11-2017

# A metadata reporting framework (FRAMES) for synthesis of ecohydrological observations

Danielle S. Christianson

Charuleka Varadharajan

Bradley O. Christoffersen

Matteo Detto

Boris Faybishenko

*See next page for additional authors*

## Authors

Danielle S. Christianson, Charuleka Varadharajan, Bradley O. Christoffersen, Matteo Detto, Boris Faybishenko, Bruno O. Gimenez, Val Hendrix, Kolby J. Jardine, Robinson Negron-Juarez, and Gilberto Z. Pastorello

Submission to *Ecological Informatics* as an Original Research Paper

Title: A metadata reporting framework (FRAMES) for synthesis of ecohydrological observations

Running title: FRAMES: Metadata reporting for ecohydrological observations

Danielle S. Christianson[a,b], Charuleka Varadharajan[a,*], Bradley Christoffersen[c], Matteo Detto[d,e], Boris Faybishenko[a], Val Hendrix[b], Kolby J. Jardine[a], Robinson Negron-Juarez[a], Bruno O. Gimenez[f], Gilberto Z. Pastorello[b], Thomas L. Powell[a], Megha Sandesh[b], Jeffrey M. Warren[g], Brett T. Wolfe[d], Jeffrey Q. Chambers[a], Lara M. Kueppers[a,h], Nathan G. McDowell[c], Deb Agarwal[b]

[a] Earth and Environmental Science Area, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

[b] Computational Research Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA

[c] Earth and Environmental Sciences, Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, NM, 87545, USA

[d] Smithsonian Tropical Research Institute, Unit 9100, Box 0948, DPO AA 34002-9998, USA

[e] Department of Ecology and Evolutionary Biology, 106A Guyot Hall, Princeton University, Princeton, NJ, 08554, USA

[f] National Institute of Amazonian Research (INPA), Ave. Andre Araujo 2936, Campus II, Building LBA, Manaus, AM 69.080-97, Brazil

[g] Climate Change Science Institute & Environmental Science Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA

[h] Energy and Resources Group, University of California, 310 Barrows Hall, Berkeley, CA 94720, USA

* Corresponding author. Tel: 510-495-8890. Email: cvaradharajan@lbl.gov
M/S: 74R316C, 1 Cyclotron Road, Berkeley CA 94720, USA.

**Keywords**

metadata; data management system; model-data integration; data synthesis; data preservation; informatics

**Abbreviations**

FRAMES = Framework for Reporting dAta and Metadata for Earth Systems

FATES = Functionally Assembled Terrestrial Ecosystem Simulator

QA/QC = Quality Assurance / Quality Control

ENSO = El Nino Southern Oscillation

STRI = Smithsonian Tropical Research Institute

CTFS = Center for Tropical Forest Study

BADM = Biological, Ancillary, Disturbance, and Metadata

ISCN = International Soil Carbon Network

**Abstract**

Metadata describe the ancillary information needed for data preservation and independent interpretation, comparison across heterogeneous datasets, and quality assessment and quality control (QA/QC). Environmental observations are vastly diverse in type and structure, can be taken across a wide range of spatiotemporal scales in a variety of measurement settings and approaches, and saved in multiple formats. Thus, well-organized, consistent metadata are required to produce usable data products from diverse environmental observations collected across field sites. However, existing metadata reporting protocols do not support the complex data synthesis and model-data integration needs of interdisciplinary earth system research. We developed a metadata reporting framework (FRAMES) to enable management and synthesis of observational data that are essential in advancing a predictive understanding of earth systems. FRAMES utilizes best practices for data and metadata organization enabling consistent data reporting and compatibility with a variety of standardized data protocols. We used an iterative scientist-centered design process to develop FRAMES, resulting in a data reporting format that incorporates existing field practices to maximize data-entry efficiency. Thus, FRAMES has a modular organization that streamlines metadata reporting and can be expanded to incorporate additional data types. With FRAMES's multi-scale measurement position hierarchy, data can be reported at observed spatial resolutions and then easily aggregated and linked across measurement types to support model-data integration. FRAMES is in early use by both data originators (persons generating data) and consumers (persons using data and metadata). In this paper, we describe FRAMES, identify lessons learned, and discuss areas of future development.

## 1. Introduction

Current earth systems research challenges, like understanding and predicting carbon cycling in tropical forests under a changing climate, require synthesis of complex and diverse earth system observations. Researchers use synthesized data products to understand the controls and rates of environmental processes, as well as constrain, parameterize, and benchmark process-rich models (e.g., Medlyn et al. 2005). Data synthesis refers to the process of connecting diverse observations collected across field sites and a wide range of spatial and temporal scales to answer a science question or to generate model inputs. Prior to synthesis, each observation must be quality checked, processed (e.g., units transformed, gap-filled, erroneous data flagged or removed), and organized in standardized, comparable formats (e.g., variable names, units). An example of a synthesized data product is the FLUXNET2015 dataset, which includes data collected at sites from a network of single-locale, eddy covariance towers that monitor an ecosystem over many years (FLUXNET 2016). In addition to ecosystem and global scale datasets, earth system science requires syntheses of individual-based measures like point observations of leaf carbohydrate content, continuous tree sap flow, and demography censuses (e.g., Walker et al. 2014). Physical measures, such as meteorological observations, measurements of soil water content, and 3D structural representations (e.g., LiDAR), are also needed (e.g., Powell et al. 2013, Hunter et al. 2015).

Metadata are essential to describe the different approaches taken to obtain, process, and report diverse ecohydrological and biogeochemical observations and the resulting data products (Michener et al. 1997; Michener 2006; Papale et al. 2010; Kervin et al. 2013). Metadata allow for interpretation and integration of heterogeneous data obtained from different measurement approaches across disparate study sites, which occur even in well-organized science projects. Additionally, metadata are often critical for quality assurance and quality control (QA/QC). For example, particular equipment can have biases under certain conditions, or events such as power outages or equipment maintenance can affect data quality. Metadata that describe the location and time period of the observations or data products are used for aggregation both in time and space. Furthermore, metadata also describe the people who conducted the work, which is important for provenance (record of data credits) and proper attribution to data originators. Given its broad range of utility, metadata can describe many aspects of observations or data products, including descriptions of the measurement setting (e.g., measurement location and approach), the data reported (e.g., measurement variable and units), and the datasets (e.g., data processing level and details).

Due to data management requirements from federal funding agencies, a variety of data collection repositories now exist, each with their own metadata requirements (e.g., KBase 2016; KNB 2016; NOAA NCEI 2016; USGS 2016). Over the last several years, the digital preservation community has developed a general consensus around best practices for metadata that define how to reliably ingest data into these data repositories, track provenance, build and maintain metadata, and enable future consumers to independently access and use the data. For example, the Open Archival Information System (OAIS) reference model describes the concept of information packages as a collection of content and metadata. The metadata is further delineated as 1) content metadata, 2) descriptive metadata that enable search and retrieval of the content, 3) preservation description metadata necessary for long-term archiving such as provenance, checksums and unique identifiers, and 4) other ancillary metadata needed to define and hold the package together (OAIS / ISO 14721:2012). Some data repositories provide tools for data originators to prepare and submit a *Submission Information Package* (hence referred to as "data package") containing content data and all the metadata, and for data consumers to download a *Dissemination Information Package* containing citation information in addition to the content data and metadata.

Several standards and formats currently exist to describe data collection, processing, and reporting for environmental data and promote interoperability between data repositories. Examples include the Open Geospatial Consortium "Observation and Measurements" standard for observations and sampling features (OGC 2013, ISO/DIS 19156:2010), International Standards Organization/ Federal Geographic Data Committee standards for geospatial (FGDC 1998, ISO 19115-1:2014) and temporal metadata (ISO 8601), netCDF formats for climate and forecast metadata (Unidata 2016), and the Ecological Metadata Language (EML; Michener et al. 1997; EML Project 2009). Data information models built upon these standards describe content data and metadata standard formats and relationships, and are easily converted to searchable relational databases (Horsburgh et al. 2016). Data information models suitable for environmental data include Morpho (NCEAS 2015) that is designed to interface smoothly with EML, and the Observational Data Model 2 (ODM2; Horsburgh et al. 2016). Data information models support a wide range of data types and enable data search, discovery, and synthesis. However, these models still require that additional standard data collection and naming protocols be defined and that metadata for both observations as well as modeled products be collected in a standardized way before it can be ingested into the searchable database. Moreover, these models require the data originator to be proficient in data science terminology or concepts, and to expend significant additional effort into translating their data and notes into the required formats.

In contrast, other domain-specific templates and accompanying databases have been developed to enable easier reporting of data and metadata by data originators for ecophysiology, hydrology, and meteorology datasets. These efforts include forest plot inventories that collect forest census data like taxa identification, locations, causes of mortality, and size (e.g., Smithsonian Tropical Research Institute - Center for Tropical Forest Study (STRI-CTFS; Condit et al. 2014), CTFS-ForestGEO (CTFS Forest Global Earth Observatories; Anderson-Teixeira et al. 2014) and the Amazon Forest Inventory Network (RAINFOR; Malhi et al. 2002; Peacock et al. 2007)). The AmeriFlux / Biological, Ancillary, Disturbance and Metadata (BADM) protocol has been developed and implemented across several flux-based networks (e.g., AmeriFlux, FLUXNET, ICOS) (Law et al. 2008; AmeriFlux 2016). Ameriflux / BADM reporting templates focus primarily on ecosystem-level observations often aggregated in space and time to describe the area within a flux tower footprint. A variety of frameworks support regional and global data repositories, such as Biofuel Ecophysiological Traits and Yields (BETYdb) Database (LeBauer et al. 2010), Sapfluxnet (Poyatos et al. 2016), and International Soil Carbon Network (ISCN) (ISCN 2016). These frameworks are designed to capture metadata specific to their respective measurement types. However, the reporting templates do not necessarily conform to published standards, and are sometimes unstructured, making data synthesis, search within the data, and integration into a database difficult.

Thus, the existing data informational models are too complex for ecohydrological data originators to use directly, and none of the existing standardized data/metadata templates have the necessary structure to support reporting of the diverse observations required for earth system modeling. To bridge this gap between data information models and domain-specific data/metadata reporting templates, we developed a new metadata reporting framework, FRAMES (A Framework for Reporting dAta and Metadata for Earth Science). FRAMES is a set of templates that standardizes reporting of diverse ecohydrological data for synthesis across a range of spatiotemporal scales, and ultimately enables ingestion into a searchable data information model.

We conducted this work as part of an interdisciplinary team-based project whose overarching goal is "to develop a predictive understanding of how tropical forest carbon balance and climate system feedbacks will respond to changing environmental drivers over the 21$^{st}$ Century" (NGEE Tropics 2016). By employing an iterative scientist-centered design approach, we identified and implemented features into FRAMES that support not only environmental process understanding but also earth system model development. These features include 1) standardization and organization of metadata according to best data science practices, 2) a modular design that can expand to accommodate diverse measurements, 3) data entry formats that facilitate efficient metadata reporting, 4) a multiscale hierarchy that links observations across spatiotemporal scales, and 5) collection of metadata needed for model-data integration. Although extensible to various earth system data types, the first version of FRAMES described here is focused on primarily automated measurements collected by permanently located sensors, including sap flow (tree water use), leaf surface temperature, soil water content, dendrometry (stem diameter growth increment), and solar radiation. In addition to describing FRAMES, we discuss key challenges, solutions, lessons learned, and areas for future development that are broadly applicable to team-based projects and science networks.

## 2. Methods

Our team-based project supports a dedicated data team that is tightly integrated with an interdisciplinary group of earth scientists. The data team encompasses responsibilities of data manager and data distributor, and refers to persons assisting data originators in metadata and data reporting, preserving data, and making data available to consumers (Peng et al. 2016). The data team led the development of FRAMES by working closely with data originators (the empiricists collecting the observations), as well as data consumers (the empiricists and also modelers using the data and metadata).

We developed FRAMES to support the project's first coordinated data collection effort centered around tree responses to drought conditions in Central and South America during the El Nino Southern Oscillation (ENSO) event of 2015-2016. Prior to developing FRAMES, we identified relevant aspects of existing protocols and standards to use as design foundations including ISO standards (ISO 8601, ISO 19115-1:2014), FGDC standards (FGDC 1998), Ameriflux/BADM templates (AmeriFlux 2016), ISCN reporting templates (ISCN 2016), STRI-CTFS protocols (Anderson-Teixeira et al. 2014; Condit et al. 2014), RAINFOR-GEM protocols (Marthews et al. 2014; RAINFOR 2016), and Sapfluxnet (Poyatos et al. 2016).

The approach we used to develop FRAMES involved a combination of agile development principles and scientist-centered design (Ramakrishnan et al. 2014). Agile development uses short incremental development cycles with reassessment of priorities and solicitation of feedback after each cycle. The scientist centered-design process works closely with a group of researchers (data originators and data consumers) that provide direction and feedback throughout product development to define the desired end products. The process begins with extensive interviews to understand each participant's standard processes and workflows. It works to 1) understand data sources, QA/QC needed, and development priorities; 2) develop data algorithms, and 3) build products that enable the science goals.

Based on requests from members of the project's science team, we focused our efforts on collecting metadata necessary to provide interpretation, cross-site comparison, and QA/QC for a prioritized list of ENSO observations. These observations were primarily automated measurements collected by permanently located sensors, including sap flow (tree water use), leaf surface temperature, soil water content, dendrometry (stem diameter growth increment), and solar radiation. Working closely with data originators and data consumers, we addressed one or two measurement types at a time, building out FRAMES as we added additional measurement types. Initial template designs were based on existing data collection protocols and informational interviews conducted with data originators to understand the measurement procedure, identify existing metadata collection methods, and discuss additional metadata collection. Through discussions with data originators and consumers as well as our expertise in data management, required metadata were distinguished from optional metadata based on which information was needed to interpret data, perform cross-site comparison, and conduct QA/QC assessment.

FRAMES was designed to fit as seamlessly as possible into the existing data collection processes of the data originators. We iteratively tested FRAMES with data originators, incorporating additional measurement types and feedback based on field metadata entry trials. Once we had tested FRAMES with four of the ENSO measurement types as well as location and equipment information, we solicited feedback from modelers (data consumers). We also conducted informational interviews with other data originators and consumers of anticipated measurement types (primarily sample-based observations including leaf water potential, gas exchange, and non-structural carbohydrates) to check for compatibility with FRAMES. To minimize the effort of data originators, we transferred information already submitted in previous versions of FRAMES to the newer versions throughout the iterative development.

Finally, FRAMES was designed to facilitate submission to data repositories, including the NGEE Tropics Archive, the project's data repository. The NGEE Tropics Archive has a web portal that allows data originators to upload and download data packages. The Archive is supported by a programmatic REST API built on top of Django Python web framework with an easy-to-use web user interface built with Foundation (Zurb 2016) front-end framework. The Foundation front-end framework is flexible, highly customizable and provides support for responsive, light-weight HTML for mobile application support. Django is a fully featured open-source Python web application framework that supports rapid development. Django makes the low-level framework decisions so that the development is primarily focused on the application domain rather than composing the framework features. NGEE Tropics Archive manages the data package by storing the data package metadata in a Postgres database and the data files on the local file system.

In general, completeness and accuracy of metadata submitted via FRAMES templates are considered to be the responsibility of the data originator, although the data team manually inspects data package submissions via the NGEE Tropics Archive portal. The peer-review process enabled by data-sharing provides input to data originators to make corrections to their data.

## 3. Results: A Framework for Reporting dAta and Metadata for Earth Science (FRAMES)

### *3.1 Key requirements and characteristics of FRAMES*

Through initial interviews, we identified key requirements of a metadata framework that would enable multisite comparisons of tree response to drought and testing of spatially explicit models. First, the framework had to support a variety of measurement types and data processing levels that were anticipated to be made and used throughout the project. Many of these measurement types shared similar metadata while some metadata was measurement specific. Secondly, the framework had to enable efficient data entry in recognition of the fact that metadata reporting is time consuming and can add significant overhead to a data originator's field collection and data reporting duties. Additionally, scientists needed the ability to use the data reported at various scales. For example, they wanted, on smaller scales to track multiple, co-located measurement types on a specific tree for assessment of plant trait co-variation, and on larger scales to track relationships across study sites. Finally, the framework had to support integration of data into carbon cycle models, which was identified as a top project priority.

Thus, FRAMES was designed to address these requirements, resulting in the following key characteristics: 1) Standardization and organization of metadata according to best data science practices (Section 3.2), 2) A modular organization in which data originators can report information about data file contents, measurement settings for a variety of observations, and high-level data descriptions and citation information (Section 3.3), 3) Reporting formats designed to match existing data collection practices for efficient and streamlined metadata entry (Section 3.4), 4) The concept of a multiscale measurement position hierarchy to enable data aggregation and usage across scales (Section 3.5), and 5) Incorporation of additional data and metadata fields that would normally not be collected as part of a field measurement, but were required for model-data integration (Section 3.6).

*3.2 Standardization and organization of metadata according to best data science practices*

FRAMES uses concepts and terminology from preexisting standards, templates and databases, to support compatibility with external data formats and protocols. First, for sites with a pre-existing, widely-used identifier such as an AmeriFlux/FLUXNET Site ID (AmeriFlux 2016), we used the existing ID, to enable standardization with a global network of sites and cross-database search. Other site and plot metadata, including location information and descriptions, were collected directly from site leads or data originators (see Appendix B). The FGDC standard (FGDC 1998) was supported for reporting spatial location metadata in different reference systems including geographic coordinates (for latitude/longitude representation), planar coordinates (for coordinate or distance/bearing representations), and vertical coordinates (for heights). All dates and timestamps had to be reported in ISO formats (ISO 8601), and a UTC offset specified. The Ameriflux/BADM reporting templates (Ameriflux 2016) were used as a starting point for determining fields for equipment information, installation, and maintenance, as well as for the multiscale measurement position hierarchy (Section 3.5).

We also supported compatibility of certain domain-specific standard terminology when applicable. For example, we have largely adopted the taxa identification protocol and based our tree characteristics on the censusing protocols of STRI-CTFS (Anderson-Teixeira et al. 2014; Condit et al. 2014). Additionally, we leveraged RAINFOR-GEM's tree assessment protocols for the measurement of tree height and canopy illumination indices (Marthews et al. 2014; RAINFOR 2016). For sap flow measurements, we consulted the AmeriFlux / BADM and Sapfluxnet protocols (AmeriFlux 2016, CREAF 2016; Poyatos et al. 2016). For soil water content and other soil-related observations, we consulted the Ameriflux/BADM and ISCN data reporting templates (Law et al. 2008; AmeriFlux 2016, ISCN 2016).

Besides the use of preexisting standards, FRAMES also incorporates other best data science practices including 1) standardization of variable names and file structure to enable automation of metadata extraction via scripts, 2) use of controlled vocabularies in drop down menus to facilitate comparability and search across sites, 3) use of descriptive data filenames and definition of data file contents, for example using header lines describing variables, and 4) tabular, row-based data entry templates with consistent column types (e.g. Borer et al. 2009, Hook et al. 2010, Tenopir et al. 2011).

### *3.3. Modular Metadata Organization*

FRAMES is organized into three main groups of related metadata: 1) descriptive information about a data package, 2) content information about the data file organization, and 3) content information about the data collection process and measurement settings. Physically, FRAMES comprises a set of Microsoft (MS) Excel spreadsheet files to describe file contents and measurement setting metadata, and package-level descriptive metadata reported in a web form (spreadsheet templates included in Appendix A, web screenshots included in Appendix E). The metadata are bundled with data files into a data package and submitted to a data repository (e.g. the NGEE Tropics Data Archive) via a web form. The data reporting workflow is illustrated in Figure 1, and an overview of FRAMES with relationships between the templates is illustrated in Figure 2. With this combination of metadata files submitted to a data repository, FRAMES enables digital preservation of the entire data history, including digital reporting of critical information from field notes and raw data files generated by data loggers, to enable reproducibility of scientific analyses.

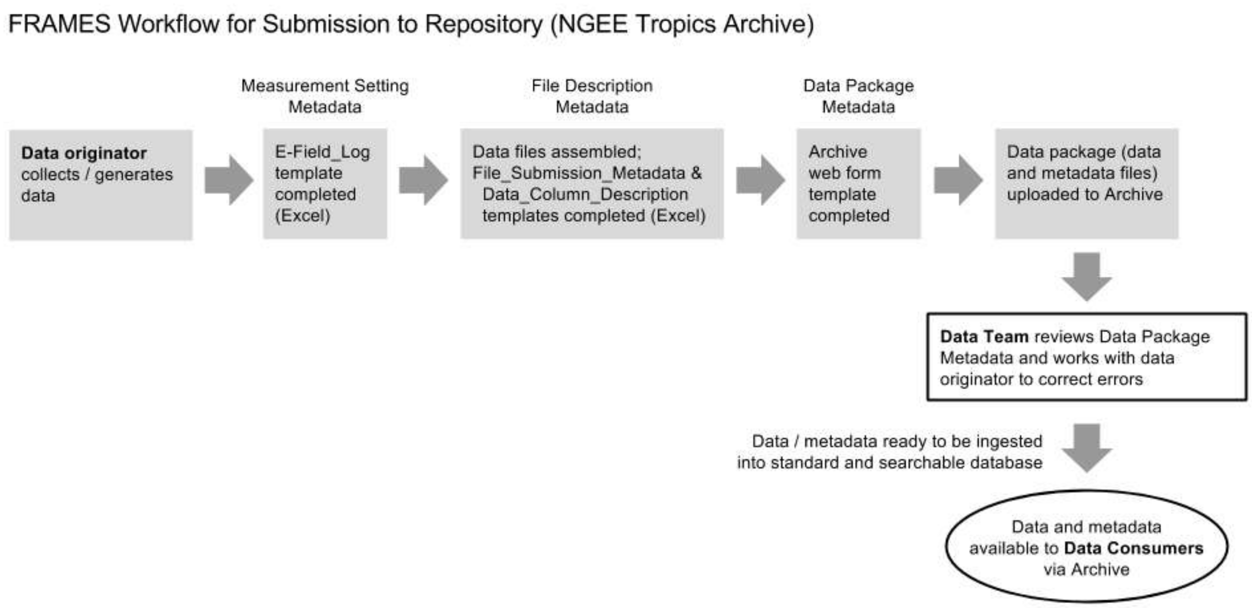


Figure 1: FRAMES metadata and data package workflow. The Data Originator (grey boxes) collects / generates data and completes FRAME metadata templates (Section 3.3) that are included with data in a data package for submission to a repository (e.g., NGEE Tropics Archive). The Data Team (outlined box) reviews the data package before it is available to Data Consumers (outlined oval) via the Archive.

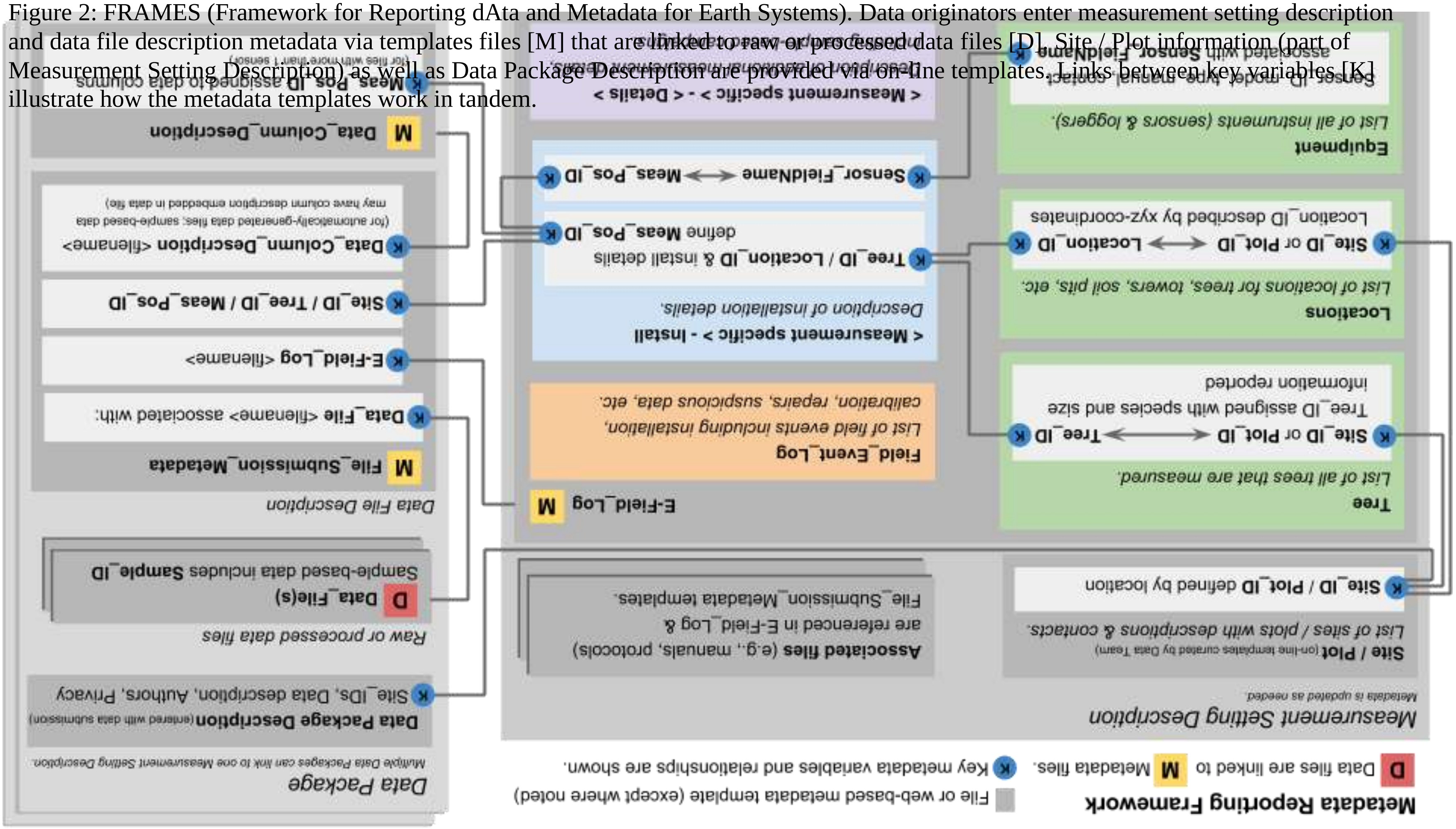


Figure 2: FRAMES (Framework for Reporting dAta and Metadata for Earth Systems). Data originators enter measurement setting description and data file description metadata via templates files [M] that are linked to raw or processed data files [D]. Site / Plot information (part of Measurement Setting Description) as well as Data Package Description are provided via on-line templates. Links between key variables [K] illustrate how the metadata templates work in tandem.

### *3.3.1 Data Package Description*

FRAMES utilizes the concept of data packages, in which data originators bundle their content (data files) and corresponding content metadata information together for submission to a repository. A data package is often determined by a common theme or activity. Within our project, data packages are typically assembled to support an experiment or set of sensor observations, a data synthesis product, a publication, or a field campaign. A data package may contain many types of data associated with the theme or activity.

The data package description is a set of basic metadata fields that describe its contents and includes information necessary to obtain a unique Digital Object Identifier (DOI), as well as other information needed to identify the package for search and retrieval in the future. These metadata include data package names and descriptions, Site ID and Plot ID, authors, institutions, citations, acknowledgements, and funding sources, as well as QA/QC status (Appendix E). The metadata collected also describes access permissions for data usage. Required fields for the data package description were determined as the minimum set of information needed to obtain a DOI from Datacite (Datacite 2016) via the U.S. Department of Energy's Office of Science and Technical Information (OSTI).

For the NGEE Tropics project, data are archived using the project's data repository NGEE Tropics Archive, which allows users to upload and access data packages. Currently, data originators can create, save, edit, and submit draft data packages via a web portal (Appendix E). Data originators provide descriptive metadata about the data package in a web form and can upload a single data file of any type (zipped file types allow for upload of multiple files). The web form enables data originators to reuse certain information, such as field site and plot information and person (name, email, institution) information to minimize inconsistent or erroneous data entry. For example, data originators only have to select the site name/ID for all related site information to be auto-populated, including spatial coordinates (numerically and via google maps), PI (principal investigator) information, and general site descriptions.

Once submitted, data package descriptions and data files are manually reviewed for completeness and accuracy as part of the project's archival approval processes. After approval, data packages with appropriate citation information are made available via the web portal to data consumers who are assigned access privileges.

### *3.3.2 Data File Descriptions (File Submission Metadata and Data Column Description)*

For each data file submitted, data originators report the following metadata in the MS Excel template "File Submission Metadata:" 1) Tree ID or other Location ID if applicable, 2) time period of the data and timestamp details (e.g., time zone and whether the timestamp is at the start, middle, or end of the sampling period), 3) data processing level with related processing approaches (e.g. raw, translated/processed, data originator QA/QC, project-level QA/QC), 4) references to the measurement setting description (e.g., E-Field Log file)—this information is essential because it links the data to additional metadata reported in the separate templates described in 3.2.3 (see Figure 2)—, and 5) references to data file descriptions (Data Column Description).

Additionally, for every data file, a corresponding "Data Column Description" template provides the information necessary to understand the data file. This is a semi-standardized template that includes information on header rows (e.g., those automatically generated by instrumentation), column names, units, data averaging (e.g., instantaneous or a mean / standard deviation over the sampling period), measurement type, and a location identifier (e.g., Tree ID, Measurement Position ID, or Sample ID) if multiple measurement positions are recorded in the same file. The location identifier is critical because it links the observations to installation details and other events affecting data quality that are described in the measurement setting templates. Data originators can configure the Data Column Description as a series of tabs in a single MS Excel file, a standalone file, or as a separate tab within the data files (if data file is MS Excel).

### *3.3.3 Measurement Setting Description (E-field Log)*

The measurement setting description contains information related to observations: 1) location; 2) equipment details, installation, and maintenance history; 3) approach and technicians; 4) events affecting data quality. We developed a standardized digital format for this information to which data originators could transfer their field notes. Because this information is complex and often hierarchical, we organized the information into a series of templates implemented as tabs in a MS Excel file "E-Field Log" (Table 1). Key variables that link the templates together are shown in Figure 2 (See Appendix C for full relational framework). All variables within each template are described in Appendix B. Examples of measurement setting description variables are illustrated for sap flow in Figure 3.

Table 1: Measurement setting description template groupings included in the E-Field Log file.

| E-Field Log Template | Template Description |
|---|---|
| Tree | Description of observed trees, including species identification and an initial assessment of size and light environment. We include this information because our framework is designed for research in tropical forests. Long-term demographic (census) data is reported elsewhere. |
| Locations | Location (relative or absolute) information, geomorphology description, and contact information for features where observations are made. Example features include trees, towers, cranes, pits, and random observations points. |
| Equipment | Description of equipment used to make observations, including make, model, contact personnel, and reference to manuals. |

| | |
|---|---|
| Field Event Log | Description of field events that affect data collection and quality. Event examples include equipment installation, maintenance, calibration, and removal, as well as broad categories like “Suspicious Data” that capture events such as power outages or animal interference. |
| Measurement-specific Install | Detailed description of installation events specific to each measurement type that requires (semi-)permanently installed equipment. For example, a sap flow sensor installation event is recorded on the Field Event Log and the details of that installation, such as sensor height and probe depth illustrated in Figure 4, are recorded on the SAP-Install template. |
| Measurement-specific Details | Detailed description of measurement specific information. These templates are designed to capture various types of measurement specific information not recorded on the Field Event Log. For example, leaf gas exchange and leaf water potential observations are conducted in campaigns. Details of the campaign are captured on the Leaf-Campaign template. |

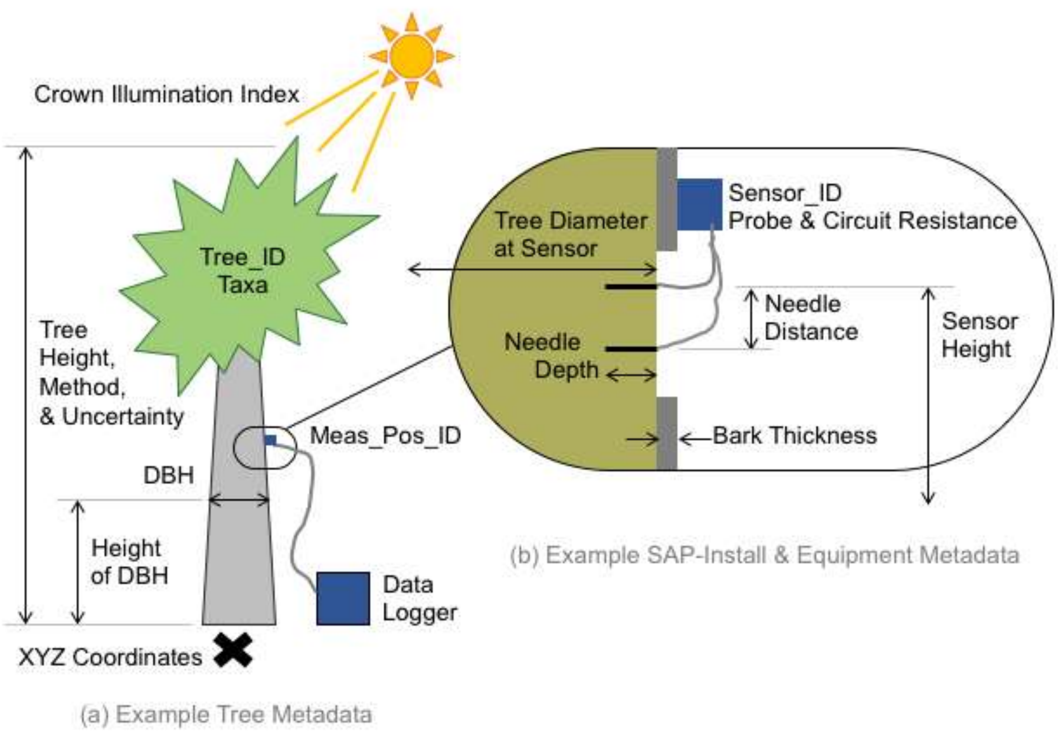


Figure 3: Examples of (a) Tree and (b) SAP-Install and Equipment metadata variables that are reported as part of the measurement setting description. SAP = Sap flow; Meas_Pos_ID = Measurement position ID; DBH = diameter at breast height.

*3.4 Design features that maximize metadata reporting efficiency and data/metadata reuse*

To maximize efficiency of reporting metadata and data reuse, we implemented several design features based on data originator interviews and observations of originators entering metadata on beta template versions.

FRAMES enables efficient data entry by being closely aligned with existing field practices as follows. The modular organization of FRAMES (Section 3.3) facilitates co-located entry of related metadata relevant to multiple measurement types or field sites/locations. One example occurs in the web form that data originators use to submit data packages to the project's repository. Data originators are allowed to submit multiple data files associated with any number of sites and variables. Thus, originators can submit several related data files, for example those associated with a field campaign, in one data package, minimizing time spent on entering metadata and uploading files. As another example, in the measurement setting description spreadsheet ("E-field Log" file), details about measured trees as well as equipment specifications are reported once in the Tree and Equipment templates respectively. Co-location of the measurement setting templates in a single file allows for quick reference between location and equipment metadata when describing installation and other field events. Data originators can also report events that affect multiple measurements in a single entry in the E-Field Log file. For example, a power outage affecting soil moisture and sap flow measurements can be reported as suspicious data in one line on the "Field Event Log" template with location and/or sensor identifiers indicated. Through translation of such suspicious data information—automated if desired—, data quality flags can be assigned to the affected data values.

We also intentionally separated the measurement setting description (E-Field Log) from metadata describing the data package and data files to allow any data originator to link multiple data files to a single set of metadata templates in the E-Field Log. Thus, data originators can submit the E-field log as a separate data package into the data repository. This structure allows for the data and the measurement setting metadata to be maintained independently of each other, as the latter are typically updated on an infrequent basis. Furthermore it enables reuse of certain metadata across research studies and field sites. For example, two research groups collecting different observations at one or multiple sites can both reference the same E-field log record in the data repository to share tree, location or equipment information. Finally, multiple types of data, for example raw, processed, or cross-site data synthesis products, can all be linked to the appropriate metadata templates.

Finally, we embedded instructional text and formatting cues to facilitate metadata entry. Within FRAMES, short instructions, metadata variable descriptions, and example entries are provided. Templates within the E-Field_Log MS Excel file are color coded to indicate similar types of metadata: infrequently changing lists relevant to multiple measurement types, infrequently changing measurement-specific installation templates, and the Field Event Log that is updated at various frequencies. These colors matched highly visual instructional documentation (See Appendix A).

*3.5 Multiscale Measurement Position Hierarchy*

We developed a multiscale measurement position hierarchy to account for the diverse spatial scales that observations represent and to reduce redundancies in reporting of various location identifiers (Figure 4). In this hierarchy, a “Site” is the largest unit of study, and is assigned a unique Site ID. We impose no limit on the physical size of a site, which can range from individual locales to regional areas and the entire globe; however, we anticipate most sites to be individual locales on the order of kilometers squared. Smaller “Plot” areas can occur within a site, and each plot has a unique Plot ID. Within a site or plot, a feature located in *x*-*y* space, including trees, towers, measurement pits, etc., is assigned a Location ID. Observations occurring repeatedly at a sub-location, e.g., at a specific height or bearing, are assigned a unique Measurement Position ID. Alternatively, observations obtained from a sample of the feature are assigned a unique Sample ID, which may have specific sub-location spatial information.

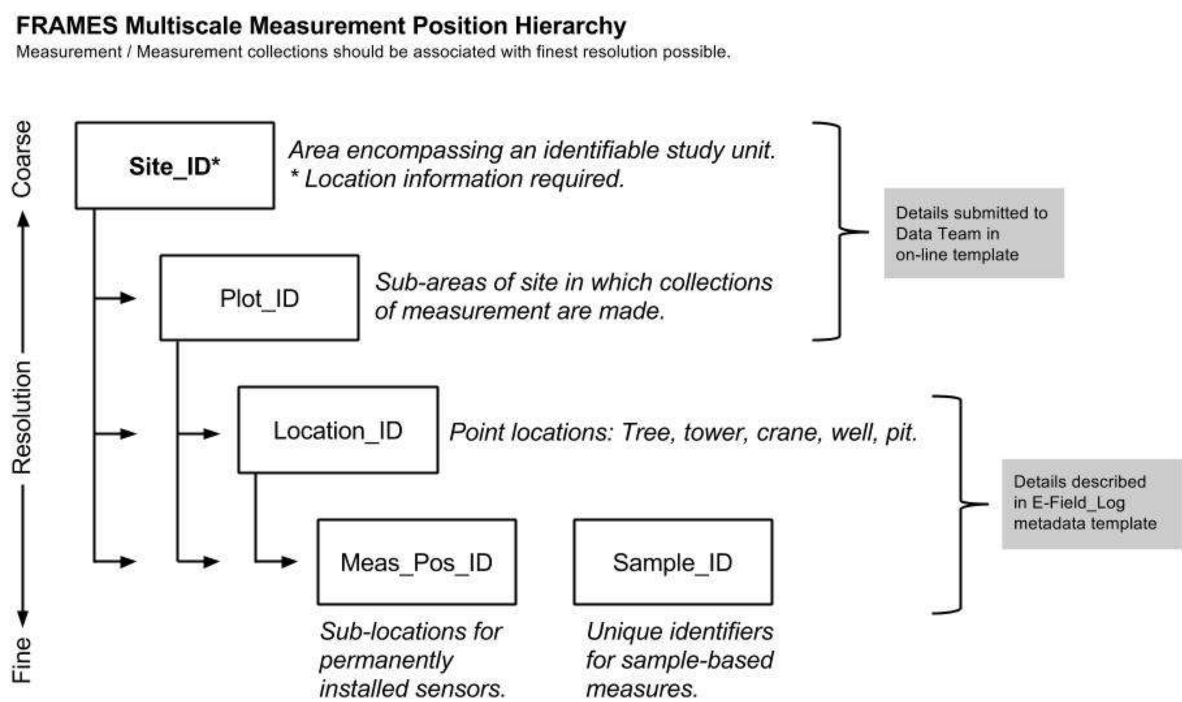


Figure 4: FRAMES Multiscale Measurement Position Hierarchy. Observations including time series are associated with a unique measurement position identifier that may be at any hierarchy level. Any finer level identifier must be linked with at least a Site ID. Within our project focused on forest system, Tree ID is a type of Location ID.

Observations are linked to a unique spatial identifier in the hierarchy and inherit location information from the coarser levels to which that ID is linked. Aggregation to coarser resolutions is thus facilitated by combining all spatial identifiers that are linked to a particular coarser level location. For example, to aggregate individual sensors in a given Plot ID, all measurement position IDs associated with the Plot ID are combined. If multiple levels of locations are defined, an observation or observation time series is associated with the finest resolution spatial identifier defined; however, only Site ID is required. In this measurement position approach, sensors, either permanently installed or mobile, are linked to the appropriate spatial position identifier. Once Site or Plot metadata is collected, it is bundled with Location and Tree metadata (Section 3.3.3) for data originator and consumer reference.

### *3.6 Integration of field observations for model development*

Integration of data with models requires translation of empirical observations into the units and time periods required for model inputs or for direct comparison with model output. For example, meteorological time series data, such as air temperature, solar radiation, precipitation, and vapor pressure deficit, are used as boundary conditions to drive earth system models at each time step. In model parameterization, functional characteristics, ideally based on field observations, are assigned to plant functional types (PFT), soil types, and other model components. These functional characteristics, or traits, such as photosynthetic capacity, minimum leaf water potential, and soil organic matter content, may vary with climate conditions, other site characteristics or plant functional traits, component age, or spatial position (e.g., canopy level or depth). For model benchmarking, model predictions through time — for example, size distributions and relative abundance of PFTs, sap flow, and soil water content — are compared to field observations. Field observations are also used to provide insight into modeled ecosystem, ecophysiological, and hydrological processes.

To support model-data integration, we designed FRAMES to capture model-relevant metadata, which are sometimes not collected as part of the data originator's field efforts. In particular, we focused on information to support parametrization and benchmarking of the Functionally Assembled Terrestrial Ecosystem Simulator (FATES) model, which is based on Community Land Model with Ecosystem Demography (CLM(ED); Fisher et al. 2015) and ED (Moorcroft et al. 2001). FATES is a vegetation model that is being developed and used by the project's modelers. In FATES, plant demography (birth, growth, and mortality processes of related plants within a defined area) is modeled with size- and plant functional type-specific responses to environmental conditions. By requiring that the tree height and species information be reported, FRAMES provides input data, like photosynthetic capacity, for FATES to model plant responses, like sap flow and leaf gas exchange. These modeled plant responses are then benchmarked against observed responses made on similar trees under similar environmental conditions. FRAMES ensures that modeled plant responses can be compared to observed responses by linking the measurements to required tree characteristic metadata via the Tree ID. For example, crown illumination index and tree height, which are typically not collected or reported with leaf-level or plant-level response measurements, are required metadata for each measured tree. FRAMES has formalized communication between field scientists and modelers by ensuring that critical information is collected in a standardized, usable way for FATES and similar earth system models, such as ED2 (Medvigy et al. 2009).

## 4. Discussion: FRAMES applications in interdisciplinary team-based earth science

### *4.1 Linking complex and diverse observations across spatiotemporal scales for data synthesis*

Linking observations across spatiotemporal scales is necessary for earth system process understanding as well as model parameterization and benchmarking (Dietze et al. 2013). FRAMES enables such linkages via the multi-scale measurement position hierarchy, its modular structure, and metadata standardization.

As a spatial example, sap flow is measured at the sub-tree level (Figure 5a). Sap flow observations at multiple positions on the tree are used to determine the radial profile of sap flow within the sapwood and at different heights along a tree (e.g., trunk or branch). Integrating these measures yields an understanding of water use for a whole tree (Figure 5b). The plant hydraulic functionality of FATES predicts tree water use for each combination of tree size and plant functional type. These model predictions can be benchmarked with whole tree water use of similar trees, as estimated from sap flow radial profile observations. Further aggregation at the site and regional scale enables benchmarking of site and regional model configurations, respectively (Figure 5c-d). Synthesizing sap flow dynamics within a tree, for the whole tree, for groups of functionally related trees, and across the pantropical region enables improved understanding of ecohydrological processes in hyper-diverse tropical forests (Goldstein et al. 1998; Meinzer et al. 2001; Meinzer et al. 2004; Meinzer et al. 2005; Bell et al. 2015). The multiscale measurement position hierarchy facilitates such spatially extensive analyses because observations are defined by their position on the landscape and are linked by unique measurement position, tree (location), and site identifiers. Additionally, the modular structure and standardization of FRAMES has enabled a pantropical sapflow synthesis effort involving several field sites (and hence many data packages). A data consumer independently automated 1) metadata ingestion from the templates, 2) integration of the metadata with the data files, and 3) additional data processing like removing duplicate timestamps (see Appendix F for R code).

Similarly, integration of observations across temporal scales is fundamental to understanding ecosystem processes (e.g., Detto et al. 2012). Furthermore, models that predict processes well across temporal scales remain elusive, i.e., models that perform well at fine scales (hourly or daily) often perform poorly at coarser scales (Dietze et al. 2011). Using FRAMES’s description of data collection time resolutions and methods (e.g. discrete data or data averaged over a time intervals with the timestamp indicating the start, end or middle of the averaging time period), data consumers can temporally aggregate observations as required. For example, FATES predicts plant water flux dynamics from sub-hourly to seasonal and inter-annual timescales, as driven by interactions between various plant hydraulic traits and environmental variation (as in Christoffersen et al. 2016). Using FRAMES these hydrodynamics may be benchmarked with sap flow data collected across project field sites at different sampling frequencies (10-, 15-, or 30-minutes) by aggregating to the desired model output time frequency (e.g. see Appendix F for R code that uses FRAMES metadata to automate this).

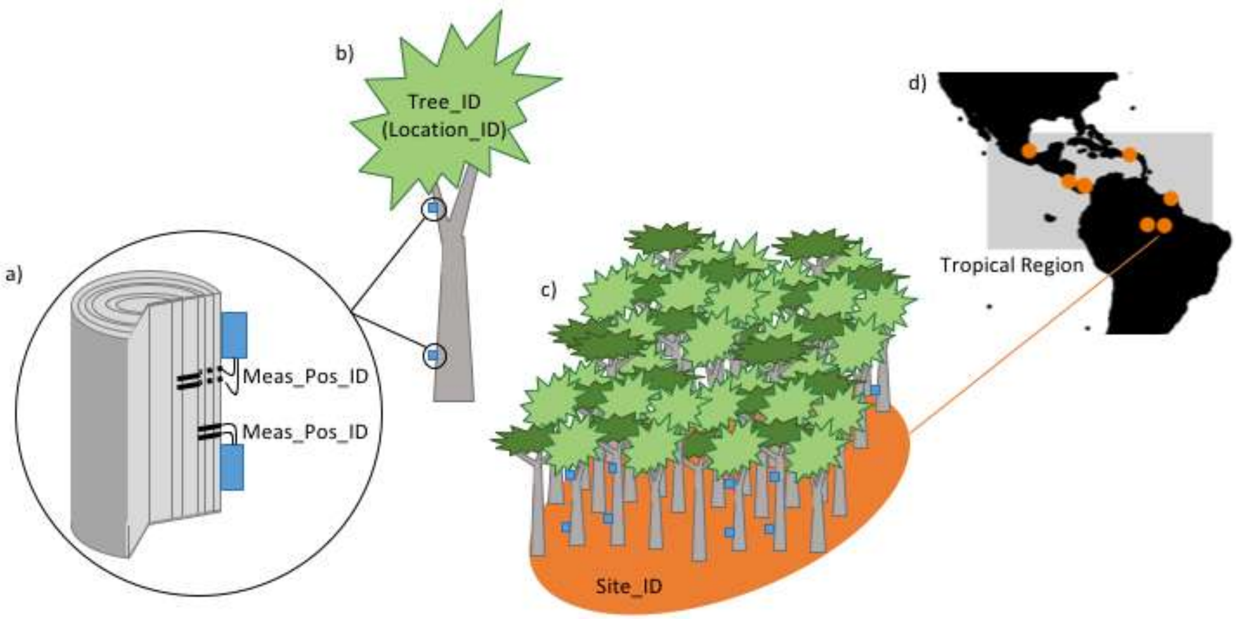


Figure 5: Spatial scaling of sap flow measurement using multiscale measurement position hierarchy. Using measurement position identifiers that are linked to a common tree identifier, individual sap velocity measurements (a) made at multiple depths and positions on the tree can be processed with sapwood area or dendrometry measurements to the characterize sap flow for the entire tree (b). (c) Aggregation across individuals within a single species or plant functional type (light or dark green trees separately) or across an entire site (light and dark trees combined) is enabled by tree identifiers that are linked to species / plant functional types and site identifiers. (d) Regional sap flow characterization can be synthesized by aggregating across site identifiers. As an example, Meas_Pos_IDs 00002A and 00002B (a) are linked to Tree_ID 00002 (b) which in turn is linked to Site_ID BR-Ma2 (c). If the tropical region (d) includes site BR-Ma2, then observations from Meas_Pos_IDs 00002A and 00002B or for Tree_ID 00002 would be easily accessed for regional aggregation.

Alternatively, FATES hydrodynamic predictions can be benchmarked by tracking sub-hourly extremes like daily maximum sap flow, the timing of which is not known a priori, over periods of gradual declines in water availability, which occurred during the ENSO measurement campaign. Thus, by providing data at the finest resolution collected with the corresponding metadata to describe it, modelers have flexibility to customize model benchmarking to best assess a specific process. Additionally, analyses to understand covariation between sap flow and leaf surface temperature are highly sensitive to mismatches or drift in the timestamp. In conjunction with the description of time resolutions and methods, FRAMES includes a consistent reporting method for tracking timestamp drift by tracking the logger and CPU timestamps at data download events (Field Event Log in E-Field Log Excel file in Appendix A).

Time-series data collected by different sensors at the same measurement position can be easily linked using the measurement position identifiers at any hierarchical level. For example, continuously measured leaf surface temperature is easily compared with sample-based measured leaf water potential measurements observed on the same tree via the tree identifier. Additionally, location information reported in FRAMES allows for linkages of spatially-explicit measures. For example, sap flow, leaf temperature, leaf water potential, and dendrometry measured on a specific tree can be simultaneously correlated with representative soil moisture conditions.

### *4.2 Expandability of FRAMES to accommodate diverse data*

Data needed for earth system science, are not only diverse but also change as models and measurement techniques advance. Thus, the metadata reporting framework for such data must accommodate a variety of existing and new measurement types and approaches. FRAMES is modular to enable expansion to additional measurement types, beyond the few ecohydrological observations for which we have currently defined it.

A key aspect of the modular organization is separation of metadata reporting into three types of descriptions: data package, data file, and measurement setting. The data package description includes a minimal set of generic information, such as site identifier(s), data owner, and privacy settings. Similarly, the data file description is applicable to a wide variety of data types because it also contains generic metadata, like time step and data processing information. Data originators are not restricted to predefined measurement types and formats because the semi-open ended data file column description can describe the content of almost any type of data file. The modular organization of the measurement setting description also readily accommodates new measurement types because the core set of reporting templates (Tree, Equipment, Location, and Field Event Log) describe information relevant to most measurement types in earth system science. New measurement types utilize some or all of these core description templates, and if necessary, a measurement-specific template can be developed to report additional measurement-specific information (see Appendix D for an example of how to add a new measurement to FRAMES).

The modular expandability of FRAMES is similar and compatible with ODM2 (Horsburgh et al. 2016), in that metadata is bundled in related groups. The difference is that ODM2 is a database structure for standardized metadata and data protocols. FRAMES operationalizes such a data structure as a reporting mechanism. In other words, data reported via FRAMES can be translated to a standardized format for assimilation into a database. This pre-database, standard-compatible flexibility differentiates FRAMES from other existing frameworks such as AmeriFlux / BADM, ISCN, and Sapfluxnet, which collect metadata and data in a standardized protocol designed for direct database assimilation.

We took this flexible approach for two reasons. First, it accommodates the needs of data originators by removing barriers to metadata and data sharing, such as the effort required to convert data to specific units and formats. Secondly, via the Data Column Description template which accommodates most types of data files, the flexible approach allows for archiving of raw data directly from loggers. Archiving unaltered data in its original format provides the full history of a data product for repeatability and data quality assessment (measurement errors as well as data processing errors). Archiving the entire data history is not only good science practice (Dietze et al. 2013, Michener 2015), but is also important for synthesizing data across sites and approaches because common and transparent processing approaches facilitate comparability. An additional advantage of this flexible approach is that data originators and consumers can assimilate data into variety of databases. A key component of this flexibility is achieved by separating the data column description from the data file description so that the data column description can be customized to the specific data file.

*4.3. Lessons learned and future development*

FRAMES has supported data package reporting for six core NGEE Tropics field sites in Brazil, Panama, and Puerto Rico across six measurement types. Portions of the templates have also been used broadly in additional data reporting. Information about sensors, approaches, and installation details have informed development of a common sap flow processing approach for a synthesis of sap flow data across nine study sites. Additionally, the uniformity of the reported data enabled a data consumer to, on his own, automate processing of sap flow measurements for model benchmarking (see Appendix F for R code).

The use of FRAMES for the initial NGEE Tropics data collection effort has enabled us to gather feedback regarding what is working and what is not. The most valuable feedback was the effort that six data originators were willing to exert in using FRAMES to archive their data in the project's repository within a few months after the templates were finalized. We attribute this success largely to the scientist-centered design approach, which allowed us to identify data collection processes and design FRAMES to match the scientific goals and practices of both data originators and consumers. Anecdotally, data originators have reported FRAMES useful in organizing their field data. Subsequent data analyses, for example assessing co-dependent physiological responses measured from different sensors on the same tree, has been facilitated by the fact that all relevant information regarding the measurements is organized centrally within the metadata templates and that the tree ID clearly identifies measurements made on the same tree. Furthermore, FRAMES helped data originators to collect important ancillary information (e.g., tree height, diameter, crown illumination index) in conjunction with scheduled field activities rather than requesting the information at a later time, which would require additional field site visits if the measurement could still be made.

Developing an adaptable and efficient reporting framework was necessary for data synthesis across diverse observations, but its complexity has disadvantages. Understanding the modular templates and linkages seemed overwhelming at first to several of our data originators. Thus, further investigation of the instructional features is needed to ascertain and improve their efficacy. We found that the majority of time costs were upfront due to learning the structure of the framework and entering the measurement setting descriptions. However, since most measurement setting information remains fairly static and is entered in a single template, maintaining the measurement setting description required minimal effort because only infrequent updates were required. For example, once the metadata for equipment and trees were entered, they remained the same over large periods of time, as observations were accumulated and/or new measurements were added.

A potential limitation to the framework is due to the efficient reporting mechanism designed to make reporting easier for data originators. FRAMES does not specify data variable names, units, or formats, which are required for database assimilation. Using FRAMES, reported data can be translated into a standardized protocol for database assimilation, as exemplified by similar case of automation of sap flow processing by a data consumer. The outstanding questions are 1) whether this reporting approach will ultimate result in improved availability of data with accompanying high quality metadata, and 2) what the tradeoffs are in terms of person-hours and who bears that cost—the data originator or dedicated data team personnel. We prioritized reporting formats in FRAMES to maximize reporting efficiency because although improving, the generally low quantity of shared data and poor quality of metadata is problematic in the earth sciences (Tenopir et al. 2011; Kervin et al. 2014; Michener 2015).

Finally, we implemented several templates in MS Excel because of its ubiquity, operating system neutrality (i.e., it runs on Macs and PCs), copy / paste functionality, and off-line access for remote areas with poor Internet. However, MS Excel is not ideal for selection from a controlled vocabulary menu, collaborative data entry, customization of measurement types, real-time automated data quality verification, and machine readability. The use of MS Excel also makes it cumbersome to release new versions of the templates and ensure backwards compatibility with previous files that were submitted. Additionally, separation of metadata in template files currently requires that the data consumer manage separate sources of metadata information and download different data packages for synthesis efforts. The standardization of metadata alleviates some aspects of this limitation by enabling the data consumer to programmatically link the data and metadata (Section 4.1). As others have reported, new software tools are needed (Michener 2015), in our case, tools that merge the functionality of MS Excel and eliminate these limitations. Possibilities include web-based or mobile tools that are available offline, can be written to appropriate output formats (e.g., comma-delimited ascii, NetCDF/HDF5, EML, or JSON files), and are customizable to originator preferences and measurement types (e.g., Jones et al. 2007; McIntosh et al. 2007). In the future, we intend that the metadata and data be ingested into a relational database (using a framework like ODM2) to facilitate programmatic data integration, searchability and easy data manipulation, such as sub-setting and aggregation.

## 5. Conclusions

We developed FRAMES, a set of online web forms and Excel-based metadata templates that position data and metadata for easier entry into an operational data repository. FRAMES is designed to facilitate and improve capture of desired metadata for ecohydrological observations, including information about how measurements were conducted, data file contents, and high-level descriptive metadata for citation and attribution. Thus, FRAMES enables synthesis of diverse ecohydrological and biogeochemical observations for study of earth system processes and for integration with predictive earth system models.

The overarching challenges for synthesizing diverse earth system observations were 1) developing a metadata framework that allowed experts to share data with team members from other disciplines, and 2) collecting sufficient metadata to organize and process data comparably across sites and measurement methods. FRAMES incorporates several key features that addresses these challenges and supports interdisciplinary team-based earth system science, including 1) compatibility with standard data protocols, and conformance with data science best practices that enable data interpretation, comparison of observations across sites and approaches, and QA/QC, 2) a modular design that accommodates diverse data types and can expand as required by measurement and model advancement, 3) compatibility of existing field practices to maximize data and metadata reporting efficiency, 4) a multi-scale measurement position hierarchy and comprehensive time step descriptions that facilitate spatiotemporal aggregation and linkage of measurement types for synthesis, and 5) targeted metadata collection that enables model-data integration.

To date, FRAMES templates have been used, in whole or in part, for several submissions to the NGEE Tropics Data repository. An iterative scientist-centered design was central to the successful use of FRAMES within our project, where the goal is to improve a predictive understanding of carbon cycling in tropical forests under climate change. As an interdisciplinary data team of ecologist, hydrologists, and data scientists working closely with data originators and consumers throughout the development process, we were able to identify features critical to the project’s science needs and develop pragmatic solutions.

This integrated data science approach will underpin further improvement to FRAMES, and we recommend it as a model for harnessing complex and diverse data inherent in team-science and observational networks.

Additionally, FRAMES promotes good data management practices that benefits both data originators and consumers by 1) digitally preserving data with adequate metadata documentation, 2) enabling sharing with the broader community with appropriate citation and attributions, 3) facilitating interoperability with other databases, and 4) broadening data use and reuse for purposes that stretch beyond the initial intentions of the data collection effort (particularly for use in earth system models). Next steps involve making improvements to FRAMES based on data originator and consumer feedback, and extraction of information in data packages into a queryable database that enables programmatic search, discovery, and processing of data.

## 6. Appendices

Appendix A: FRAMES reporting templates and instructional materials

Appendix B: Description of FRAMES metadata variables

Appendix C: FRAMES relational diagram

Appendix D: Example of measurement addition to FRAMES

Appendix E: Screenshots of NGEE Tropics Archive

Appendix F: R code for merging data with FRAMES metadata


## Acknowledgments

We thank the larger Next Generation Ecosystem Experiments-Tropics (NGEE-Tropics) team for helpful feedback throughout the development process. Additionally, we thank Cory Snavely for background on digital data preservation concepts and terminologies. This research was supported as part of NGEE-Tropics, funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under contract no. DE-AC02-05CH11231. We acknowledge support from the Central Office of the Large Scale Biosphere Atmosphere Experiment in Amazonia (LBA) and the National Institute of Amazonia Research (INPA).

## References

AmeriFlux, 2016. http://ameriflux.lbl.gov/data/badm-data-templates/, accessed November 18, 2016.

Anderson-Teixeira, K.J., Davies, S.J., Bennett, A.C., Gonzalez-Akre, E.B., Muller-Landau, H.C., Joseph Wright, S., Abu Salim, K., Almeyda Zambrano, A.M., Alonso, A., Baltzer, J.L., Basset, Y., Bourg, N.A., Broadbent, E.N., Brockelman, W.Y., Bunyavejchewin, S., Burslem, D.F.R.P., Butt, N., Cao, M., Cardenas, D., Chuyong, G.B., Clay, K., Cordell, S., Dattaraja, H.S., Deng, X., Detto, M., Du, X., Duque, A., Erikson, D.L., Ewango, C.E.N., Fischer, G.A., Fletcher, C., Foster, R.B., Giardina, C.P., Gilbert, G.S., Gunatilleke, N., Gunatilleke, S., Hao, Z., Hargrove, W.W., Hart, T.B., Hau, B.C.H., He, F., Hoffman, F.M., Howe, R.W., Hubbell, S.P., Inman-Narahari, F.M., Jansen, P.A., Jiang, M., Johnson, D.J., Kanzaki, M., Kassim, A.R., Kenfack, D., Kibet, S., Kinnaird, M.F., Korte, L., Kral, K., Kumar, J., Larson, A.J., Li, Y., Li, X., Liu, S., Lum, S.K.Y., Lutz, J.A., Ma, K., Maddalena, D.M., Makana, J.-R., Malhi, Y., Marthews, T., Mat Serudin, R., McMahon, S.M., McShea, W.J., Memiaghe, H.R., Mi, X., Mizuno, T., Morecroft, M., Myers, J.A., Novotny, V., de Oliveira, A.A., Ong, P.S., Orwig, D.A., Ostertag, R., Ouden, den, J., Parker, G.G., Phillips, R.P., Sack, L., Sainge, M.N., Sang, W., Sri-ngernyuang, K., Sukumar, R., Sun, I.-F., Sungpalee, W., Suresh, H.S., Tan, S., Thomas, S.C., Thomas, D.W., Thompson, J., Turner, B.L., Uriarte, M., Valencia, R., Vallejo, M.I., Vicentini, A., Vrška, T., Wang, X., Wang, X., Weiblen, G., Wolf, A., Xu, H., Yap, S., Zimmerman, J., 2014. CTFS-ForestGEO: a worldwide network monitoring forests in an era of global change. Global Change Biology 21, 528–549. doi:10.1111/gcb.12712

Bell, D.M., Ward, E.J., Oishi, A.C., Oren, R., Flikkema, P.G., Clark, J.S., 2015. A state-space modeling approach to estimating canopy conductance and associated uncertainties from sap flux density data. Tree Physiology 35, 792–802. doi:10.1093/treephys/tpv041

Borer, E.T., Seabloom, E.W., Jones, M.B., Schildhauer, M., 2009. Some Simple Guidelines for Effective Data Management. The Bulletin of the Ecological Society of America 90, 205–214. doi:10.1890/0012-9623-90.2.205

Christoffersen, B.O., Gloor, M., Fauset, S., Fyllas, N.M., Galbraith, D.R., Baker, T.R., Kruijt, B., Rowland, L., Fisher, R.A., Binks, O.J., Sevanto, S., Xu, C., Jansen, S., Choat, B., Mencuccini, M., McDowell, N.G., Meir, P., 2016. Linking hydraulic traits to tropical forest function in a size-structured and trait-driven model (TFS v.1-Hydro). Geoscientific Model Development 9, 4227–4255. doi:10.5194/gmd-9-4227-2016

Condit, R., Lao, S., Singh, A., Esufali, S., Dolins, S., 2014. Data and database standards for permanent forest plots in a global network. Forest Ecology and Management 316, 21–31. doi:10.1016/j.foreco.2013.09.011

CREAF (Centre for Research on Ecology and Forestry Applications) 2016, https://github.com/sapfluxnet/sapfluxnet-public/wiki, accessed June 22, 2016.

Datacite 2016. https://www.datacite.org/, accessed July 2016.

Detto, M., Molini, A., Katul, G., Stoy, P., Palmroth, S., Baldocchi, D., 2012. Causality and Persistence in Ecological Systems: A Nonparametric Spectral Granger Causality Approach. The American Naturalist 179, 524–535. doi:10.1086/664628

Dietze, M.C., Vargas, R., Richardson, A.D., Stoy, P.C., Barr, A.G., Anderson, R.S., Arain, M.A., Baker, I.T., Black, T.A., Chen, J.M., Ciais, P., Flanagan, L.B., Gough, C.M., Grant, R.F., Hollinger, D., Izaurralde, R.C., Kucharik, C.J., Lafleur, P., Liu, S., Lokupitiya, E., Luo, Y., Munger, J.W., Peng, C., Poulter, B., Price, D.T., Ricciuto, D.M., Riley, W.J., Sahoo, A.K., Schaefer, K., Suyker, A.E., Tian, H., Tonitto, C., Verbeeck, H., Verma, S.B., Wang, W., Weng, E., 2011. Characterizing the performance of ecosystem models across time scales: A spectral analysis of the North American Carbon Program site-level synthesis. J. Geophys. Res. Biogeosci. 116, G04029. doi:10.1029/2011JG001661

Dietze, M.C., Lebauer , D.S., Kooprt, R., 2013. On improving the communication between models and data. Plant Cell and Environment 36, 1575–1585. doi:10.1111/pce.12043

EML (Ecological Metadata Language) Project 2009, EML version 2.1.1, https://knb.ecoinformatics.org/#external//emlparser/docs/index.html, accessed 19 June 2016.

Federal Geographic Data Committee. FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee. Washington, D.C.

Fisher, R.A., Muszala, S., Verteinstein, M., Lawrence, P., Xu, C., McDowell, N.G., Knox, R.G., Koven, C., Holm, J., Rogers, B.M., Spessa, A., Lawrence, D., Bonan, G., 2015. Taking off the training wheels: the properties of a dynamic vegetation model without climate envelopes, CLM4.5(ED). Geoscientific Model Development 8, 3593–3619. doi:10.5194/gmd-8-3593-2015

FLUXNET 2016. http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/, accessed 20 December 2016.

Goldstein, G., Andrade, J.L., Meinzer, F.C., Holbrook, N.M., Cavelier, J., Jackson, P., Celis, A., 1998. Stem water storage and diurnal patterns of water use in tropical forest canopy trees. Plant, Cell & Environment 21, 397–406. doi:10.1046/j.1365-3040.1998.00273.x

Hook, L.A., Santhana-Vannen, S., Beaty, T.W., Cook, R.B., 2010. Best Practices for Preparing Environmental Data Sets to Share and Archive, Oak Ridge National Laboratory Distributed Active Archive Center.

Horsburgh, J.S., Aufdenkampe, A.K., Mayorga, E., Lehnert, K.A., Hsu, L., Song, L., Jones, A.S., Damiano, S.G., Tarboton, D.G., Valentine, D., Zaslavsky, I., Whitenack, T., 2016. Observations Data Model 2: A community information model for spatially discrete Earth observations. Environmental Modelling and Software 79, 55–74. doi:10.1016/j.envsoft.2016.01.010

Hunter, M.O., Keller, M., Morton, D., Cook, B., Lefsky, M., Ducey, M., Saleska, S., de Oliveira, R.C., Schietti, J., 2015. Structural Dynamics of Tropical Moist Forest Gaps. PLoS ONE 10, e0132144. doi:10.1371/journal.pone.0132144

ISCN (International Soil Carbon Network) 2016. http://iscn.fluxdata.org/data/dataset-information/, accessed April 18, 2016.

ISO (International Organization for Standards) 19156:2011 Geographic information – Observations and measurements. 2011. doi:10.13140/2.1.1142.3042.

Jones, C., Blanchette, C., Brooke, M., Harris, J., Jones, M., Schildhauer, M., 2007. A metadata-driven framework for generating field data entry interfaces in ecology. Ecological Informatics 2, 270–278. doi:10.1016/j.ecoinf.2007.06.005

KBase (Department of Energy Systems Biology Knowledgebase), http://kbase.us, accessed July 2016.

Kervin, K., Michener, W., Cook, R., 2013. Common Errors in Ecological Data Sharing. JESLIB 1–15. doi:10.7191/jeslib.2013.1024

KNB (The Knowledge Network for Biocomplexity), https://knb.ecoinformatics.org, accessed April 2016.

Law, B.E., Arkebauer, T., Campbell, J.L., Chen, J., Sun, O., 2008. Terrestrial carbon observations: Protocols for vegetation sampling and data submission. FAO.

LeBauer, D., Dietze, M., Kooper, R., Long, S., Mulrooney, P., Rohde, G.S., Wang, D., 2010. Biofuel Ecophysiological Traits and Yields Database (BETYdb), Energy Biosciences Institute, University of Illinois at Urbana-Champaign. doi:10.13012/J8H41PB9

Malhi, Y., Phillips, O.L., Lloyd, J., Baker, T., Wright, J., Almeida, S., Arroyo, L., Frederiksen, T., Grace, J., Higuchi, N., Killeen, T., Laurance, W.F., Leaño, C., Lewis, S., Meir, P., Monteagudo, A., Neill, D., Núñez Vargas, P., Panfil, S.N., Patiño, S., Pitman, N., Quesada, C.A., Rudas Ll, A., Salomão, R., Saleska, S., Silva, N., Silveira, M., Sombroek, W.G., Valencia, R., Vásquez Martínez, R., Vieira, I.C.G., Vinceti, B., 2002. An international network to monitor the structure, composition and dynamics of Amazonian forests (RAINFOR). Journal of Vegetation Science 13, 439–450. doi:10.1111/j.1654-1103.2002.tb02068.x

Marthews, T.R., Riutta, T., Oliveras Menor, I., Urrutia, R., Moore, S., Metcalfe, D., Malhi, Y., Phillips, O., Huaraca Huasco, W., Ruiz Jaén, M., Girardin, C., Butt, N., Cain, R., and colleagues from the RAINFOR and GEM networks, 2014. Measuring Tropical Forest Carbon Allocation and Cycling: A RAINFOR-GEM Field Manual for Intensive Census Plots (v3.0). Manual, Global Ecosystems Monitoring network, http://gem.tropicalforests.ox.ac.uk/.

McIntosh, A.C.S., Cushing, J.B., Nadkarni, N.M., Zeman, L., 2007. Database design for ecologists: Composing core entities with observations. Ecological Informatics 2, 224–236. doi:10.1016/j.ecoinf.2007.07.003

Medlyn, B.E., Robinson, A.P., Clement, R., McMurtrie, R.E., 2005. On the validation of models of forest CO2 exchange using eddy covariance data: some perils and pitfalls. Tree Physiology 25, 839–857. doi:10.1093/treephys/25.7.839

Meinzer, F.C., Goldstein, G., Andrade, J.L., 2001. Regulation of water flux through tropical forest canopy trees: Do universal rules apply? Tree Physiology 21, 19–26. doi:10.1093/treephys/21.1.19

Meinzer, F.C., James, S.A., Goldstein, G., 2004. Dynamics of transpiration, sap flow and use of stored water in tropical forest canopy trees. Tree Physiology 24, 901–909. doi:10.1093/treephys/24.8.901

Meinzer, F.C., Bond, B.J., Warren, J.M., Woodruff, D.R., 2005. Does water transport scale universally with tree size? Funct Ecology 19, 558–565. doi:10.1111/j.1365-2435.2005.01017.x

Medvigy, D., Wofsy, S.C., Munger, J.W., Hollinger, D.Y., Moorcroft, P.R., 2009. Mechanistic scaling of ecosystem function and dynamics in space and time: Ecosystem Demography model version 2. J. Geophys. Res. 114, G01002. doi:10.1029/2008JG000812

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., Stafford, S.G., 1997. Nongeospatial Metadata for the Ecological Sciences. Ecological Applications 7, 330–342. doi:10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2

Michener, W.K., 2006. Meta-information concepts for ecological data management. Ecological Informatics 1, 3–7. doi:10.1016/j.ecoinf.2005.08.004

Michener, W.K., 2015. Ecological data sharing. Ecological Informatics 29, 33–44. doi:10.1016/j.ecoinf.2015.06.010

Moorcroft, P.R., Hurtt, G.C., Pacala, S.W., 2001. A Method for Scaling Vegetation Dynamics: The Ecosystem Demography Model (ED). Ecological Monographs 71, 557–585. doi:10.2307/3100036

NCEAS 2015. Morpho 1.11.0 User Guide, https://knb.ecoinformatics.org/software/dist/MorphoUserGuide.pdf, accessed 13 April 2016.

NGEE Tropics 2016. http://eesa.lbl.gov/ngee-tropics/, accessed 20 December 2016.

NOAA NCEI (National Centers for Environmental Information), https://www.nodc.noaa.gov, accessed November 2016.

OGC 2013. Open Geospatial Consortium (OGC) Observations and Measurements v2.0 OGC Document 10-004r1 http://www.opengis.net/doc/AS/OM/2.0 (also published as ISO/DIS 19156:2010, Geographic information — Observations and Measurements)

Papale, D., Agarwal, D.A., Baldocchi, D., Cook, R.B., Fisher, J.B., van Ingen, C., 2012. Database Maintenance, Data Sharing Policy, Collaboration, in: Aubinet, M., Vesala, T., Papale, D. (Eds.), Eddy Covariance: a Practical Guide to Measurement and Data Analysis. Springer Netherlands, Dordrecht, pp. 399–424. doi:10.1007/978-94-007-2351-1

Peacock, J., Baker, T.R., Lewis, S.L., Lopez Gonzalez, G., Phillips, O.L., 2007. The RAINFOR database: monitoring forest biomass and dynamics. Journal of Vegetation Science 18, 535–542. doi:10.1111/j.1654-1103.2007.tb02568.x

Peng, G., Ritchey, N.A., Casey, K.S., Kearns, E.J., Privette, J.L., 2016. Scientific stewardship in the Open Data and Big Data era—Roles and responsibilities of stewards and other major product stakeholders. D-Lib Magazine 13, 1–25. doi:10.1080/02757259509532294

Powell, T.L., Galbraith, D.R., Christoffersen, B.O., Harper, A., Imbuzeiro, H.M.A., Rowland, L., Almeida, S., Brando, P.M., da Costa, A.C.L., Costa, M.H., Levine, N.M., Malhi, Y., Saleska, S.R., Sotta, E., Williams, M., Meir, P., Moorcroft, P.R., 2013. Confronting model predictions of carbon fluxes with measurements of Amazon forests subjected to experimental drought. New Phytol 200, 350–365. doi:10.1111/nph.12390

Poyatos, R., Granda, V., Molowny-Horas, R., Mencuccini, M., Steppe, K., Martínez-Vilalta, J., 2016. SAPFLUXNET: towards a global database of sap flow measurements. Tree Physiology 36, 1449–1455. doi:10.1093/treephys/tpw110

RAINFOR (Amazon Forest Inventory Network), Liana and Canopy Index Protocol, http://www.rainfor.org/upload/ManualsEnglish/crown%20liana%20protocols_Sep%202014_EN.pdf, accessed March 2014.

Ramakrishnan, L., Poon, S., Hendrix, V., Gunter, D., Pastorello, G.Z., Agarwal, D., 2014. Experiences with User-Centered Design for the Tigres Workflow API, Presented at the 2014 IEEE 10th International Conference on e-Science (e-Science), IEEE, pp. 290–297. doi:10.1109/eScience.2014.56

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., Frame, M., 2011. Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6, e21101–21. doi:10.1371/journal.pone.0021101

Unidata 2016. https://www.unidata.ucar.edu/software/netcdf/, accessed Mar 2016.

USGS Science Data Catalog, http://data.usgs.gov/, accessed November 2016.

Walker, A.P., Hanson, P.J., De Kauwe, M.G., Medlyn, B.E., Zaehle, S., Asao, S., Dietze, M.C., Hickler, T., Huntingford, C., Iversen, C.M., Jain, A.K., Lomas, M., Luo, Y., McCarthy, H.R., Parton, W.J., Prentice, I.C., Thornton, P.E., Wang, S., Wårlind, D., Weng, E., Warren, J.M., Woodward, F.I., Oren, R., Norby, R.J., 2014. Comprehensive ecosystem model-data synthesis using multiple data sets at two temperate forest free-air $CO_2$ enrichment experiments: Model performance at ambient $CO_2$ concentration. Journal of Geophysical Research: Biogeosciences 119, 937–964. doi:10.1002/(ISSN)2169-8961

Zurb 2016. http://foundation.zurb.com/, accessed October 2016.