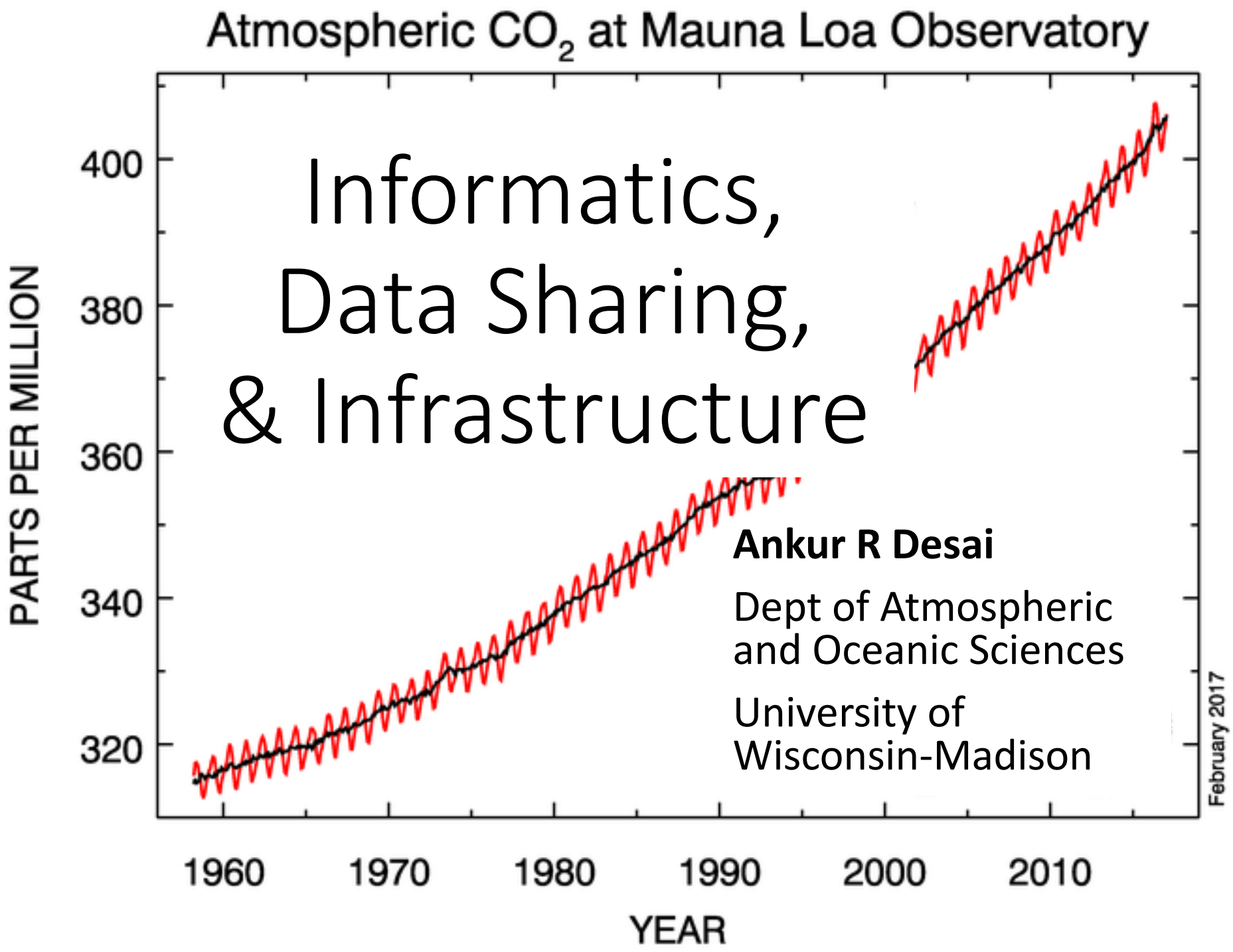


Informatics, Data Sharing, & Infrastructure

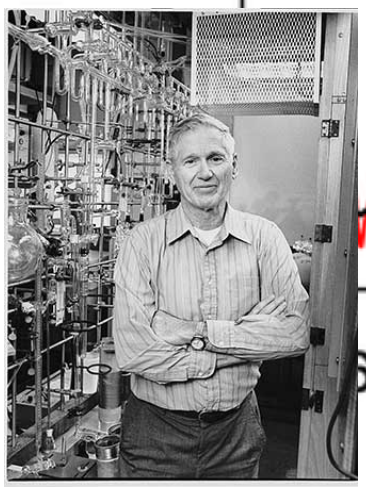
Ankur R Desai

Dept of Atmospheric and Oceanic Sciences

University of Wisconsin-Madison

Take Homes

- Big data is not just about data volume
  - Data/code diversity, accessibility, and metadata matter
- Tackling challenges in informatics is a key to solving the scientific reproducibility crisis
  - Big data is really about the people, ethics, networks
- UW is well-positioned to be a leader here, especially in agricultural/environmental big data challenges
  - If it invests in it



PARTS PER MILLION

400

380

360

340

CO_2 CONC. (p.p.m.)

315

310

● MAUNA LOA
△ AIRCRAFT (HIGHER AIR TEMPERATURE)
× AIRCRAFT (LOWER AIR TEMPERATURE)
○ ICE FLOE STATIONS

9 10 11 12 1 2 3 4 5 6 7 8 9 10 11 12 1 2 3 4 5 6 7 8 9 10 11 12 1 2 3

1957 1958 1959 1960

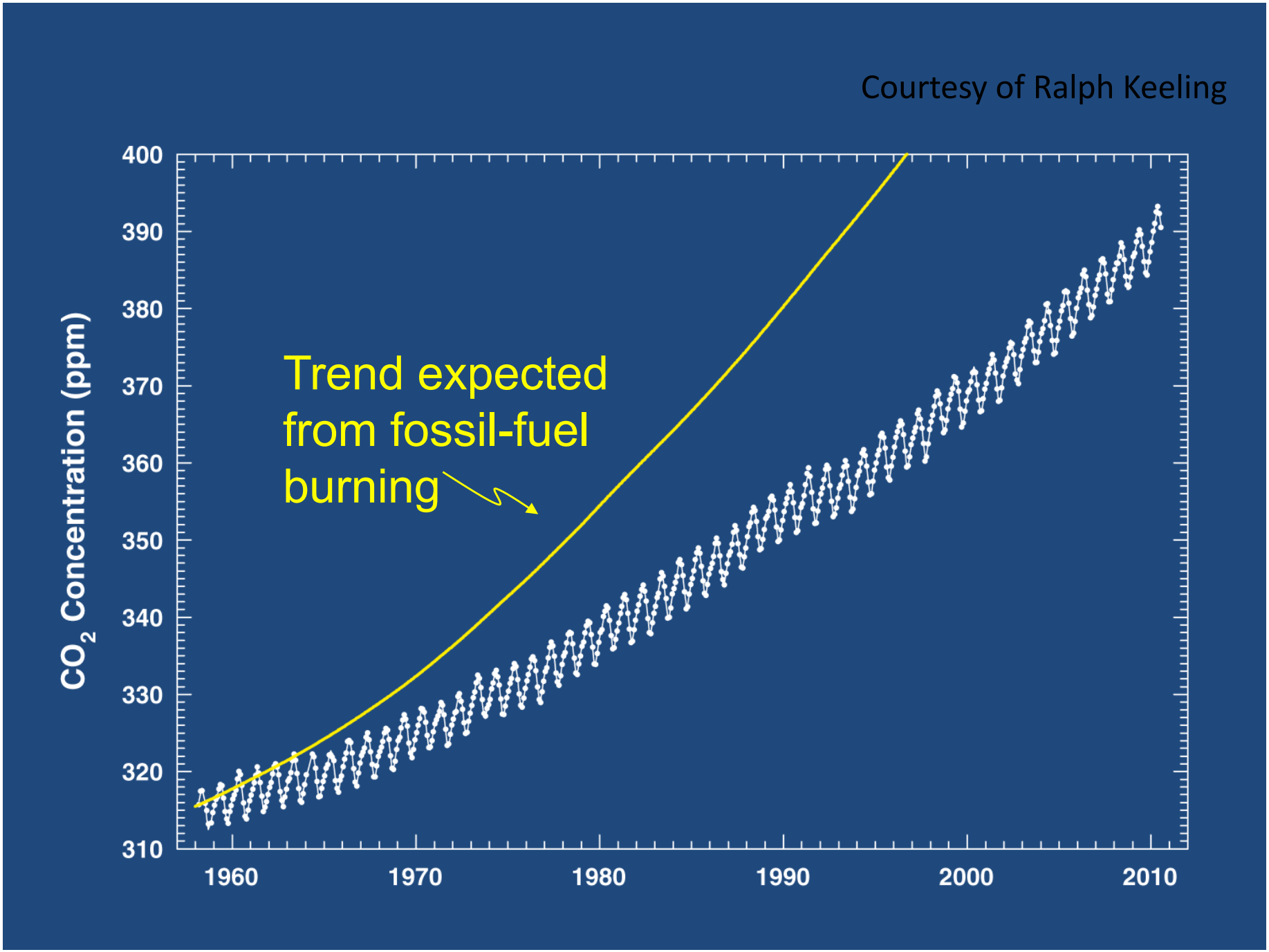
MONTH AND YEAR

60 1970 1980 1990 2000 2010

YEAR

February 2017

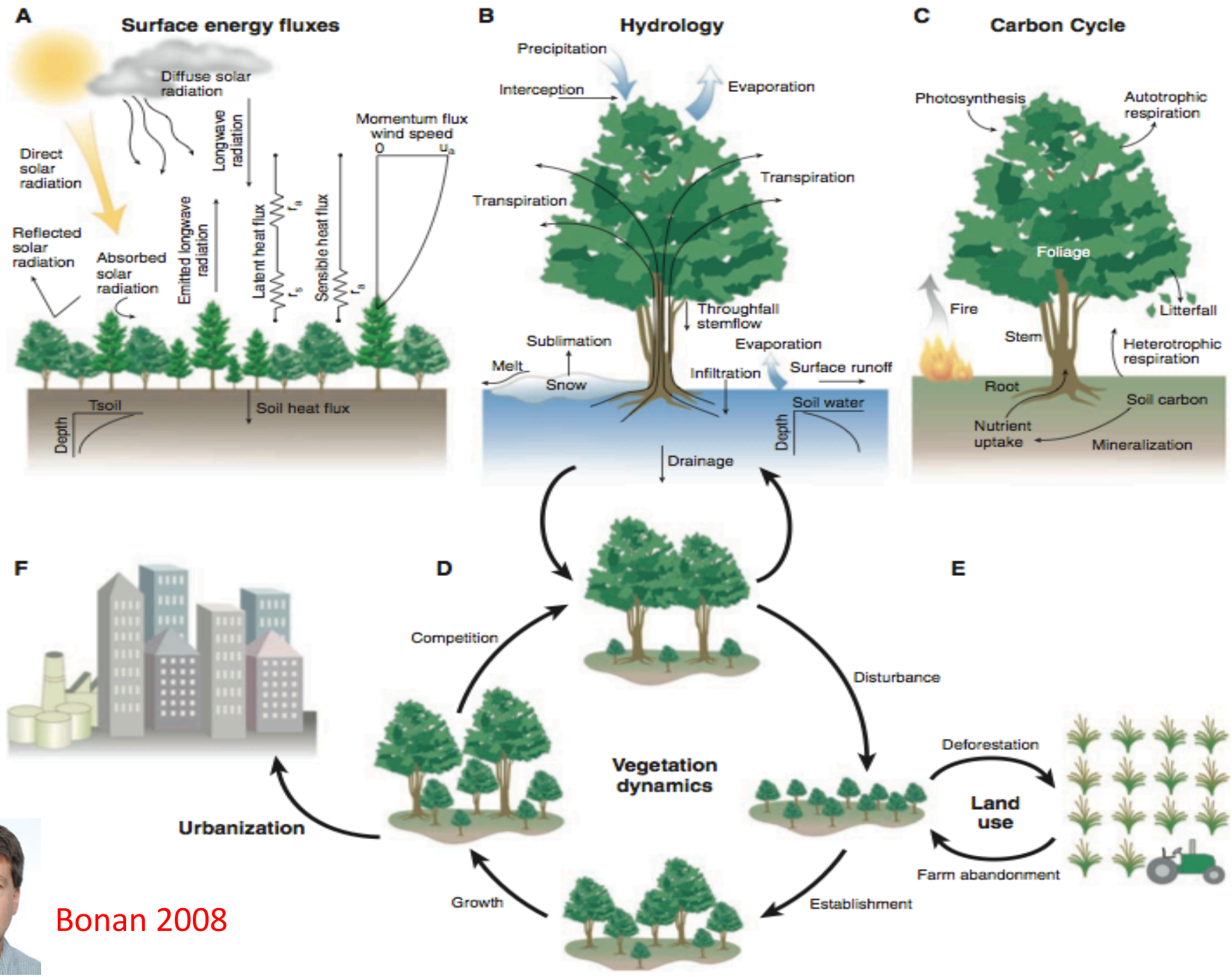
Courtesy of Ralph Keeling



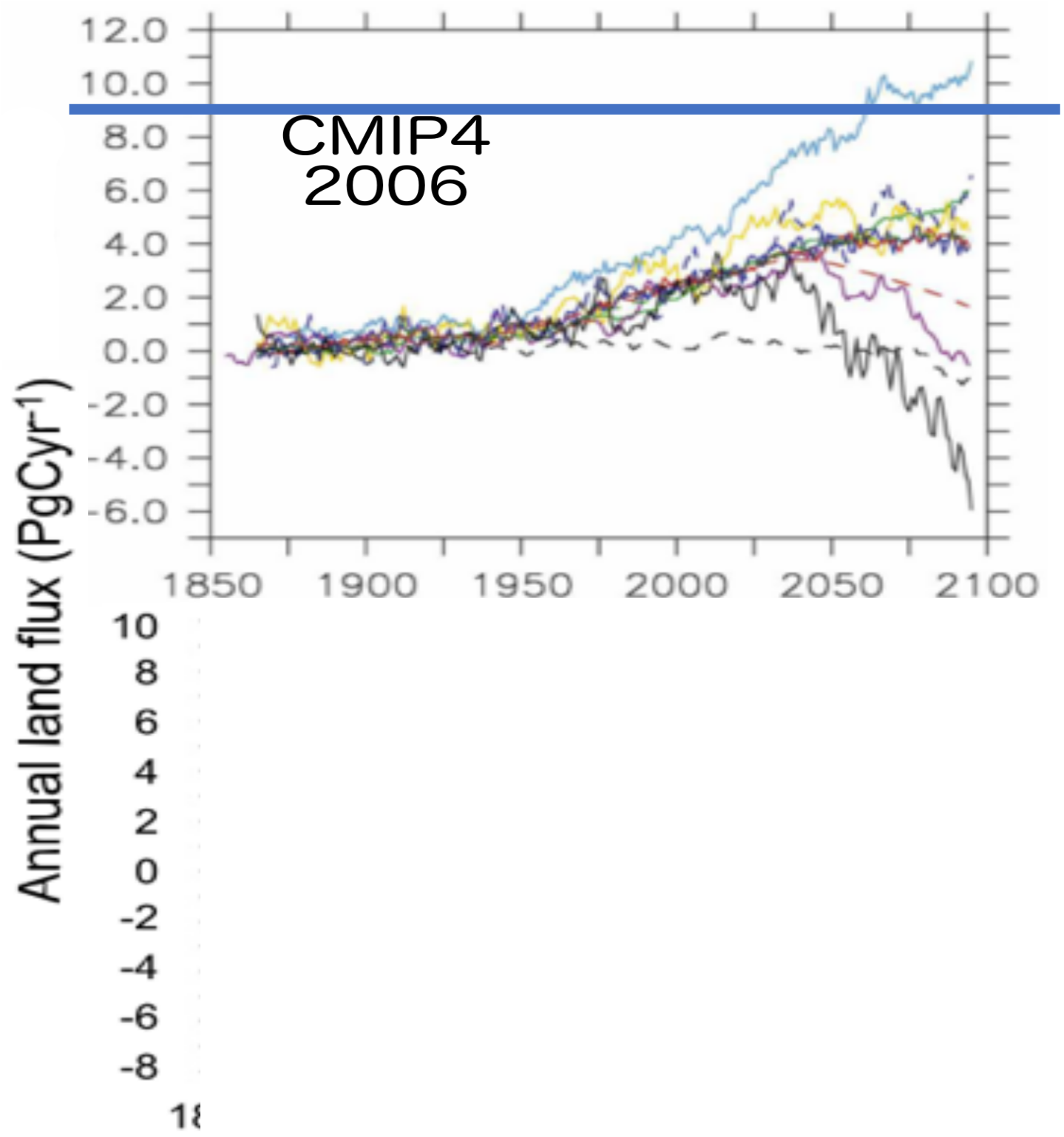
What is this data good for?

- Understand, measure, and predict the fate of global-warming greenhouse gases and how that influences ongoing and future climate change
  - Atmospheric and ecological theories of vegetation-climate **feedbacks**
  - Long-term, **multi-scale** observations of soil and vegetation carbon and water use
  - **Fusing** these to confront numerical models of land surface biophysics, ecosystem dynamics, and atmospheric forcing/feedbacks

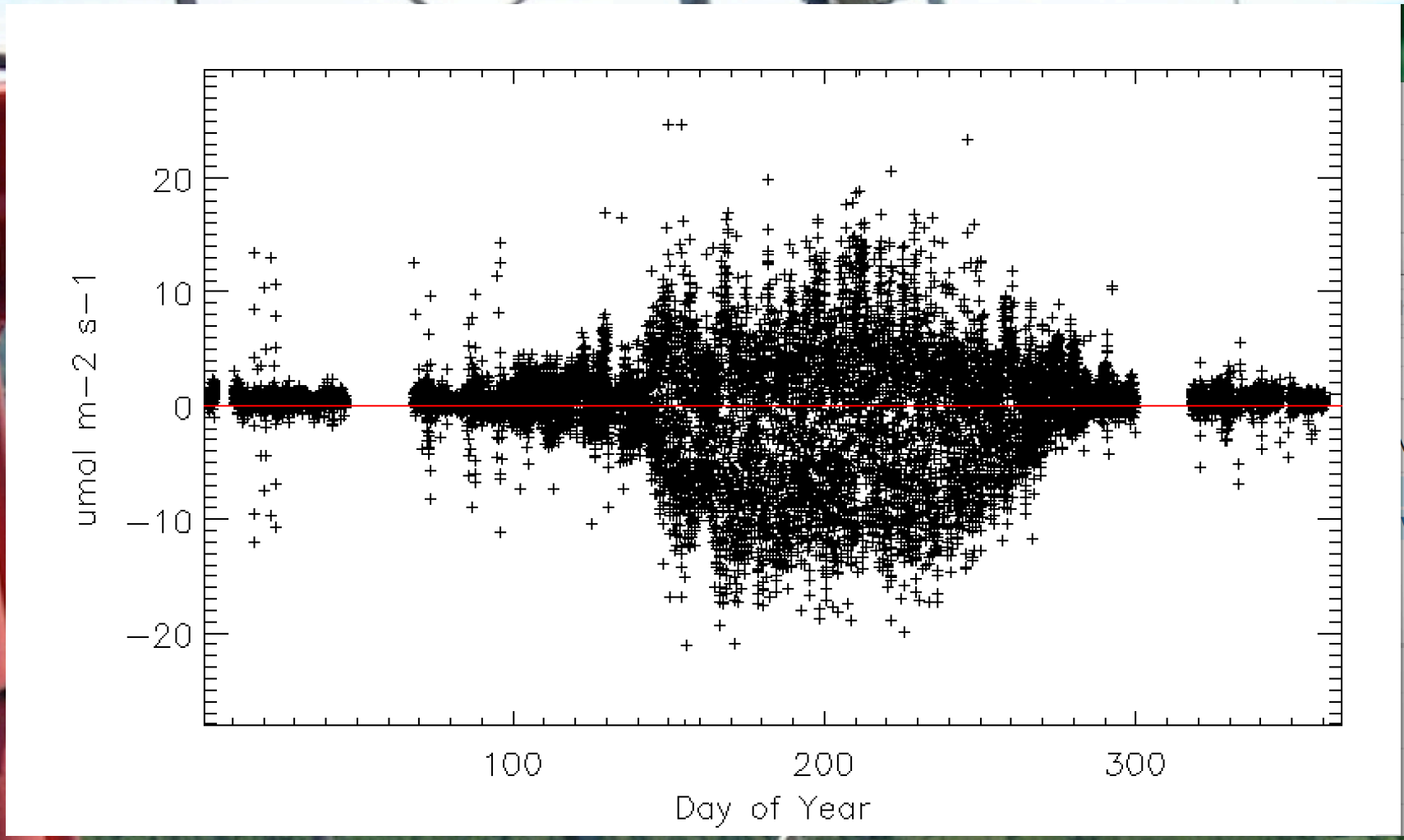
Forests in Flux

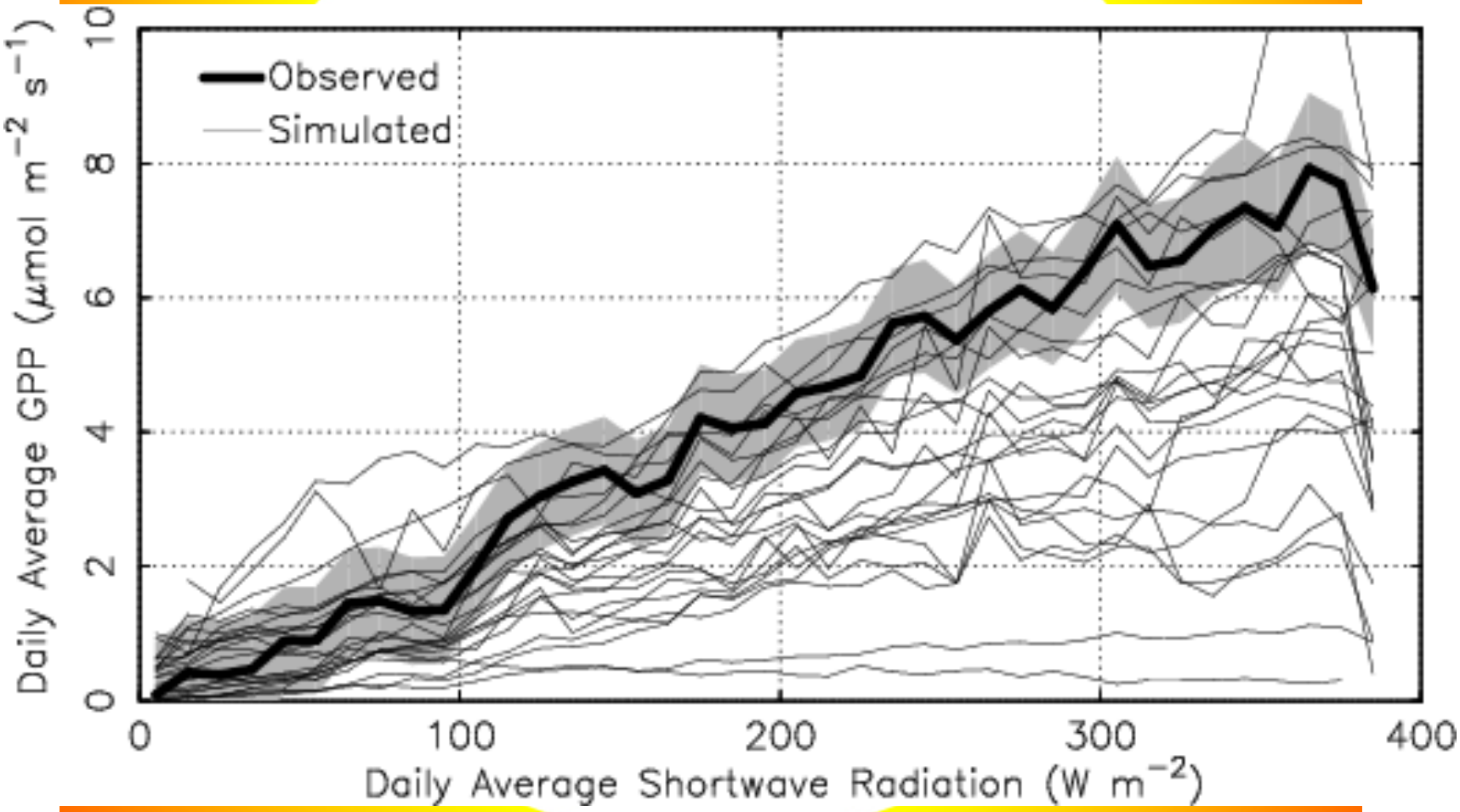


Bonan 2008


CMIP4
2006
Annual land flux (PgCyr-1)
12.0
10.0
8.0
6.0
4.0
2.0
0.0
-2.0
-4.0
-6.0
1850
1900
1950
2000
2050
2100
10
8
6
4
2
0
-2
-4
-6
-8

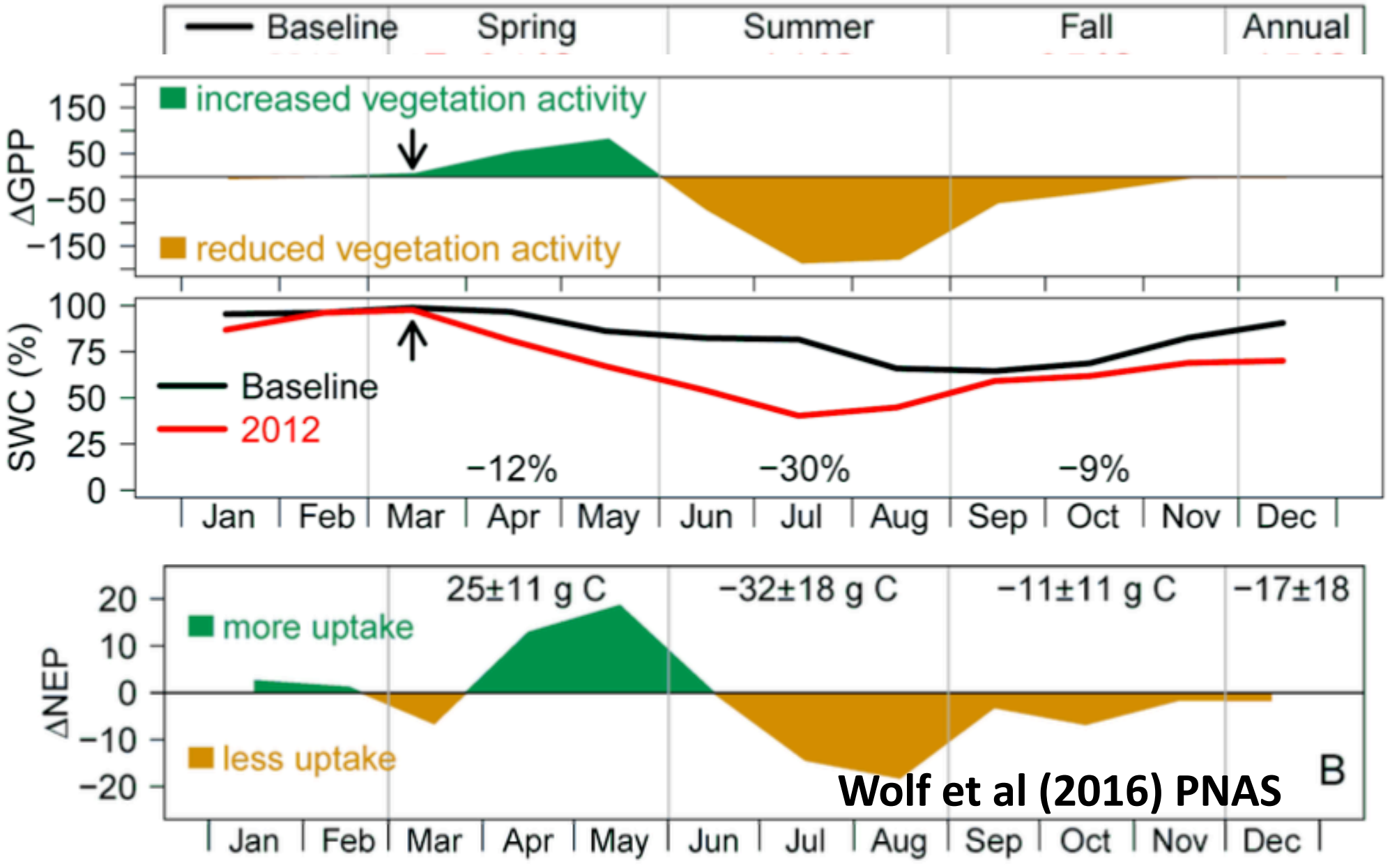
DATA!!! Om nom nom…



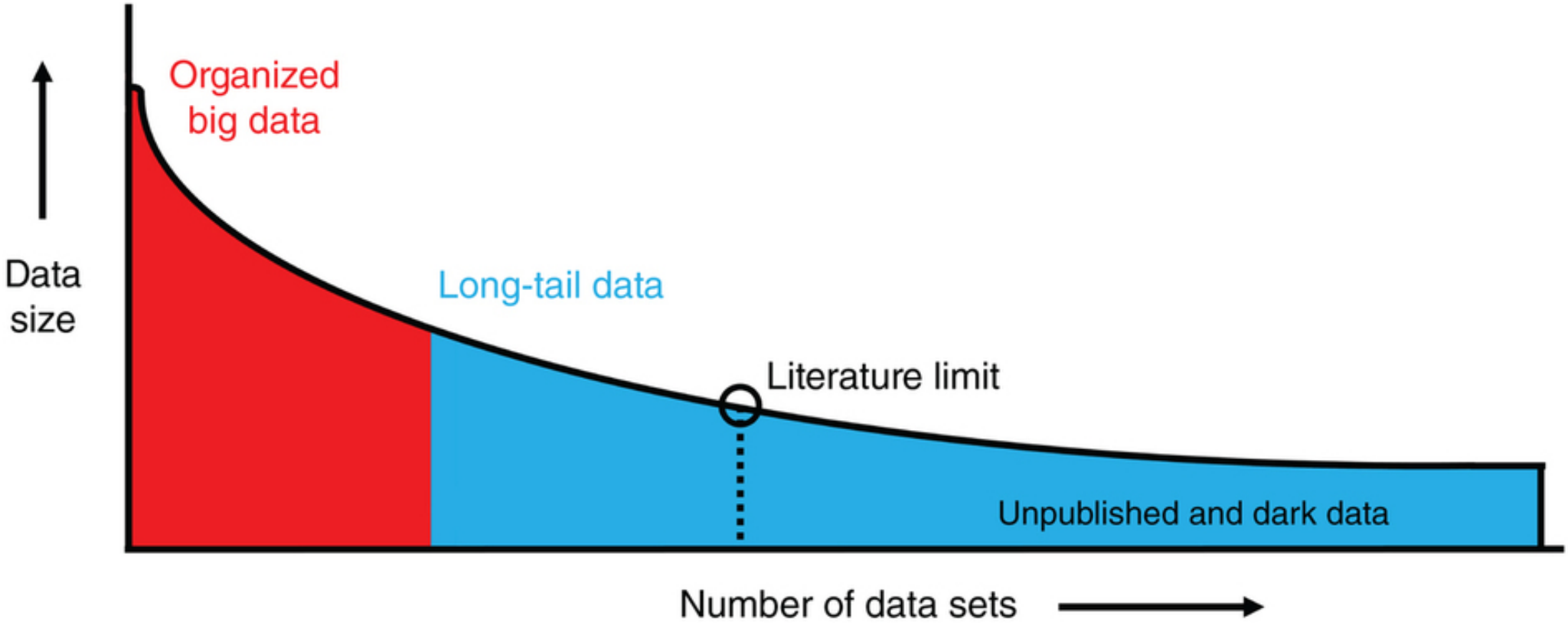


Schaefer et al., 2012

Net Carbon Uptake Anomaly @ sites (EC)



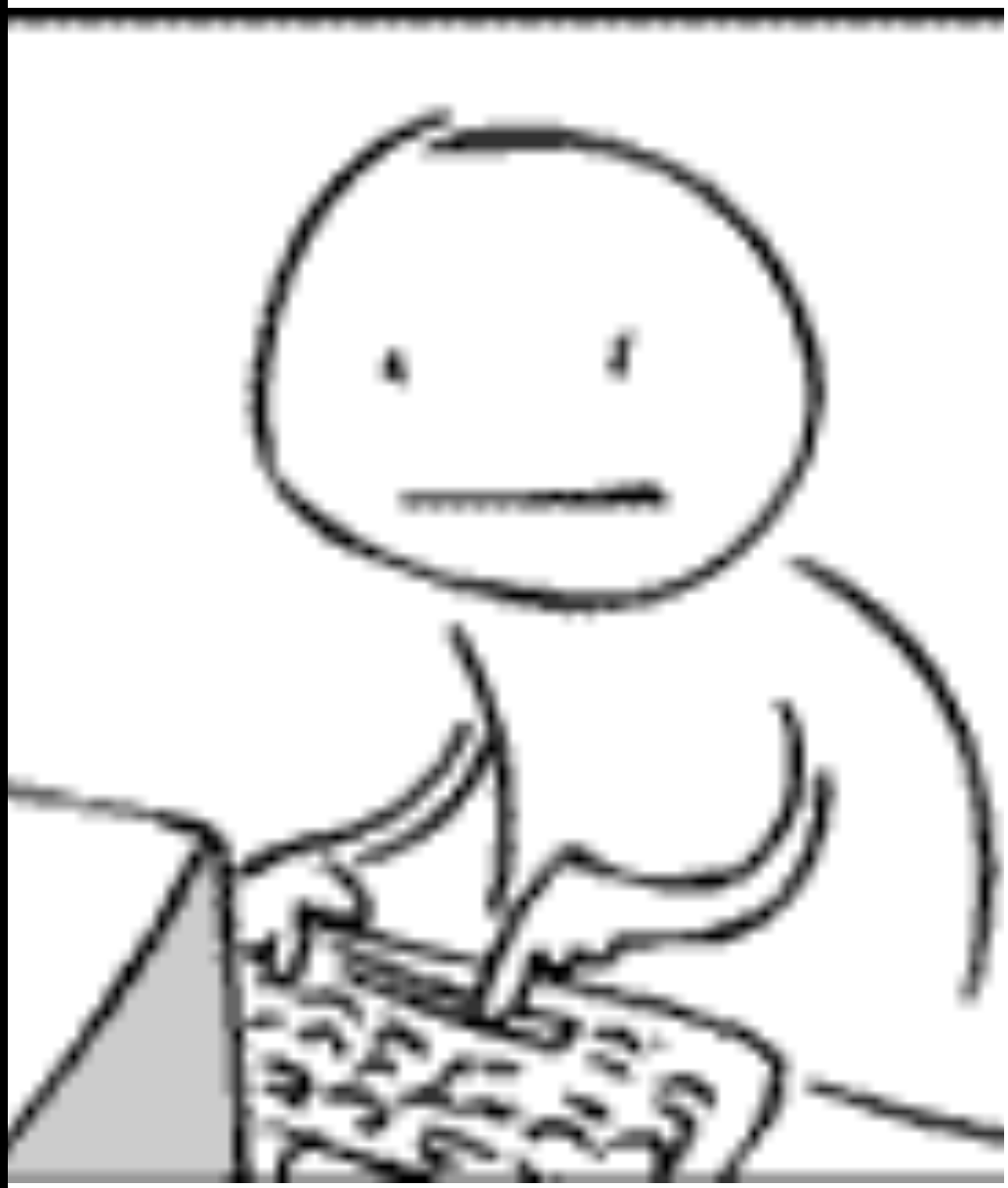
Wolf et al (2016) PNAS



Ferguson et al., 2014
Nature Neuroscience

- Data synthesis: volume, diversity
- Modeling not scalable
- Models are not accessible

NO EASY WAY FOR NON-MODELERS TO HELP

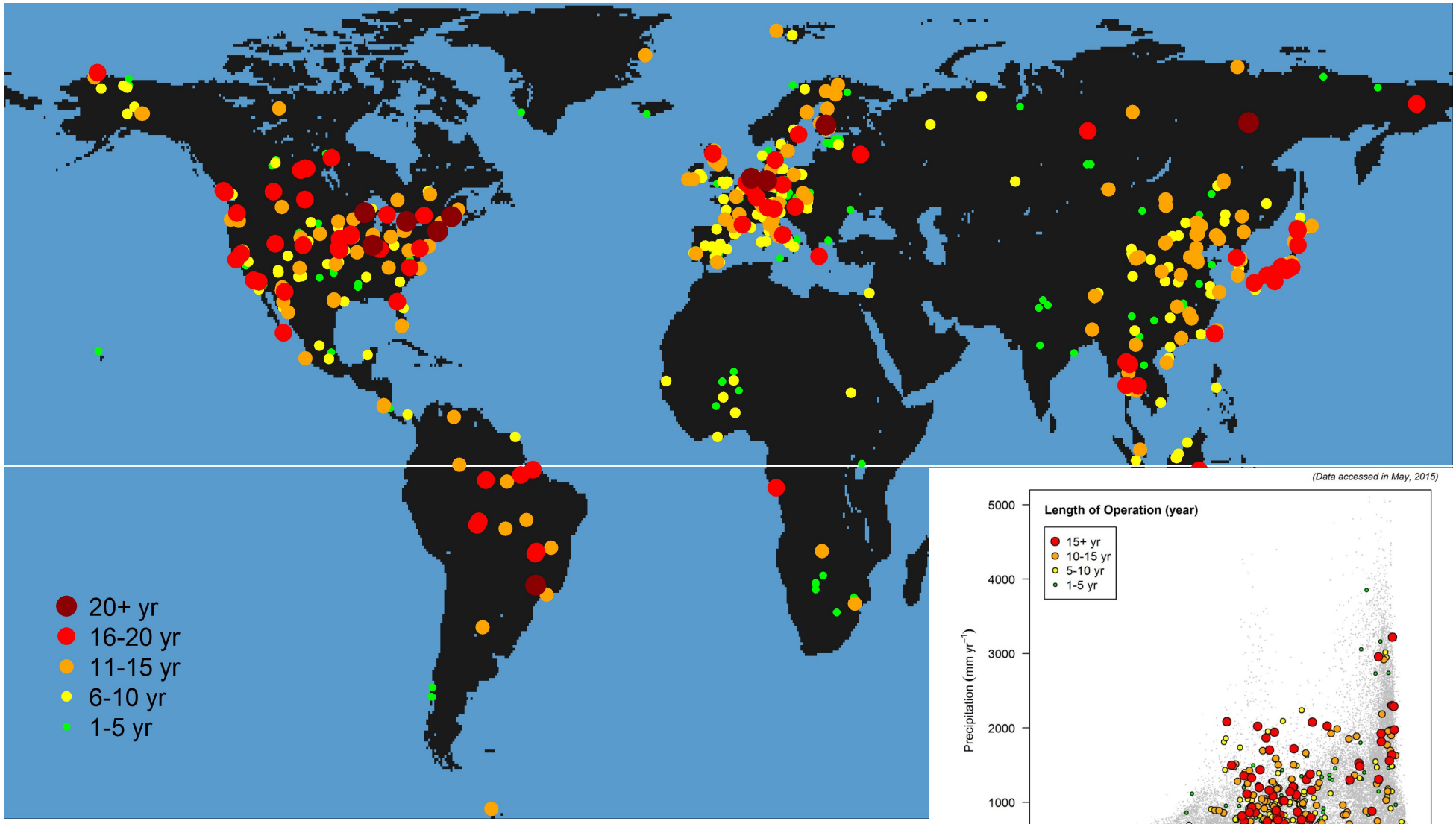


M Dietze / J Zobitz

Traits of a Positive Informatics Culture

- Open
- Collaborative
- Sharable
- Reproducible

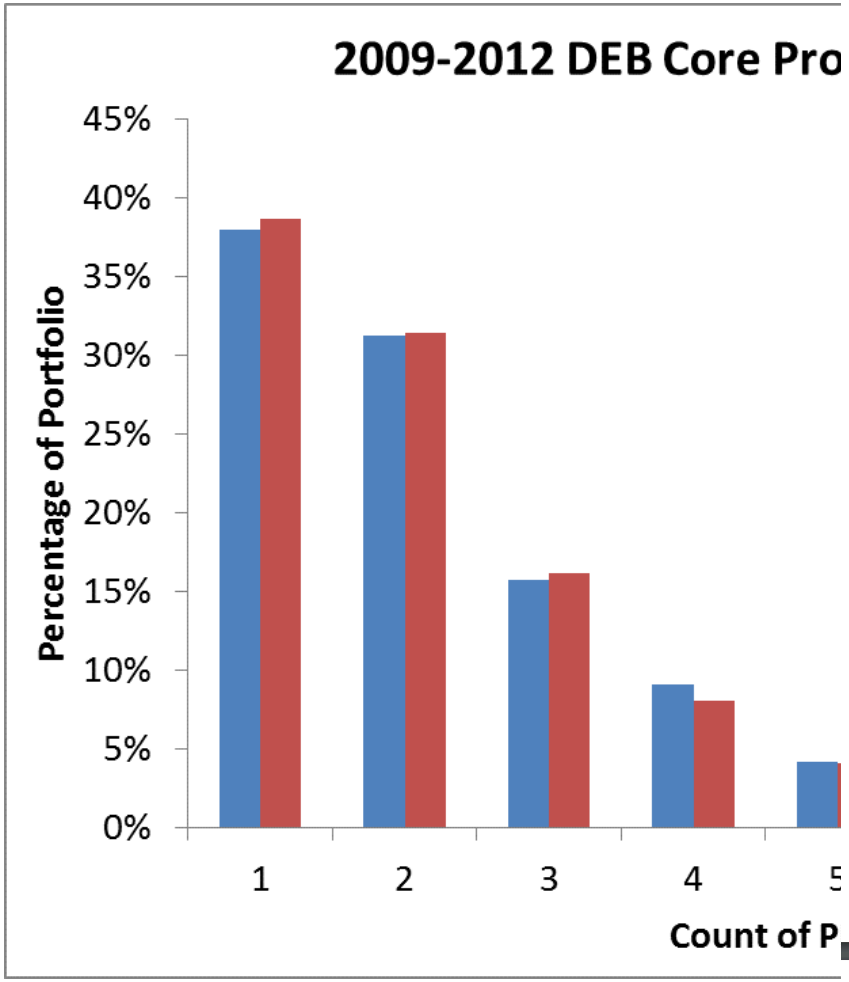
OPEN



Ameriflux: The Coalition of the Willing
Novick et al (in prep)

COLLABORATIVE

Most scientific pr
multi-PI, multi-ins



https://debblog.nsfbio.com/2013
numbers-award-size-and-duratior

IOP Publishing
Environ. Res. Lett. **5** (2010) 034007 (10pp)
Environmental Research Letters
doi:10.1088/1748-9326/5/3/034007

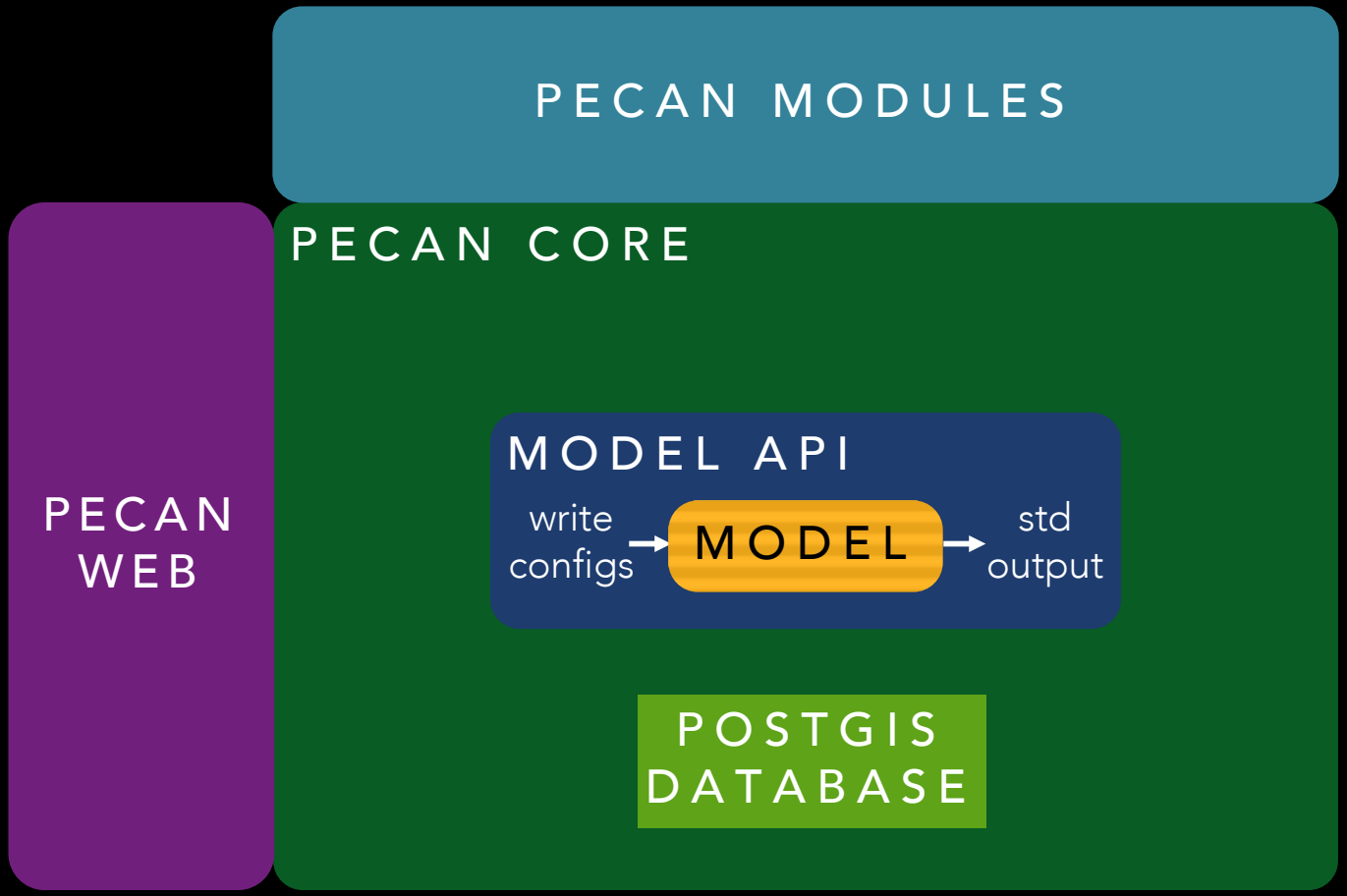
Climate control of terrestrial carbon exchange across biomes and continents

Chuixiang Yi[1], Daniel Ricciuto[2], Runze Li[3], John Wolbeck[1], Xiyan Xu[1], Mats Nilsson[4], Luis Aires[5,117], John D Albertson[6,117], Christof Ammann[7,117], M Altaf Arain[8,117], Alessandro C de Araujo[9,117], Marc Aubinet[10,117], Mika Aurela[11,117], Zoltán Barcza[12,117], Alan Barr[13,117], Paul Berbigier[14,117], Jason Beringer[15,117], Christian Bernhofer[16,117], Andrew T Black[17,117], Paul V Bolstad[18,117], Fred C Bosveld[19,117], Mark S J Broadmeadow[20,117], Nina Buchmann[21,117], Sean P Burns[22,117], Pierre Cellier[23,117], Jingming Chen[24,117], Jiquan Chen[25,117], Philippe Ciais[26,117], Robert Clement[27,117], Bruce D Cook[28,117], Peter S Curtis[29,117], D Bryan Dail[30,117], Ebba Dellwik[31,117], Nicolas Delpierre[32,117], Ankur R Desai[33,117], Sabina Dore[34,117], Danilo Dragoni[35,117], Bert G Drake[36,117], Eric Dufrêne[32,117], Allison Dunn[37,117], Jan Elbers[38,117], Werner Eugster[21,117], Matthias Falk[39,117], Christian Feigenwinter[40,117], Lawrence B Flanagan[41,117], Thomas Foken[42,117], John Frank[43,117], Juerg Fuhrer[7,117], Damiano Gianelle[44,117], Allen Goldstein[45,117], Mike Goulden[46,117], Andre Granier[47,117], Thomas Grünwald[48,117], Lianhong Gu[2,117], Haiqiang Guo[49,117], Albin Hammerle[50,117], Shijie Han[51,117], Niall P Hanan[52,117], László Haszpra[53,117], Bernard Heinesch[10,117], Carole Helfter[54,117], Dimmie Hendriks[55,117], Lindsay B Hutley[56,117], Andreas Ibrom[57,117], Cor Jacobs[38,117], Torbjörn Johansson[58,117], Marjan Jongen[59,117], Gabriel Katul[60,117], Gerard Kiely[61,117], Katja Klumpp[62,117], Alexander Knohl[21,117], Thomas Kolb[34,117], Werner L Kutsch[63,117], Peter Lafleur[64,117], Tuomas Laurila[11,117], Ray Leuning[65,117], Anders Lindroth[58,117], Heping Liu[66,117], Benjamin Loubet[23,117], Giovanni Manca[67,117], Michal Marek[68,117], Hank A Margolis[69,117], Timothy A Martin[70,117], William J Massman[43,117], Roser Matamala[71,117], Giorgio Matteucci[72,117], Harry McCaughey[73,117], Lutz Merbold[74,117], Tilden Meyers[75,117], Mirco Migliavacca[76,117], Franco Miglietta[77,117], Laurent Misson,[78,117,118], Meelis Mölder[58,117], John Moncrieff[27,117], Russell K Monson[79,117], Leonardo Montagnani[80,81,117], Mario Montes-Helu[34,117], Eddy Moors[82,117], Christine Moureaux[10,83,117], Mukufute M Mukelabai[84,117], J William Munger[85,117], May Myklebust[65,117], Zoltán Nagy[86,117], Asko Noormets[87,117], Walter Oechel[88,117], Ram Oren[89,117], Stephen G Pallardy[90,117], Kyaw Tha Paw U[39,117], João S Pereira[59,117], Kim Pilegaard[57,117], Krisztina Pintér[86,117], Casimiro Pio[91,117], Gabriel Pita[92,117], Thomas L Powell[93,117], Serge Rambal[94,117], James T Randerson[46,117], Celso von Randow[95,117], Corinna Rebmann[64,117], Janne Rinne[96,117], Federica Rossi[77,117], Nigel Roulet[97,117], Ronald J Ryel[98,117], Jorgen Sagerfors[4,117], Nobuko Saigusa[99,117], María José Sanz[100,117], Giuseppe-Scarascia Mugnozza[101,117], Hans Peter Schmid[102,117], Guenther Seufert[103,117], Mario Siqueira[89,117], Jean-François Soussana[62,117], Gregory Starr[104,117], Mark A Sutton[105,117], John Tenhunen[106,117], Zoltán Tuba,[86,117,118], Juha-Pekka Tuovinen[11,117], Riccardo Valentini[107,117], Christoph S Vogel[108,117], Jingxin Wang[109,117], Shaoqiang Wang[110,117], Weiguo Wang[111,117], Lisa R Welp[112,117], Xuefa Wen[110,117], Sonia Wharton[113,117], Matthew Wilkinson[20,117], Christopher A Williams[114,117],

1748-9326/10/034007+10$30.00

Georg Wohlfahrt[50,117], Susumu Yamamoto[115,117], Guirui Yu[110,117], Roberto Zampedri[44,117], Bin Zhao[49,117] and Xinquan Zhao[116,117]


MENU
ICESab 2



Standardized inputs and outputs

Provenance: Transparent & Repeatable

Accessible interface

Reusable tools for execution, analysis, visualization

No central repository!

For code or data!

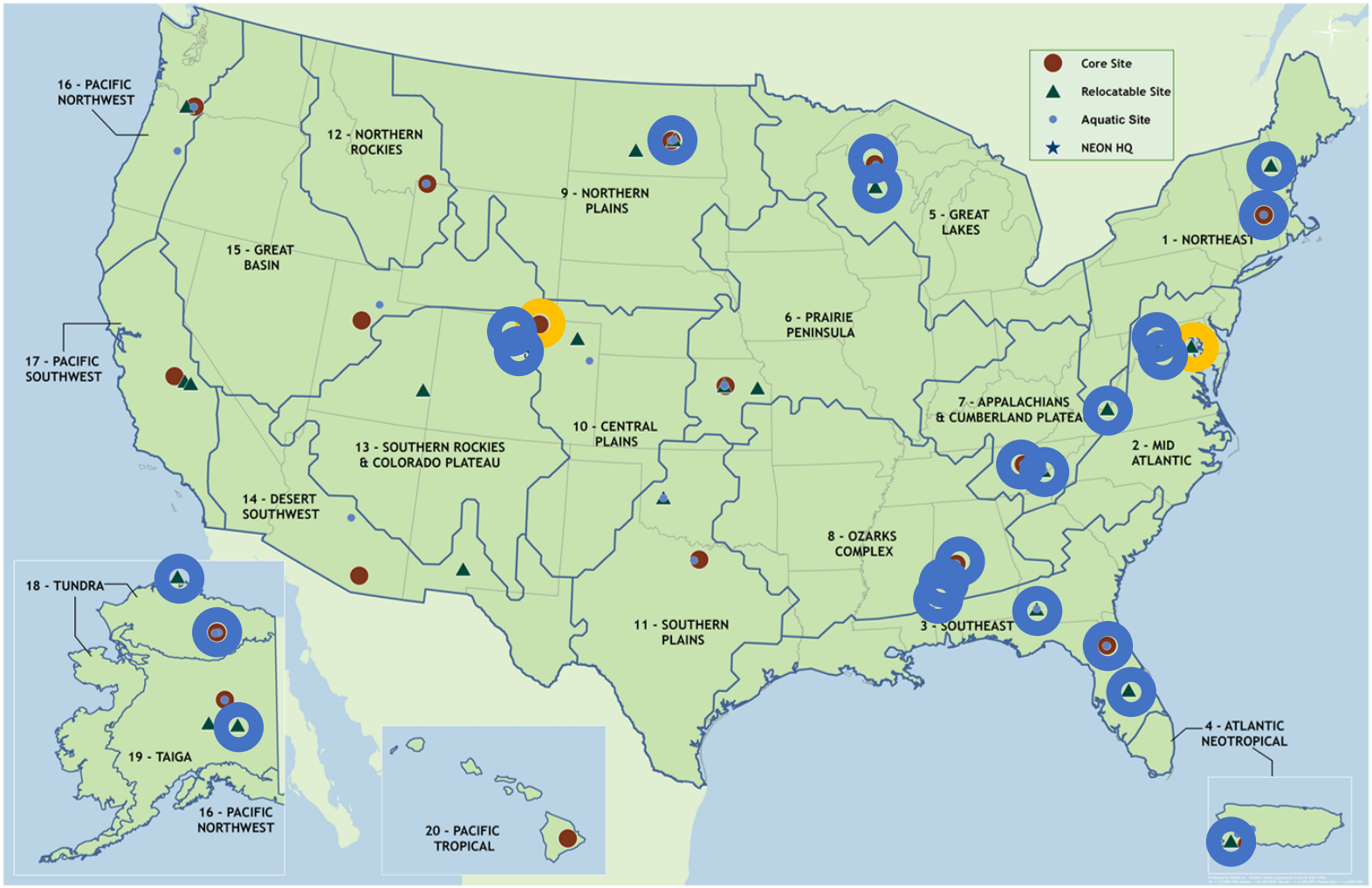
SHARABLE

Sharing is caring…

- The National Ecological Observatory Network is a $450 million NSF set of coordinated U.S. ecological observing sites to address grand challenges in global change
  - The “supercollider” of ecology
- Community resource – consistent instruments on all sites, open data, documentation for every variable REST/JSON API for access

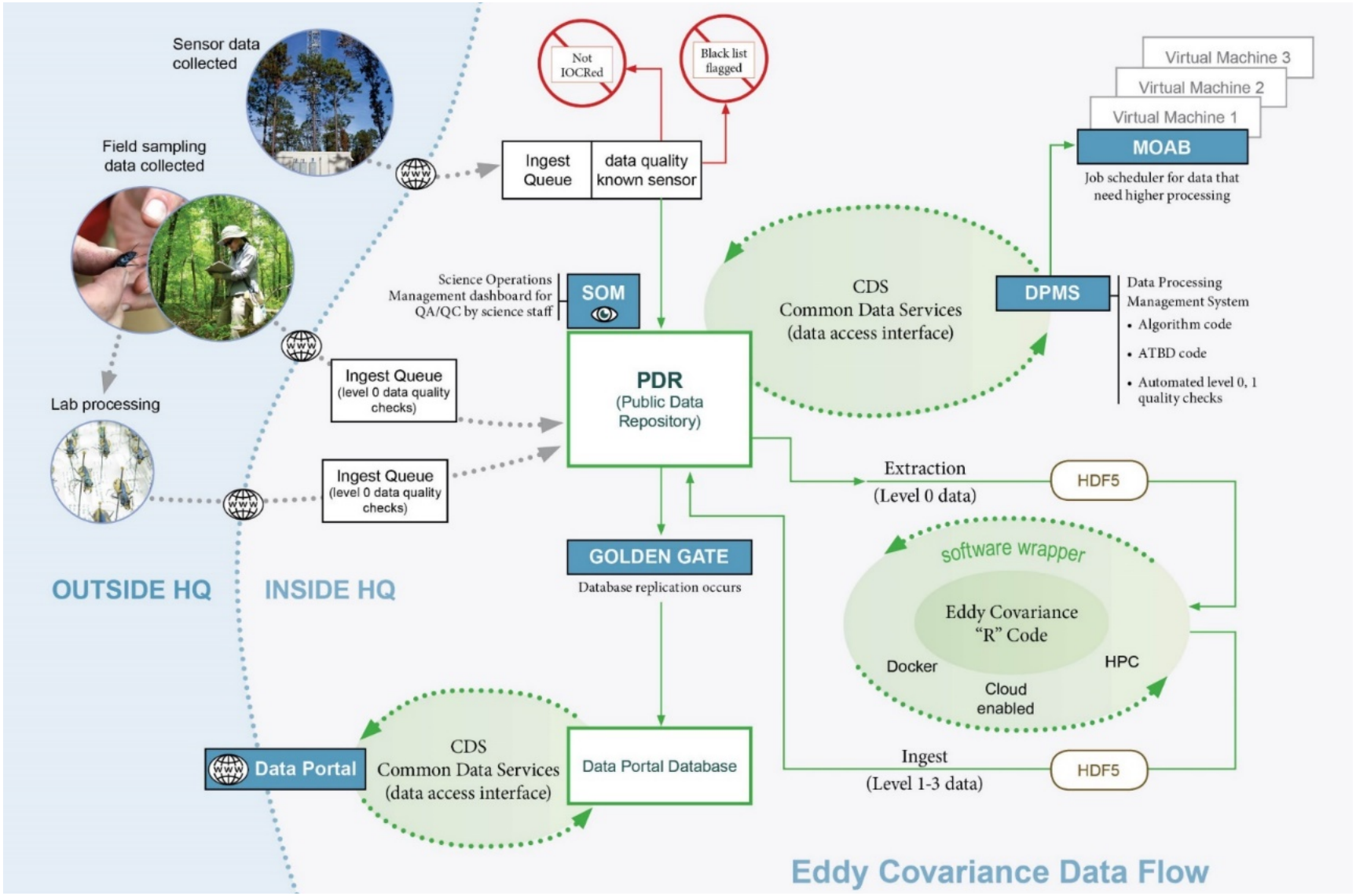
- But can this infrastructure support ecology?

eddy-covariance data products: sites and schedule



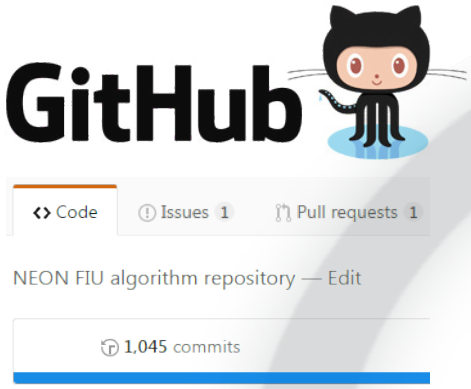
- initially: 2 sites
- +6 months: 25 sites
- +12 months: all 47 sites
- provisional data until first versioning (mid-2019)

NEON data processing pipeline



eddy-covariance usability tools: DevOps cycle

(3)



(4)

(5) neon CI

(2) neon Science

(1) Science Community

eddy-covariance usability tools: eddy4R-Docker image

- Docker = shipping container system for code

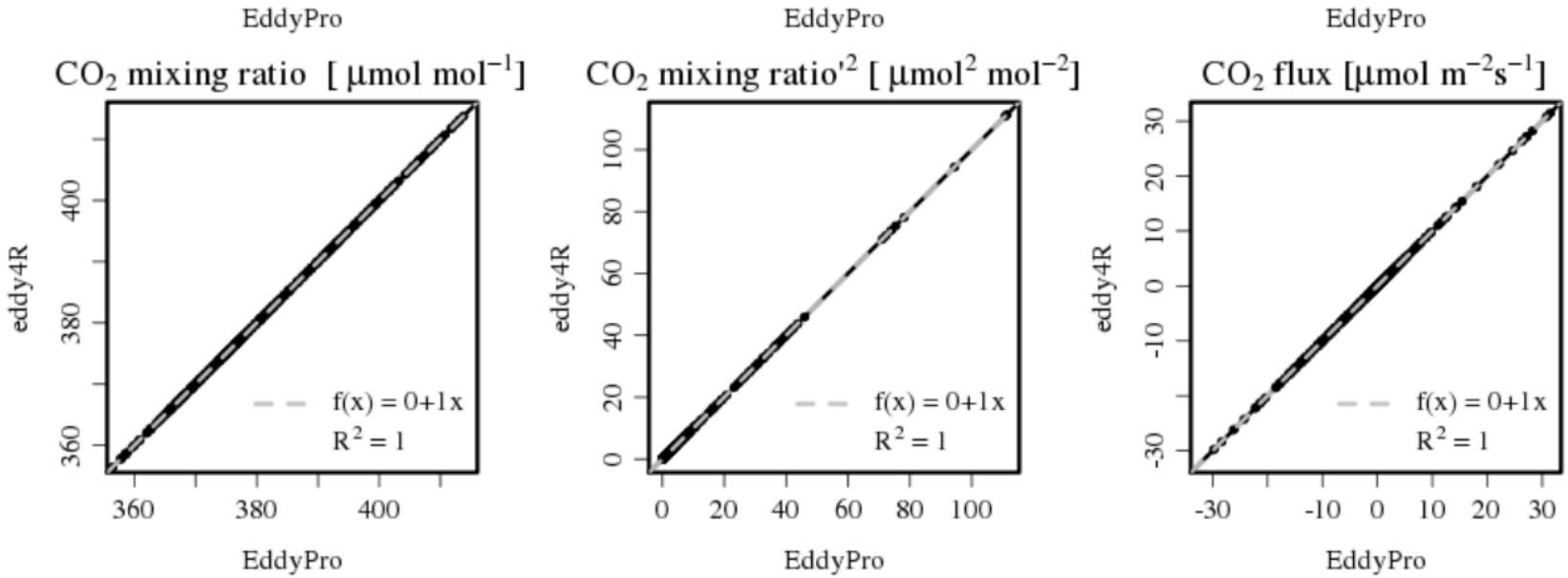
- "containers wrap a piece of software in a complete filesystem that contains everything needed to run: code, runtime, system tools, system libraries"
  - efficient: shares host operating system (OS) instead of guest OS emulation
  - reproducible: same results, regardless of the host operating system
  - lightweight, distributed via a web-based portal (hub.docker.com)
  - deployable at scale, from laptop to massively parallel applications

Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-318, 2017
Manuscript under review for journal Geosci. Model Dev.
Published: 1 February 2017

Geoscientific Model Development Discussions
Open Access
EGU

eddy4R: A community-extensible processing, analysis and modeling framework for eddy-covariance data based on R, Git, Docker and HDF5

Stefan Metzger[1], David Durden[1], Cove Sturtevant[1], Hongyan Luo[1], Natchaya Pingintha-Durden[1], Torsten Sachs[2], Andrei Serafimovich[2], Jörg Hartmann[3], Jiahong Li[4], Ke Xu[5], Ankur R. Desai[5]

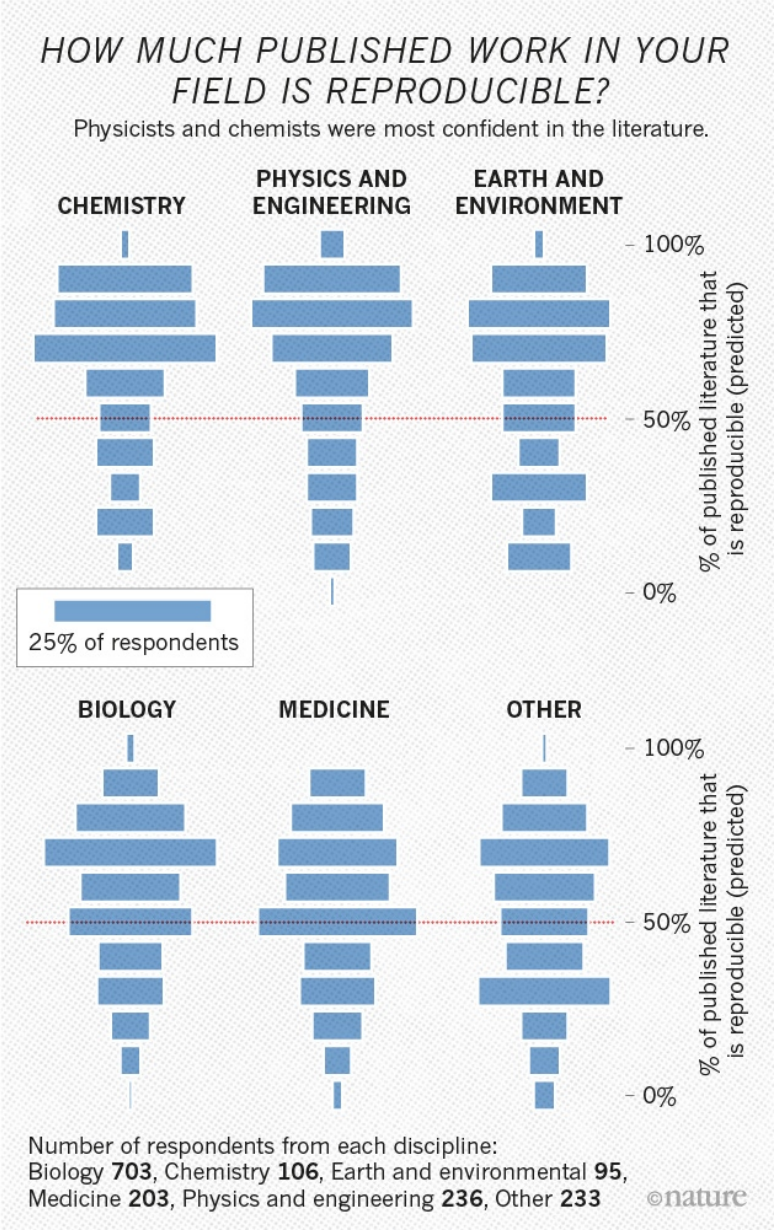


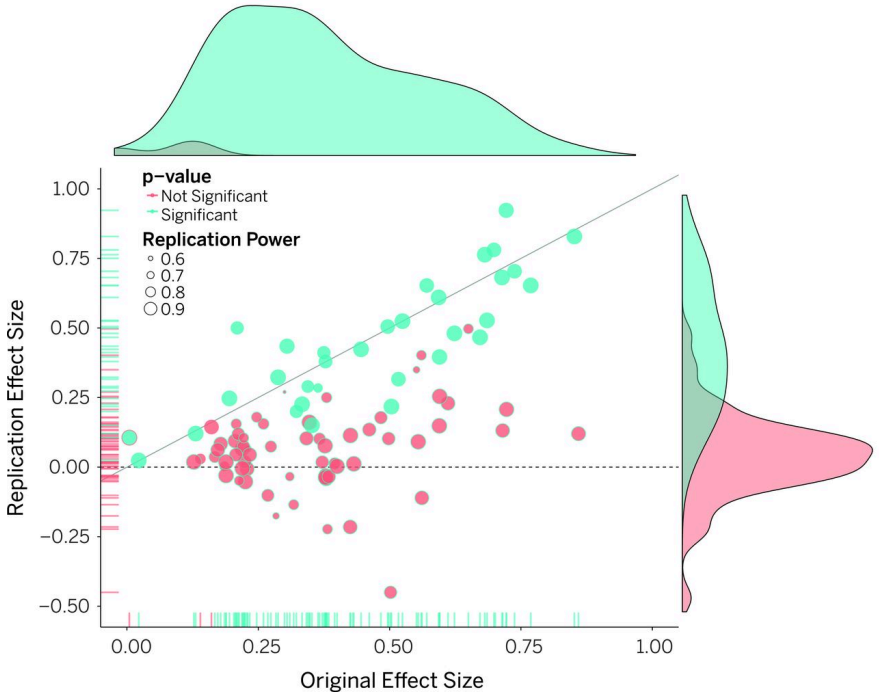
Big data is not open, collaborative nor sharable if…

- Code to generate/analyze is not reusable by others
  - Github, Docker, DevOps cycle is key to making “big science” happen
- Data lack open, common APIs to access by machines
  - THREDDS, JSON/XML
- Data formats are non-standard, not machine-readable
  - NetCDF, Unidata CF convention as an example in meteorology
  - Ecological Metadata Language (EML)
- Data requires complex authentication methods to access or repositories don’t have multiple points of entries, distributed nodes
  - Kill the password!
- Data/code sharing policies limit what you can do
  - Important to set this out by community, be open to ideas beyond intended use
- Data quicklooks, comparisons, documentation on variable names, time steps, units are not easy to find
  - Simple tables, online, vignettes, forums/chat rooms

REPRODUCIBLE

We have a reproducibility crisis…





Estimating the reproducibility of psychological science

Open Science Collaboration*,†
+ See all authors and affiliations

Science 28 Aug 2015:
Vol. 349, Issue 6251,
DOI: 10.1126/science.aac4716

http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Lack of full metadata is an issue

- Protocol
- Code
- Data
- Filtering and tests
- Experiments and vignettes

- #openexperiment

Environ. Res. Lett. **11** (2016) 084004 doi:10.1088/1748-9326/11/8/084004

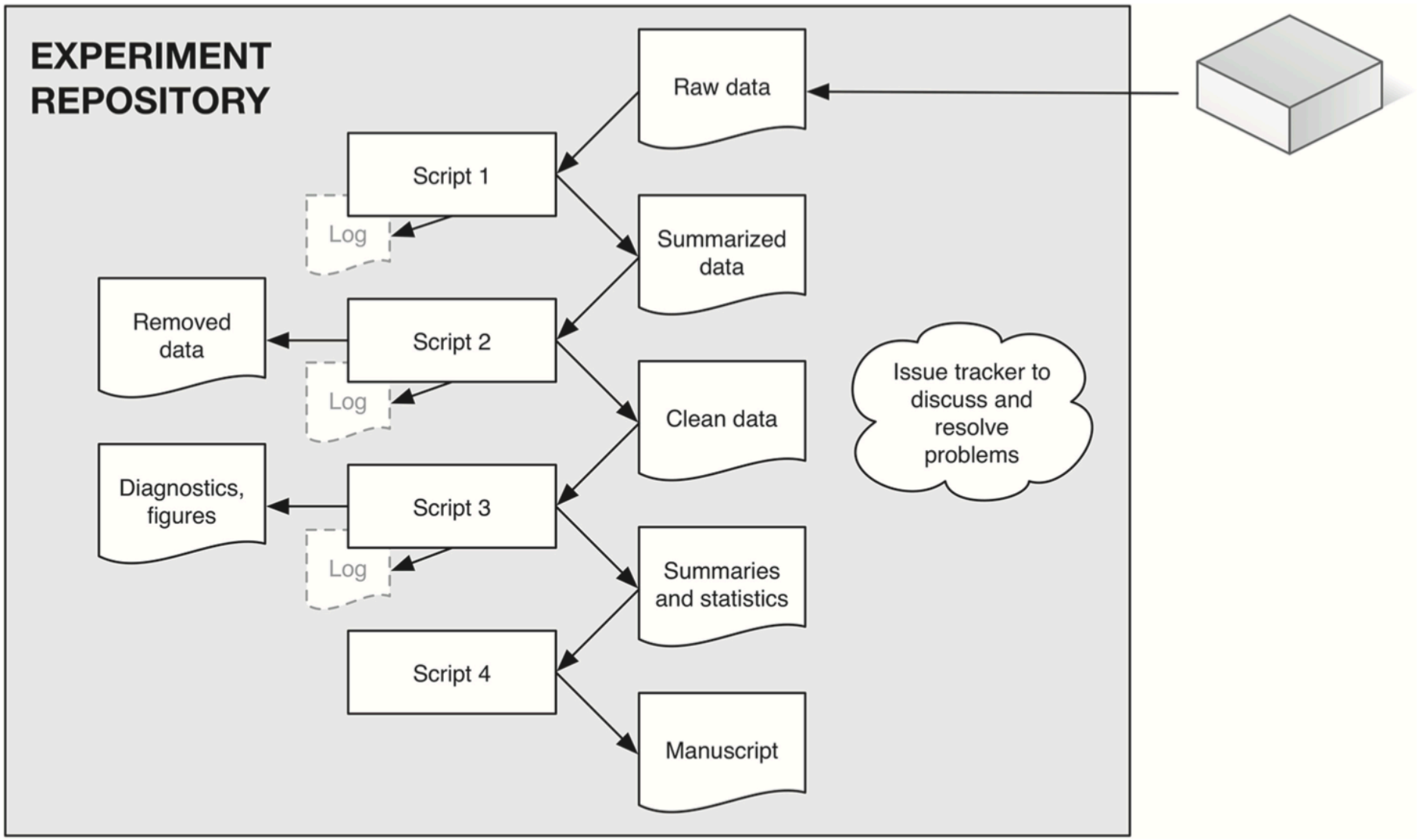
Environmental Research Letters

LETTER

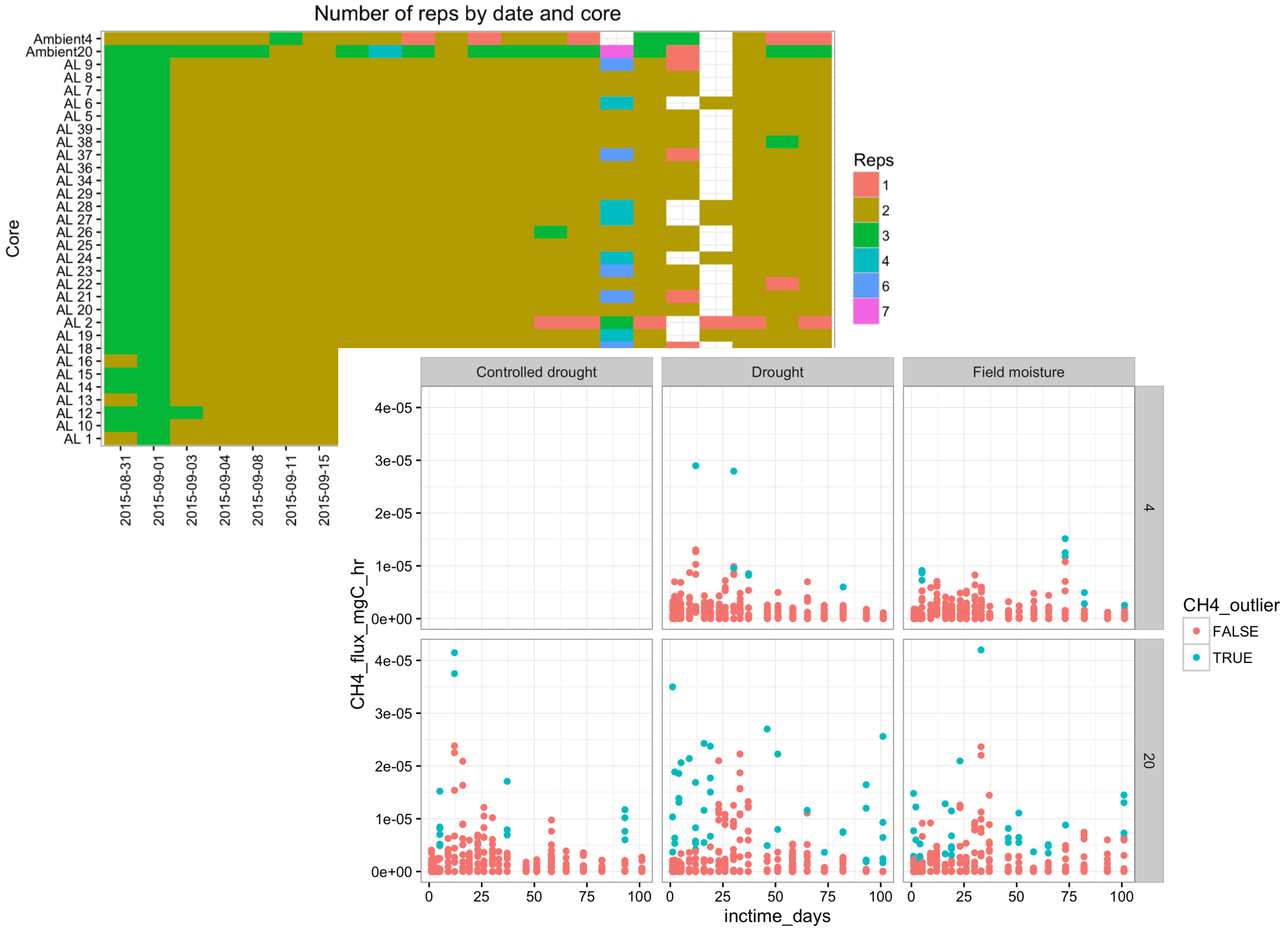
Running an open experiment: transparency and reproducibility in soil and ecosystem science

Ben Bond-Lamberty[1], A Peyton Smith[2] and Vanessa Bailey[2]

https://github.com/bpbond/cpcrw_incubation


EXPERIMENT
REPOSITORY
Raw data
Script 1
Log
Summarized
data
Removed
data
Script 2
Log
Clean data
Issue tracker to
discuss and
resolve
problems
Diagnostics,
figures
Script 3
Log
Summaries
and statistics
Script 4
Manuscript

Missing/problematic data



How do we encourage and support informatics culture at UW and elsewhere?

- Training for graduate students
  - Seminars taught by academic staff?
- Pilot projects linking ACI/HTPC, CS, CALS, L&S
  - Budget models that encourage collaborative grants
- Funding support for data archival and informatics
  - Digitization/generation of metadata for long-tail data
  - The mantra does NOT have to be centralization
- … what else?

THANK YOU!

Ankur Desai, desai@aos.wisc.edu, http://flux.aos.wisc.edu

Funding: NSF Advances in Biological Informatics (**ABI-1457897 , ABI-1062205)**
National Ecological Observatory Network, Inc.

Collaborators to this talk:
Mike Dietze, Boston Univ.
Stefan Metger, NEON/Battelle
Ben Bond-Lamberty, DOE PNNL