Informatics, Data Sharing, & Infrastructure

Atmospheric CO$_2$ at Mauna Loa Observatory

Ankur R Desai
Dept of Atmospheric and Oceanic Sciences
University of Wisconsin-Madison

February 2017
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
Trend expected from fossil-fuel burning

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
CMIP4
2006
DATA!!! Om nom nom nom...
Net Carbon Uptake Anomaly @ sites (EC)

Wolf et al (2016) PNAS
• Data synthesis: volume, diversity
• Modeling not scalable
• Models are not accessible

NO EASY WAY FOR NON-MODELERS TO HELP
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 projects are multi-PI, multi-institutional.

2009-2012 DEB Core Program

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

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

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
Environmental Research Letters

LETTER

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

Ben Bond-Lamberthy, A Peyton Smithz and Vanessa Baileyz

https://github.com/bpbond/cpcrw_incubation
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

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

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