for site data
3. Multiscale Global Modeling
   - Computing
   - Source, docs, reproducibility
   - Archival & Serving Model output

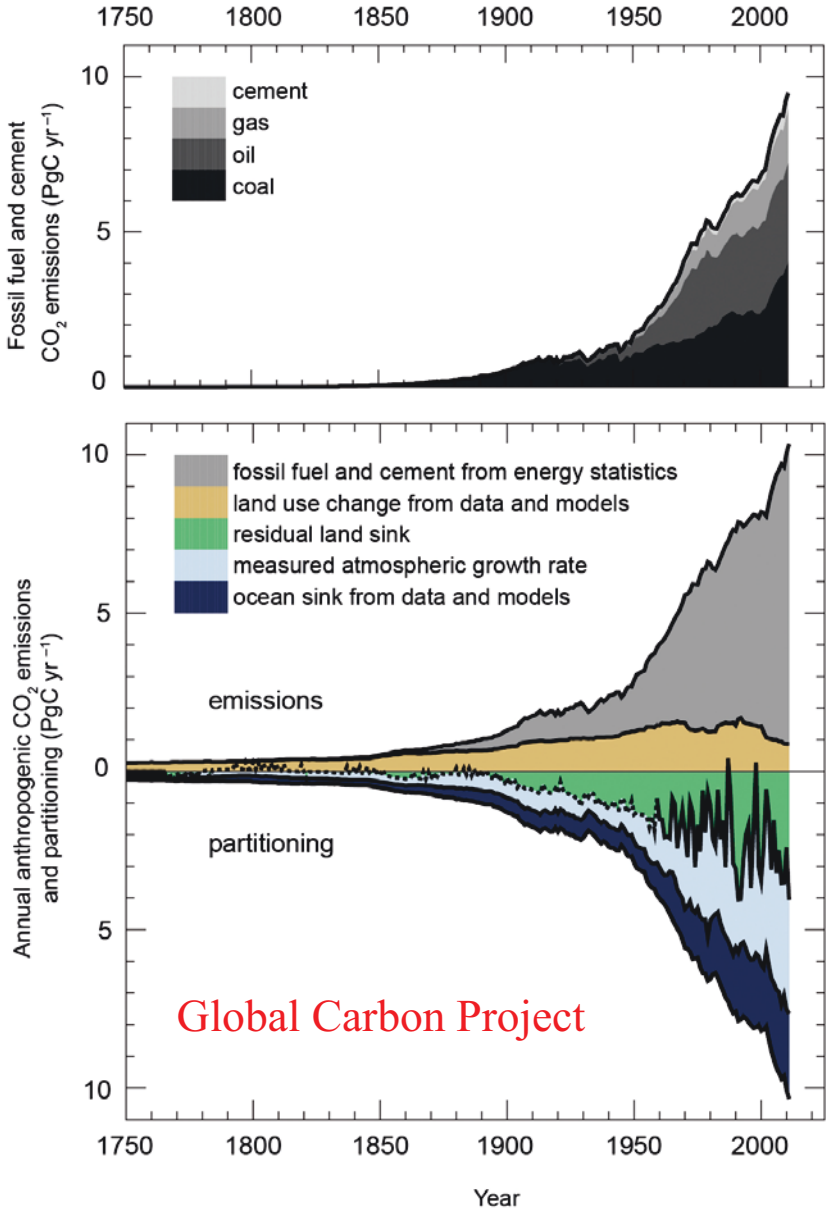
The Global Carbon Cycle

About half the CO_2 released by humans is absorbed by oceans and land

"Missing" carbon is hard to find among large natural fluxes



Carbon Sources and Sinks



- Half the carbon from fossil fuels remains in the atmosphere
- The other half goes into land and oceans
- Land sink was unexpected is very noisy, and remains unreliable in future
- Future of carbon sinks is much harder to predict than temperatures

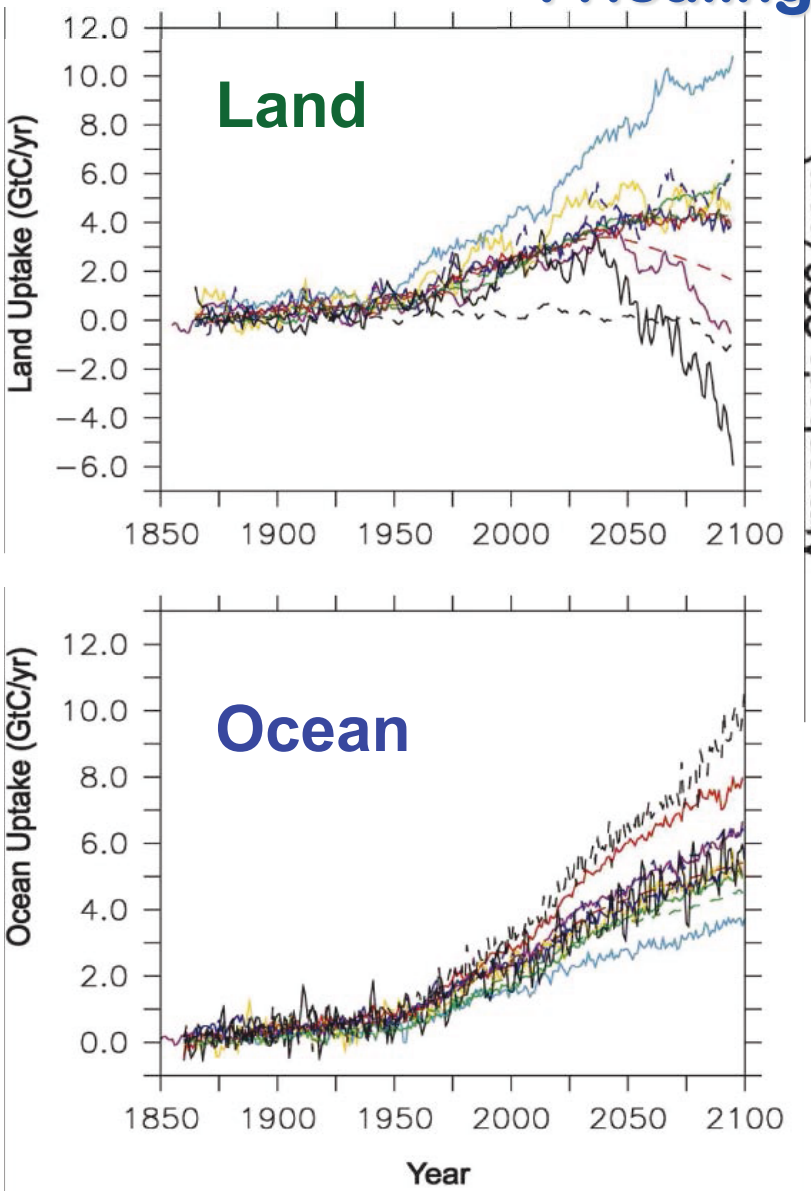
Where Has All the Carbon Gone?

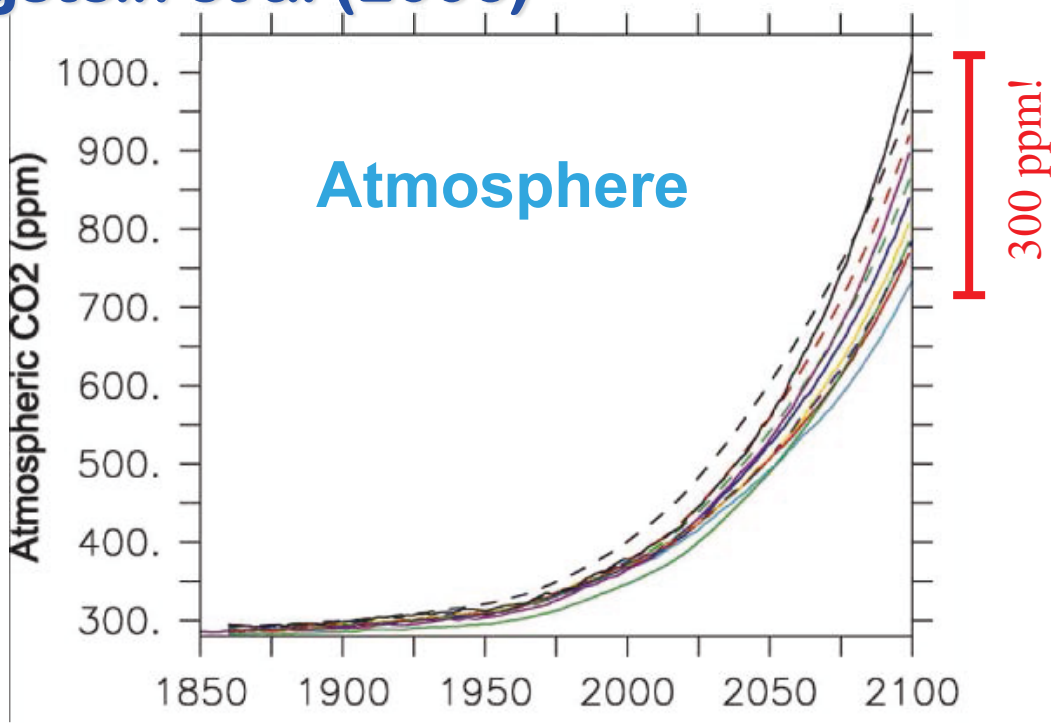
- Into the **oceans**
  - **Solubility pump** (CO_2 very soluble in cold water, but rates are limited by slow physical mixing)
  - **Biological pump** (slow "rain" of organic debris)
- Into the **land**
  - **CO_2 Fertilization** (plants eat CO2 … is more better?)
  - **Nutrient fertilization** (N-deposition and fertilizers)
  - **Land-use change** (forest regrowth, fire suppression, woody encroachment … but what about Wal-Marts?)
  - Response to **changing climate** (e.g., Boreal warming)

Carbon-Climate Feedback

Carbon-Climate Futures

Friedlingstein et al (2006)





- Coupled simulations of climate and the carbon cycle (CMIP3, C4MIP)
- Given nearly identical human emissions, different models project dramatically different futures!
- Mostly depends on CO_2 fert & temp

Outline

1. Carbon & climate
2. Eddy Covariance
   - Data collection, analysis, reduction
   - Network is bigger than collection of sites
   - Credit for site data
3. Multiscale Global Modeling
   - Computing
   - Source, docs, reproducibility
   - Archival & Serving Model output

Turbulence

Rhymes within Rhymes

- “Big whorls have little whorls,
  Which feed on their velocity;
  And little whorls have lesser whorls,
  And so on to viscosity”
  - Lewis Richardson, *The supply of energy from and to Atmospheric Eddies* 1920

- “Great fleas have little fleas
  Upon their backs to bite 'em,
  And little fleas have lesser fleas,
  And so, ad infinitum”
  - Augustus De Morgan
    (19th century mathematician, parodying Jonathon Swift, 1733)

Mass Conservation Equation for CO_2

$$\frac{d\bar{c}}{dt} = \frac{\partial \bar{c}}{\partial t} + \bar{u}\frac{\partial \bar{c}}{\partial x} + \bar{v}\frac{\partial \bar{c}}{\partial y} + \bar{w}\frac{\partial \bar{c}}{\partial z}$$

local rate of change

advection (3D)

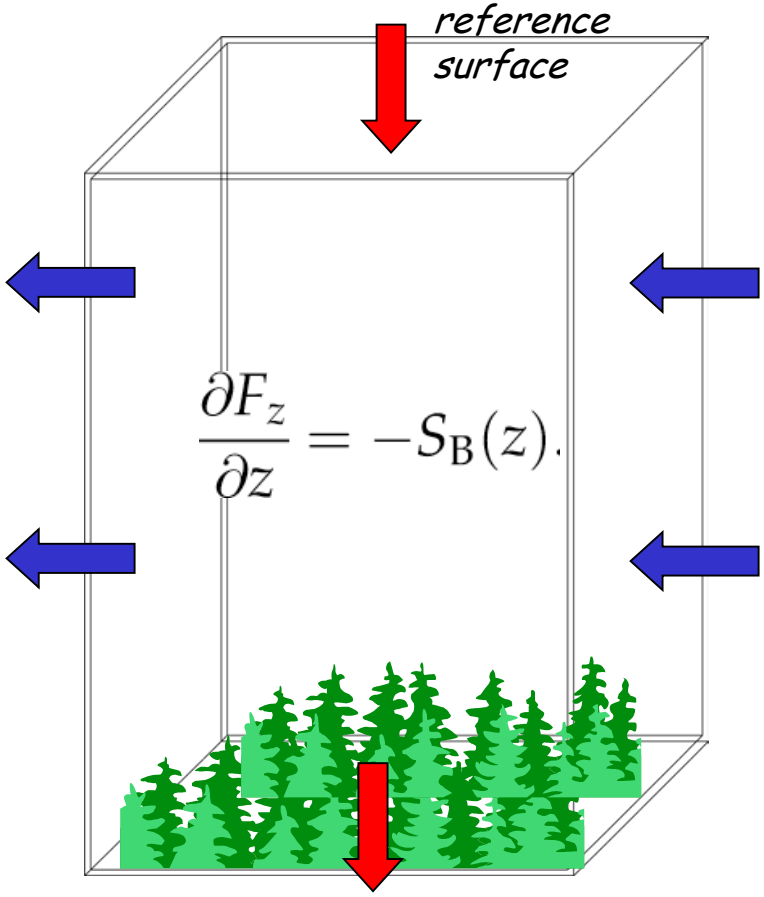
$$= -\left(\frac{\partial F_z}{\partial z} + \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + S_B(x,y,z)\right)$$

flux divergence (3D)

biological source/sink

- c is mass mixing ratio (density of CO_2/density of air)
- (u, v, w) are components of vector wind (positive toward east, toward north, and upward)
- (F_x, F_y, F_z) are components of vector mass flux of CO_2

Consider Idealized Conditions



$$\frac{d\bar{c}}{dt} = \frac{\partial \bar{c}}{\partial t} + \bar{u}\frac{\partial \bar{c}}{\partial x} + \bar{v}\frac{\partial \bar{c}}{\partial y} + \bar{w}\frac{\partial \bar{c}}{\partial z}$$

local rate of change — advection (3D)

$$= -\left(\frac{\partial F_z}{\partial z} + \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + S_B(x,y,z)\right)$$

flux divergence (3D) — biological source/sink

$$F_z(h) = F_z(0) - \int_0^h S_B(z)\,dz$$

- No change in CO_2 with time (term 1 ~ 0)
- Wind is steady
- Surface is horizontally homogeneous
- Terrain is flat
- Advection ~ 0
- Horizontal fluxes don't diverge … term 3 ~ 0

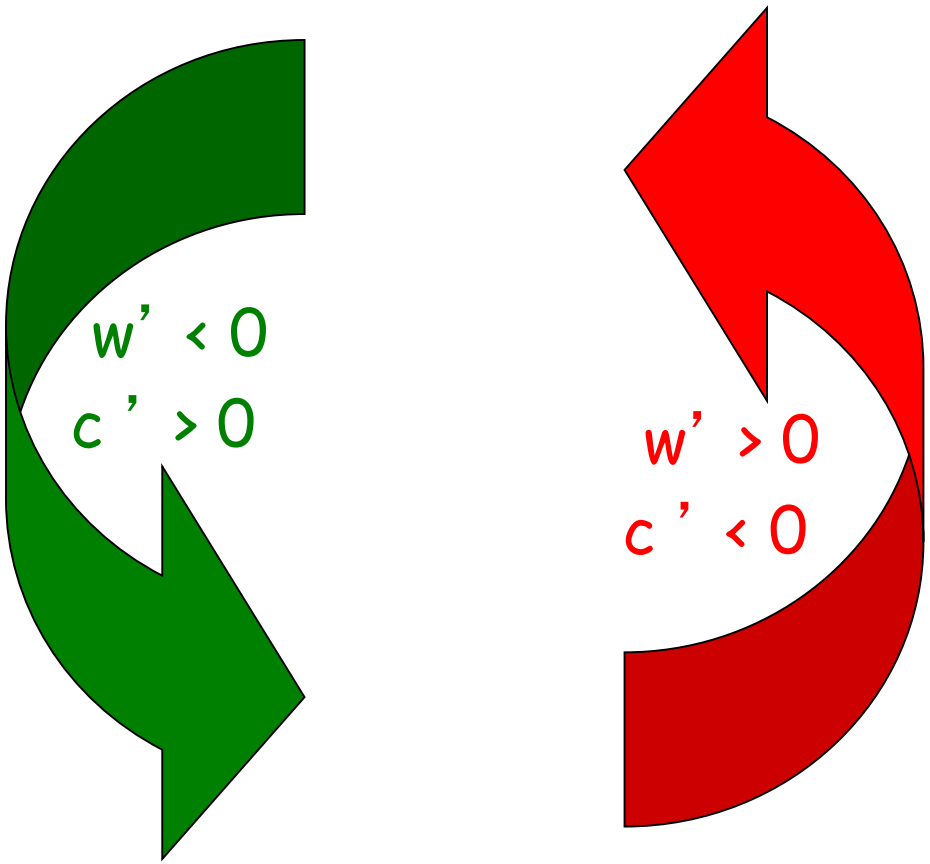
Sonic Anemometer

- Measures elapsed time for sound pulses to cross air in 3D
- Speed of sound is a known function of temperature
- Relative motion determined accurately in 3D
- Very fast instrument response time

Eddy Covariance

Turbulent CO_2 Fluxes

$$w \equiv \overline{w} + w' \qquad c \equiv \overline{c} + c'$$



photosynthesizing ecosystem

- Imagine a turbulent eddy over an active ecosystem
- Updrafts are systematically depleted in CO_2 relative to downdrafts

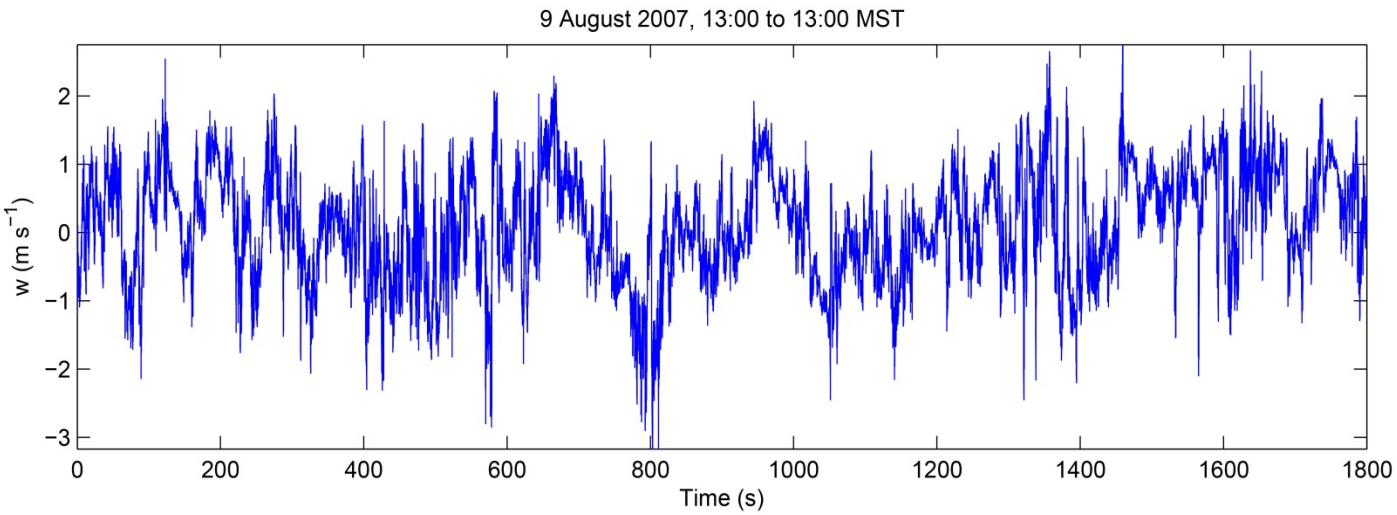
- Updraft: $\overline{w'c'} < 0$
- Downdraft: $\overline{w'c'} < 0$

Average over eddy:

$$\overline{w'c'} < 0$$

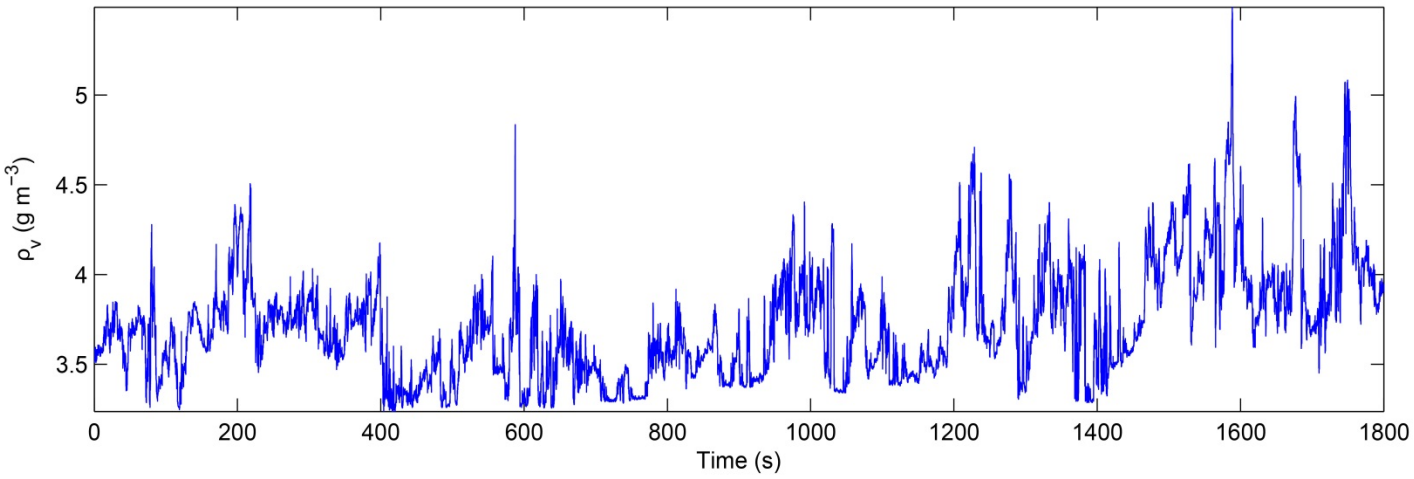
Real EC Data

vertical wind



20 measurements per second

gas concentration



36,000 pairs per half-hour flux

"Data Reduction"

$$\left\{\int_0^z \overline{\frac{\partial \varrho_c}{\partial t}}\, dz' - \overline{\chi}_c(z)\int_0^z \overline{\frac{\partial \varrho_d}{\partial t}}\right\} +$$

$$\left\{\int_0^z \nabla_H \bullet (\overline{\mathbf{u}\varrho_d}\,\overline{\chi}_c + \overline{\varrho}_d\,\overline{\mathbf{u}'\chi_c'})\, dz' - \overline{\chi}_c(z)\int_0^z \nabla_H \bullet (\overline{\mathbf{u}\varrho_d})\, dz'\right\} + \overline{\varrho}_d(z)\,\overline{w'\chi_c'}(z) =$$

$$\left\{\int_0^z \frac{\overline{S}_c}{m_c}\, dz' + \overline{J}_c(0)\right\} - \overline{\chi}_c(z)\left\{\int_0^z \frac{\overline{S}_d}{m_d}\, dz' + \overline{J}_d(0)\right\} +$$

$$\overline{\varrho}_d(0)\,\overline{w'\chi_c'}(0) + \overline{w\varrho_d}(0)[\overline{\chi}_c(0) - \overline{\chi_c}(z)]$$

The terms of the last expression are easily identified. They are:

$\left\{\int_0^z \overline{\frac{\partial \varrho_c}{\partial t}}\, dz' - \overline{\chi}_c(z)\int_0^z \overline{\frac{\partial \varrho_d}{\partial t}}\right\}$ = 'Effective storage'.

$\left\{\int_0^z \nabla_H \bullet (\overline{\mathbf{u}\varrho_d}\,\overline{\chi}_c + \overline{\varrho}_d\,\overline{\mathbf{u}'\chi_c'})\, dz' - \overline{\chi}_c(z)\int_0^z \nabla_H \bullet (\overline{\mathbf{u}\varrho_d})\, dz'\right\}$ = 'Horizontal advection'.

$\overline{\varrho}_d(z)\,\overline{w'\chi_c'}(z)$ = 'Eddy covariance flux' or 'Turbulent surface exchange flux'.

$\left\{\int_0^z \frac{\overline{S}_c}{m_c}\, dz' + \overline{J}_c(0)\right\}$ = 'Net Ecosystem Exchange' or 'NEE' if CO_2 is the constituent.

$\overline{\chi}_c(z)\left\{\int_0^z \frac{\overline{S}_d}{m_d}\, dz' + \overline{J}_d(0)\right\}$ = 'Dry air source/sink term' or 'Dry air source correction term'.

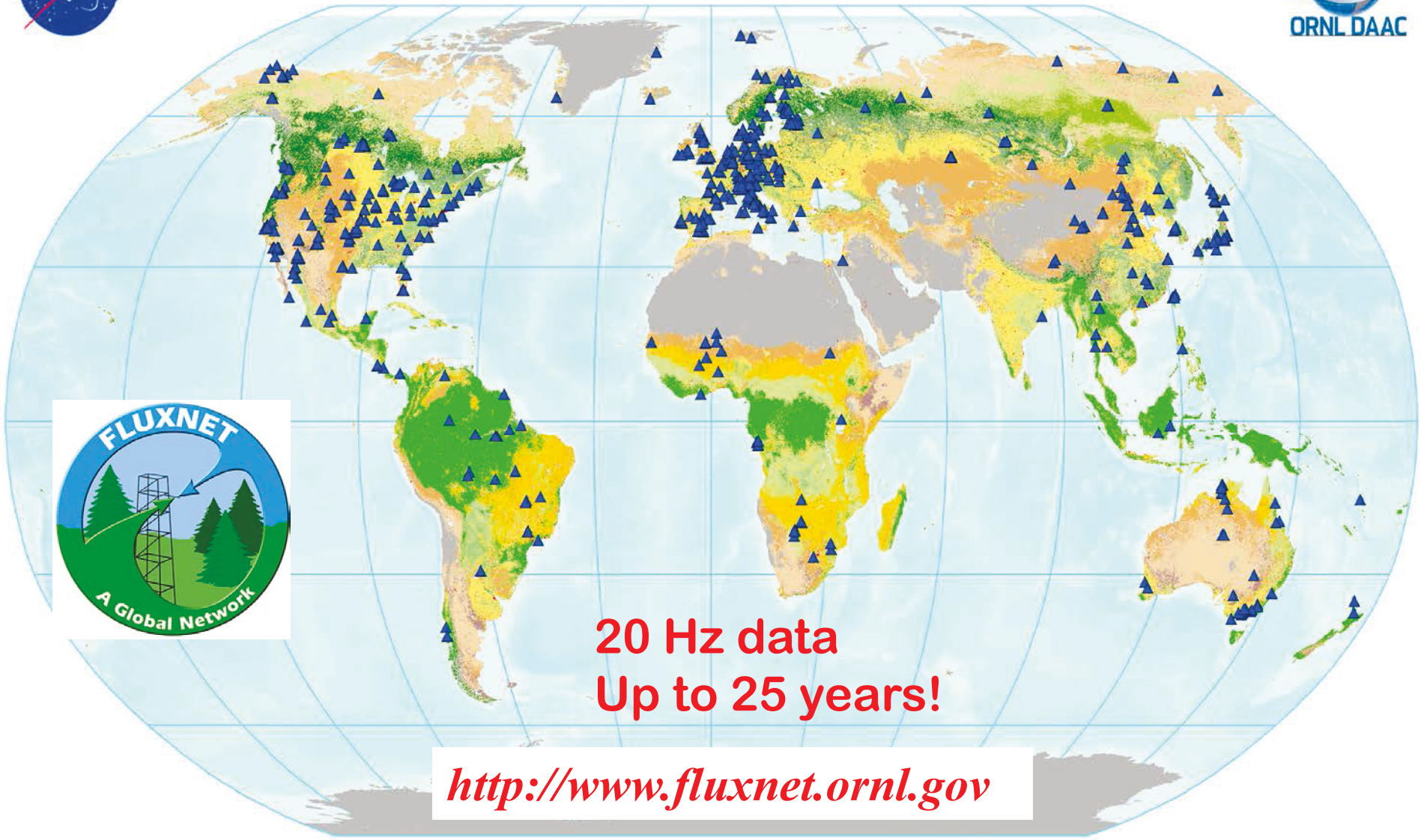
$\overline{\varrho}_d(0)\,\overline{w'\chi_c'}(0)$ = 'Enhanced soil diffusion term' or 'Pressure pumping term'.

$\overline{w\varrho_d}(0)[\overline{\chi}_c(0) - \overline{\chi_c}(z)]$ = 'Dry air flux lower boundary condition'.

FLUXNET

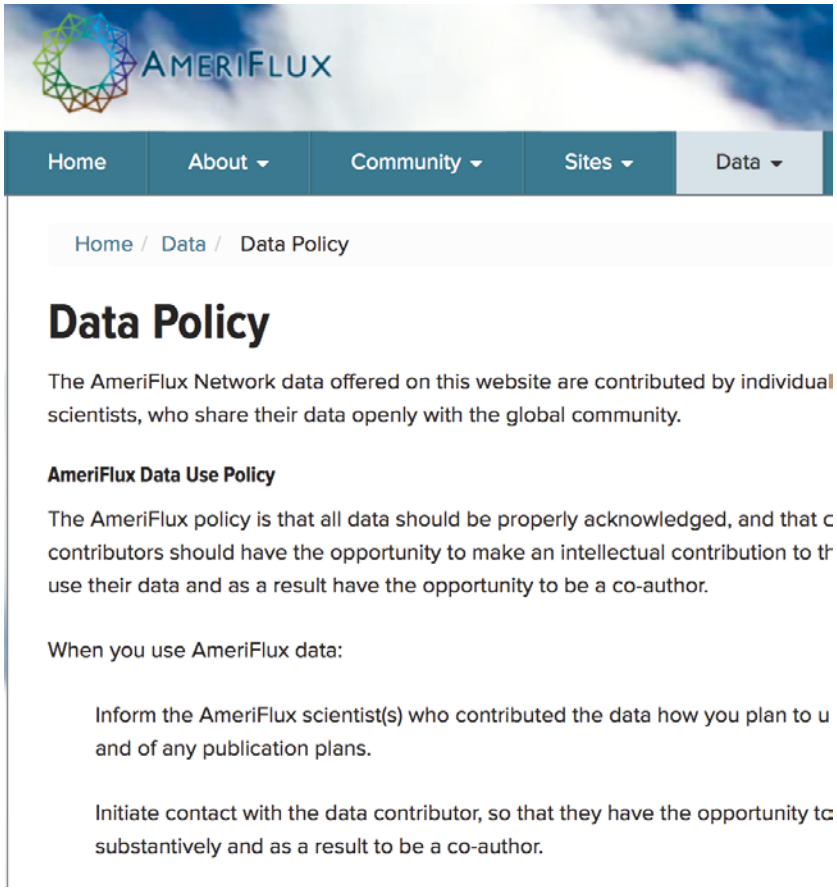
July 2015

723 Sites



AmeriFlux Data Policy

- Included verbatim at the head of every ASCII file on the AmeriFlux ftp site.
- Downloading data form the site also includes an "I agree" button that automatically notifies site PI of download



AmeriFlux Data Policy

The AmeriFlux Network data offered on this website are contributed by individual AmeriFlux scientists, who share their data openly with the global community.

The AmeriFlux policy is that all data should be properly acknowledged, and that data contributors should have the opportunity to make an intellectual contribution to the papers that use their data and as a result have the opportunity to be a co-author.

AmeriFlux Data Policy

When you use AmeriFlux data:

- Inform the AmeriFlux scientist(s) who contributed the data how you plan to use the data and of any publication plans.
- Initiate contact with the data contributor, so that they have the opportunity to contribute substantively and as a result to be a co-author.
- Acknowledge AmeriFlux data by citing the relevant DOI or paper(s), and/or acknowledging funding for the site support. If the data download was not accompanied by the preferred acknowledgment language, ask the site principal investigator.
- Acknowledge the AmeriFlux data resource as "funding for AmeriFlux data resources was provided by the U.S. Department of Energy's Office of Science."

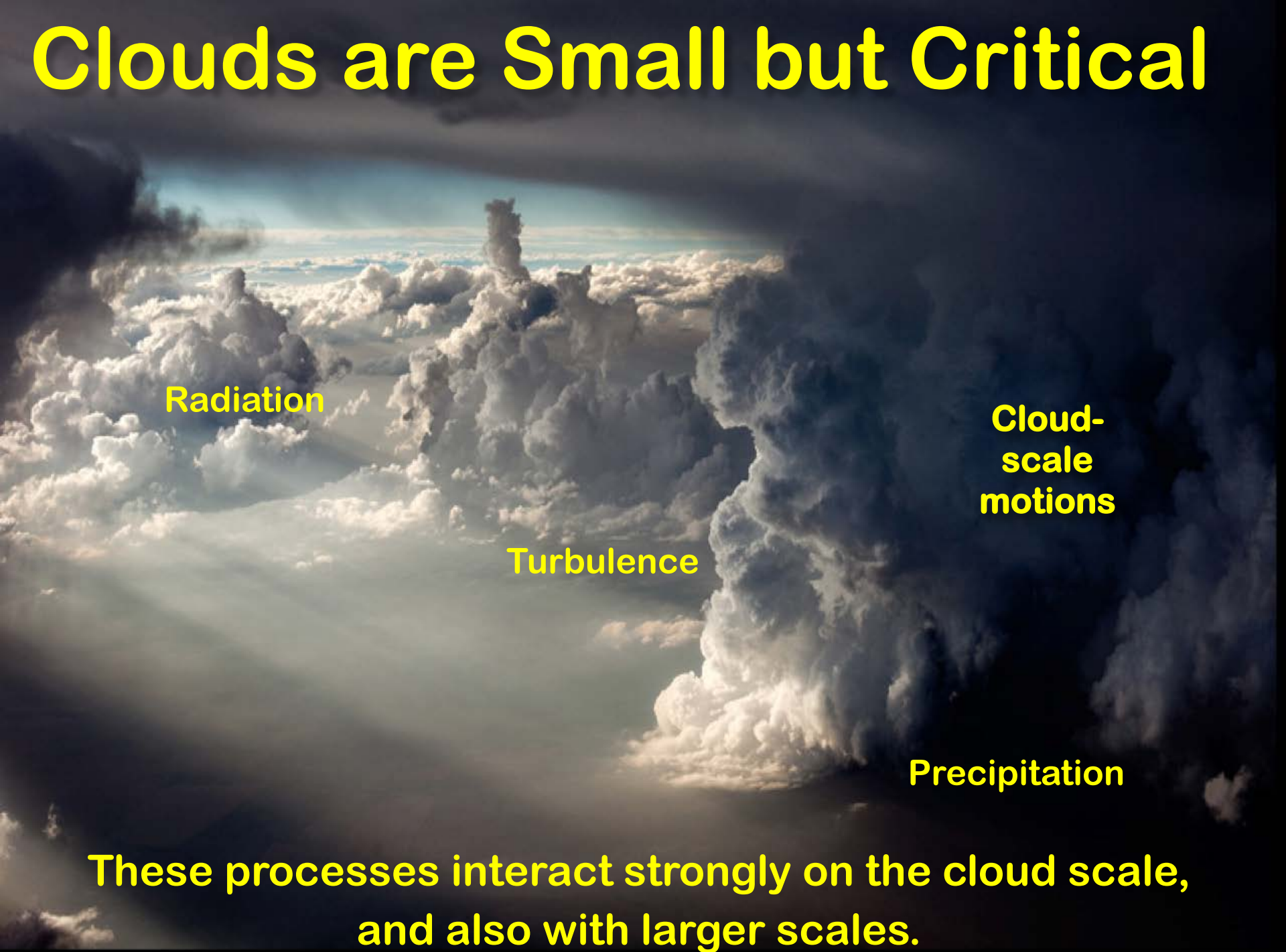
Outline

1. Carbon & climate
2. Eddy Covariance
   - Data collection, analysis, reduction
   - Network is bigger than collection of sites
   - Credit for site data
3. **Multiscale Global Modeling**
   - **Computing**
   - **Source, docs, reproducibility**
   - **Archival & Serving Model output**

Climate-Carbon Models

- Reproducibility, Transparency, Traceability
  - Document all model algorithms
  - Document source code
  - Input data, libraries, compilers, executables?
- Evaluate model against real-world observations
- Publish results
- *Archive & serve model output (forever?)*

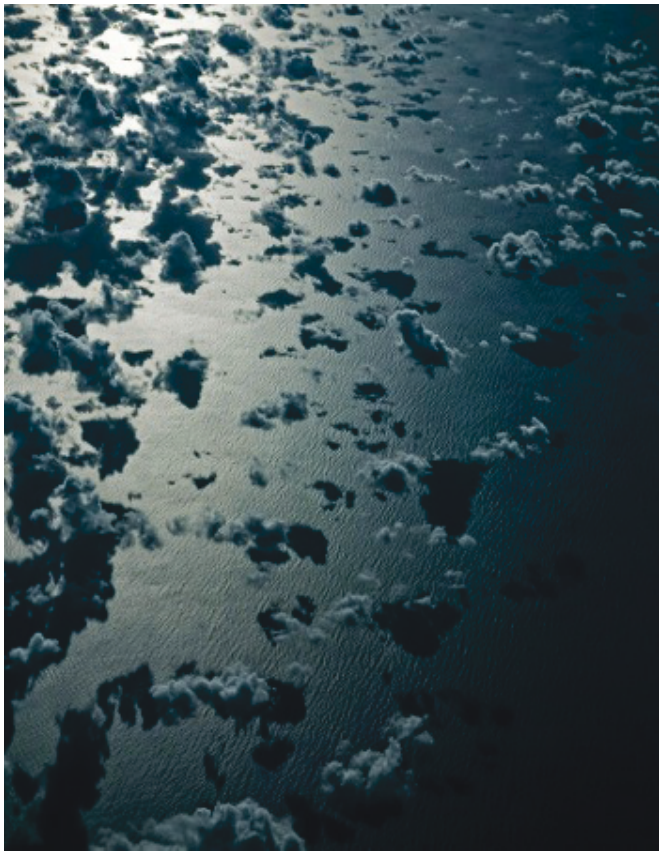
Clouds are Small but Critical



These processes interact strongly on the cloud scale, and also with larger scales.

Resolve clouds?

- Modest increases in resolution don't help. A big increase is needed.
- Global cloud-resolving models are still too expensive
- Super-parameterization is a relatively affordable alternative approach.



would need a factor of ~ 1 million x more computing!

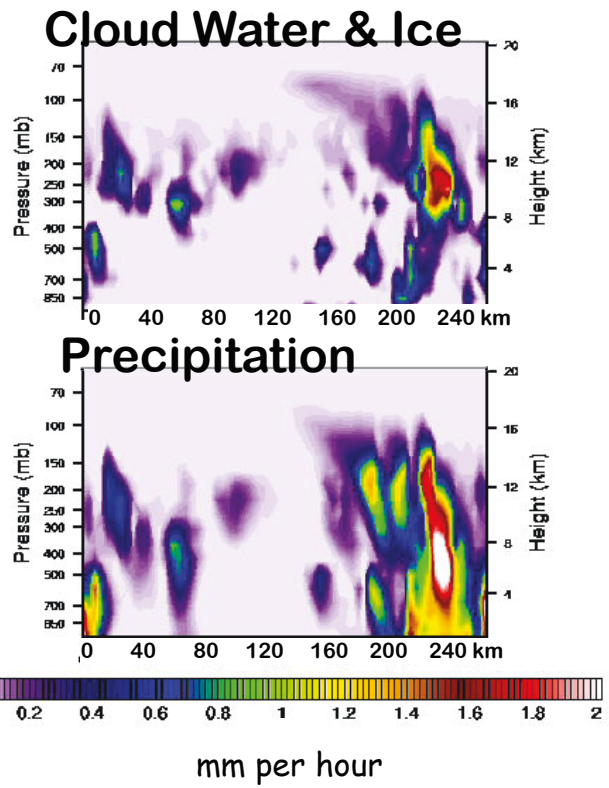
SuperParameterization (SP)

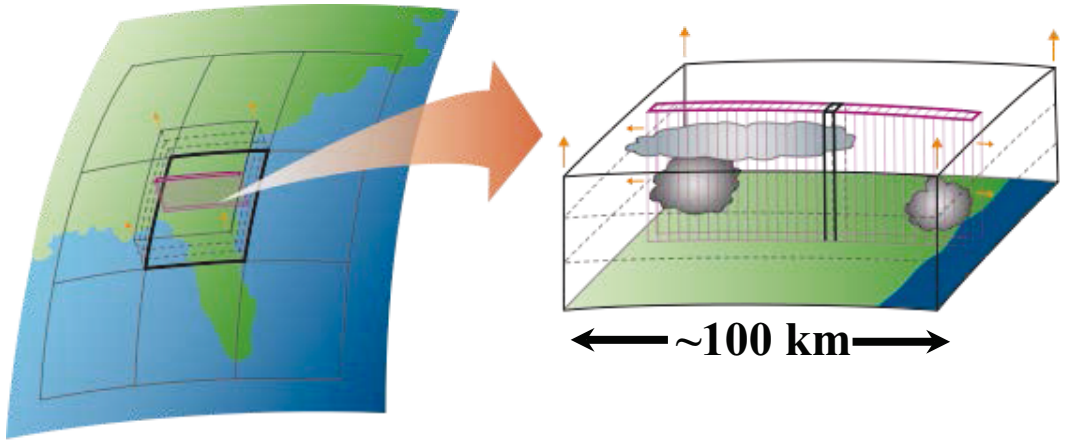
Replace

$$y = mx + b$$

with

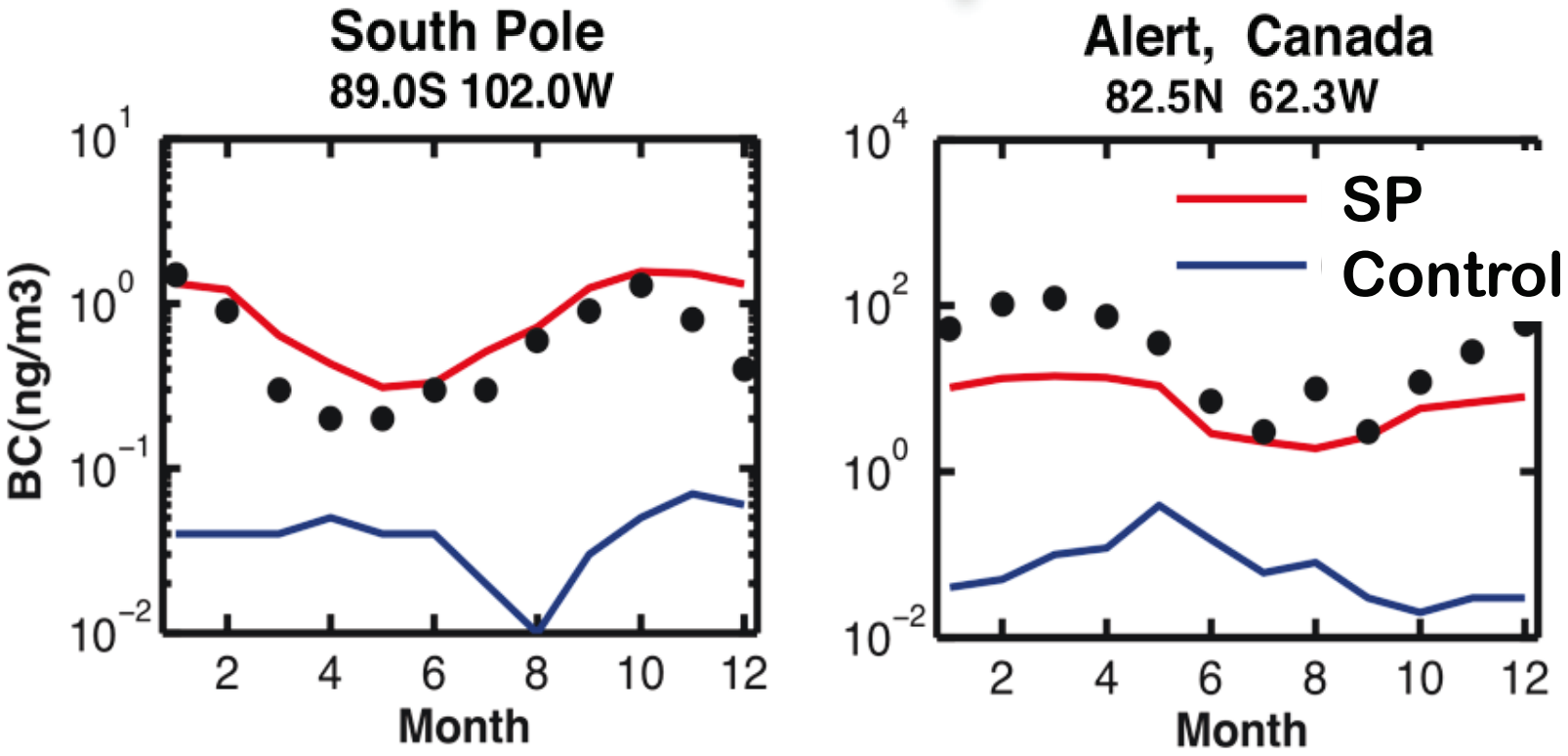
$$F = m\,a$$





- Completely remove all subgrid-scale parameterizations of clouds, precipitation, radiation, turbulence, from GCM
- Run a separate cloud-resolving model (CRM) in every column instead
- All subgrid-scale processes happen in the CRMs
- All communication among CRMs happens in the GCM

Black carbon (soot) near the poles



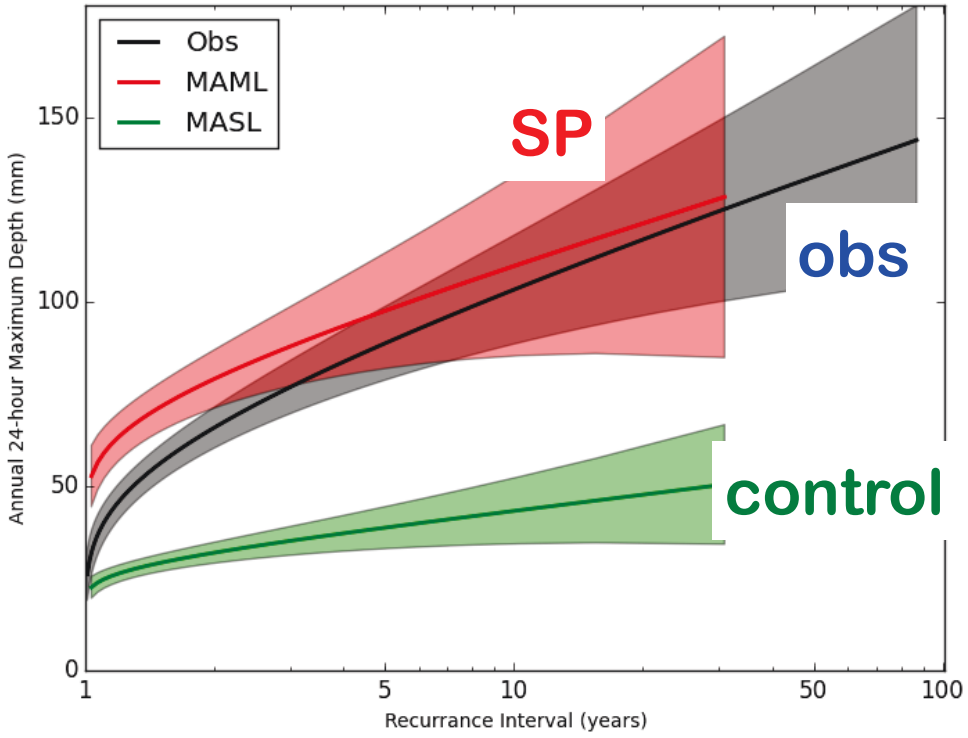
NOTE LOG SCALE!

SP is dramatically more realistic!

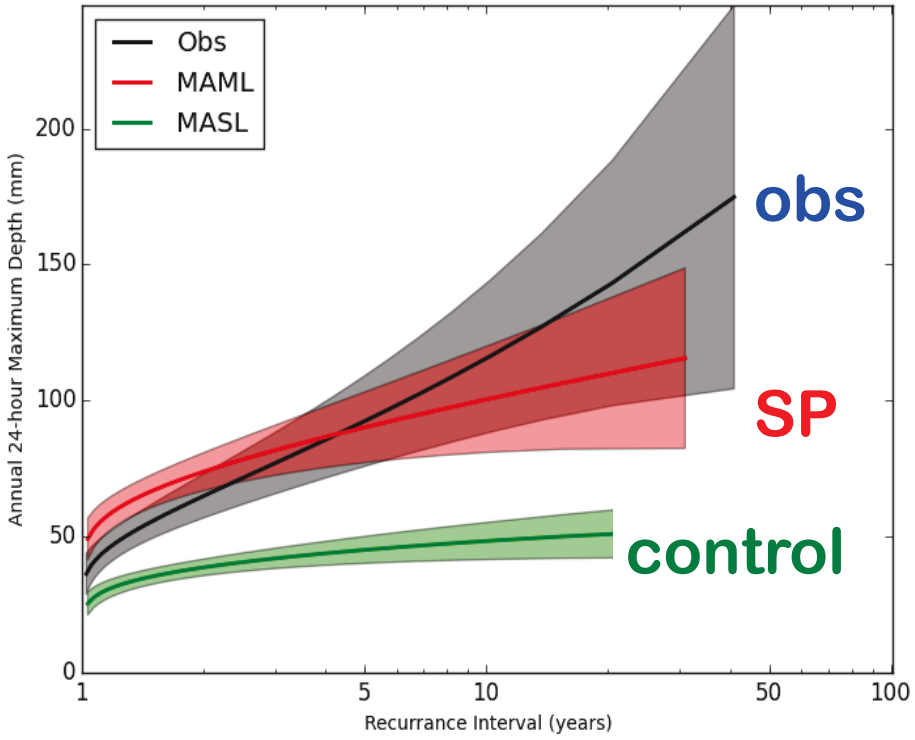
Slide from Minghuai Wang of PNNL

More Intense Rainfall

Chicago



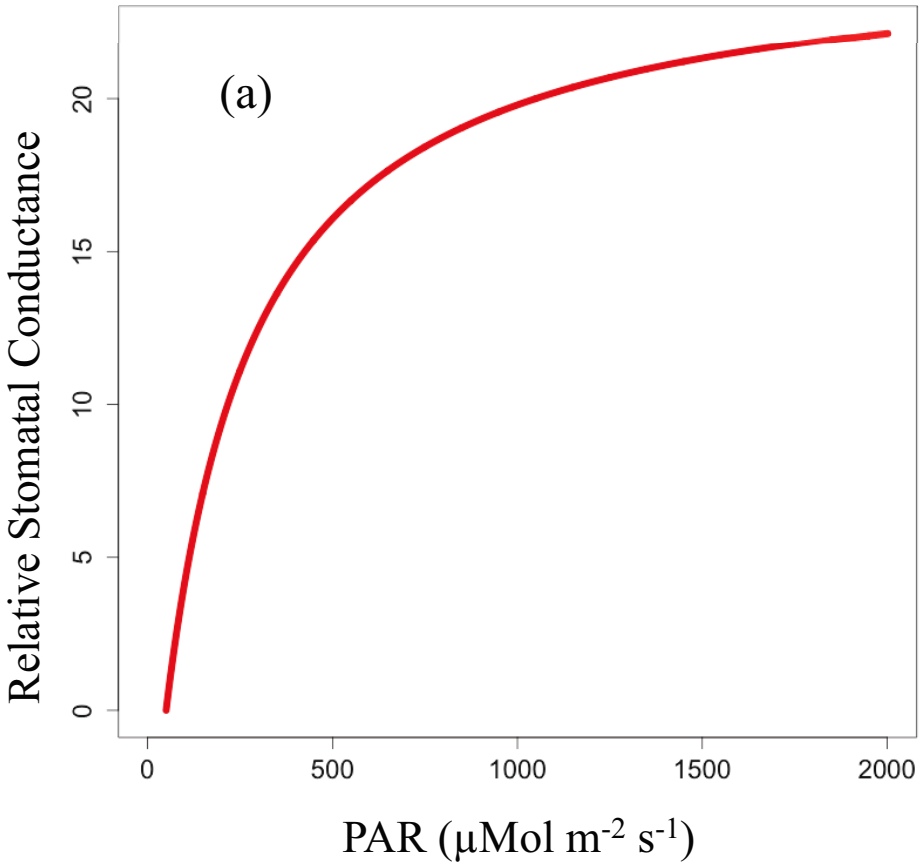
Atlantic City



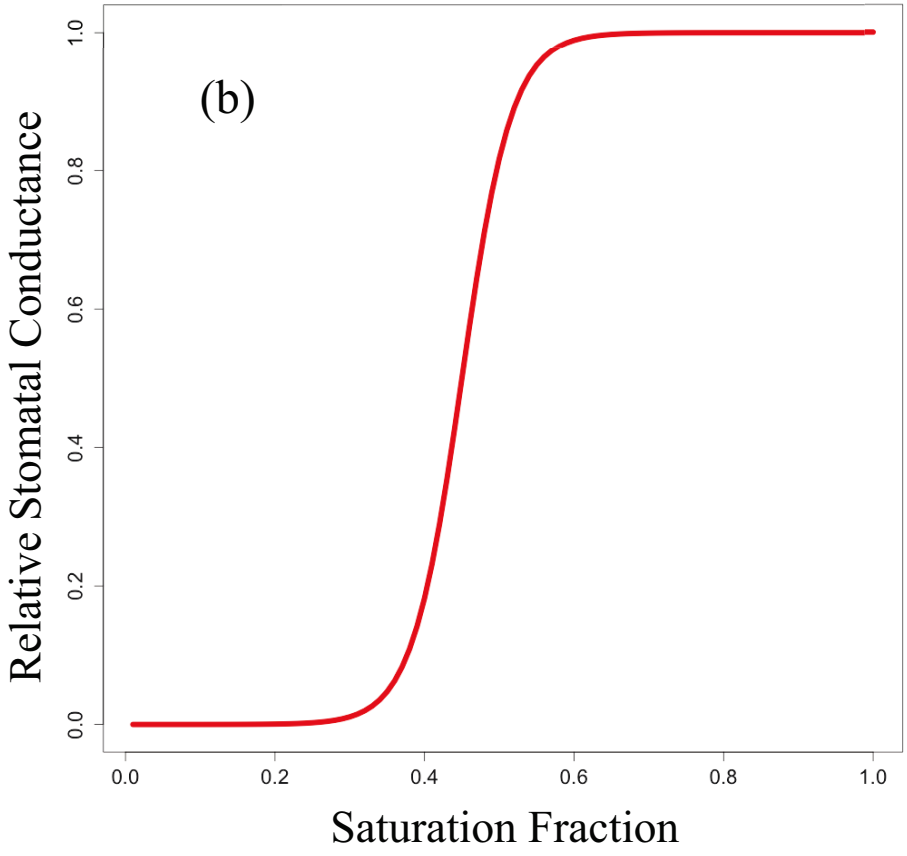
Sampling the physics produces dramatically more realistic result

Nonlinear Plants

Photosynthetic Light Response

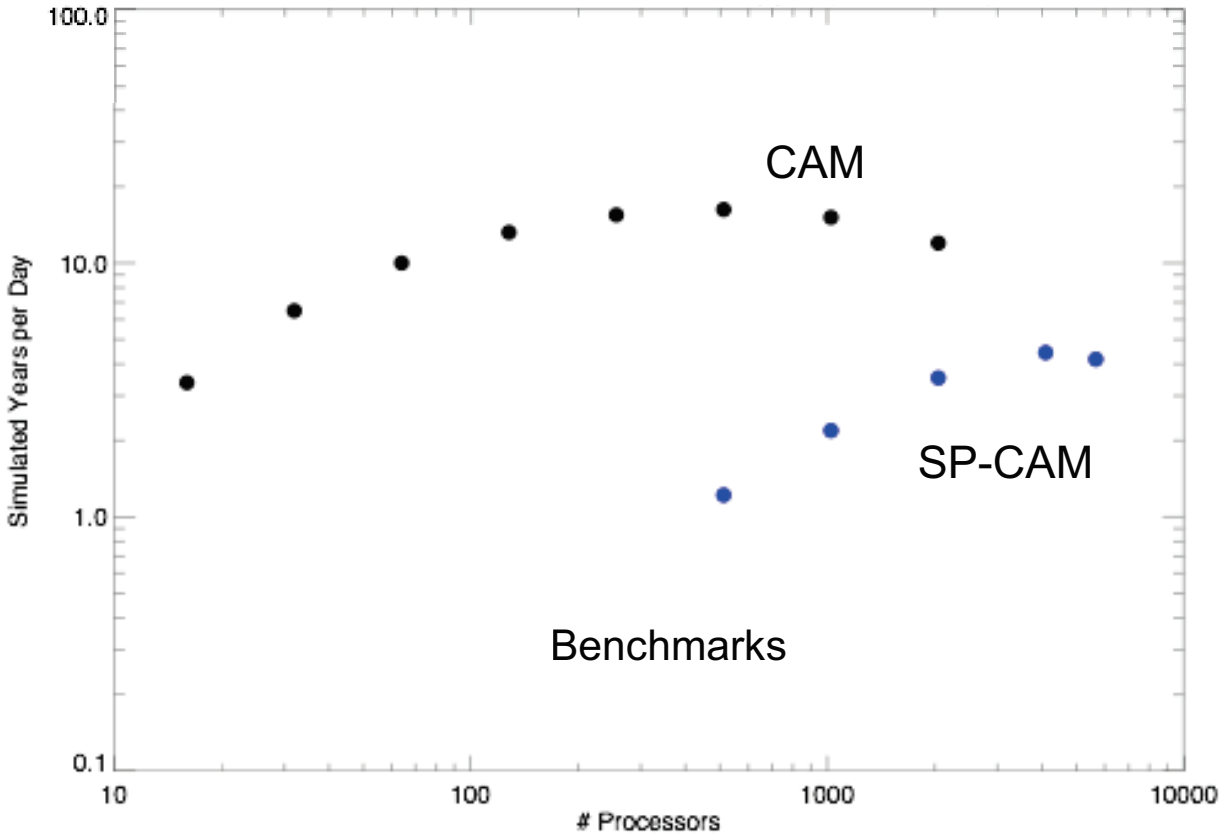


Soil Moisture Stress



$$f(\bar{x}) \neq \overline{f(x)}$$

Multiscale Computing



Supercomputer centers

NSF, DOE, NASA

- **Bad news: 200x more computing than standard model!**
- **Good news: 200 is a lot less than 1,000,000**
- **Embarrassingly parallel — excellent scaling**

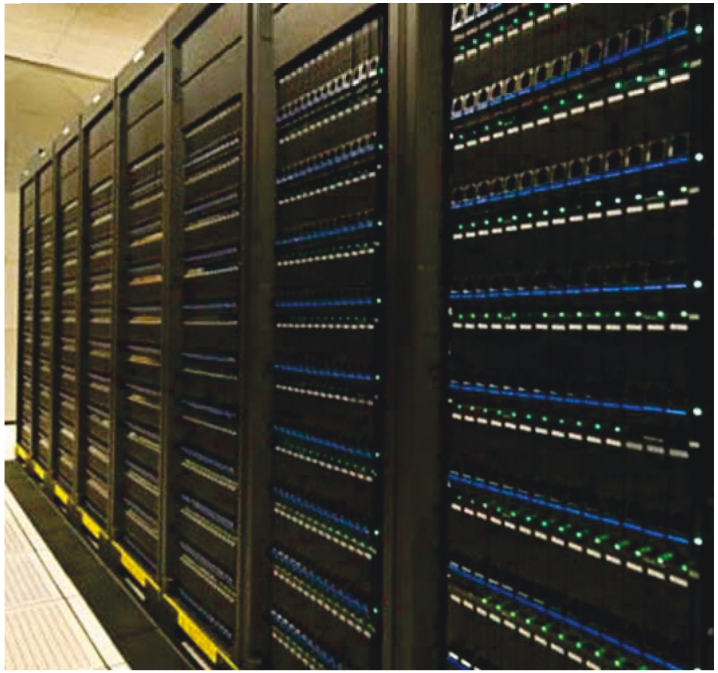
Reproducibility & Transparency

In principle: Save and document everything needed to reproduce the experiment

- Source code
- Input data
- Parameter settings
- Compilers and compiler options
- Operating system?
- Hardware?

Archive & *Serve* Model Output?

- Save ~ 100 variables
- Grid = 1° x 1° x 47 levels = 3 M cells
- 32 subgrid-scale cloud-resolving columns
- 10 second timestep (CFL stability)
- 100 year climate simulation
- *= 22 exabytes! (uh oh)*



How to Reduce Output?

- Average over time and space!
  - But be careful! $f(\overline{x}) \neq \overline{f(x)}$
- Select times, places, variables of interest
- Save states in "restart" files to enable re-runs of short periods with more complete diagnostics
- Save entire package (code, input data, parameters, compilers?) to allow reproducibility

Model Output "Service"

- Large storage requirement
- Backup?
- Bandwidth (fiber? FedEx?)
- Standardization and self-documentation of output files (netcdf, etc) & analysis software
- Bring analysis to data instead of vice versa
- Documentation
- Discovery!
- Persistence

Earth System Grid
at NCAR

Climate Data at the National Center for Atmospheric Research

Find and download climate data and analysis tools

Popular Global Climate Models

Community Earth System Model CESM

Community Earth System Model (CESM/CCSM4)

CESM1 CAM5 BGC 20C + RCP8.5 Large Ensemble

CESM1 CAM5 BGC RCP4.5 Medium Ensemble

CESM1 Last Millennium Ensemble

Summary

- Interactions between carbon and climate are complex and important for the future
- Observations are critical, hard to collect & synthesize, effort must be acknowledged
- Modeling across spatial scales very technically challenging, making progress!
- Unique problems of transparency, reproducibility, discovery, archival, & access