Strategic Data Management to Advance Environmental Systems Science

Conveners:
Joan Damerow, Lawrence Berkeley National Laboratory, ESS-DIVE JoanDamerow@lbl.gov
Christin Buechner, Lawrence Berkeley National Laboratory, Ameriflux Cbuechner@lbl.gov

2024 ESS PI Meeting
April 16, 2024

What is FAIR data?

F **Findable:** Metadata and data should be easy to find for both humans and computers.

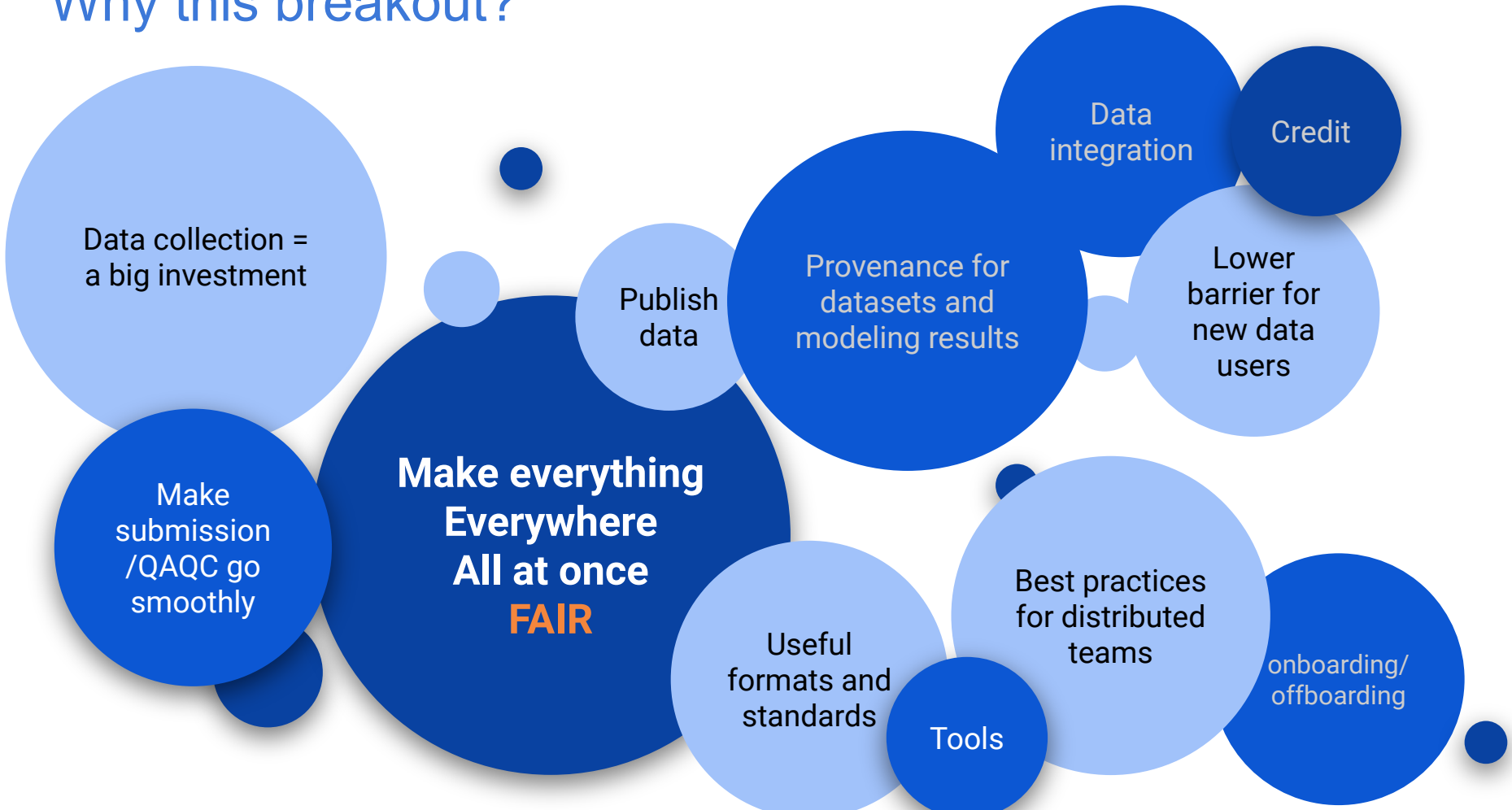
A **Accessible:** The exact conditions under which the data is accessible should be provided in such a way that humans and machines can understand them.

I **Interoperable:** The (meta)data should be based on standardized vocabularies, ontologies, thesauri etc. so that it integrates with existing applications or workflows.

R **Reusable:** Metadata and data should be well-described so that they can be replicated and/or combined in different research settings.

Why this breakout?

Why this breakout?



Session Agenda

Presentations on ESS Data Management and Use






- Data Repositories
  - Joan Damerow (ESS-DIVE)
  - Housen Chu (Ameriflux Management Project)
- Project Strategies for Data Management
  - Ben Bond-Lamberty - COMPASS-FME
  - Amy Goldman - River Corridor and Biogeochemistry SFA
- Data Integration and Use (NOAA)
  - Youmi Oh (NOAA)

Q&A with the panel

Group Brainstorming: community solutions for data management and integration challenges

ESS-DIVE Data Repository Overview

Joan Damerow

ESS PI Meeting 2024, April 16

What is ESS-DIVE?

Data Repository to

preserve

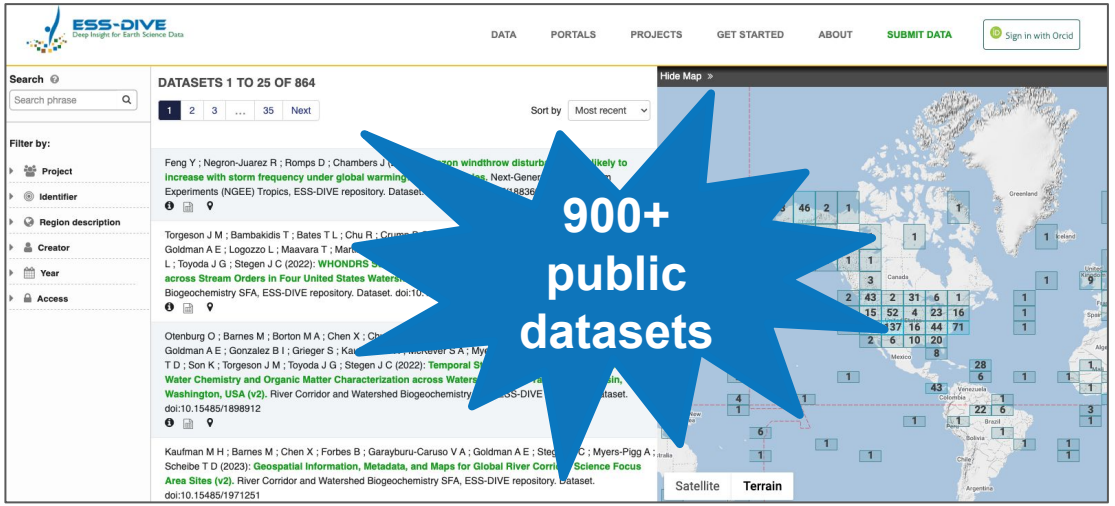
expand **access** to

and improve **usability** of

ESS data

Findable Accessible Interoperable Reusable

data.ess-dive.lbl.gov



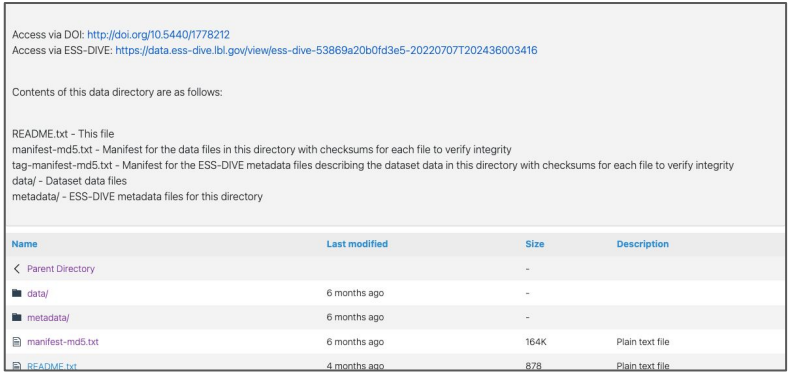
Use of ESS-DIVE is mandatory for ESS project data management plans

Scope of ESS-DIVE Data

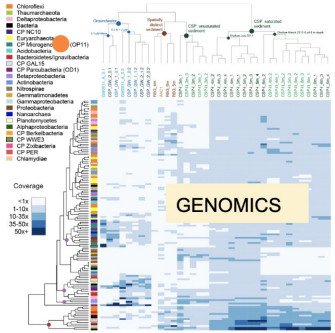
- Data from projects funded by or related to DOE Environmental System Science Program
- Observational, experimental, and modeling activities in the Earth sciences.

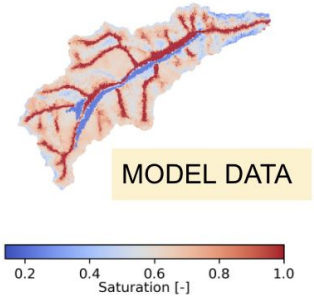
Support for Large Files and Hierarchical Datasets

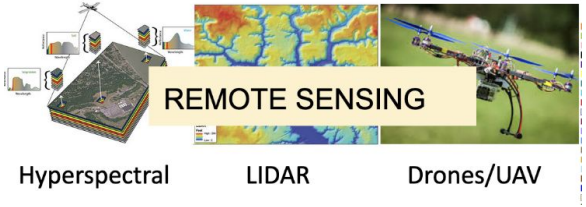
- Large Data: 500GB - multiple TB range
- Browsable Folder Hierarchy
- Globus and HTTP Access
- Email ess-dive-support@lbl.gov if you need to store large data

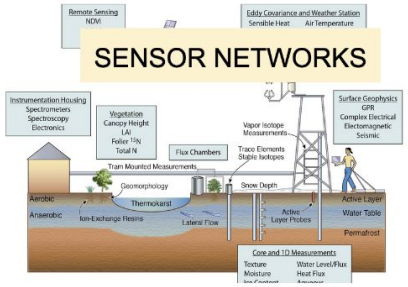


ESS-DIVE Tier 2 storage interface with UAS data example







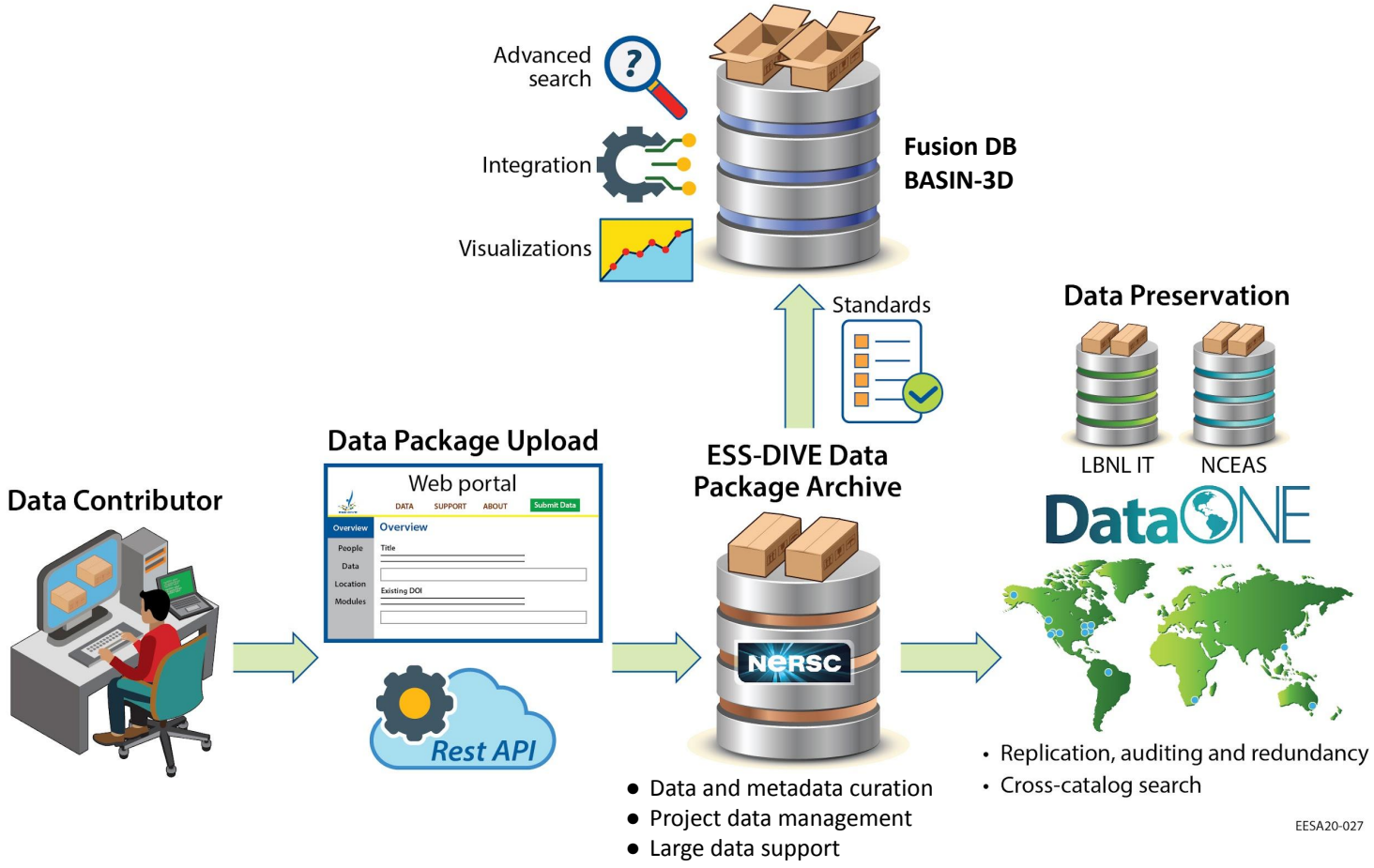


ESS-DIVE At-a-glance

Authorized users	Public datasets	Private datasets	Total file size	Number of Files	Contributing projects
383 (498%)	909 (144%)	266 (1673%)	2.82TB	16,762	90 (233%)
Public dataset updates	**Temporal coverage**	**Public Portals**	**File downloads**	**Data views**	**Support Requests**
102 (1600%)	1841 to 2024	15 (650%)	429.1K+ (504%)	753.8K+ (510%)	1333 (256%)

Growth statistics measured over the period from May 2020 - March 2024

Easy Data Submission & Broad Data Availability

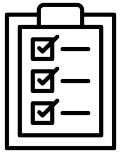


ESS-DIVE provides …

Project data management features for teams

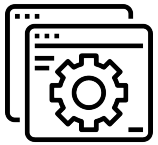
Community reporting formats for several ESS data types to enable advanced search

Guidance related to data packaging and authorship



Curation of metadata and ensures file readability

Digital Object Identifier and citation text for published data

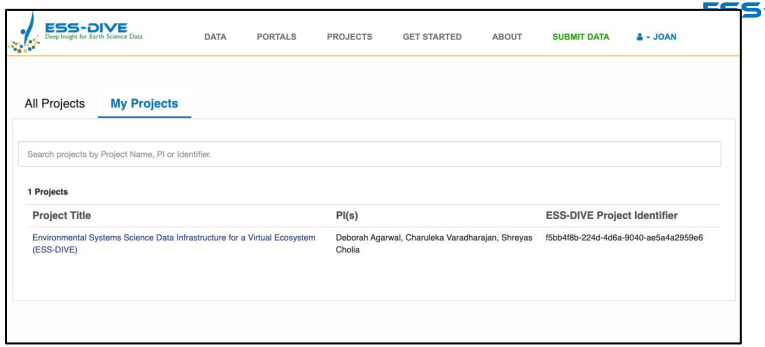


Long-term data search and download availability

ESS-DIVE is the DOE BER ESS location for long-term data preservation

Project Data Management Tools

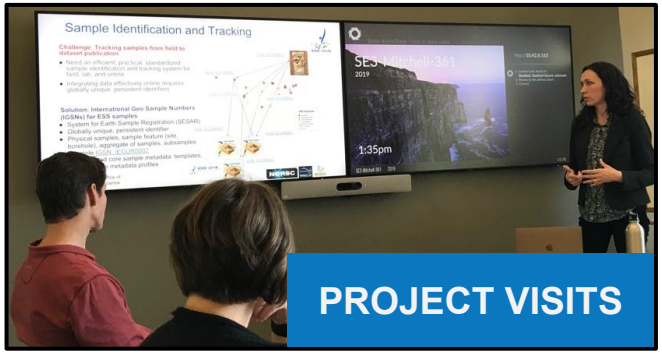
- **Project teams** for internal sharing, collaboration and curation
- **Bulk data uploads, updates** using API
- Project-centric data **search** and **portals**
- **COMING SOON!! Dataset tracking and reserving DOIs**
- **NEED INPUT!!** Project data **citations and metrics**

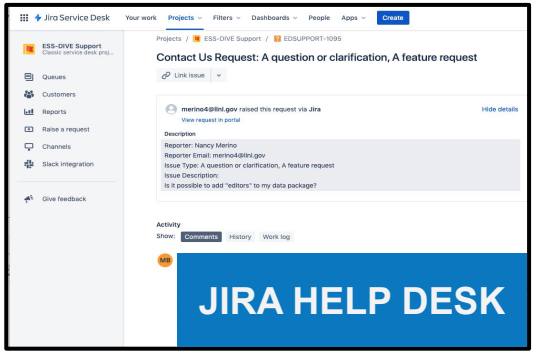


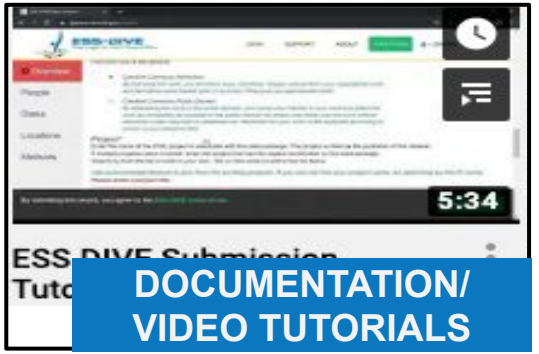
Lunchtime Breakout: ESS-DIVE Tutorial for PIs and Data Managers
Wed., April 17th; 12:30-1:30pm EST

PIs will walk away with a To-Do List

Community Engagement Activities Guided Feature Development and Identified Challenges







Example Challenge: Connect Interdisciplinary Sample Data and other Research Outputs

Tracking Samples and resulting data that was sent to numerous collaborators and labs, for a variety of bio and environmental analyses, and then compiled, analyzed, and published in numerous files and data systems.

Current Practices:

- Sample records are rarely published with standard metadata and identifiers
- SESAR provides persistent identifiers, metadata, management
- Use of sample PIDs rare, inconsistent

Challenges:

- The process of submitting Samples to different data systems and labs, and then compiling the resulting data is currently inefficient and even prone to error
- Sample data and associated information disconnected
- Currently no infrastructure to track provenance, metrics, etc

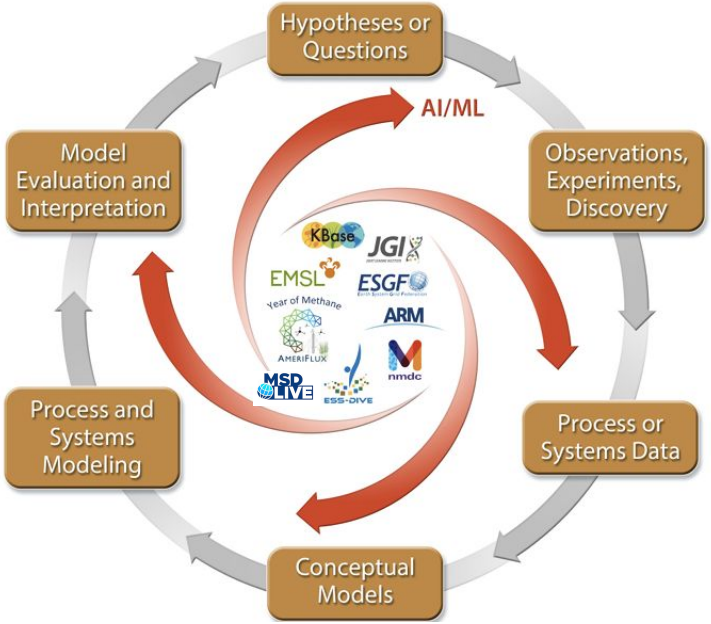
BER Data Systems and Project Use Cases

Collaborative paper

1. Scientific use cases to link and exchange (meta)data
2. Identify community practices and infrastructure needs

Approach:

- 8 BER Data Systems, 10+ project use cases
- Connect related field to lab to model (meta)data
- Samples, ‘omics, sensors, remote sensing, model outputs, processed data, papers



Connecting BER Data for ModEx
Figure from AI4CH4 workshop report (US DOE 2023)
https://ess.science.energy.gov/ai4ch4/

ESS-DIVE Community Funds

- **$1M funds** available in FY24
- Submit **white paper** proposals by May 15 (max 2 pages)
  - Email with details and template will be sent later this week
- Preference given to **Priority Topics**
  - **New reporting formats**
  - **New versions of existing reporting formats:** Improve machine readability and compatibility with FusionDB, BASIN-3D etc., model data archiving for large outputs/ML
  - **Community data curators**: Provide guidance to ESS community on best practices for submitting data to ESS-DIVE & help with adoption of reporting formats
  - **Data integration**: Tools to integrate ESS-DIVE and other BER data
  - **Community data products:** Products using ESS-DIVE & BER data for broad scientific use
- Can propose other ideas that have clear value for the ESS community

AMERIFLUX
MANAGEMENT PROJECT

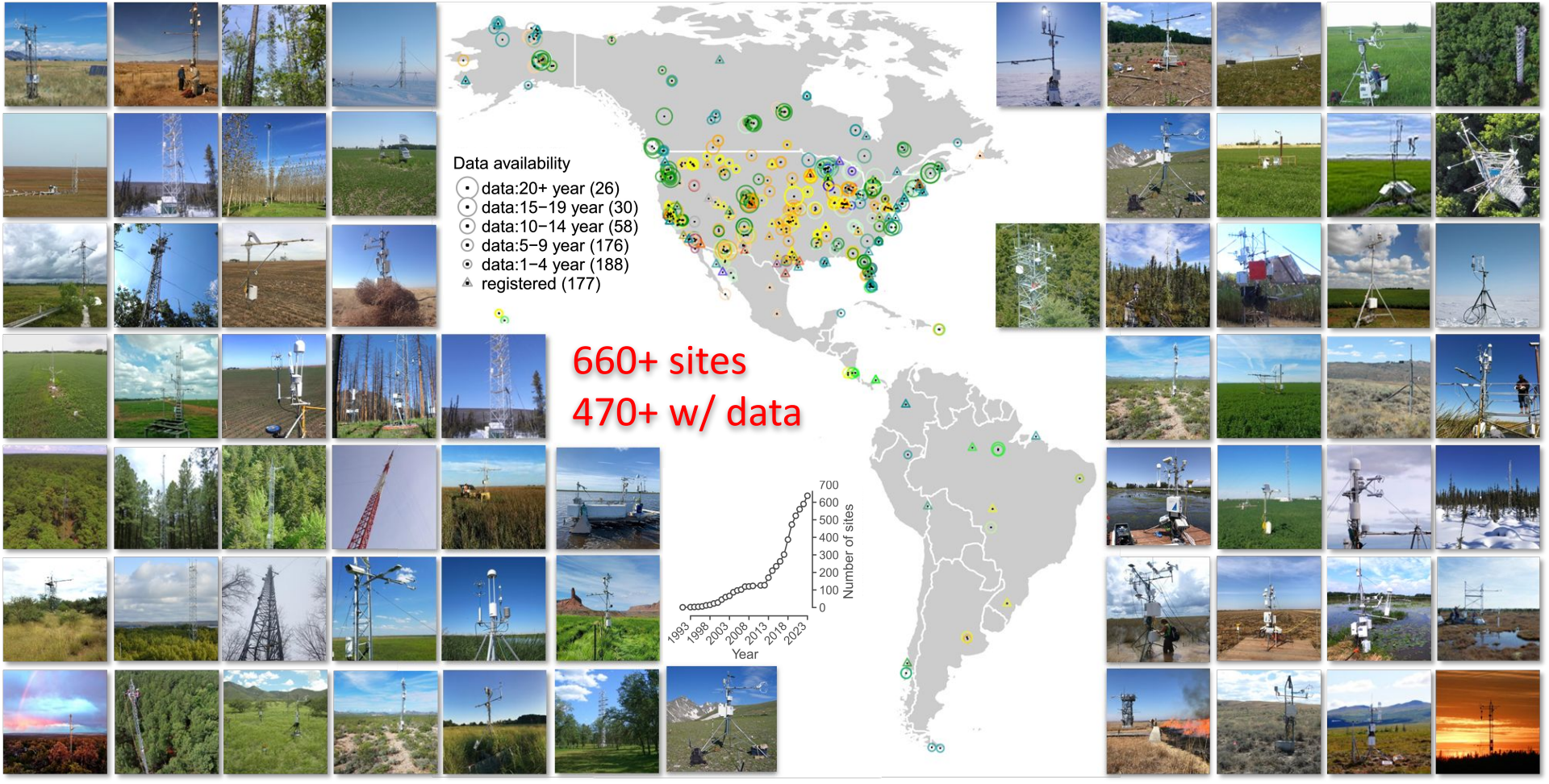
AmeriFlux Data Pipeline, QA/QC

Housen Chu
Lawrence Berkeley National Lab

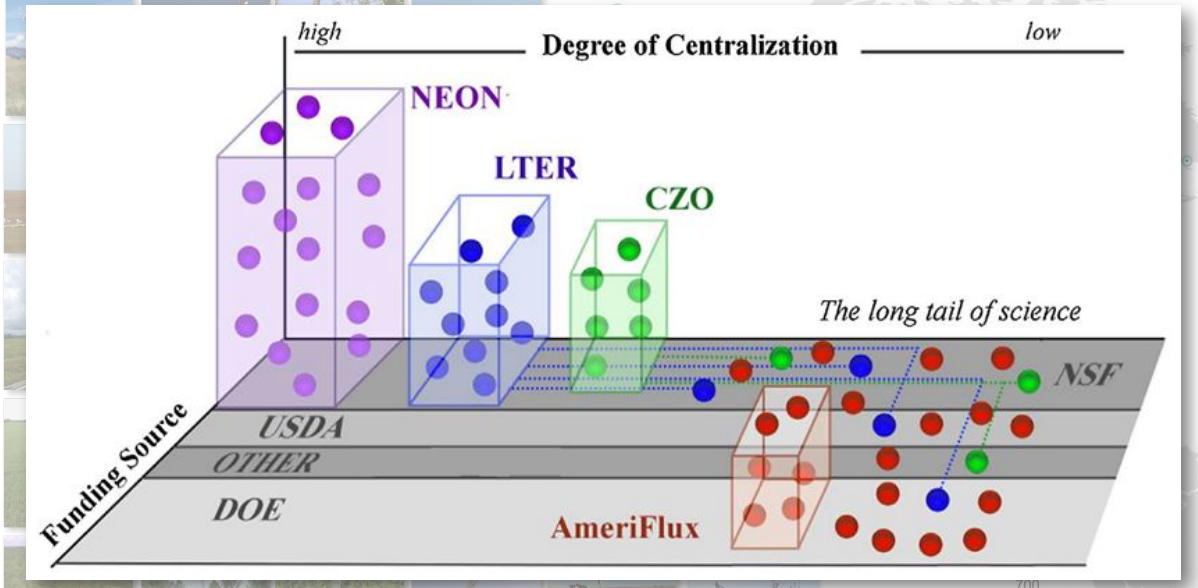
Margaret Torn, You-Wei Cheah, Sébastien Biraud, Trevor Keenan, Dennis Baldocchi, Stephen Chan, Housen Chu, Brian Wang, Sigrid Dengel, Rachel Hollowgrass, Fianna O'Brien, Gilberto Pastorello, Danielle Svehla-Christianson, André Luís Diniz dos Santos, and Koong Yi, Sy-Toan Ngo, Chad Hanson, Dario Papale, Christine Buechner

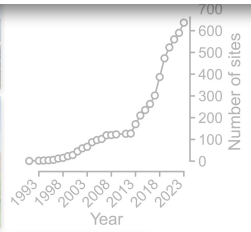
AmeriFlux – A network of flux towers

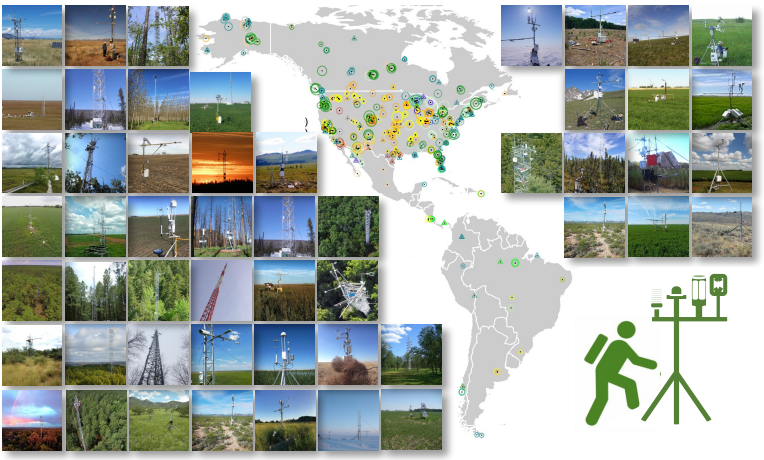
AMERIFLUX



AmeriFlux – A network of flux towers





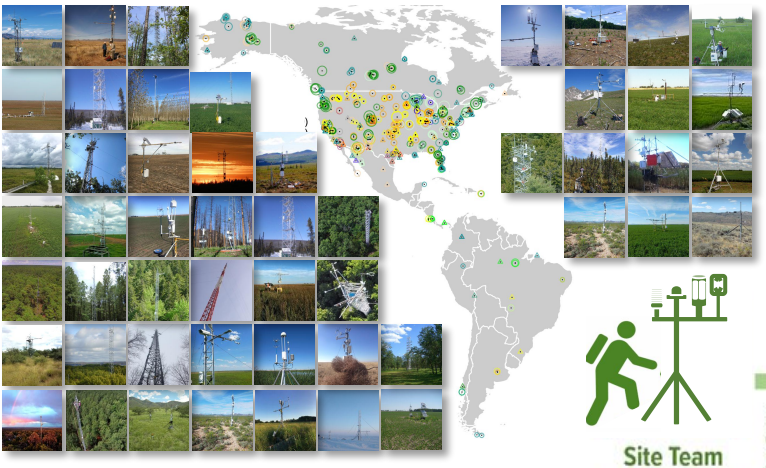


AMERIFLUX
MANAGEMENT PROJECT

Technical Support & QA/QC	Data Support & QA/QC
Biraud, Chan, Chu, Dengel	Agarwal, Cheah, Christianson, Chu, Pastorello
Core Site Support	**Outreach & Coordination**
Torn, Biraud, Buechner	Keenan, Torn, Christianson, Buechner

Core Sites | **AmeriFlux Network** | **Other Stakeholders**

Available to Data Users



AMERIFLUX MANAGEMENT PROJECT

Site Team
Data Provider

Submit BADM

Submit Flux/Met Data

AMP
(AmeriFlux Management Project)

Biological
Ancillary
Disturbance
Metadata

Iteration with site team to fix issues

Format QA/QC → Data QA/QC → BASE Publish → ONEFlux Processing → FLUXNET Publish

AmeriFlux Data Processing Pipeline

BASE Data Product

- QA/QC on general data quality
- Versioned / site package DOI
- >75% under CC-BY-4.0 policy

470+ sites
3400+ site-years

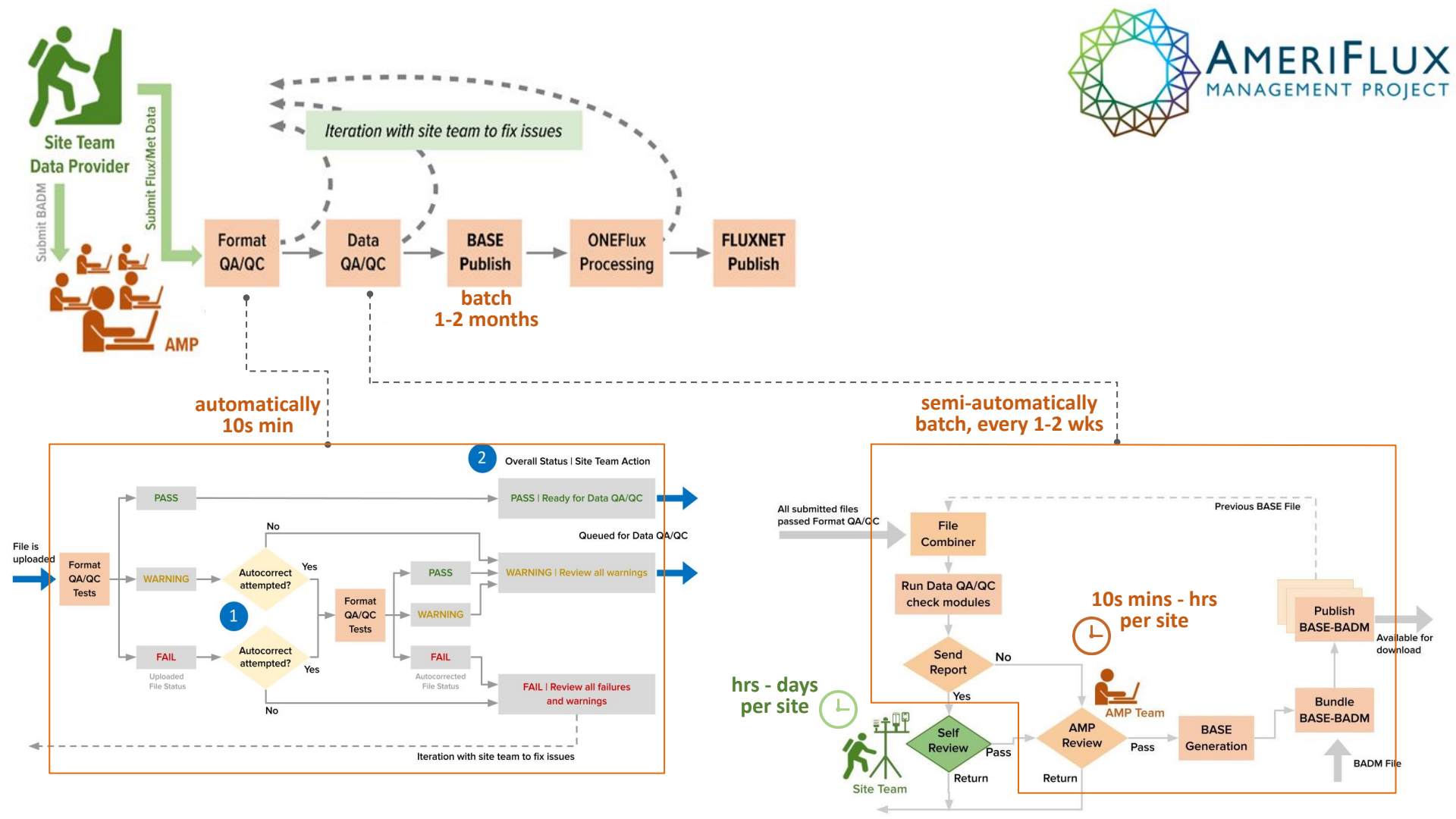
Available to Data Users

7000+ users

FLUXNET Data Product

- Value-added data product
- Versioned / site package DOI
- All under CC-BY-4.0 policy

226+ sites
1600+ site-years


AMERIFLUX
MANAGEMENT PROJECT
Site Team
Data Provider
Submit BADM
Submit Flux/Met Data
AMP
Iteration with site team to fix issues
Format QA/QC
Data QA/QC
BASE Publish
ONEFlux Processing
FLUXNET Publish
batch
1-2 months
automatically
10s min
semi-automatically
batch, every 1-2 wks
File is uploaded
Format QA/QC Tests
PASS
WARNING
FAIL
Uploaded File Status
Autocorrect attempted?
Yes
No
1
Format QA/QC Tests
PASS
WARNING
FAIL
Autocorrected File Status
2
Overall Status | Site Team Action
PASS | Ready for Data QA/QC
Queued for Data QA/QC
WARNING | Review all warnings
FAIL | Review all failures and warnings
Iteration with site team to fix issues
All submitted files passed Format QA/QC
File Combiner
Previous BASE File
Run Data QA/QC check modules
10s mins - hrs per site
Send Report
No
Yes
hrs - days per site
Site Team
Self Review
Pass
Return
AMP Team
AMP Review
Pass
Return
BASE Generation
Publish BASE-BADM
Available for download
Bundle BASE-BADM
BADM File

Format QA/QC

Format QA/QC Report

AMERIFLUX MANAGEMENT PROJECT

Overall status

Format QA/QC Summary

File link

QA/QC Report: Format

This report details results of the AmeriFlux QA/QC data processing pipeline.
For more information, see How to Read This Report, QA/QC Results Definition

PASS **Ready for Data QAQC**
No further action needed by the site team.

Uploaded File Report US-PFa_HR_201801010000_201901010000.csv

Site ID: US-PFa
Site contact:

Uploader:
Upload date: 2018-Jul-16 11:44
Uploaded filename: US-PFa_HR_201801010000_201901010000-201

Format QA/QC report summary:
All format QA/QC tests attempted. No issues were encountered. AMP w

Test	Results	A
All Format QA/QC tests passed.	✔ PASS	

Variable names found in the file:
TIMESTAMP_START, TIMESTAMP_END, CO2_1_1_1, CO2_1_2_1, C
CH4_1_1_1, CH4_1_2_1, CH4_1_3_1, FC_1_1_1, FC_1_2_1, FC_1_3
SCH4_1_1_1, H, H_1_1_1, H_1_2_1, H_1_3_1, LE, LE_1_1_1, LE_1_
SLE_1_2_1, SLE_1_3_1, WD_1_1_1, WD_1_2_1, WD_1_3_1, WD_F_
USTAR_1_1_1, USTAR_1_2_1, USTAR_1_3_1, USTAR_F_1_3_1, PA_
VPD_F_1_3_1, SWC_1_1_1, PPFD_IN_1_1_1, P, NEE, NEE_F, NEE_

Processing code version: 0.4.19
Processing log file: http://ameriflux-data.lbl.gov/QAQCLogs/QAQC_re

QA/QC Report: Format

This report details results of the AmeriFlux QA/QC data processing pipeline.
For more information, see How to Read This Report, QA/QC Results Definitions, F

WARNING **Review all warnings**
If autocorrected file is OK, no action is needed by the site tear

Autocorrected File Report US-MOz_HH_200501010000_200601010000.csv

Site ID: US-MOz
Site contact:

Uploader: AMP Data Team (original file uploaded by Format QAQC Pipeli
Upload date: 2018-Aug-15 17:27
Uploaded filename: US-MOz_HH_200501010000_200601010000-20180

Format QA/QC report summary:
All format QA/QC tests attempted. Issues were encountered. AMP attempt
warnings below. If autocorrected file is OK, no action is needed by the site

Test	Results	Add
AMP made these autocorrections.	? WARNING	• F ti
Any Variables suspected gap-fill?	? WARNING	The no r
Any Variables with ALL Data Missing?	? WARNING	The H_1 be c

Variable names found in the file:
TIMESTAMP_START, TIMESTAMP_END, P_1_1_1, PPFD_IN_1_1_1, PP
LW_IN_1_1_1, LW_OUT_1_1_1, NETRAD_1_1_1, TA_1_1_1, RH_1_1_1,
USTAR_1_1_1, TS_1_1_1, SWC_1_1_1, G_1_1_1, PA_1_1_1, FC_1_1_1,
NEE, NEE_F

Processing code version: 0.4.23
Processing log file: http://ameriflux-data.lbl.gov/QAQCLogs/QAQC_repor

Uploaded File Report US-MOz_HH_200501010000000_200601010000000.csv

⚠ **Consider** revising your file preparation for future submissions by opening and

QA/QC Report: Format

This report details results of the AmeriFlux QA/QC data processing pipeline.
For more information, see How to Read This Report, QA/QC Results Definitions, FAQ, and Upload Format Instructions

FAIL **Review failures and warnings**
Upload a corrected replacement file.

Uploaded File Report Tonzi-understory-2016.dat Report ID: 8

Site ID: US-Ton
Site contact:

Uploader:
Upload date: 2017-Mar-22 14:26
Uploaded filename: Tonzi-understory-2016-2017011912065058.dat

Format QA/QC report summary:
All format QA/QC tests attempted. Issues were encountered. Please correct issues and upload a replacement file.

Test	Results	Additional Information
Are Timestamp variables present?	⚠ FAIL	Expected timestamp variable(s) **TIMESTAMP_START, TIMESTAMP_END** is / are missing.
Timestamp problem encountered.	⚠ FAIL	Filename Matches File Contents, Timestamp Column Resolutio Timestamp Row Resolution, Timestamp Duplicates
Issues that cannot be autocorrected.	⚠ FAIL	Unable to repair timestamps. AutoRepair FAILED.
Is Filename Format valid?	⚠ FAIL	These filename components are not in the standard AmeriFlux format: **extension is not csv**
Are Timestamp variables as expected?	⚠ FAIL	These unexpected variables were found in columns 1 & 2 inste TIMESTAMP_START and TIMESTAMP_END: **yr, day**
Is Filename Format valid?	? WARNING	These filename components are not in the standard AmeriFlux format: **incorrect number of components (expect timestam errors)**
Are Data Variable names in correct format?	? WARNING	These variable names are not in standard AmeriFlux format: **yr day, endhour, endmin, DOY, FC_WPL_2D, fc_flag, WC_2D, CO2_LI7500, RHOC, CO2_var, CO2_skewness, CO2_kurtos RHOQ, q_var, q_skewness, q_kurtosis, Tsonic, Tsonic_var Tsonic_skewness, Tsonic_kurtosis, Wind_Direction, Wind_Velocity, Friction_Velocity, stdw, wbar, w_var, w_kurtosis, u2D_var, v2D_var, Tair, absolute_humidity, Vapor_pressure_deficit, Relhumidity, Pressure, TSOIL2, TSOIL4, TSOIL8, TSOIL16, TSOIL32, soil_moisture_00cm, soil_moisture_20cm, soil_moisture_50cm, precipitation**. Tl will not be included in the standard AmeriFlux data products.

AMP-Summarized Data Issues

Format QA/QC → Data QA/QC →

Data QA/QC and Report

AMERIFLUX MANAGEMENT PROJECT

[AmeriFlux] (QAQC-2405) Data Results | CC-XXX HH 20020101 - 20180803 | Using uploads through Sep 19, 2018

Inbox ×

AMF Data Team (AMF-JIRA) via berkeley.edu
to dschristianson

AMF Data Team **commented** on QAQC-2405

Re: Data Results | CC-XXX HH 20020101 - 20180803 | Using uploads through Sep 19, 2018

Dear Tower Team

Thank you for your data submissions for CC-XXX site.

In the context of the new processing for AmeriFlux data products, we are applying a new Data QA/QC scheme that follows the independent analysis of your data and help identify potential issues in data formats and contents earlier in the pipeline.

Briefly, Data QA/QC includes the inspection of sign conventions, ranges, diurnal-seasonal patterns, and potential outliers of var SW_IN) are also analyzed to detect potentially erroneous data. The comparison of measured radiation (e.g., PPFD_IN, SW_IN) t for a given location (i.e., SW_IN_POT) is also analyzed to check the timestamp specification and alignment.

In analyzing your data, we have the following questions where we request your expert opinion and suggestion. Please note that some i your site. Please verify, clarify, or correct the following issues before we can make your data available as an AmeriFlux BASE data prod please upload files using https://ameriflux.lbl.gov/data/upload-data/.

[Data QA/QC]

- Issue #1 .....
- Issue #2 .....
- Issue #3 .....

We hope that this will not take too much time from your work, but it will help to make your data more robust and clear. You can view th https://ameriflux.lbl.gov/qaqc-reports-data-team/.

Please reply to this email with any questions. You can track communications on this Data QA/QC at QAQC-2405 using your AmeriFlux account I

Best regards and thanks for the collaboration,
AmeriFlux Data Team

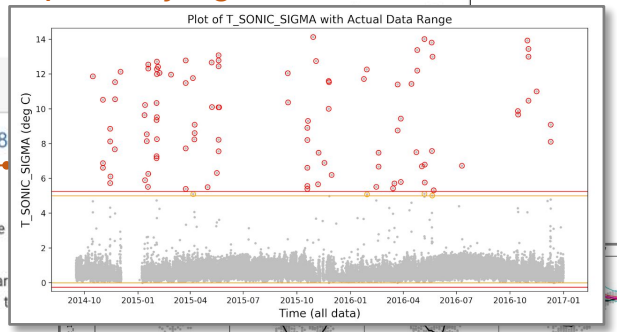
FTP link to Data QA/QC, where you can access all figures and intermediate files generated during Data QA/QC:
ftp://ftp.fluxdata.org/.ameriflux_downloads/data/CC-XXX/######/####/output

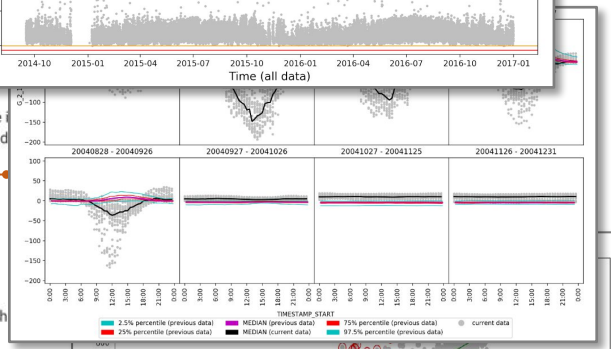
Format QA/QC reports associated with this Data QA/QC, where you can glance at the file sources used in this Data QA/QC:
http://ameriflux.lbl.gov/qaqc-report/?site_id=CC-XXX&report_id=######

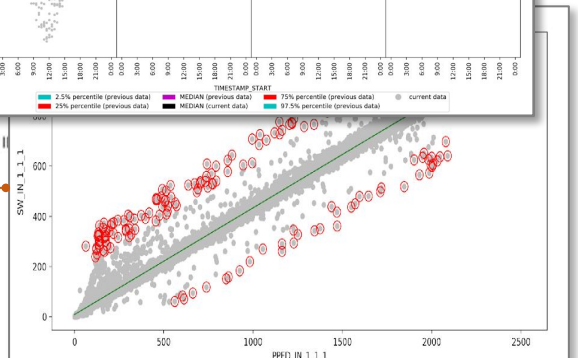
Data QA/QC Summary

Additional Links

Explanatory Figures

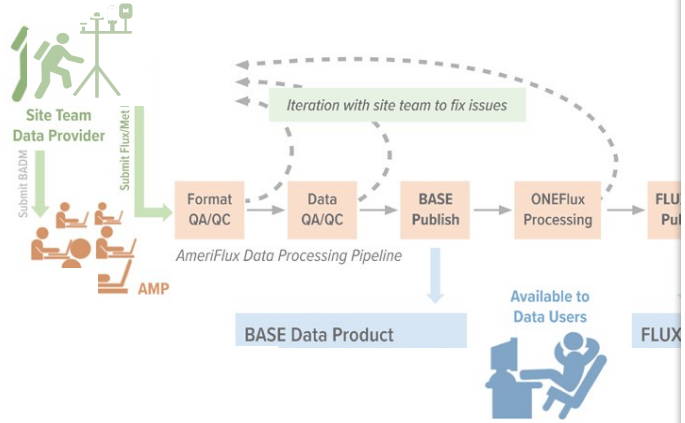


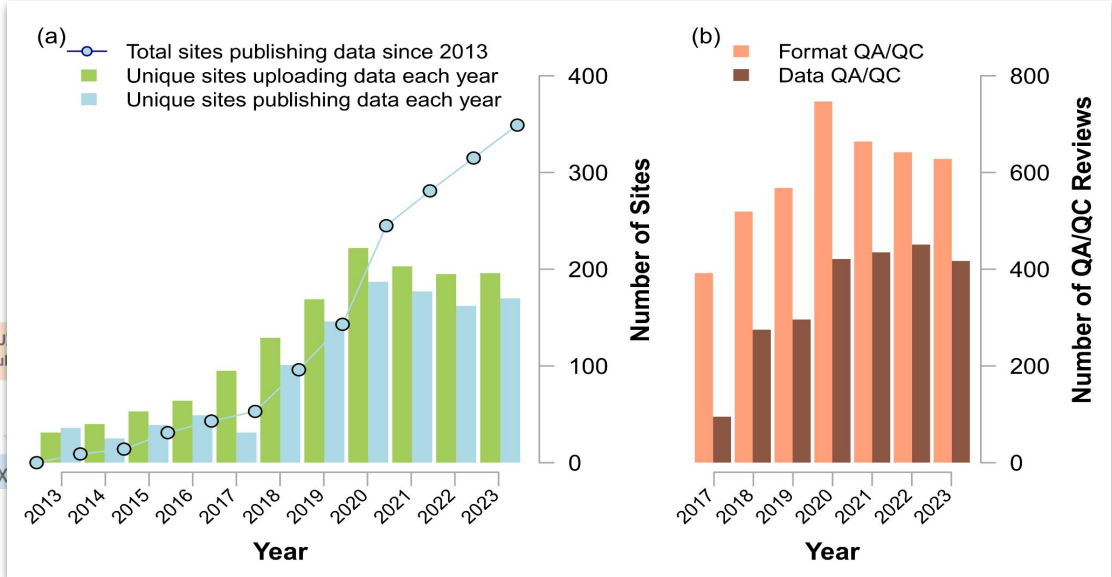




Challenge

- Keep up with network growth / continuous data streams
- Communicate, build trust with data providers / users
- Hard to automate certain parts of the data pipeline
- Reduce latency time from data collection to data use
- A wide variety of metadata to support the flux science
- Data citation for using hundreds of sites (site-package DOIs)





Pacific Northwest NATIONAL LABORATORY

April 15, 2024

How COMPASS-FME manages sensor data to accelerate environmental science

Ben Bond-Lamberty
PNNL
on behalf of the COMPASS-FME Data Management Working Group

U.S. DEPARTMENT OF ENERGY

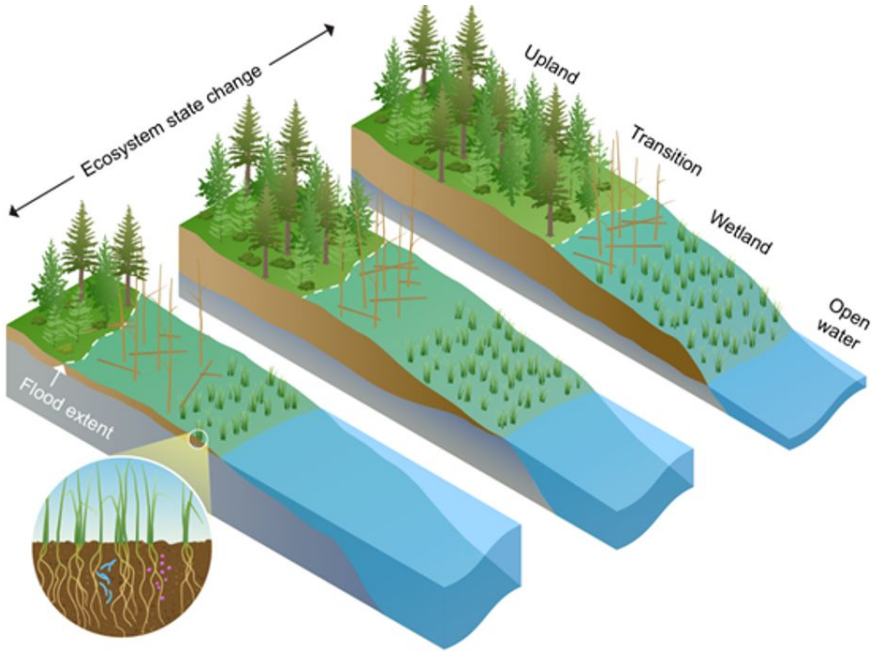
COMPASS is a multi-institutional program led by PNNL and funded by the Earth and Environmental Systems Science Division of the U.S. Department of Energy's Office of Science

COMPASS | FIELD, MEASUREMENTS, AND EXPERIMENTS

Coastal Observations, Mechanisms, and Predictions Across Systems and Scales

Pacific Northwest NATIONAL LABORATORY · Argonne NATIONAL LABORATORY · BERKELEY LAB · Los Alamos NATIONAL LABORATORY · OAK RIDGE National Laboratory · Smithsonian Environmental Research Center · THE UNIVERSITY OF TOLEDO

COMPASS-FME produces a variety of data types



Ecosystem Domain	Data Product Type
Water	Water level and quality; Anion and cation speciation
Soil	C/N concentration, sources, composition; Soil properties (moisture, temperature, etc.); Plant & microbial community composition
Soil, Water	Multi-omics (DNA, RNA, proteins)
Soil	Greenhouse gas fluxes and pathways (e.g. CO2 flux)
Plant	Water, nutrient, and carbon relations
Plant	Forest health (DBH, species, litterfall, etc.)

COMPASS

We have 500+ real-time sensors installed across 7 field sites

Weather

Temperature, barometric pressure, rain, wind speed, relative humidity

Vegetation

Sapflow

Soil

Temperature, volumetric water content, electrical conductivity, dissolved oxygen

Groundwater

Water level, dissolved oxygen, specific conductivity, pH

Open Water

Dissolved oxygen, pH, temperature, FDOM, specific conductivity, TSS

COMPASS

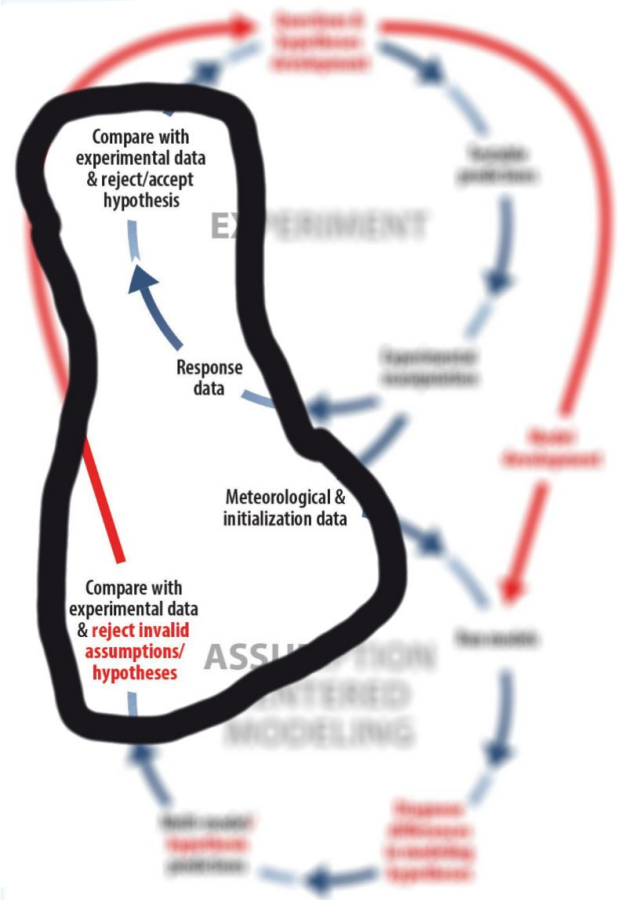
COMPASS-FME sensors generate

3+ million

observations per month

Data are crucial—we need a performant data processing pipeline

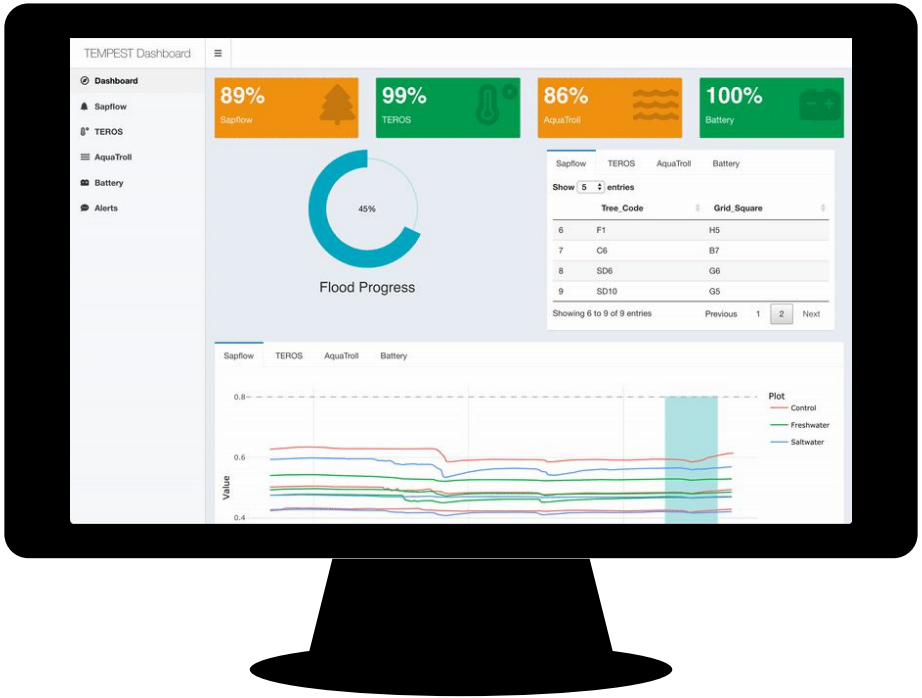
- **Rapid QA/QC**
  - TEMPEST
- **Experimentalists**
  - Hypothesis testing
- **Models**
  - Parameterization, forcing, and benchmarking



Hanson and Walker (2019)

COMPASS

Informing scientific decision-making in real time



COMPASS

COMPASS-FME data processing principles

- **Lightweight,** open source software;
- **Self-documentation** of processing
- **Different levels** of data for different uses;
- Traceability (**provenance**) of data;
- **Extensive metadata** for the entire dataset, as well as different sites and data types
- Straightforward future **integration with ESS-DIVE** (https://ess-dive.lbl.gov/)

Different data processing levels for different uses

	Raw	L0	L1	L2
Structure	Wide form from datalogger	Long form with unique IDs	Long form, split by site, month	Long form, split by type, month
Audience	Data managers	Data managers	Technicians and expert analysts	All data users
QA/QC	⊗	⊗	Basic flags provided but not applied	Human and algorithmic QA/QC applied

COMPASS

Software tools and approaches

Free and open source R and Quarto for

- Data processing
- Logging
- Visualization

Workflow: L0

AUTHOR
COMPASS workflows team

PUBLISHED
2024-04-12 13:06:20-04:00

This script

- Reads in the raw data files one by one
- Extracts `Logger` and `Table` information from the header and adds them as columns
- Reshapes from wide to long; only `Logger`, `Table`, and `TIMESTAMP` don't get reshaped
- Adds a unique observation ID (using `digest::digest()`)
- Writes as CSV files with row/col/hash info in filename
- Moves the raw files to a Raw_done folder

Initializing

I see 200 files to process in data_TEST//Raw/.

Output directory is data_TEST//L0/.

Moving done files to data_TEST//Raw_done/.

HTML outfile is L0.html.

Logfile is .

Working directory is /Users/d3x290/Code/data-workflows/synoptic.

Processing

▶ Code

```
2024-04-12 13:05:57 About to L0
```

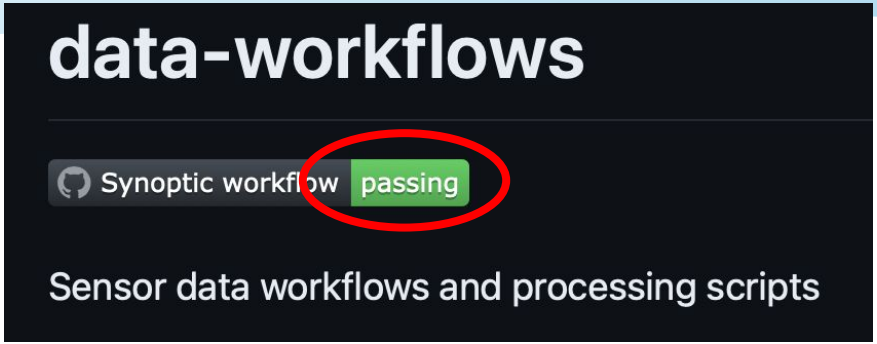
▶ Code

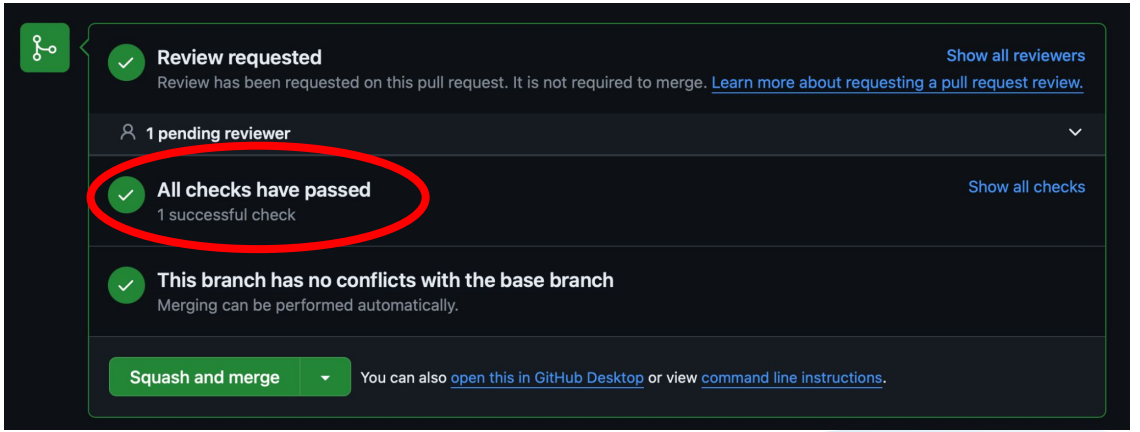
```
2024-04-12 13:05:57.959718 Processing Compass_CRC_TR_302_CheckTable_20230508000018_short.

    Original data is 6 x 25

    Pivoted data is 132 x 5

    Overwriting existing file
```


COMPASS

Software tools and approaches

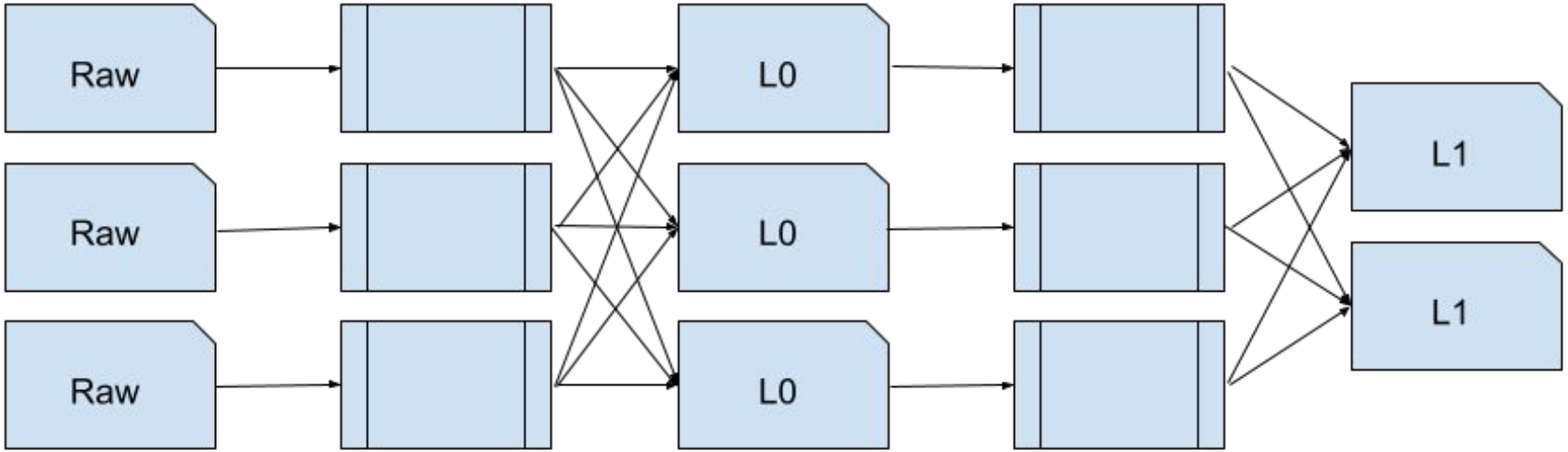
- Git is used for [version control]
- Automated tests run sample data to verify performance



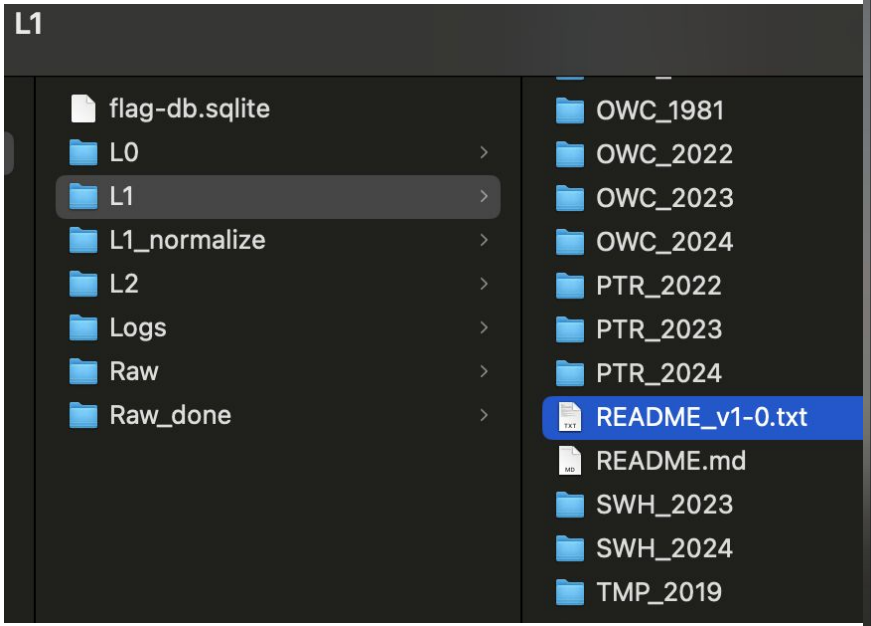


Software tools and approaches

- Processing pipeline is parallelized



Quality metadata = quality data



README_v1-0.txt

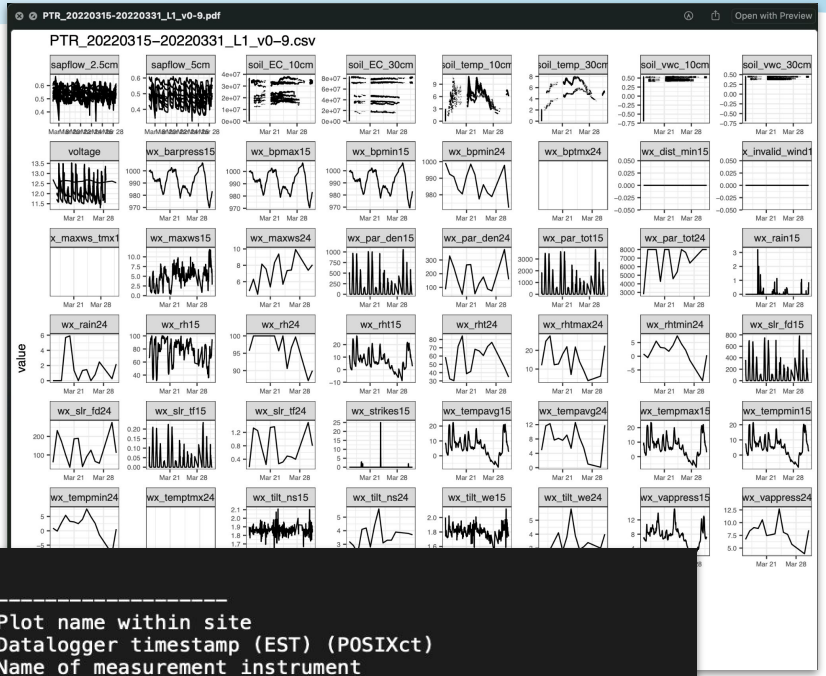
```
COMPASS-FME Level 1 data
Version: 1-0
Date: 2024-04-10
Observations: 139,922,183
Git commit: c5d0a3c

DESCRIPTION
-----------------------------------
Level 1 (L1) data are close to raw, but are units-transformed and have
out-of-instrument-bounds and out-of-service flags added. Duplicates and
missing data are removed but otherwise these data are not filtered, and
have not been subject to any additional algorithmic or human QA/QC. Any
scientific analyses of L1 data should be performed with care.

CONTACT
-----------------------------------
Project: https://compass.pnnl.gov
Data lead: Stephanie Pennington, stephanie.pennington@pnnl.gov

HOW TO CITE THESE DATA
-----------------------------------
Pennington, Bittencourt Peixoto, Cheng, LaGorga, Machado-Silva, Peresta,
Phillips, Regier, Rich, Sandoval, Stearns, Ward, Wilson, Weintraub,
Megonigal, and Bailey (2024). COMPASS-FME Level 1 Sensor Data (version
1-0 released 2024-04-10), downloaded YYYY-MM-DD,
https://compass.pnnl.gov.

CHANGELOG
-----------------------------------
Version 1-0 released 2024-04-15
* Covers late 2019 through March 2024 for TEMPEST and all synoptic sites
* Restructured for ease of use, with metadata (location, sensor ID, etc) in separa
* SWH plot naming reworked for new upland plot; mirroring TMP C to GCW UP
* Many fixes to variable units and bounds

Version 0-9 released 2024-01-22
* Preliminary release covering all synoptic site and TEMPEST data collected to dat
* Units and bounds (and thus OOB flags) are missing for some ClimaVue, AquaTROLL,
* Some research_name assignments for may incorrect for ClimaVue
* Out-of-service is working only for AquaTROLL
```

Quality metadata = quality data

TMP_2022_L1_v1-0_metadata.txt

```
COMPASS-FME Level 1 documentation file
Data version 1-0
TMP_2022
Wed Apr 10 06:20:41 2024


General information
-----------------------------------
Level 1 (L1) data are close to raw, although they have been units-transformed
and have out-of-instrument-bounds and out-of-service flags added.
Duplicates are removed but otherwise these data are not filtered,
and have not been subject to any additional algorithmic or human QA/QC.
Any use of L1 data for science analyses should be performed with care.


Site information:
-----------------------------------
TMP is TEMPEST, the COMPASS-FME ecosystem-scale flooding experiment and
has three 2000 m2 plots: control ("C"; located at 38.8747N, 76.5519W),
freshwater ("F"; 38.87403N, 76.5516W), and saltwater ("S"; 38.8744N,
76.5525W). The TEMPEST experiment is in a mid- to late-successional (~80
years old) temperate, deciduous coastal forest.

Contact for the TEMPEST site:
J. Patrick Megonigal megonialp@si.edu

Contacts for TEMPEST data streams:
Overall: P. Megonigal; Anya Hopple <anyahopple@gmail.com>
Sapflow: Stephanie Pennington stephanie.pennington@pnnl.gov; Alice Stearns <
TEROS: Evan Phillips <PhillipsE@si.edu>; S. Pennington; Peter Regier <peter.
AquaTROLL: E. Phillips; S. Pennington; P. Regier
Dataloggers: E. Phillips; A. Stearns; Roy Rich <RichR@si.edu>

Key publications:
Hopple et al.: Attaining freshwater and estuarine-water soil saturation in a
ecosystem-scale coastal flooding experiment, Environ. Monit. Assess., 195, 4
http://dx.doi.org/10.1007/s10661-022-10807-0


Files in this folder:
-----------------------------------
TMP_20220101-20220131_L1_v1-0.csv
Rows: 1617896
md5: 3276467cf13afe6e9a73946099cc5e13

TMP_20220201-20220228_L1_v1-0.csv
Rows: 1426883
md5: 82485706bd7564d329bf4986ba55cc87
```

PTR_20220315-20220331_L1_v0-9.pdf — Open with Preview



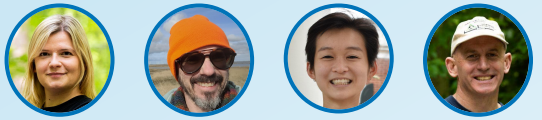
File columns:

Plot	Plot name within site
TIMESTAMP	Datalogger timestamp (EST) (POSIXct)
Instrument	Name of measurement instrument
Instrument_ID	Which instrument within plot
Sensor_ID	Individual sensor, tree, etc. being measured
Location	Location
Value	Observed value (numeric). The no-data value is ''
research_name	Measurement name (character)
ID	Observation ID (character)
F_OOB	Flag: Out of instrumental bounds (logical; 1=TRUE)
F_OOS	Flag: Sensor listed as out of service (logical; 1=TRUE)

COMPASS

This was a team project!

Thank you to everyone who helped fill in data descriptions, gave beta test feedback, acted as site/data contacts, informed QA/QC metrics, and continue to work on monitoring this data in real time.



(not pictured: the snacks needed to fuel these 2-hour meetings)

COMPASS

Pacific Northwest
NATIONAL LABORATORY

Data Management for Team Science and Reuse in the River Corridor SFA and WHONDRS

Amy Goldman, Brieanne Forbes, and Bibi Powers-McCormack

11 April 2024

DOE BER ESS PI Meeting

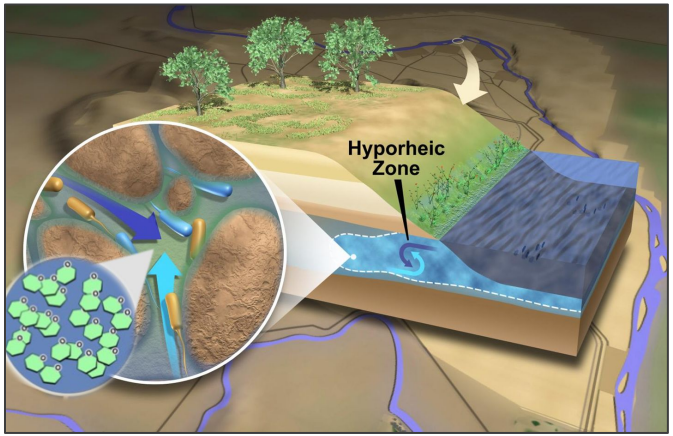
amy.goldman@pnnl.gov
https://whondrs.pnnl.gov

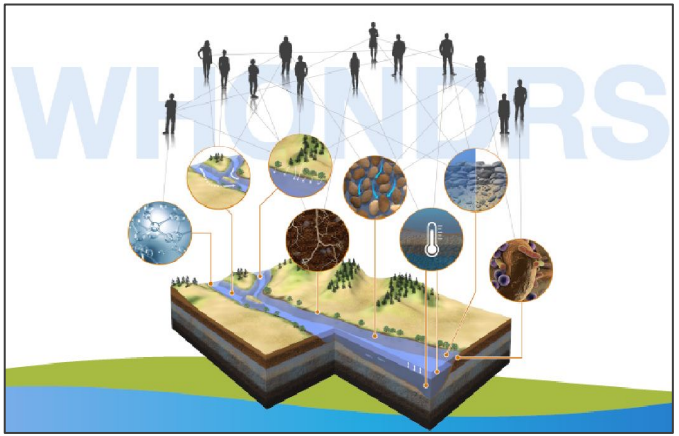
U.S. DEPARTMENT OF ENERGY BATTELLE

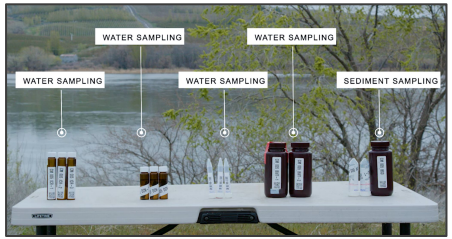
PNNL is operated by Battelle for the U.S. Department of Energy

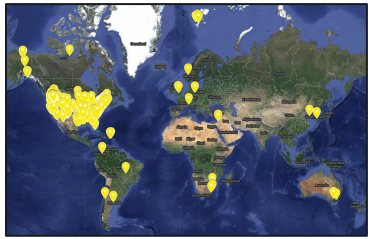
Pacific Northwest
NATIONAL LABORATORY

River Corridor SFA and WHONDRS navigate publishing dozens of data types across hundreds of collaborators and many thousands of samples









RC-SFA data on ESS-DIVE has been used by people inside and outside the project

Manuscripts using RC-SFA data led by people outside the project, including a Frontiers in Water special issue

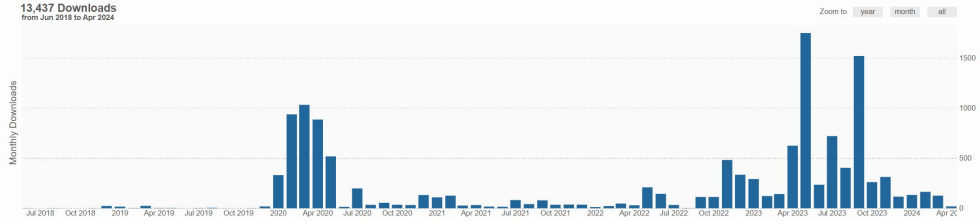
Data used for educational activities, including 2020 EMSL Summer School

RC-SFA ESS-DIVE metrics

13.4K Downloads

For all versions of the data sets in this portal, the number of times that all or part of these data sets were downloaded over time. These download counts are COUNTER compliant, meaning that downloads from some Internet robots and repeat downloads within a certain time window are excluded.

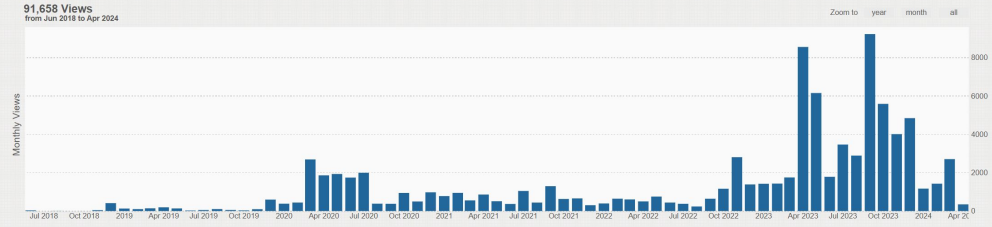
Drag the slider to visualize a specific time window for the download events.



91.7K Views

For all versions of the data sets in this portal, the number of times that all or part of these data sets was viewed over time. These view counts are COUNTER compliant, meaning that views from some Internet robots and repeat views within a certain time window are excluded.

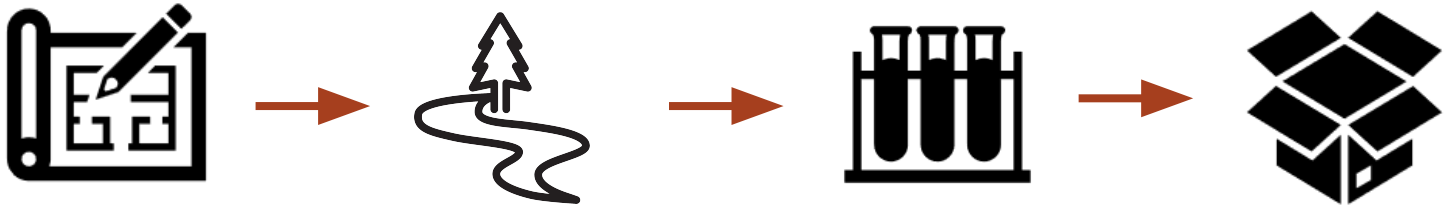
Drag the slider to visualize a specific time window for the view events.



https://data.ess-dive.lbl.gov/portals/PNNLRiverCorridorSFA

Pacific Northwest
NATIONAL LABORATORY

Data management for team science requires planning throughout the research lifecycle



Planning	Collecting	Analyzing	Publishing
Folder structure	Consistent protocol	Record methods	Reporting formats
Standardized naming	Record discrepancies	Record inconsistencies	Cross checking
Protocol & metadata	Metadata	Statistical QAQC	Consistency
Plan for digitizing	Barcode/RFID	File structure	Use published data

Automation enables data management at this scale

Planning

Collecting

Analyzing

Publishing

- Folder structure
- Standardized naming
- Protocol & metadata
- Plan for digitizing

- Consistent protocol
- Record discrepancies
- Metadata
- Barcode/RFID

- Record methods
- Record inconsistencies
- Statistical QAQC
- File structure

- Reporting formats
- Cross checking
- Consistency
- Use published data

✓ Consistency
✓ Automation
✓ Efficiency
✓ Accuracy
✓ Quality
✓ FAIR

Processes are developed with the team and frequently modified based on feedback

Planning

Collecting

Analyzing

Publishing

- Folder structure
- Standardized naming
- Protocol & metadata
- Plan for digitizing

- Consistent protocol
- Record discrepancies
- Metadata
- Barcode/RFID

- Record methods
- Record inconsistencies
- Statistical QAQC
- File structure

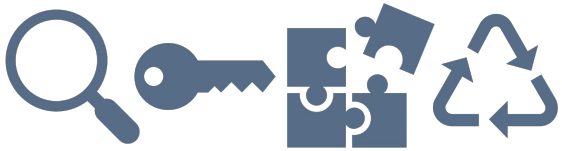
- Reporting formats
- Cross checking
- Consistency
- Use published data

✓ Consistency
✓ Automation
✓ Efficiency
✓ Accuracy
✓ Quality
✓ FAIR

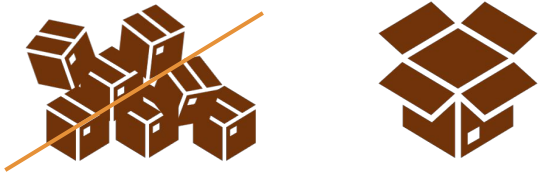
✓ Growth Mindset

Data publishing approach emphasizes the future data user

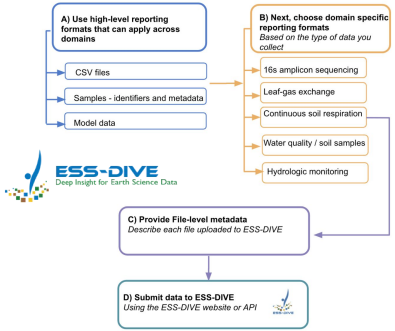
FAIR Data



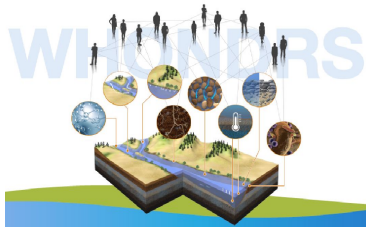
Publish Data Types Together to Facilitate Use



ESS-DIVE Reporting Formats



Publish before Data Use

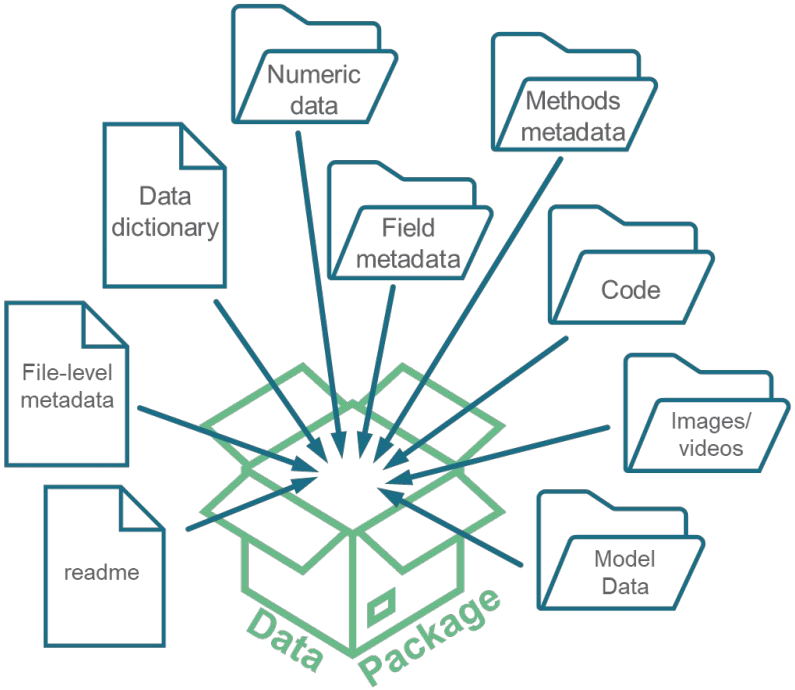


https://data.ess-dive.lbl.gov/portals/WHONDRS

Graphics: https://ess-dive.lbl.gov/data-reporting-formats/, PNNL

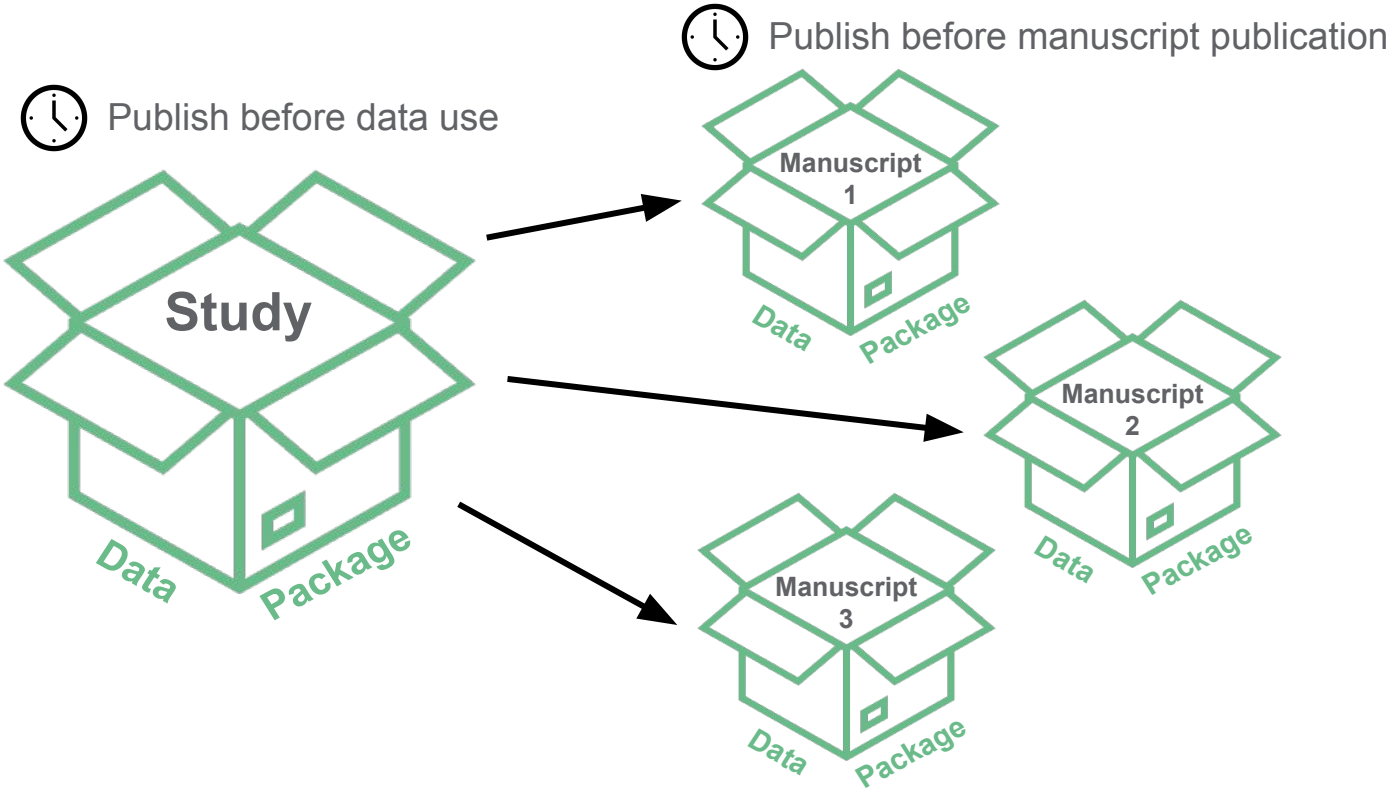
Pacific Northwest
NATIONAL LABORATORY

Data packages contain many types of (meta)data

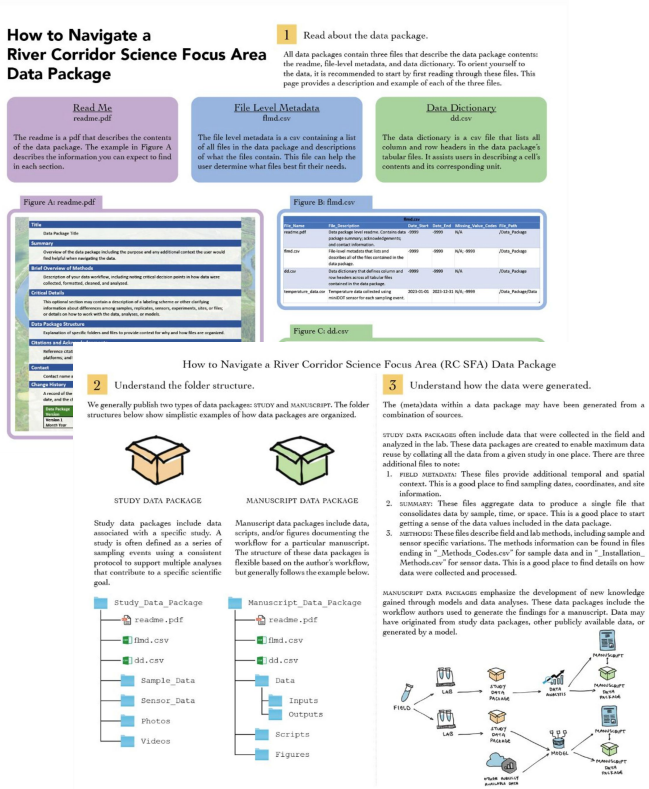


Pacific Northwest
NATIONAL LABORATORY

Publishing two types of data packages helps achieve clear data provenance

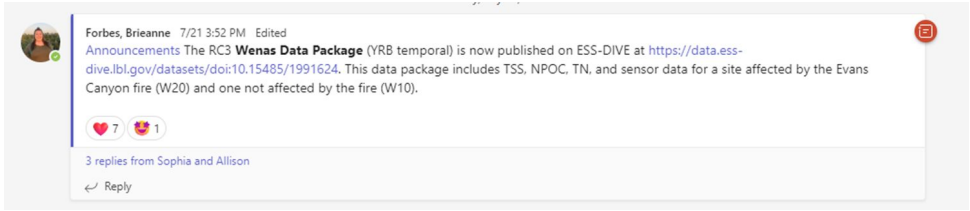


Frequent and creative communication inside and outside the project facilitates wider data use

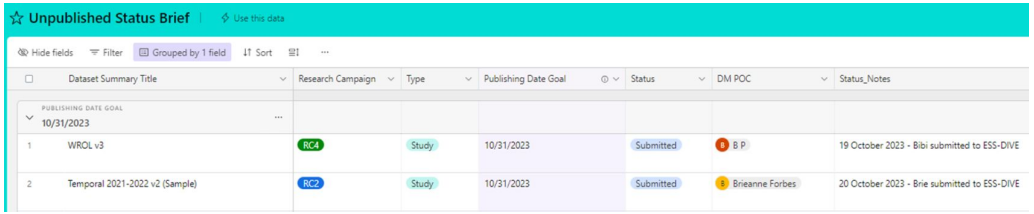


https://data.ess-dive.lbl.gov/portals/PNNLRiverCorridorSFA

MS Teams announcements



AirTable inventory of unpublished and published packages



Three-person data management team

Amy Goldman

Brieanne Forbes

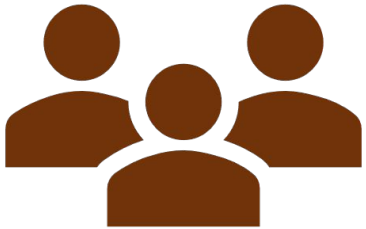
Bibi Powers-McCormack

Pacific Northwest
NATIONAL LABORATORY

Challenges focus on resources, consistency, and linking across tools

Connecting information across data archive, generators, users, and citations (ESS-DIVE, EMSL, JGI, Kbase, NMDC)

Coordination throughout the research lifecycle



Interoperability with past data and other data sources

Pacific Northwest
NATIONAL LABORATORY

PNNL's River Corridor Scientific Focus Area (SFA) project is supported by the U.S. DOE Office of Science, Office of Biological and Environmental Research (BER), Environmental System Science (ESS) program.

ESS
Environmental
System Science

WHONDRS
Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems

amy.goldman@pnnl.gov

https://www.pnnl.gov/projects/river-corridor

https://whondrs.pnnl.gov

U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

Sharing our experience as a data user of Ameriflux/FLUXNET Data

Youmi Oh[1] and Licheng Liu[2]
[1] NOAA GML & CU Boulder
[2] University of Minnesota

AmeriFlux

Home | About | Community | Sites | Data | Tech | Theme Years | Resources

********** SERVICE DISRUPTION NOTICE **********

AmeriFlux data services will be unavailable betweeen April 19th (Friday, 9am PST) to April 21st (Monday, 10am PST) due to scheduled maintenance and upgrades. Thank you for your patience!

POSTCARDS

Check out the AmeriFlux Resource for Effective and Inclusive Job Ads

Have you ever wondered whether your job postings could be more effective? Many hiring managers prefer...

FEATURED SERVICES

Upload Data | Download Data | Join Community | Register a Site

MORE POSTCARDS

Quick Sites: Recents | Favorites

UPCOMING EVENTS

MAY 26 — May 26 - May 31
Japan Geoscience Union Meeting 2024

JUL 9 — July 9 - July 10
workshop "Remote Sensing and Fluxes Upscaling for Real-world Impact"
Lawrence Berkeley National Lab
Berkeley CA

More Events...

SITE SEARCH & MAPS

Download Data

1. Select A Data Product | 2. Refine Your Selection | 3. Select Sites | 4. Agree to Policy | 5. Download Data

Want to start by selecting sites? Go to Site Search.

Select the data product that you want to download

Visit AmeriFlux Data Processing Pipelines for more information about AmeriFlux data products.

AmeriFlux FLUXNET

Features
- Continuous flux/met
- Gap-filled
- Partitioned
- Uncertainty anaylsis
- 5 temporal resolutions
- **BADM** optional

GPP + Reco

Generated by
- AMP using ONEFlux processing codes
- FP standard format
- Subset of standard FP variables

Sites
- 196 sites
- AmeriFlux sites only
- See fluxnet.org for global datasets

AmeriFlux Data Use Policy
- CC-BY-4.0: 196 sites

AmeriFlux BASE

Features
- Flux/met data
- Half-hourly / hourly
- **BADM** included

Generated by
- Site team in FP standard format
- QA/QC'ed by AMP
- All FP variables
- All levels of aggregration

Sites
- 492 sites
- AmeriFlux sites only

AmeriFlux Data Use Policy
- CC-BY-4.0: 414 sites
- Legacy: 492 sites (includes all sites with data available under CC-BY-4.0)

BADM Only

Features
- **B**iological data
- **A**ncillary data
- **D**isturbance data
- **M**etadata

Generated by
- Site team in BADM standard format
- QA/QC'ed by AMP

Sites
- 660 sites
- AmeriFlux sites only

AmeriFlux Data Use Policy
- CC-BY-4.0: 498 sites
- Legacy: 660 sites (includes all sites with data available under CC-BY-4.0)

ONEFlux code is quite hard to use…
Please make it easy!

fluxnet / ONEFlux

Code | Issues 29 | Pull requests 9 | Actions | Projects | Security | Insights

ONEFlux

Public

Watch 20 | Fork 39 | Star 74

main | 4 Branches | 3 Tags | Go to file t | Add file | Code

gilbertozp Merge pull request #23 from gilbertozp/era_options 9201beb · 2 years ago | 24 Commits

oneflux	update version to 0.4.1-rc	2 years ago
oneflux_steps	add v0.3.0-beta	6 years ago
tests	add v0.3.0-beta	6 years ago
.gitignore	add v0.3.0-beta	6 years ago
CHANGELOG.md	fix changelog typo	5 years ago
LICENSE	add v0.3.0-beta	6 years ago
Makefile	add package manager option to makefile	2 years ago
README.md	update README description and funding	4 years ago
requirements.txt	multiple bug fixes	5 years ago
runoneflux.py	remove broken import	2 years ago

About

Open Network-Enabled Flux processing pipeline

Readme
View license
Activity
Custom properties
74 stars
20 watching
39 forks
Report repository

Releases 2

v0.4.1-rc Latest
on Mar 25, 2022

+ 1 release

Download Data

1. Select A Data Product	2. Refine Your Selection	3. Select Sites	4. Agree to Policy	5. Download Data
✔ AmeriFlux FLUXNET (CC-BY-4.0)	✔ SUBSET	✔ 5 selected	✔ Agreed	✔ Files ready

The download links below will require re-generation if you navigate away from the Download Data page or Step 5.

Download Info

README Requested_Files Citations_for_Site_Data Team_Contacts_for_Site_Data AmeriFlux_CC-BY-4.0_Data_License

Multi-site Metadata (BADM)

None selected

Site Data

Click on a link below to download that site's file. Consider using a 3rd party browser tool like DownThemAll! (FireFox, Chrome) to download all the files at once.

CA-Ca1_FLUXNET CA-MA1_FLUXNET CA-NS1_FLUXNET CA-NS4_FLUXNET US-ALQ_FLUXNET

Previous

A few useful data processing tools

Tools and Software for Flux scientists

- AmeriFlux Data Visualizer (released Sept 2022): This app was developed to celebrate 3,000 site years of AmeriFlux data. It is intended to be used only for initial data visualization and exploration to give users a better understanding of data availability before downloading. It is not intended for detailed analysis or network synthesis. This application was developed by Sophie Ruehr with support from members of the AmeriFlux community and management team. Read more about the visualizer here.
- R package 'ameriflux' (released Feb 2022): an R package for querying, downloading, and handling AmeriFlux data and metadata. Please check the webpage for its installation, usage, and features, and the tutorials for site selection and data download/import.
- Additional lists of tools and software for flux related data processing, sorted by
  - Raw Data Processing and QA/QC
  - Flux/Met Data Post-Processing and QA/QC
  - Data API Interface
  - Auxiliary Data Processing

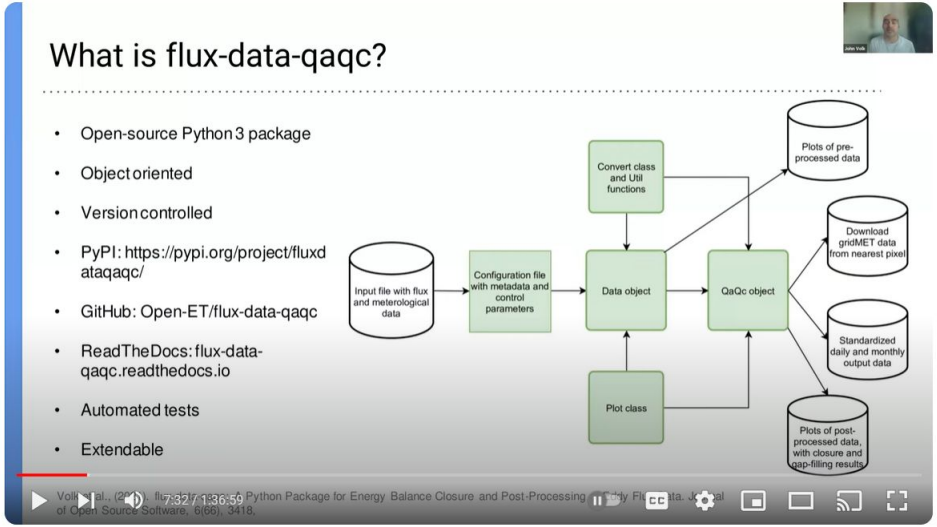
```
# get a summary table of sites with available data,
#  & grouped by data use policy & IGBP
pander::pandoc.table(sites_dt[!is.na(DATA_START), .N, by = .(IGBP, DATA_POLICY)][order(IGBP)])
```

IGBP	DATA_POLICY	N
BSV	CCBY4.0	2
BSV	LEGACY	2
CRO	CCBY4.0	49
CRO	LEGACY	14

https://ameriflux.lbl.gov/resources/resource-list/tools-and-software-for-flux-scientists/

Data tool training hosted by FLUXNET-ECN

Python package: FLUX-DATA-QAQC



FLUXNET ECN Seminar: Post-process flux data with flux-data-qaqc

AmeriFlux
576 subscribers
Subscribe
6
Share
Download

https://www.youtube.com/watch?v=7T27WiW4K1k

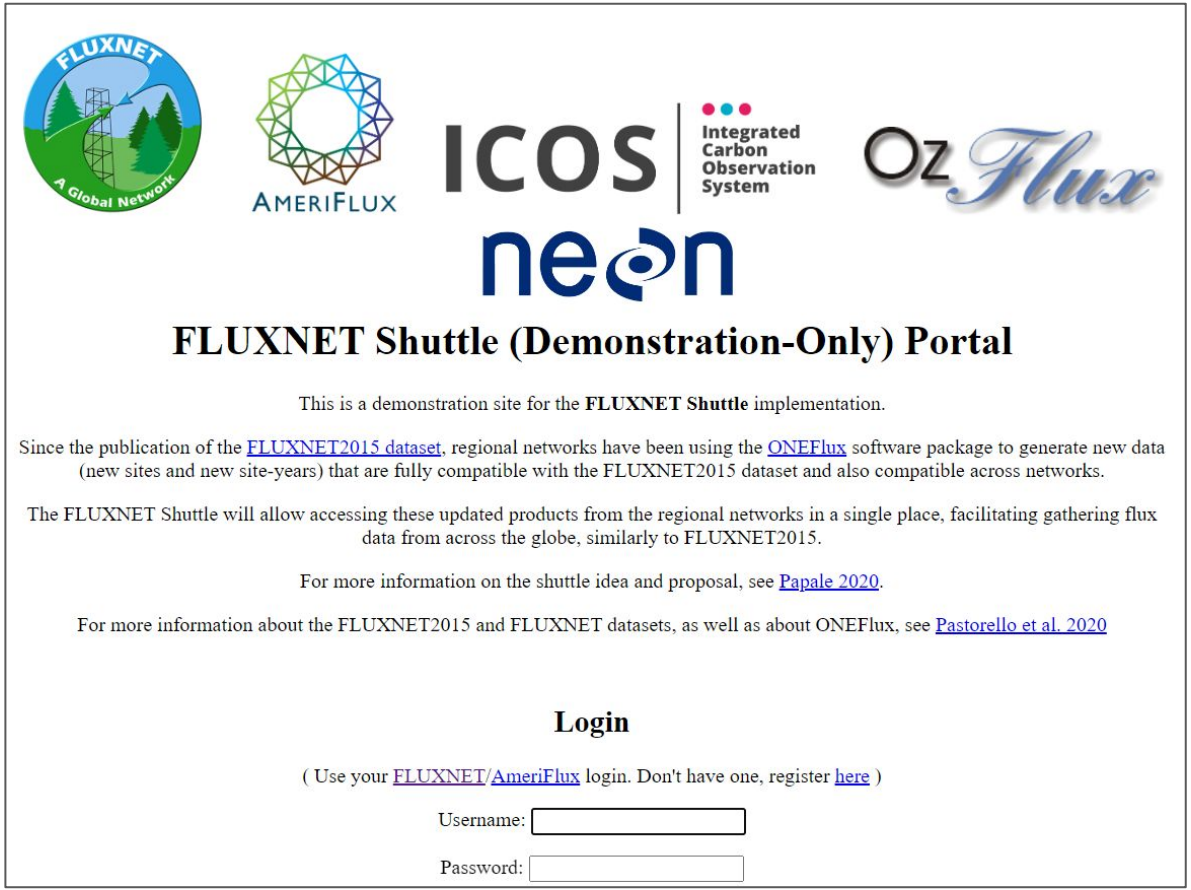
R package: REddyProc

FLUXNET-ECN: post-processing of eddy covariance flux data with REddyProc

AmeriFlux
576 subscribers
Subscribe
16
Share
Download

https://www.youtube.com/watch?v=-b0vc4u8kls

A unified data portal will be super helpful!



https://shuttle-demo.fluxnet.org/accounts/login/?next=/

Time for Questions

How Can We Address Data Management Challenges?

Make high quality data management and integration feasible for projects

2023 ESS CI Meeting Questionnaire Highlights

43 Respondents

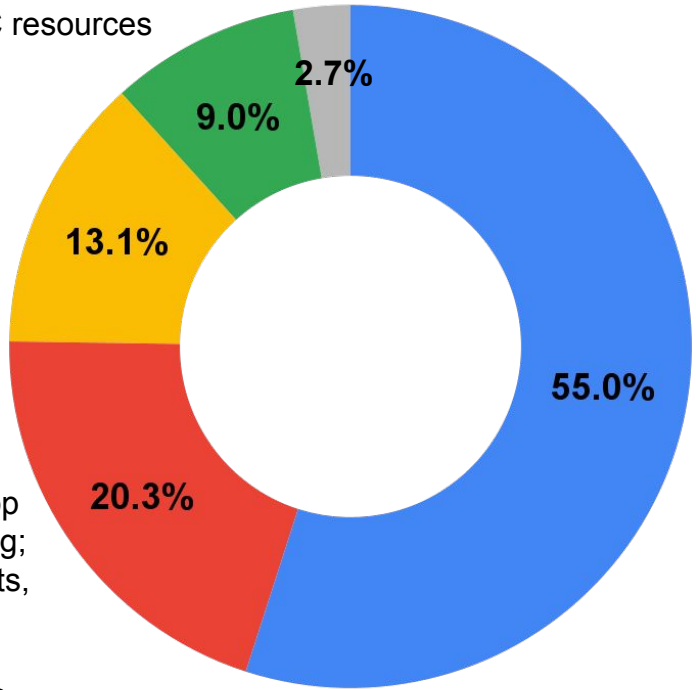
CAVEAT
These results have not been vetted as representative of the ESS Community at large.

What is the main CI challenge in your research?

Compute Infrastructure
- Creating computing capabilities responsive to researcher needs
- Access to / integration with HPC resources

ALL
Diverse data, modeling, and computing across investigators / themes

Software E&I
- Allocate time/resources between development of software and actual science application
- Developing and maintaining component models across multiple host models and machines
- Reproducibility pipelines / workflows



Data Management
- Inter-project data sharing
- Making observation/experimental data rapidly available
- Efficient large-data access | API access
- Data management, integration, QA/QC
- Barriers to high quality data publication
- Developing FAIR datasets
- Lack of standardized formats, conventions

Model-Data Integration
- Faster data to model to output loop
- Data evaluation and benchmarking; With so many similar data products, how to choose?
- Empiricist / modeler collaboration
- Model setup and mesh generation

WG Affiliation

Data Management Challenges

Metadata sufficiency

Provide enough contextual information for broad research re-use

Standards & guidance lacking (but improving)

Data normalization

(aka data synthesis or harmonization)

Compare apples to apples: same units, variables, formats

Archiving (sharing) data

Human and tech resources lacking

Guidance on data organization and documentation lacking

Need better incentives

Data Discovery and Access

Develop tools

Requires agency coordination

Dependent on other 3 challenges

Activity: design an ideal ESS data management and integration infrastructure

Think of your top priority data management challenge, and how you would like it to be solved…

Describe components of your dream/ideal system for ESS data management that could address this challenge. This may enable you to ***efficiently***:

(1) create and publish high-quality datasets AND/OR
(2) find, access, integrate, and use data from ESS-DIVE, Ameriflux, and other systems that you need/use.
[10 minutes of ideation + 10 discussion]

go.lbl.gov/StrategicData

Activity: design an ideal ESS data management infrastructure

What do you think are the biggest roadblocks to achieving an ideal data management infrastructure is? And what action would you like to see next (from ESS-DIVE, Ameriflux, etc.)?

go.lbl.gov/StrategicData