The *human* factor in big-data and eco-informatics

Ankur R Desai
University of Wisconsin-Madison
ESA 2018, SYMP 9
7 Aug 2018
New Orleans, LA

Acknowledgments

- Jack Williams, UW-Madison
- Kim Novick, University of Indiana
- Michael Dietze, Boston University
- Stefan Metzger and Wendy Gram, NEON/Battelle
- Kathleen Weathers, Cary Institute
- Ben Bond Lamberty, DOE PNNL
- Enablish FAIR Data Project
- And many contributors…
- + Support from NSF BIO (DEB, EF/MSB, ABI, RCN) and AGS, DOE TES Ameriflux, Battelle/NEON, NASA Carbon Cycle, NOAA, USGCRP

Take Homes

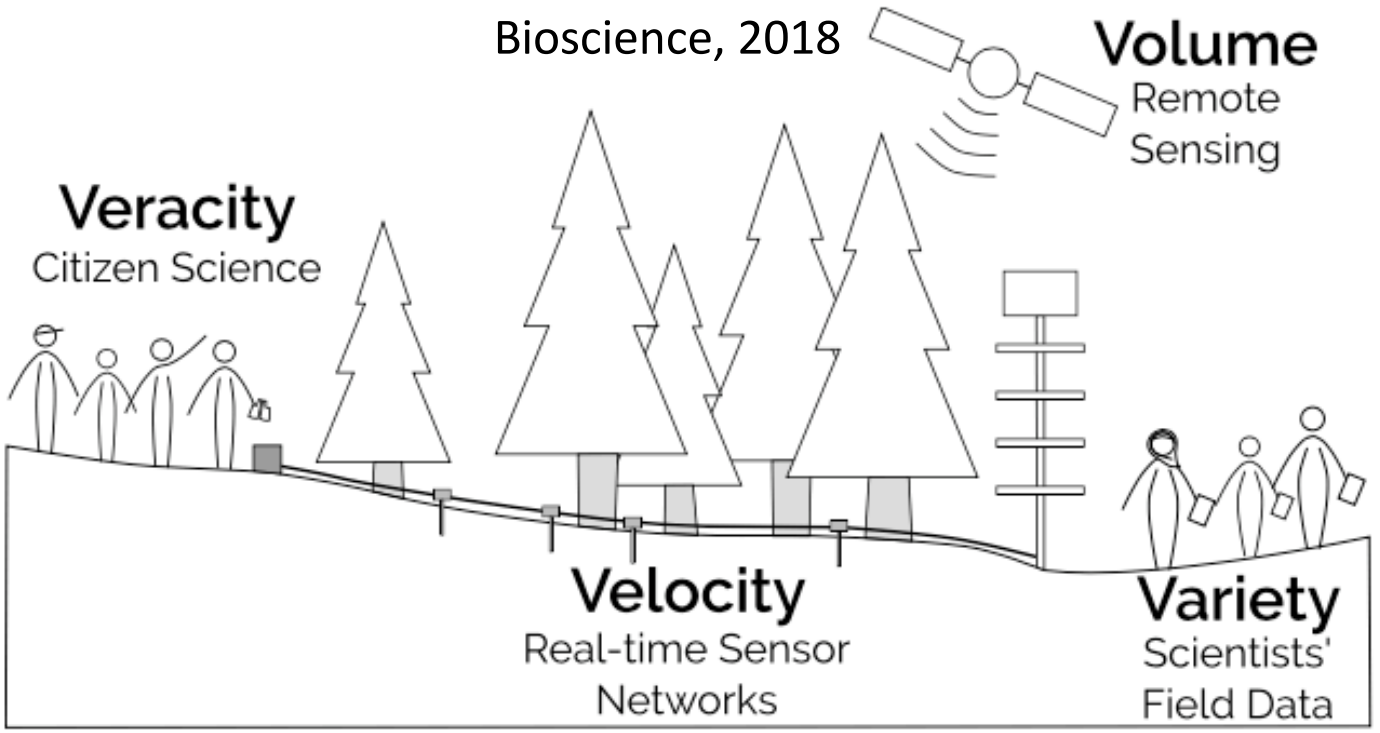
- Big data is not just about data volume
  - Data/code diversity, accessibility, and metadata matter
- Tackling challenges in informatics is a key to solving the scientific reproducibility crisis
  - Big data is really about the people, ethics, networks
- Ecologists are well-positioned to be a leader here
  - If we invest in **people** and **infrastructure**

You *all* already do
“big-data”
“fusion”
“forecasting”
and “informatics”

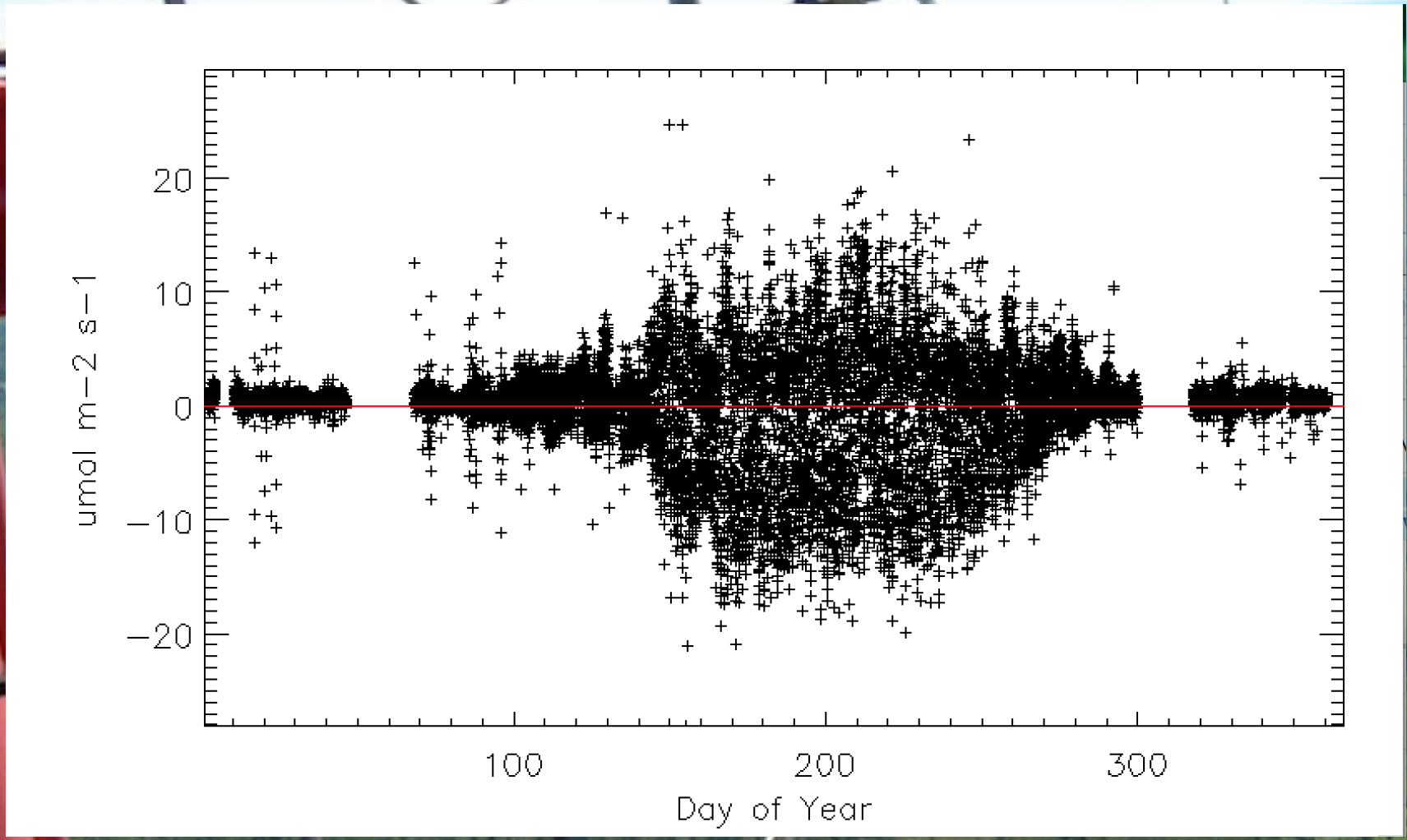
Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions

SCOTT S. FARLEY, ANDRIA DAWSON, SIMON J. GORING AND JOHN W. WILLIAMS

Bioscience, 2018



DATA!!! Om nom nom…




MERCURY
75
BACK

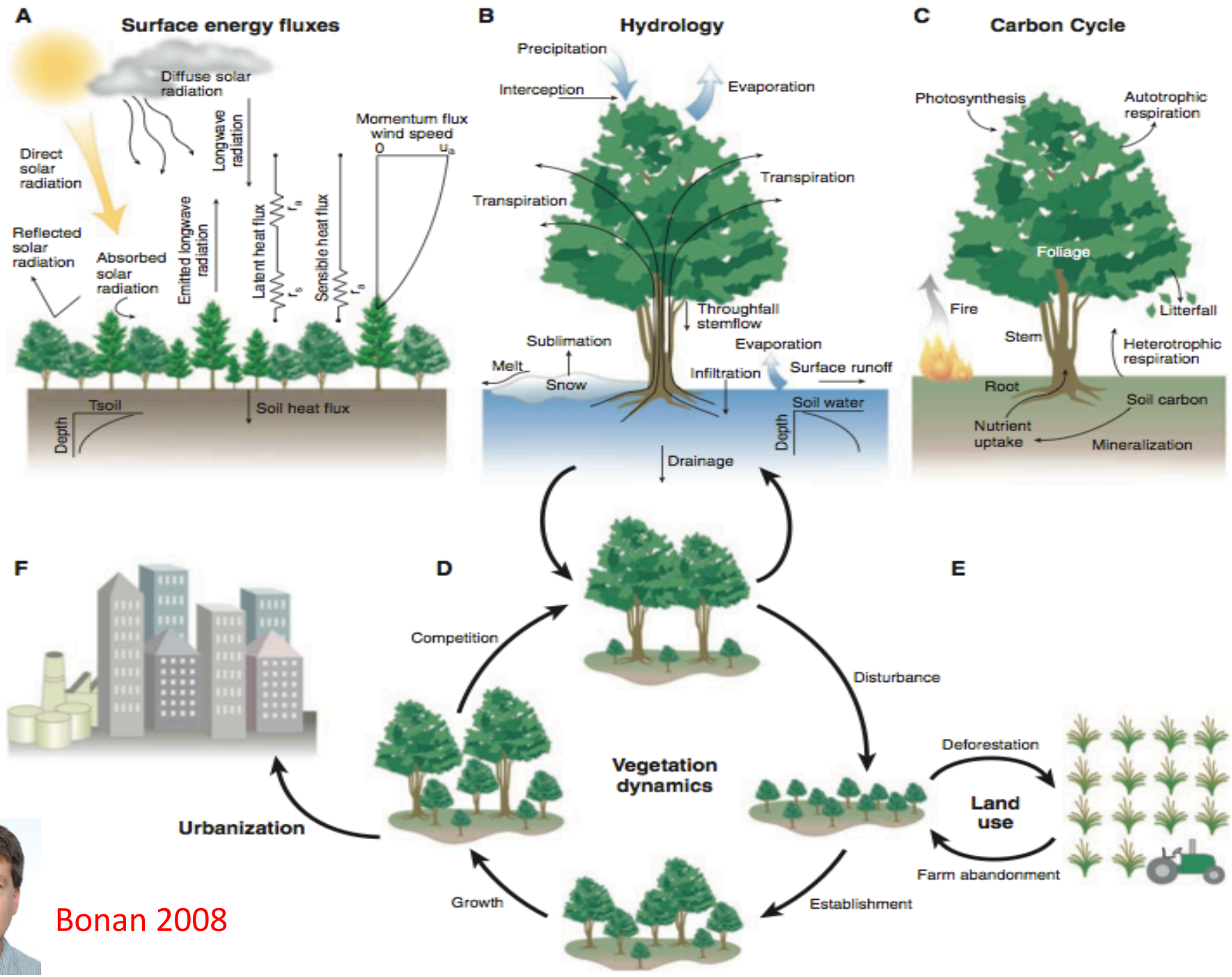
https://phenocam.sr.unh.edu/webcam/gallery/

Andrew Richardson, Northern Arizona U

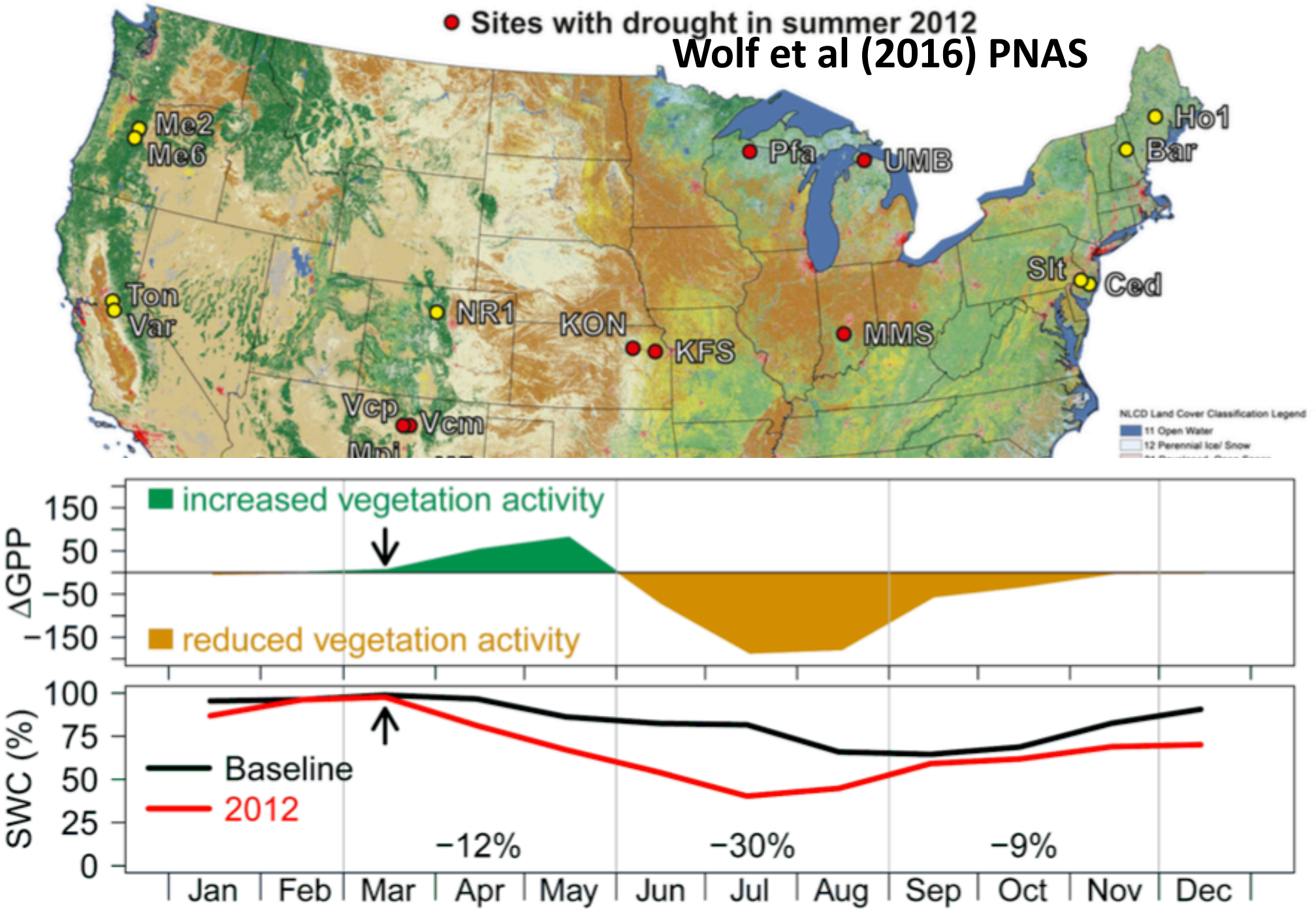
What are these data good for?

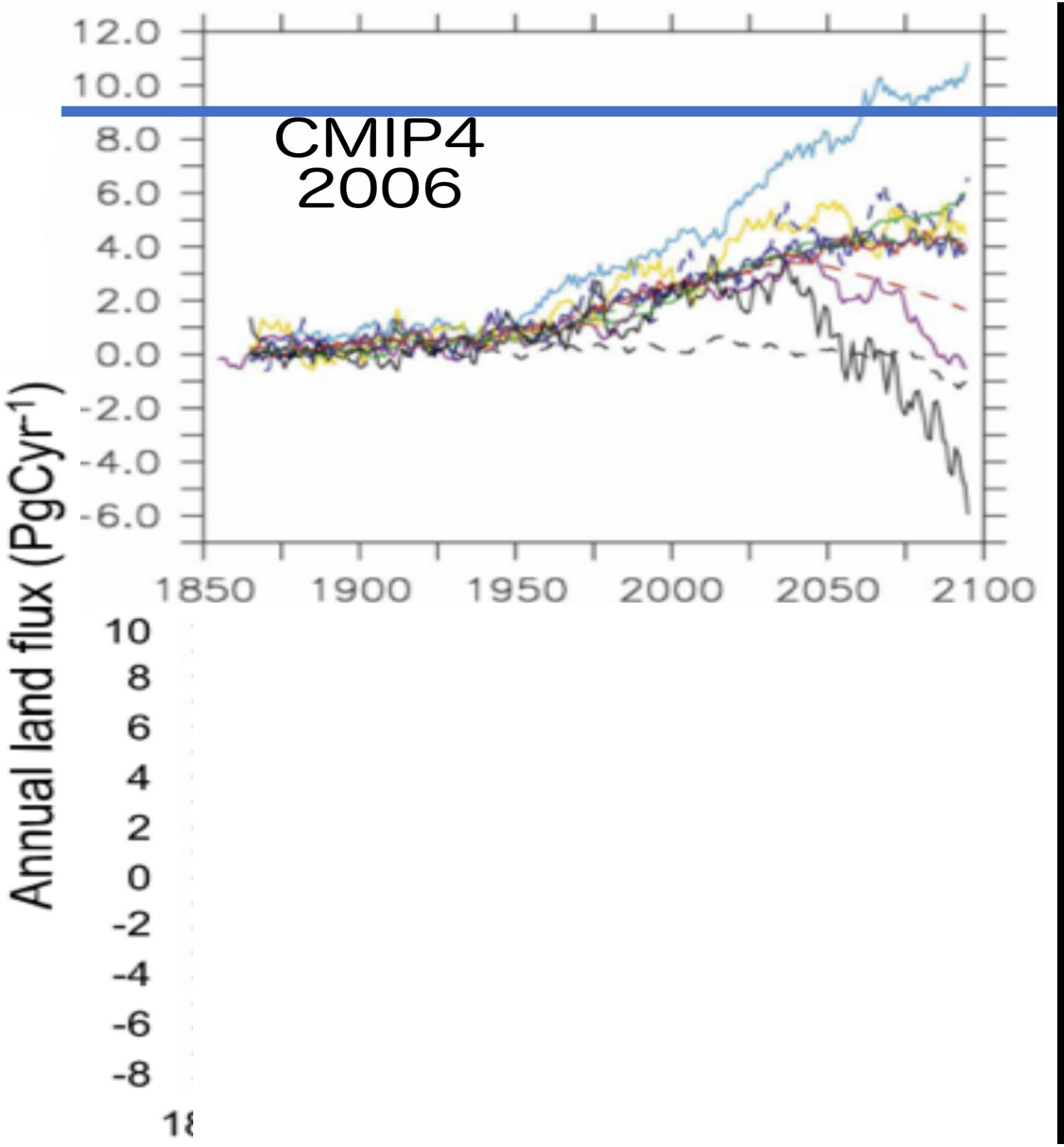
- Understand, measure, and predict the fate of global-warming greenhouse gases and how that influences ongoing and future climate change
  - Atmospheric and ecological theories of vegetation-climate **feedbacks**
  - Long-term, **multi-scale** observations of soil and vegetation carbon and water use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

Forests in Flux



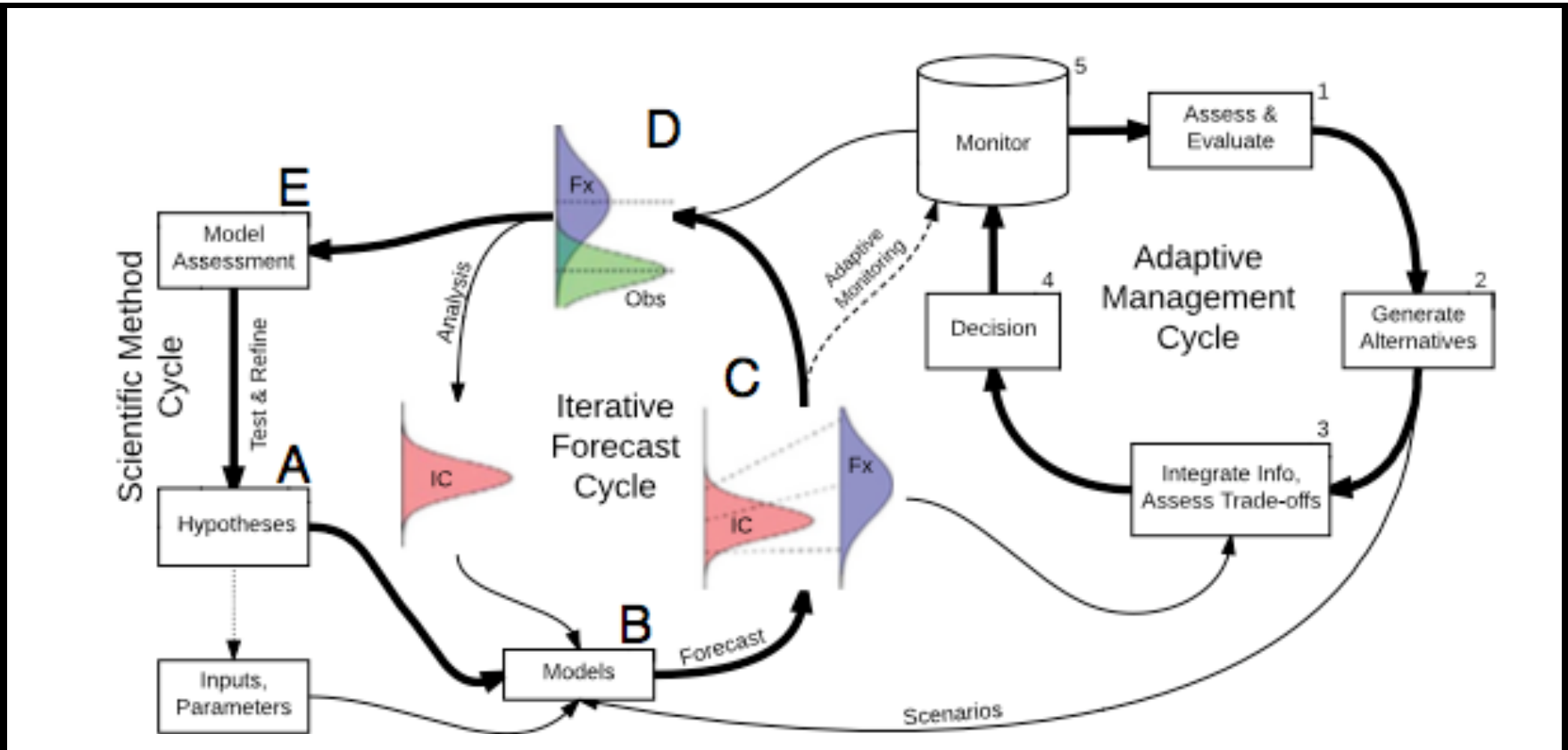
Bonan 2008


Sites with drought in summer 2012
Wolf et al (2016) PNAS
Me2
Me6
Ton
Var
NR1
KON
KFS
Vcp
Vcm
Pfa
UMB
MMS
Slt
Ced
Ho1
Bar
NLCD Land Cover Classification Legend
11 Open Water
12 Perennial Ice/ Snow
increased vegetation activity
reduced vegetation activity
ΔGPP
150
50
−50
−150
SWC (%)
100
75
50
25
0
Baseline
2012
−12%
−30%
−9%
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec


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

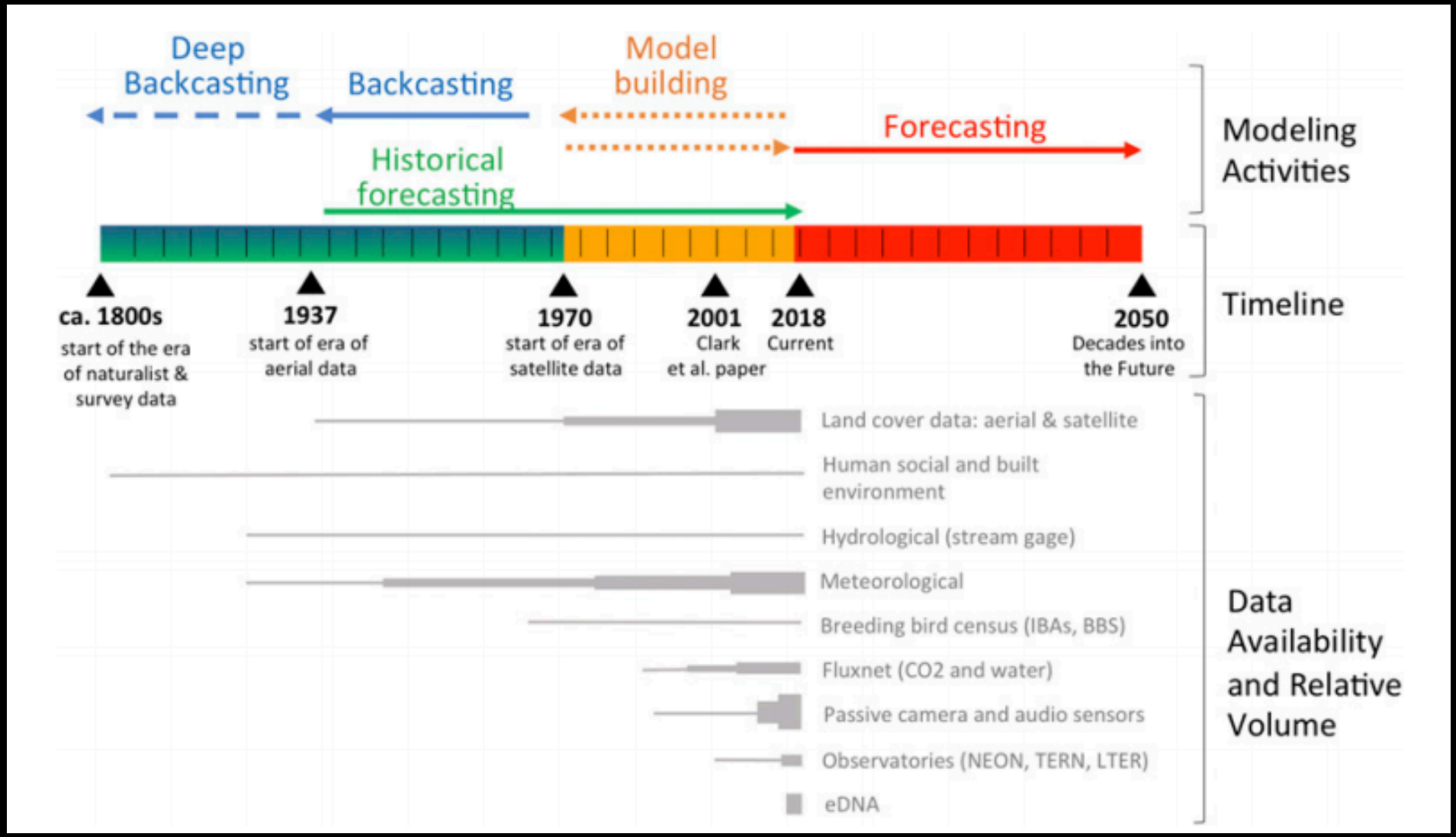
Iterative near-term ecological forecasting: Needs, opportunities, and challenges

Michael C. Dietze[a,1], Andrew Fox[b], Lindsay M. Beck-Johnson[c], Julio L. Betancourt[d], Mevin B. Hooten[e,f,g], Catherine S. Jarnevich[h], Timothy H. Keitt[i], Melissa A. Kenney[j], Christine M. Laney[k], Laurel G. Larsen[l], Henry W. Loescher[k,m], Claire K. Lunch[k], Bryan C. Pijanowski[n], James T. Randerson[o], Emily K. Read[p], Andrew T. Tredennick[q,r], Rodrigo Vargas[s], Kathleen C. Weathers[t], and Ethan P. White[u,v,w]



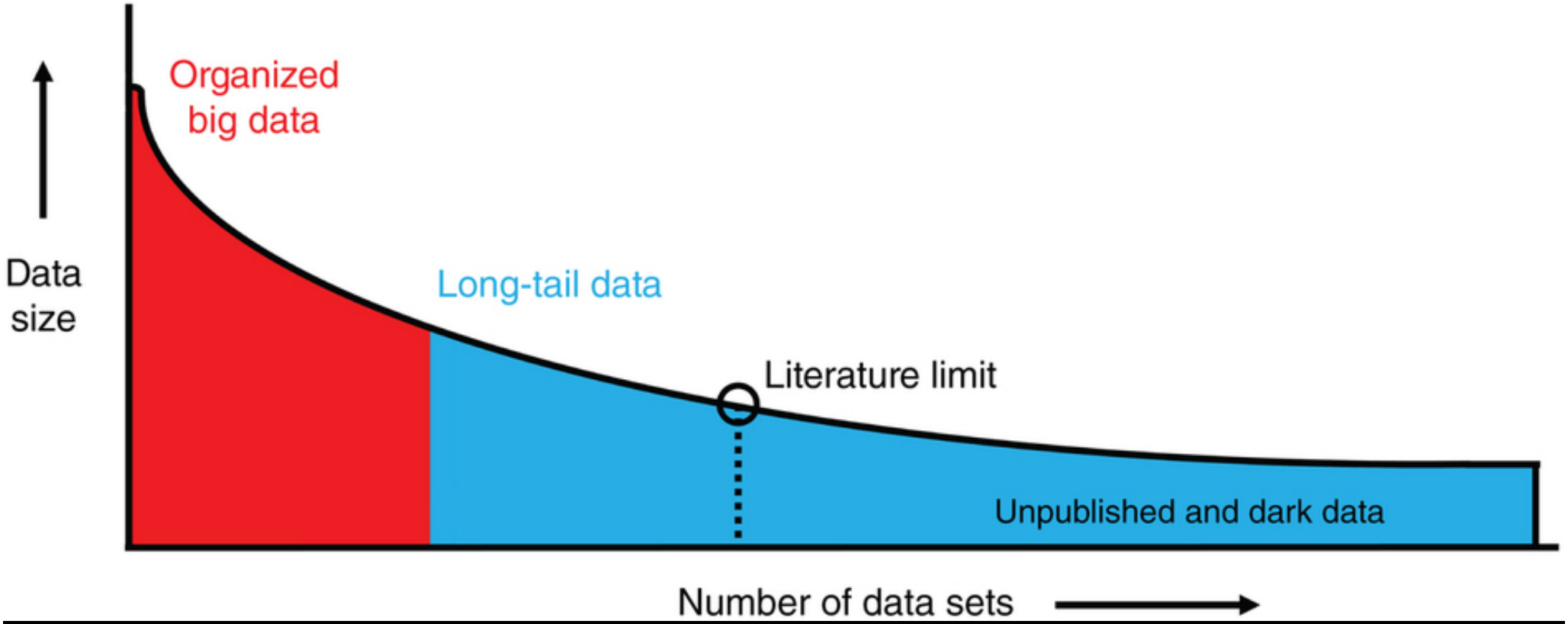
www.pnas.org/cgi/doi/10.1073/pnas.1710231115

Observations are big and long!



PNAS 2018; published ahead of print January 30, 2018, https://doi.org/10.1073/pnas.1710231115

And hard to extract from literature!



Ferguson et al., 2014
Nature Neuroscience

Agricultural and Forest Meteorology 249 (2018) 444–456

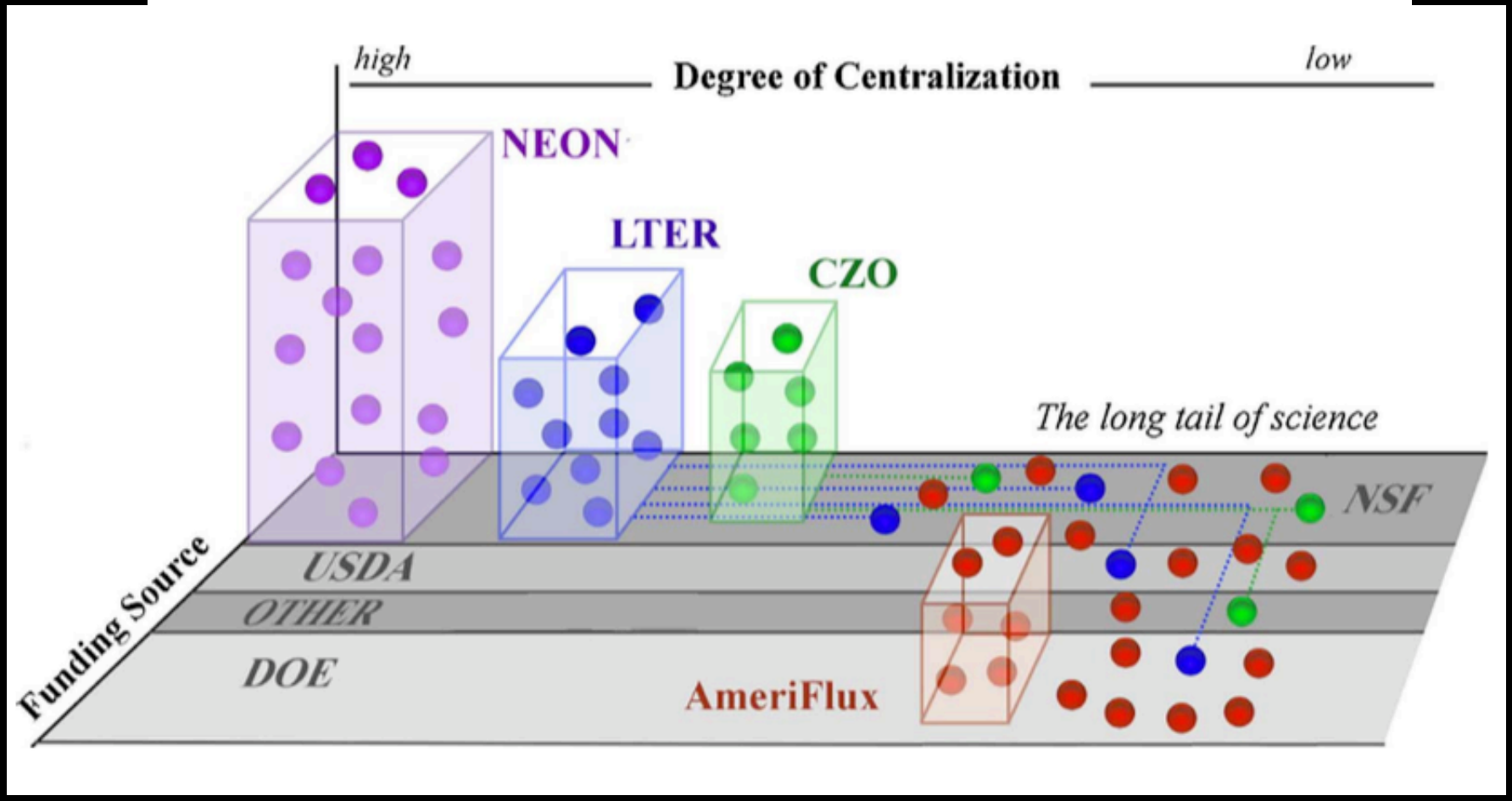
Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

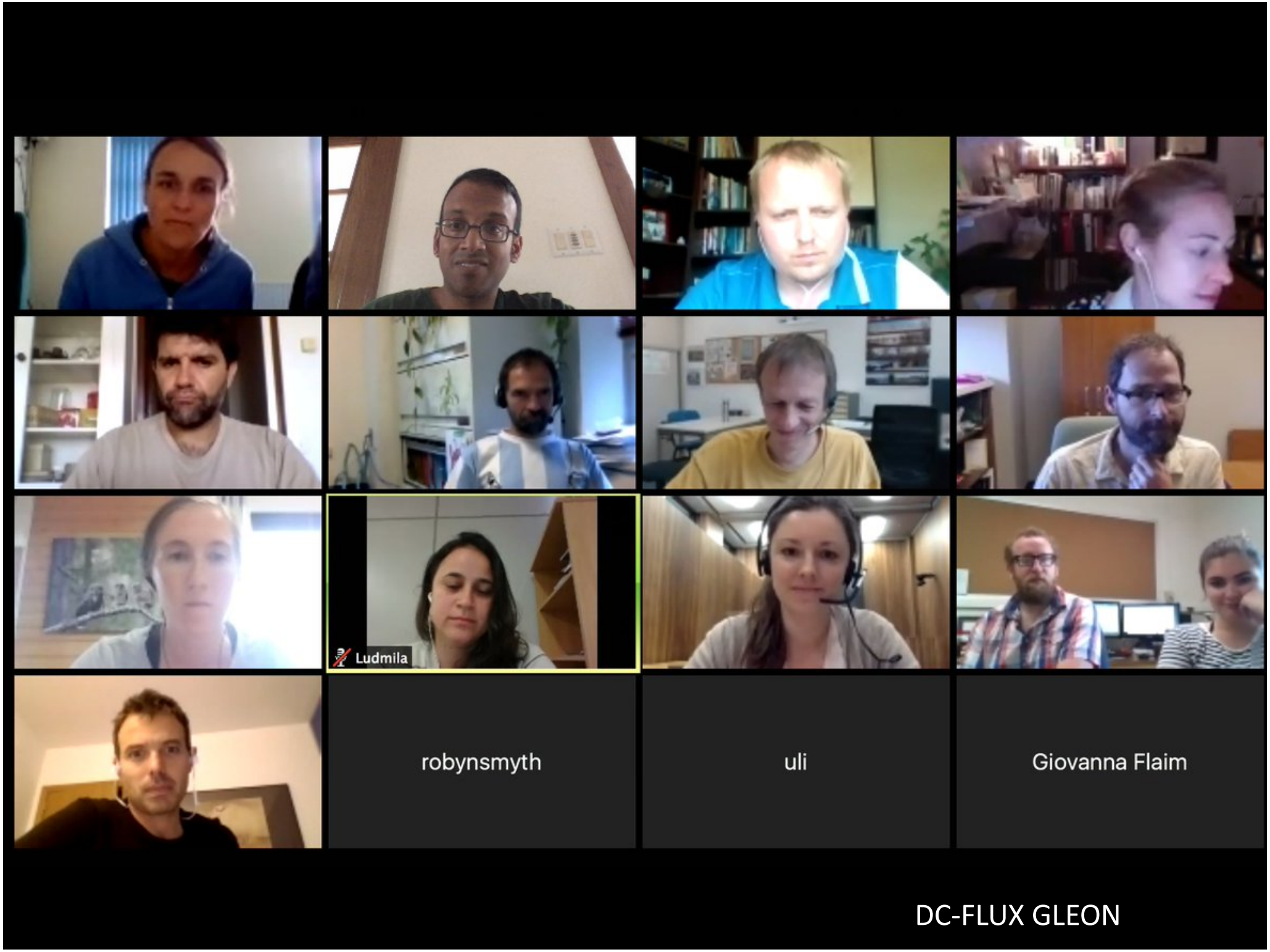
The AmeriFlux network: A coalition of the willing

K.A. Novick[a,*], J.A. Biederman[b], A.R. Desai[c], M.E. Litvak[d], D.J.P. Moore[e], R.L. Scott[b], M.S. Torn[f]




WISCONSIN

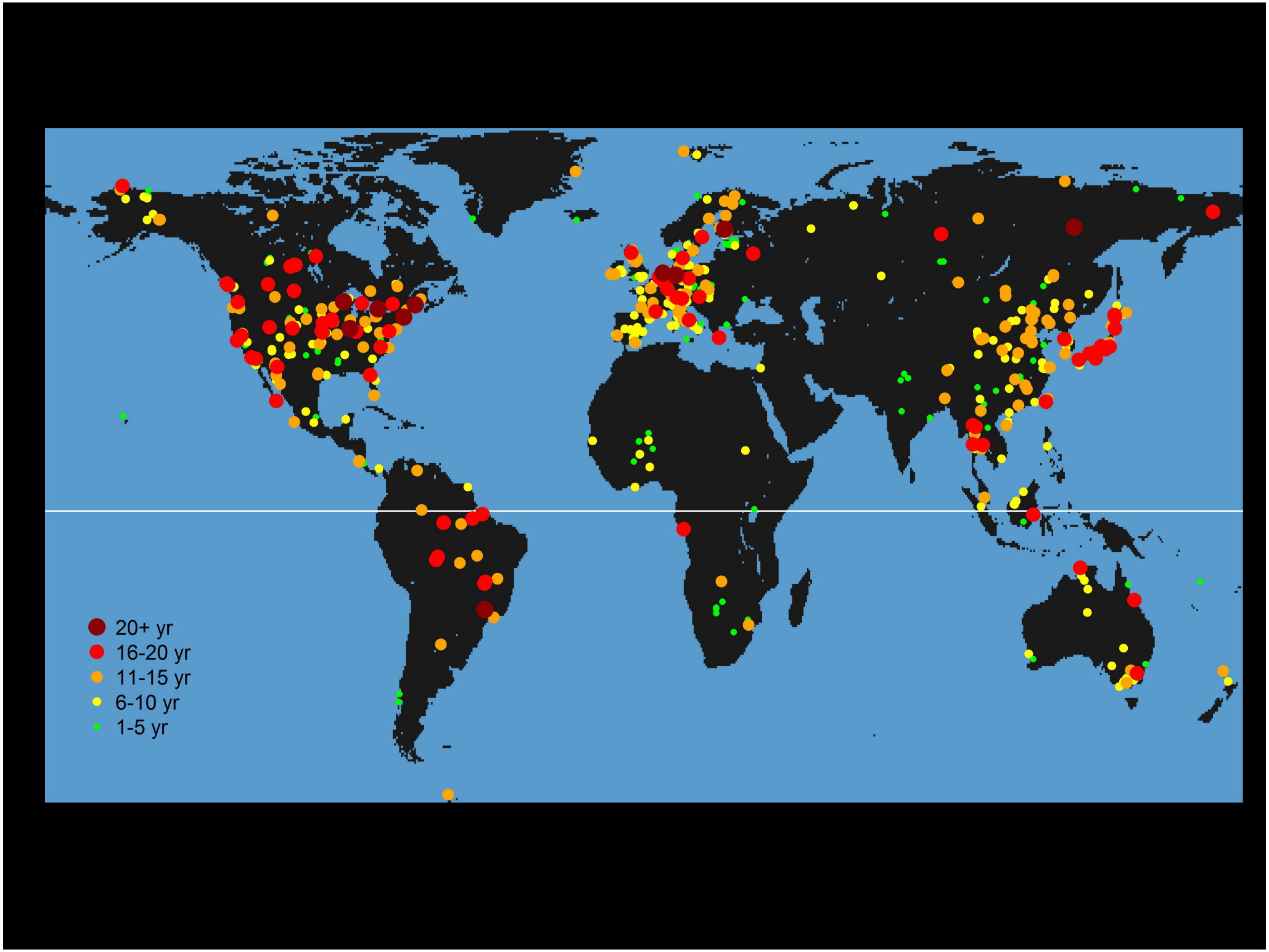

Twins



DC-FLUX GLEON

What are traits of a Positive Informatics *Culture?*


OPEN
WE ARE...
Lion Proud!
MARCH FOR SCIENCE
BOSTON


20+ yr
16-20 yr
11-15 yr
6-10 yr
1-5 yr

Agricultural and Forest Meteorology 249 (2018) 444–456

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

The AmeriFlux network: A coalition of the willing

K.A. Novick[a,*], J.A. Biederman[b], A.R. Desai[c], M.E. Litvak[d], D.J.P. Moore[e], R.L. Scott[b], M.S. Torn[f]

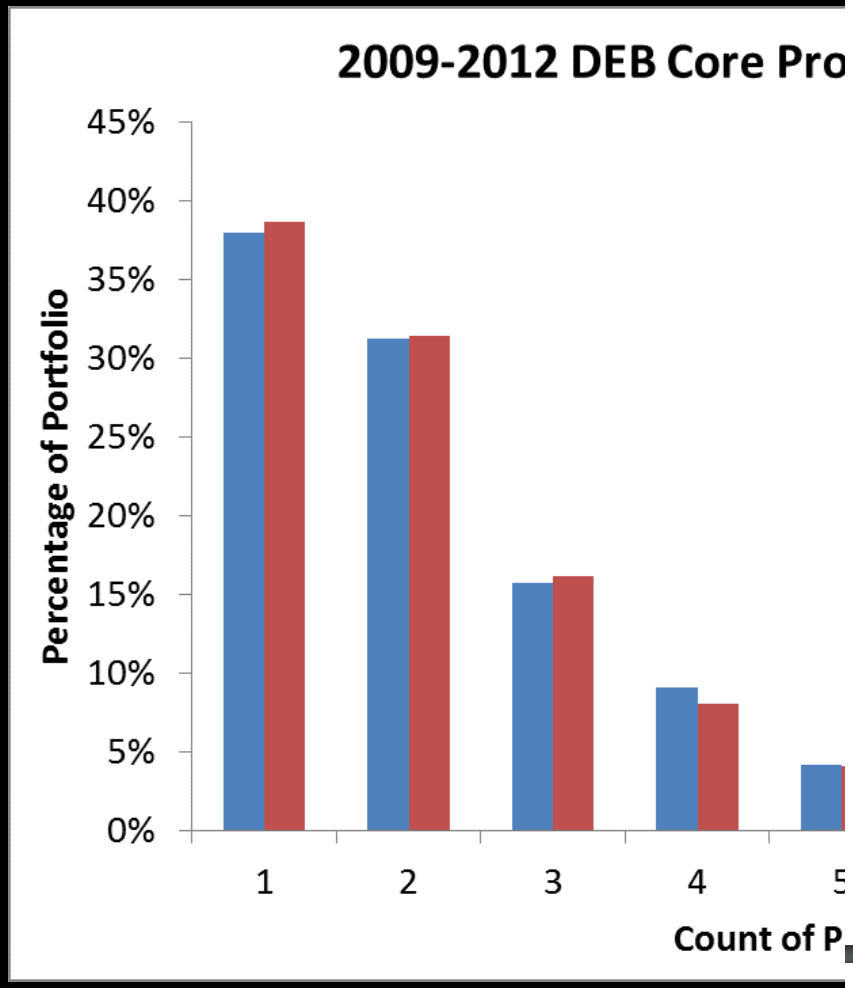
Table 3
Strengths and weaknesses of AmeriFlux's bottom-up approach.

Feature of Approach	Associated Strengths	Associated Weaknesses
Voluntary, PI-driven research; inclusive approach to network participation	Diverse research questions; interdisciplinarity; strong sense of community Good spatial and temporal representativeness of many biome types.	Lack of incentives for data sharing. Insecurity of funding for many sites. Underrepresentation of some biomes.
Lack of standardization of instrumentation and processing	Flexibility in methodological approach can advance observation theory.	Biases related to instrument design and processing can challenge cross-site syntheses.
"collaborative' data policy	Promotes cross-disciplinary perspectives; strengthens interpersonal connections within the network; promotes incentive for PIs to submit data	Large, multi-author papers are sometimes challenging to write, presenting a disincentive for network end-users.
Network oriented around a relatively few core observations (i.e. fluxes and meteorological drivers)	Few required variables makes it easier for sites to join the network	Inconsistent submission of non-biometeorological data across sites, which when present provides important ecological context for the fluxes, and guides model development.

WED COS 82
Dave Moore
Kim Novick
The all singing all dancing ecologist? Forming communities of practice by cross training graduate students in empirical and modelling approaches
3:20 pm
Rm 245


COLLABORATIVE

Most scientific pr
multi-PI, multi-ins



https://debblog.nsfbio.com/2013/07/23/deb-numbers-award-size-and-duration/

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

Climate control of terrestrial carbon exchange across biomes and continents

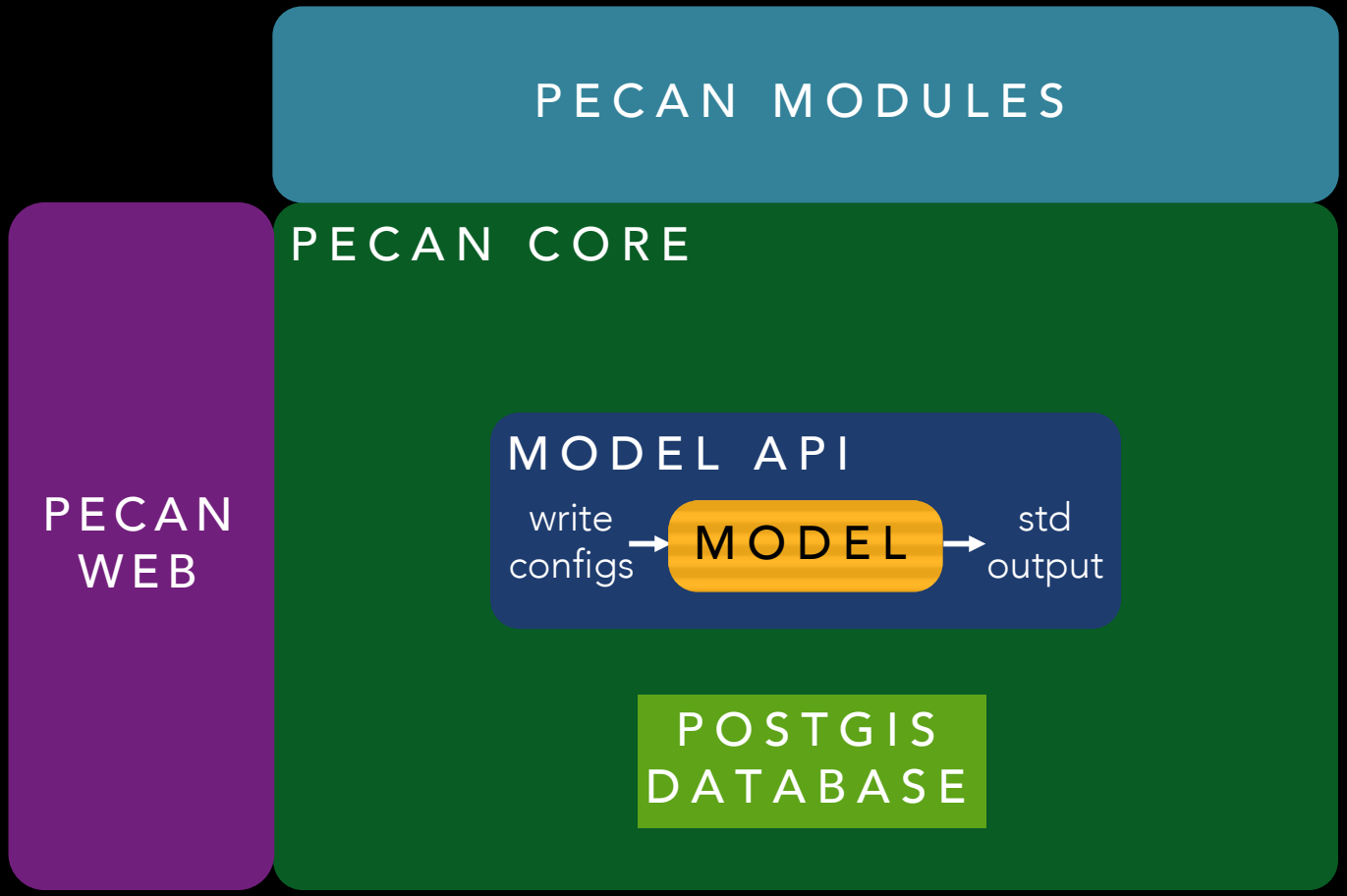
Chuixiang Yi[1], Daniel Ricciuto[2], Runze Li[3], John Wolbeck[1], Xiyan Xu[1], Mats Nilsson[4], Luis Aires[5,117], John D Albertson[6,117], Christof Ammann[7,117], M Altaf Arain[8,117], Alessandro C de Araujo[9,117], Marc Aubinet[10,117], Mika Aurela[11,117], Zoltán Barcza[12,117], Alan Barr[13,117], 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]

Le Bauer *et al.*, *Ecol. Mono.*, 2013

M Dietze
Boston U



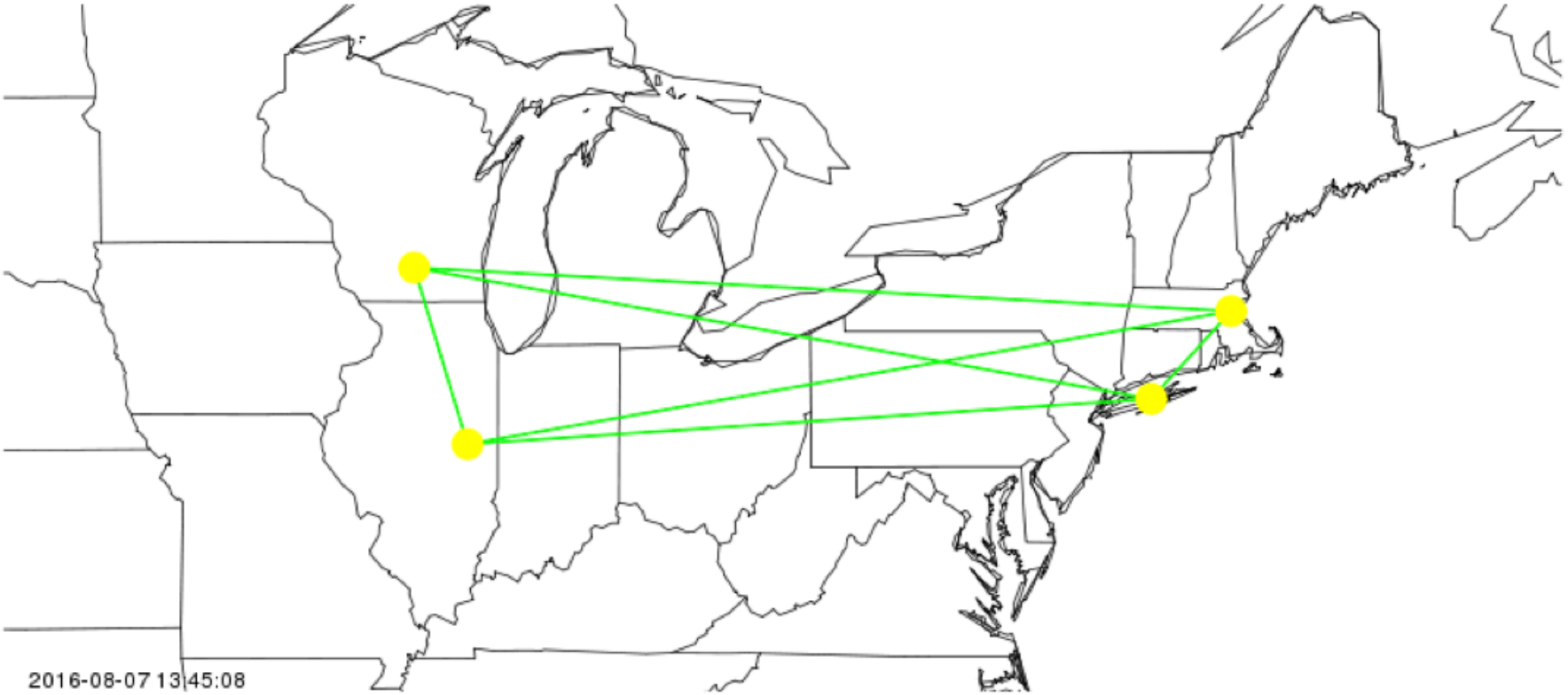
Standardized inputs and outputs

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 “space telescope” 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 usability tools: eddy4R-Docker image

- Docker = shipping container system for code

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

1 **eddy4R: A community-extensible processing, analysis and**
2 **modeling framework for eddy-covariance data based on R,**
3 **Git, Docker and HDF5**
4
5 **Stefan Metzger[1], David Durden[1], Cove Sturtevant[1], Hongyan Luo[1], Natchaya**
6 **Pingintha-Durden[1], Torsten Sachs[2], Andrei Serafimovich[2], Jörg Hartmann[3],**
7 **Jiahong Li[4], Ke Xu[5], Ankur R. Desai[5]**

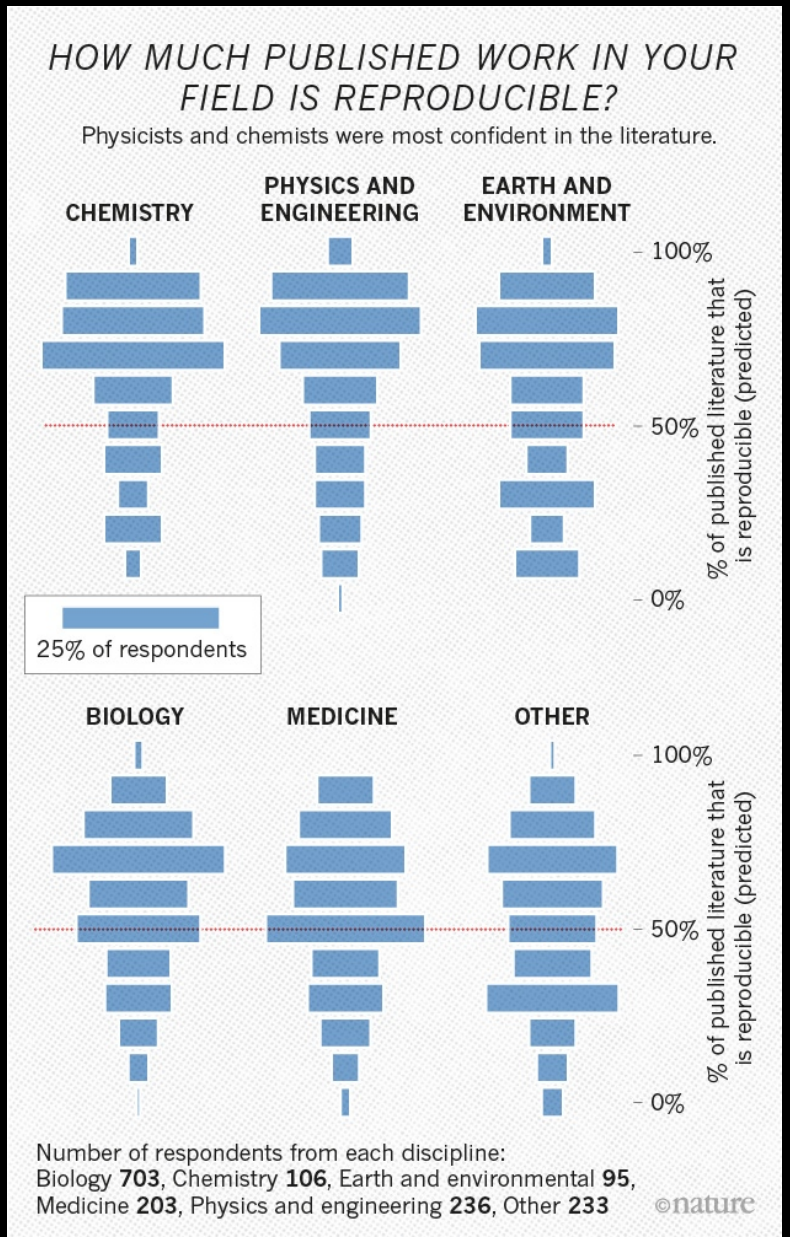
Enabling FAIR Data

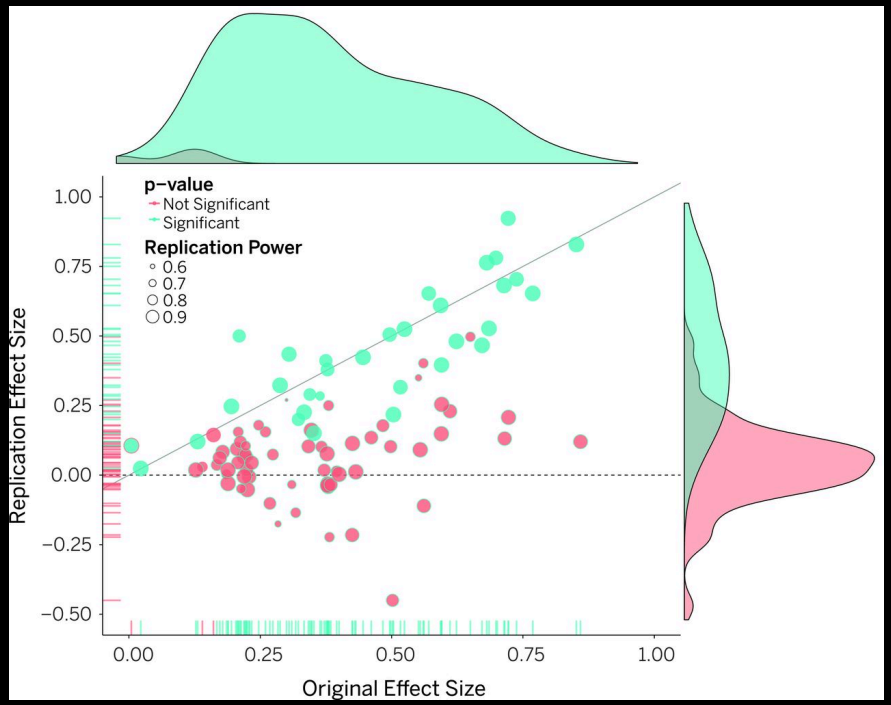
- Findable
- Accessible
- Interoperable
- Re-Usable

- https://www.force11.org/group/fairgroup/fairprinciples
- http://www.copdess.org/enabling-fair-data-project/commitment-to-enabling-fair-data-in-the-earth-space-and-environmental-sciences/


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

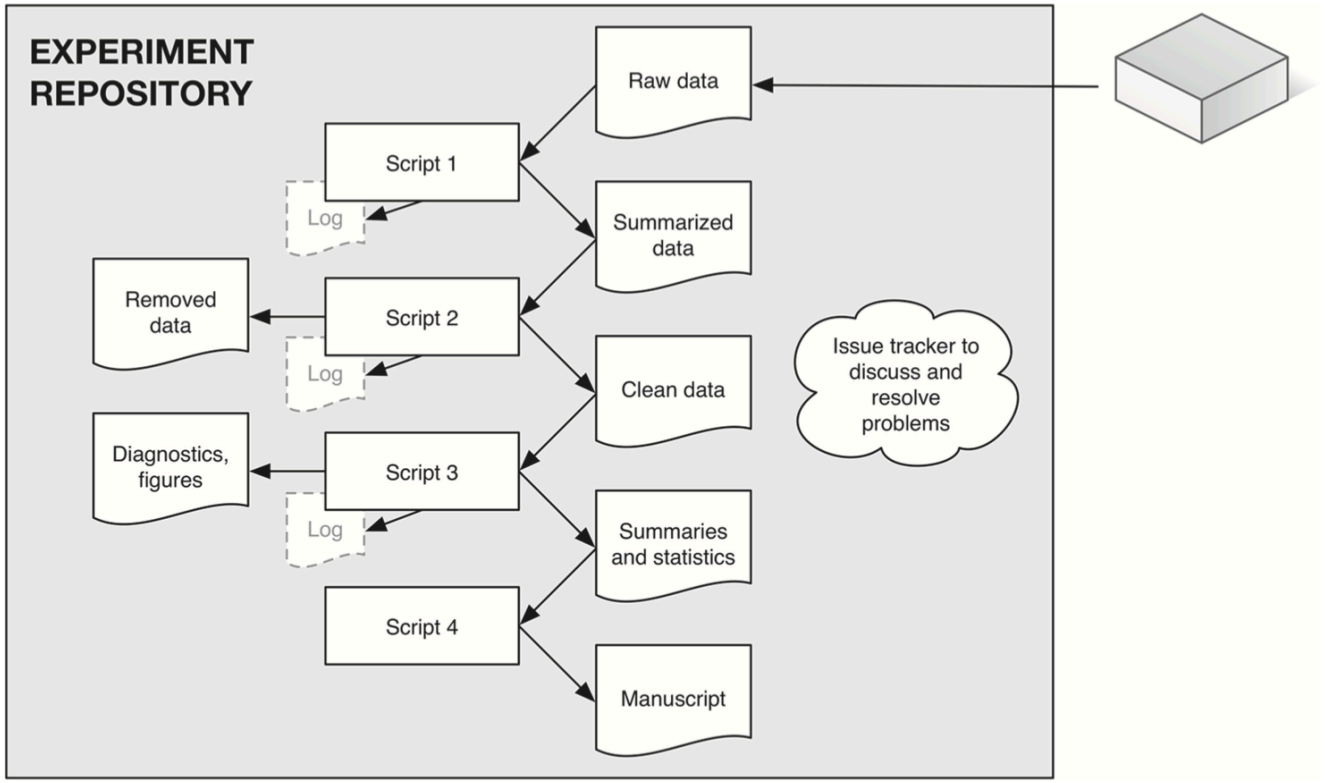
Environmental Research Letters

https://github.com/bpbond/cpcrw_incubation

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]



Missing/problematic data

[redacted]

Paul Stoy Retweeted

Ben Bond-Lamberty @BenBondLamberty · Aug 1

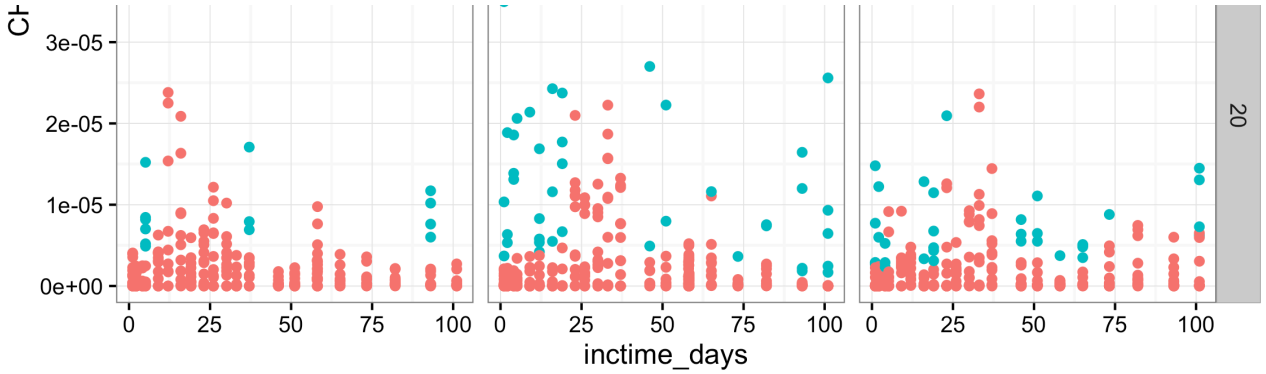
Terrible internet here in rural Ontario, but: our @Nature paper published today entirely #openscience #opendata. All results & figures reproducible, and you can look through commits to see initial idea, final tweaks, and all my curse words in between.

bpbond/rh-changes

Testing for changes in global Rh. Contribute to rh-changes development by creating an account on GitHub.

github.com

[redacted]



Big data is not open, collaborative sharable, nor reproducible 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

How do we encourage and support an *open, collaborative, sharable, reproducible* informatics culture?

- Training for students *and* us old farts
- Best practices: GLEON, LTER, Ameriflux
- Science of Team Science
  - NSF RCN MSB Grassroots Global Network Sciences
- Pilot projects for new collaboration methods
  - Will some please invent an actually usable collaborative video-conference platform?
- Funding support for data archival and informatics
  - Digitization/generation of metadata for long-tail data
  - The mantra does NOT have to be centralization
  - CyVerse
- … what else?

NIH NATIONAL CANCER INST

COLLABORATION • TEAM SCIENCE •

FIELD GUIDE

L. Michelle Bennett
Howard Gadlin
Christophe Marchand

TRUST

It is almost impossible
trust. Trust provides th
nearly impossible to su

VISION

A strong and captivati
provides a foundation
provides a focal point a
coalesce.

SELF-AWARENESS

Emotional Intelligence
the effective functioni
people greater control
improves the quality o
awareness.

LEADERSHIP

Strong collaborative le
members' strengths ar
Leadership can be der
the formal leader(s).

MENTORING

Mentoring is an indisp
mentor recognizes the
areas in which newer s
and can help coach pe
mentoring, the develo
strengthening team dy

TEAM EVOLUTION AND DYNAMICS

Research teams form and develop through critic
their highest potential (Forming, Storming, Norm
positive team dynamic sustains and further stren
team, enabling it to achieve successful outcomes

COMMUNICATION

Effective communication within and outside a re
contributes to effective group functioning. It dep
safe environment where team members can ope
discuss new scientific ideas and take research int
unconsidered directions as well as ensure that di
can take place.

RECOGNITION AND SHARING SUCCESS

Individual contributions should be recognized, re
rewarded in the context of a collaboration. Recog
of all team members should be done thoughtful
context of the team and the institution.

CONFLICT AND DISAGREEMENT

Conflict can be both a resource and a challenge–
because disagreement can expand thinking, add
to a complex scientific problem, and stimulate ne
for research. A challenge because if it is not hand
conflict impedes effective team functioning and
advancement.

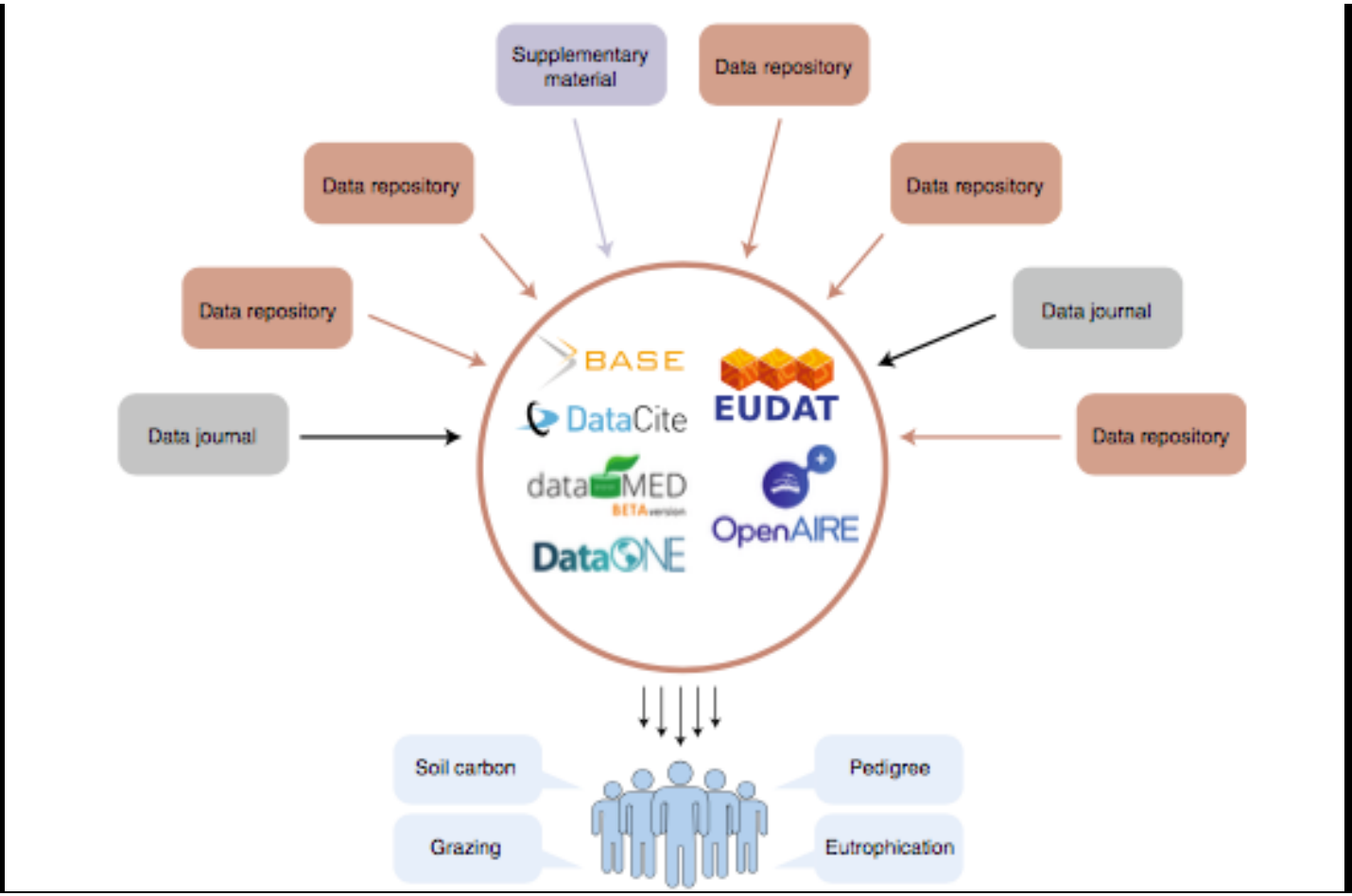
NAVIGATING AND LEVERAGING NETWORKS AND SYSTEMS

Highly collaborative teams can transcend differe
structures, extending their reach across and beyo
organization. They often function within the cont

PERSPECTIVE
https://doi.org/10.1038/s41559-017-0458-2
nature ecology & evolution

Navigating the unfolding open data landscape in ecology and evolution

Antica Culina[1*], Miriam Baglioni[2], Tom W. Crowther[1,3], Marcel E. Visser[1], Saskia Woutersen-Windhouwer[1] and Paolo Manghi[2]



THANK YOU!

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

Funding: NSF Advances in Biological Informatics (**ABI-1457897 , ABI-1062205)**
NSF MSB RCN (**EF-1702991**), NSF NTL LTER (**DEB- 1440297**), DOE Ameriflux Project