Cyberinfrastructure: Year 2 Advances

Jeffery S. Horsburgh

Amber Spackman Jones, Anthony Castronova
And the rest of the CI Team

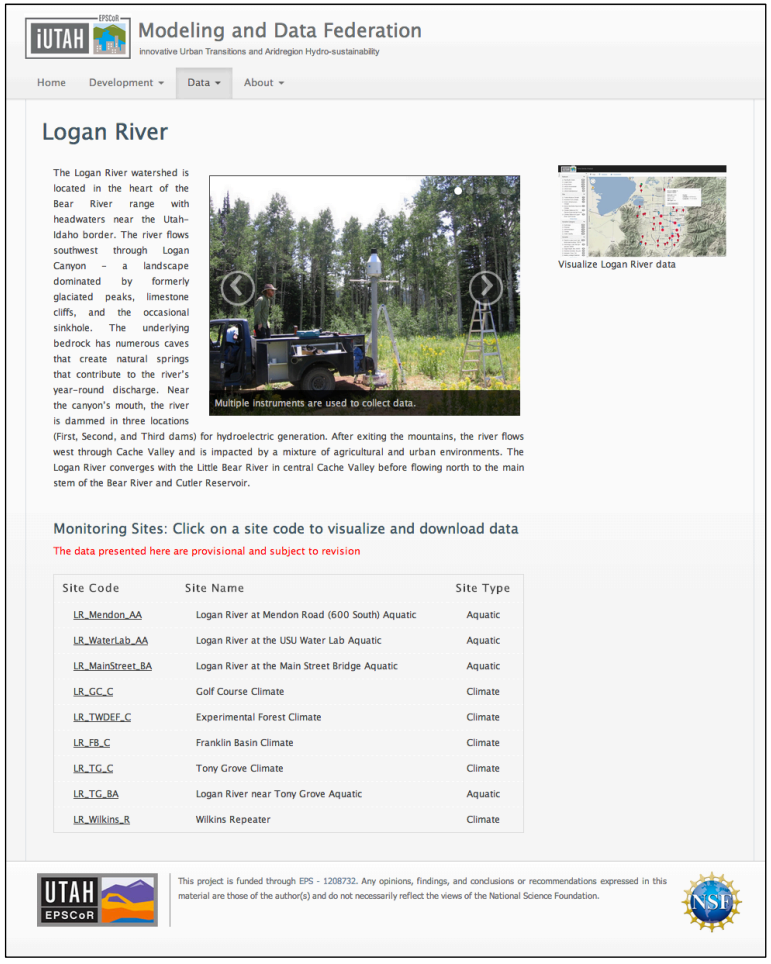
iUTAH Symposium and All Hands Meeting
7-11-2014

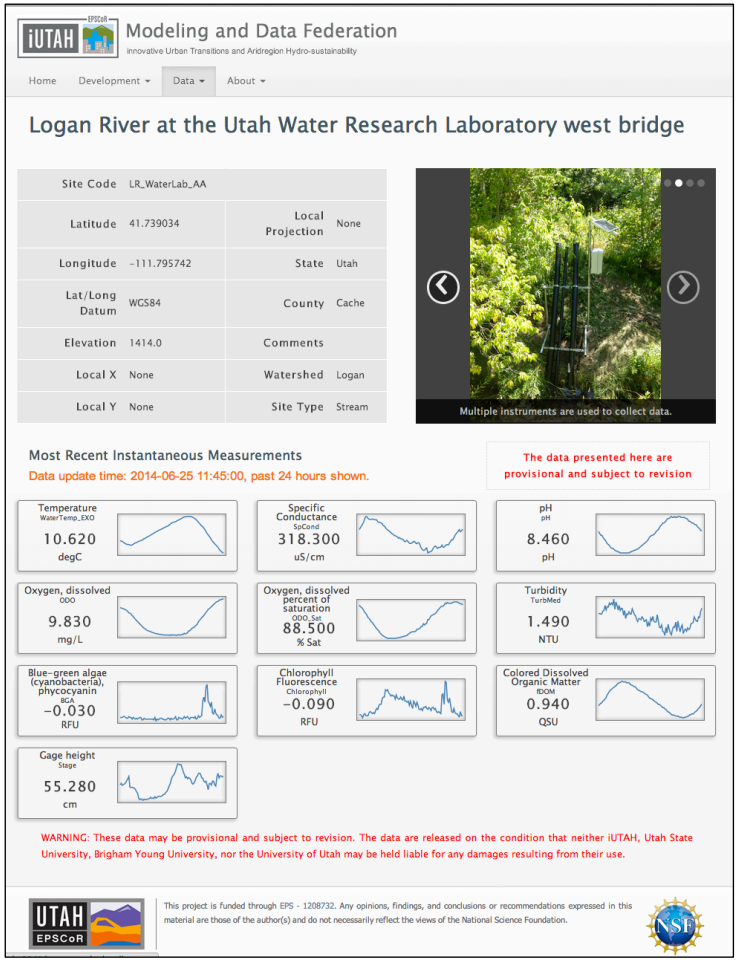
Overview

- Improved access to GAMUT Data
  - Demonstration of dynamic site pages at http://data.iutahepscor.org
  - Demonstration of web-based time series data visualization at http://data.iutahepscor.org/tsa
- iUTAH Data Policy
  (available at http://data.iutahepscor.org or http://iutahepscor.org)
  - Data Typology
  - Data Collection Plans
- Data Publication System
  - Newly released
  - Demonstration of publication and search at http://repository.iutahepscor.org
- Cyberinfrastructure for Modeling

Access to GAMUT Data

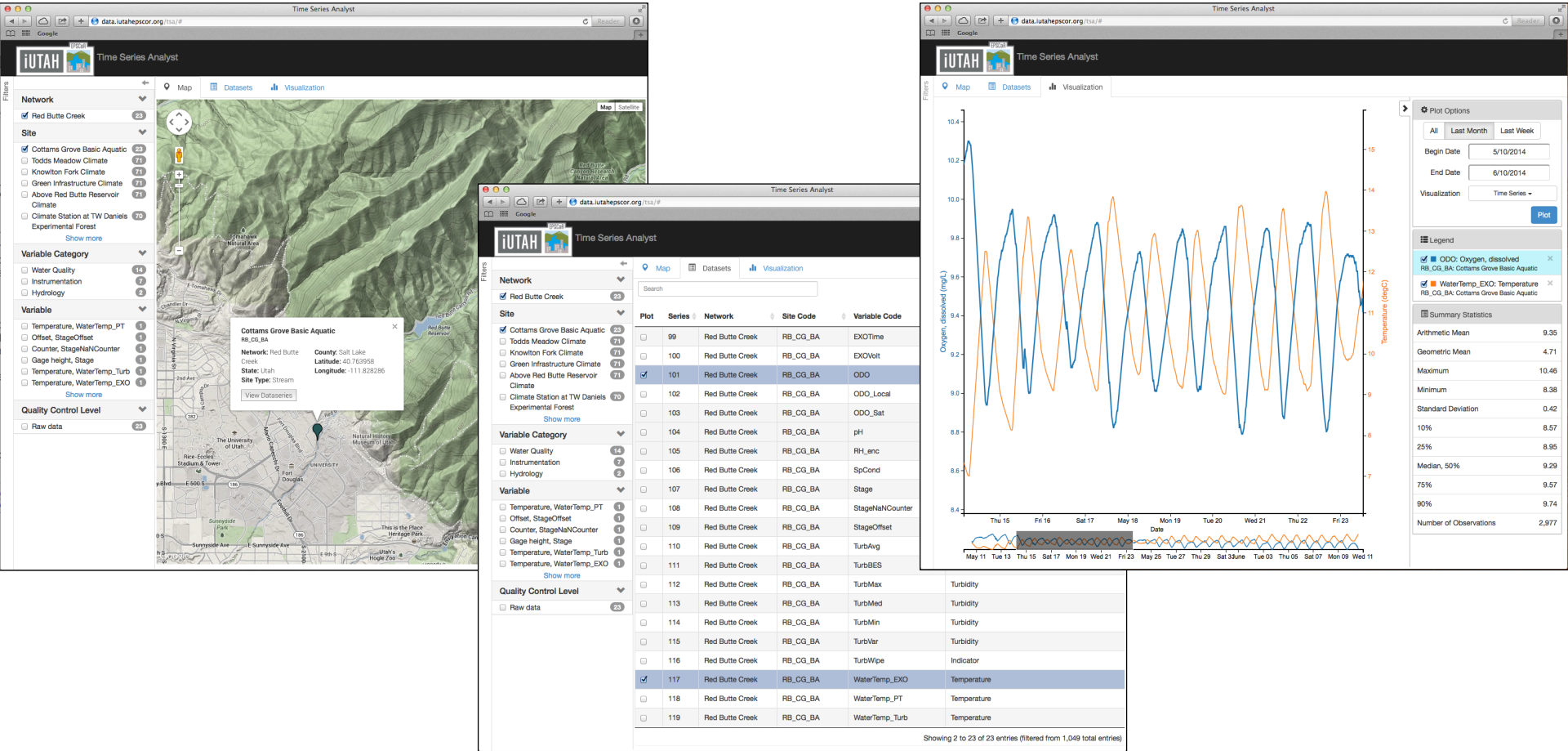
Dynamic site pages for each GAMUT site at http://gamut.iutahepscor.org with previews and links to data.





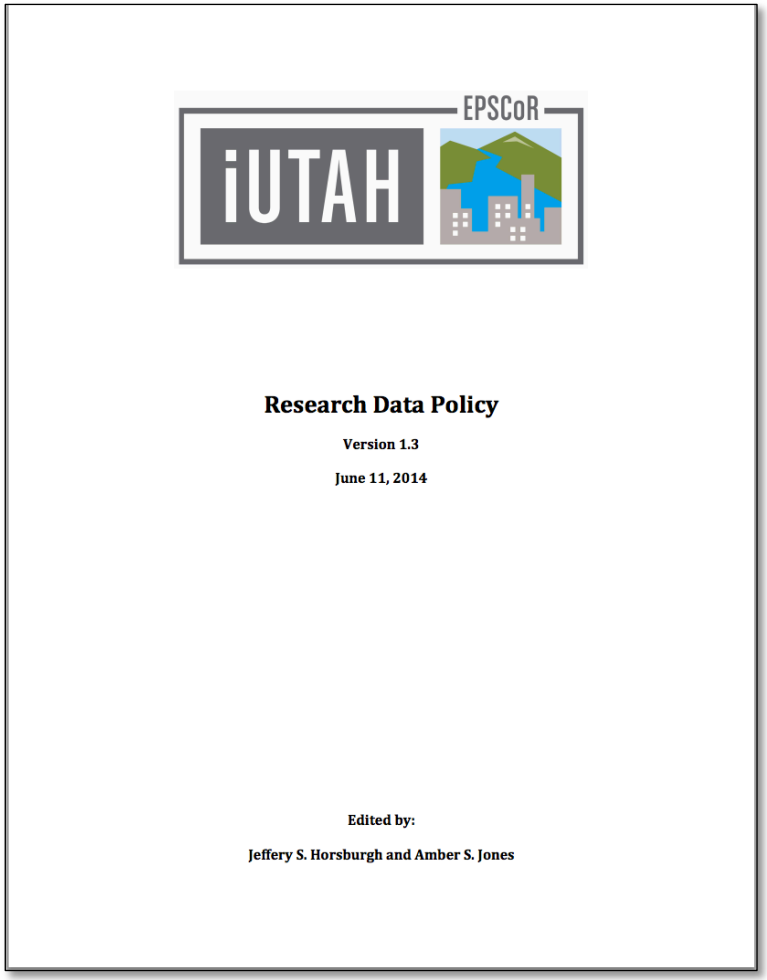
Access to GAMUT Data

- Time Series Analyst: Enhanced web access and visualization of time series data at http://data.iutahepscor.org/tsa
- Features map-based interface, faceted searching, and multiple plot types



iUTAH Data Policy

- **Applies to all datasets created or developed with support from iUTAH**
- Recommended reading for all investigators. Available at: http://iutahepscor.org/data_modeling/data_management_Policy.html
- In general, open data policy to maximize the impact and broad use of datasets collected by iUTAH research teams
- Specifies **procedures** and **timeframes** for publishing and sharing data
- Developed with input by Data Policy Committee, Leadership Team, and Management Team
- Researchers expected to adhere to policies, provide high quality datasets with sufficient metadata
- Researchers should have expectation of first rights to analyze and publish data



iUTAH Data Policy: Data Typology

Timeframes

A. Primary iUTAH datasets and research products (e.g., raw and QAQC sensor data, baseline sampling datasets, general iUTAH datasets for the iUTAH community).

→ Published as soon as results are created

B. Support from iUTAH provided, but created by a specific investigator or group to support particular research question/goal.

→ Finalized data submitted within one year of completion of data collection activities

C. Types A and B but subject to IRB restrictions.

→ Same timeframes as A and B, but may require anonymization

D. Proprietary data that may be subject to licensing, copyright, other restrictions.

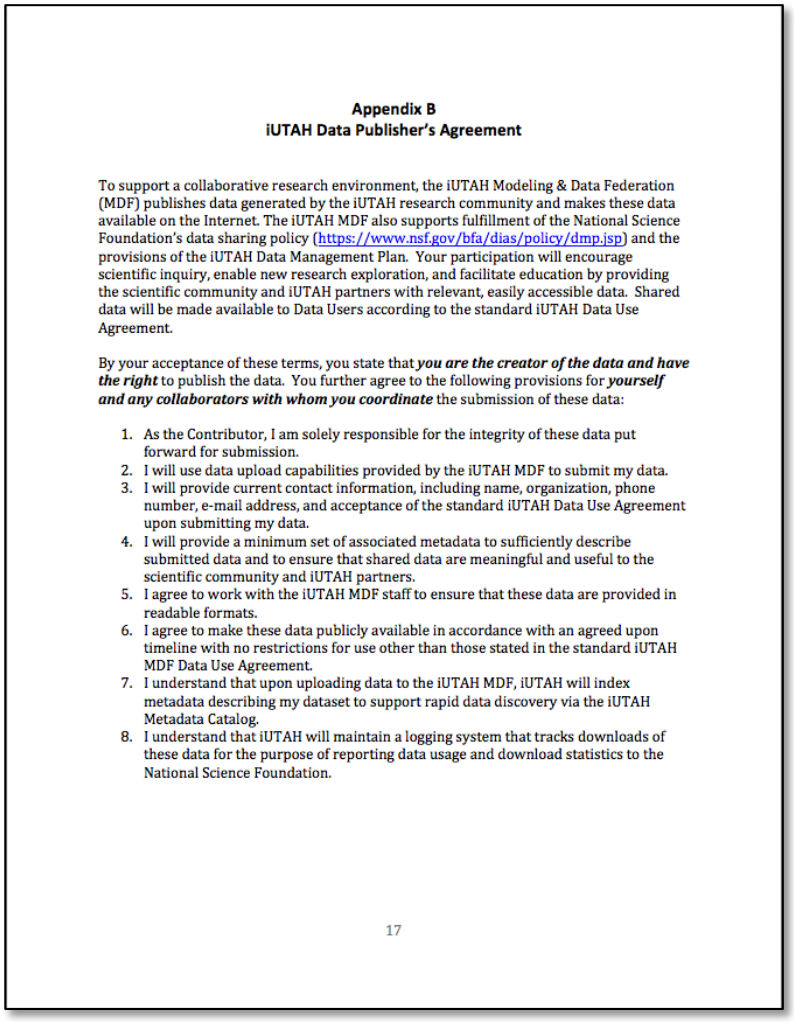
NOTE: For all datasets, a metadata record must be submitted within one month of beginning data collection. All submissions will be reviewed and approved by the CI Team.

iUTAH Data Policy: Data Collection Plans

- **ALL** data creation efforts with **ANY** funding from iUTAH (salary, travel, sampling, equipment, etc.) **MUST** submit a brief plan to the Data Policy Committee **PRIOR** to funding.
- Example plans/templates are available
- Policy is retroactive → plans need to be submitted for current efforts
- Plan should include:
  1. Identification of types of data to be collected/created
  2. Brief description of methods, data formats, and data products
  3. Identification of who will have access to preliminary data during collection
  4. How data products will be made available
  5. Information on collaborators/co-authors of data products or publications

iUTAH Data Policy: Data Publishers Agreement

- You assert that you are the creator of the data
- You are responsible for your data
- You agree to submit your metadata and data
- You agree to let iUTAH publish the data



Appendix B
iUTAH Data Publisher's Agreement

To support a collaborative research environment, the iUTAH Modeling & Data Federation (MDF) publishes data generated by the iUTAH research community and makes these data available on the Internet. The iUTAH MDF also supports fulfillment of the National Science Foundation's data sharing policy (https://www.nsf.gov/bfa/dias/policy/dmp.jsp) and the provisions of the iUTAH Data Management Plan. Your participation will encourage scientific inquiry, enable new research exploration, and facilitate education by providing the scientific community and iUTAH partners with relevant, easily accessible data. Shared data will be made available to Data Users according to the standard iUTAH Data Use Agreement.

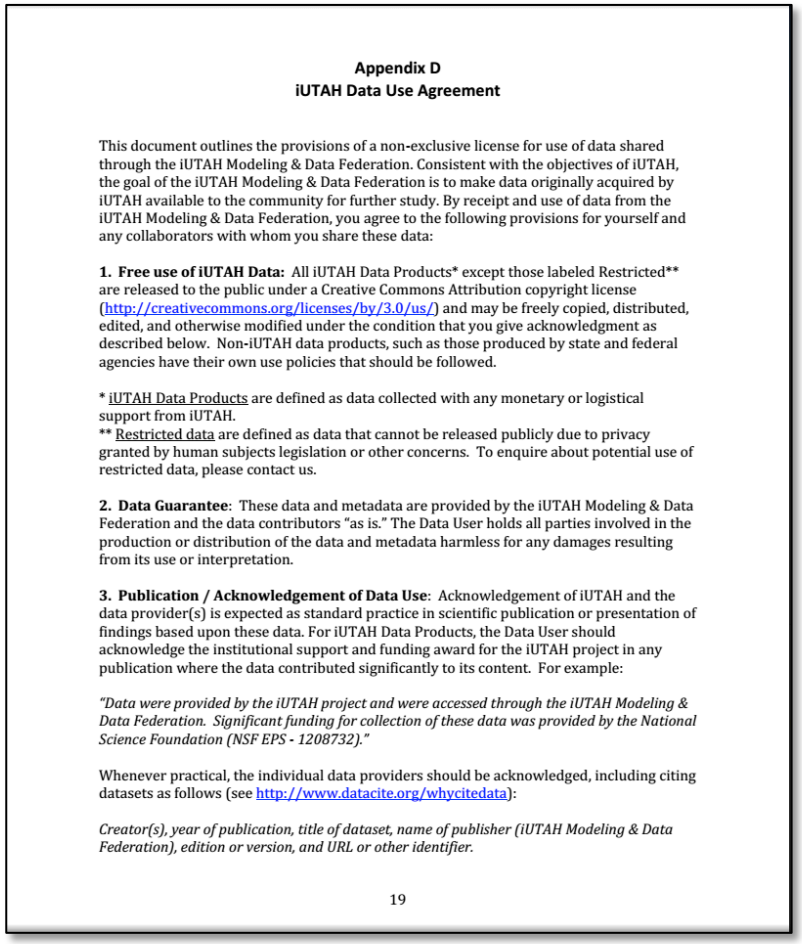
By your acceptance of these terms, you state that ***you are the creator of the data and have the right*** to publish the data. You further agree to the following provisions for ***yourself and any collaborators with whom you coordinate*** the submission of these data:

1. As the Contributor, I am solely responsible for the integrity of these data put forward for submission.
2. I will use data upload capabilities provided by the iUTAH MDF to submit my data.
3. I will provide current contact information, including name, organization, phone number, e-mail address, and acceptance of the standard iUTAH Data Use Agreement upon submitting my data.
4. I will provide a minimum set of associated metadata to sufficiently describe submitted data and to ensure that shared data are meaningful and useful to the scientific community and iUTAH partners.
5. I agree to work with the iUTAH MDF staff to ensure that these data are provided in readable formats.
6. I agree to make these data publicly available in accordance with an agreed upon timeline with no restrictions for use other than those stated in the standard iUTAH MDF Data Use Agreement.
7. I understand that upon uploading data to the iUTAH MDF, iUTAH will index metadata describing my dataset to support rapid data discovery via the iUTAH Metadata Catalog.
8. I understand that iUTAH will maintain a logging system that tracks downloads of these data for the purpose of reporting data usage and download statistics to the National Science Foundation.

17

iUTAH Data Policy: Data Use Agreement

- Free use of iUTAH data
- Data are provided 'as is'
- Users should acknowledge iUTAH support
- Data may be re-distributed
- Collaboration with original creators encouraged



Appendix D
iUTAH Data Use Agreement

This document outlines the provisions of a non-exclusive license for use of data shared through the iUTAH Modeling & Data Federation. Consistent with the objectives of iUTAH, the goal of the iUTAH Modeling & Data Federation is to make data originally acquired by iUTAH available to the community for further study. By receipt and use of data from the iUTAH Modeling & Data Federation, you agree to the following provisions for yourself and any collaborators with whom you share these data:

1. Free use of iUTAH Data: All iUTAH Data Products* except those labeled Restricted** are released to the public under a Creative Commons Attribution copyright license (http://creativecommons.org/licenses/by/3.0/us/) and may be freely copied, distributed, edited, and otherwise modified under the condition that you give acknowledgment as described below. Non-iUTAH data products, such as those produced by state and federal agencies have their own use policies that should be followed.

* iUTAH Data Products are defined as data collected with any monetary or logistical support from iUTAH.
** Restricted data are defined as data that cannot be released publicly due to privacy granted by human subjects legislation or other concerns. To enquire about potential use of restricted data, please contact us.

2. Data Guarantee: These data and metadata are provided by the iUTAH Modeling & Data Federation and the data contributors "as is." The Data User holds all parties involved in the production or distribution of the data and metadata harmless for any damages resulting from its use or interpretation.

3. Publication / Acknowledgement of Data Use: Acknowledgement of iUTAH and the data provider(s) is expected as standard practice in scientific publication or presentation of findings based upon these data. For iUTAH Data Products, the Data User should acknowledge the institutional support and funding award for the iUTAH project in any publication where the data contributed significantly to its content. For example:

"Data were provided by the iUTAH project and were accessed through the iUTAH Modeling & Data Federation. Significant funding for collection of these data was provided by the National Science Foundation (NSF EPS - 1208732)."

Whenever practical, the individual data providers should be acknowledged, including citing datasets as follows (see http://www.datacite.org/whycitedata):

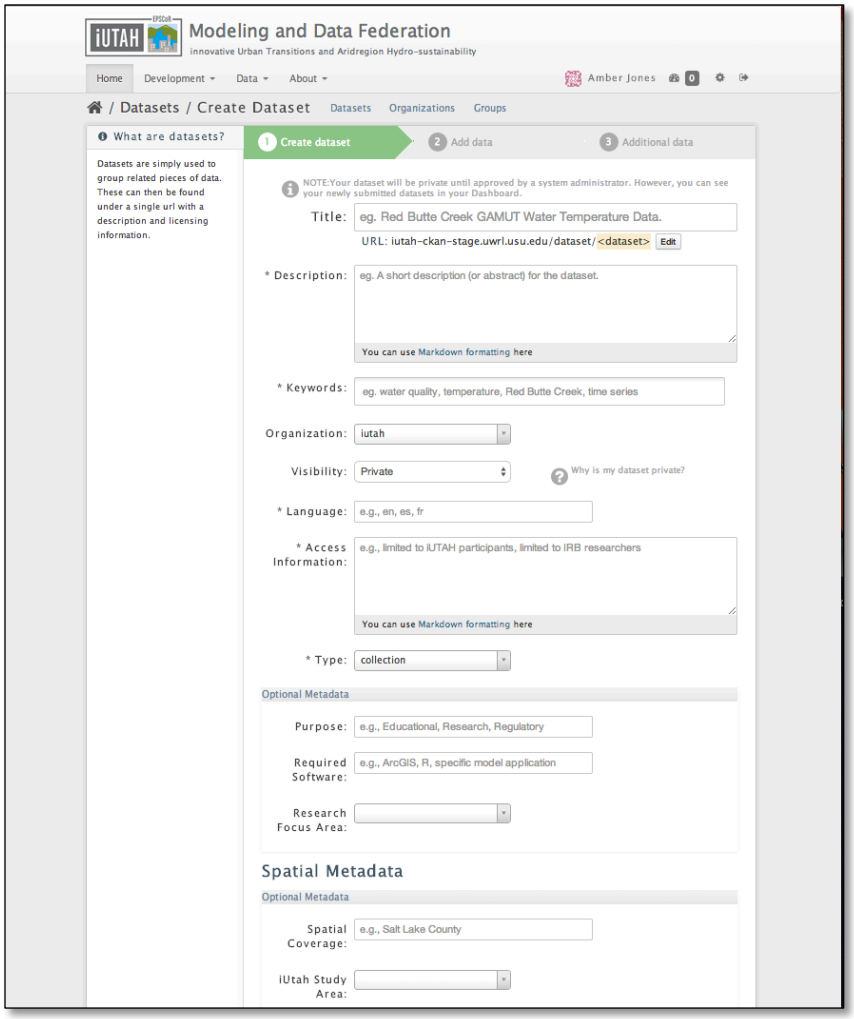
Creator(s), year of publication, title of dataset, name of publisher (iUTAH Modeling & Data Federation), edition or version, and URL or other identifier.

19

Data Publication System

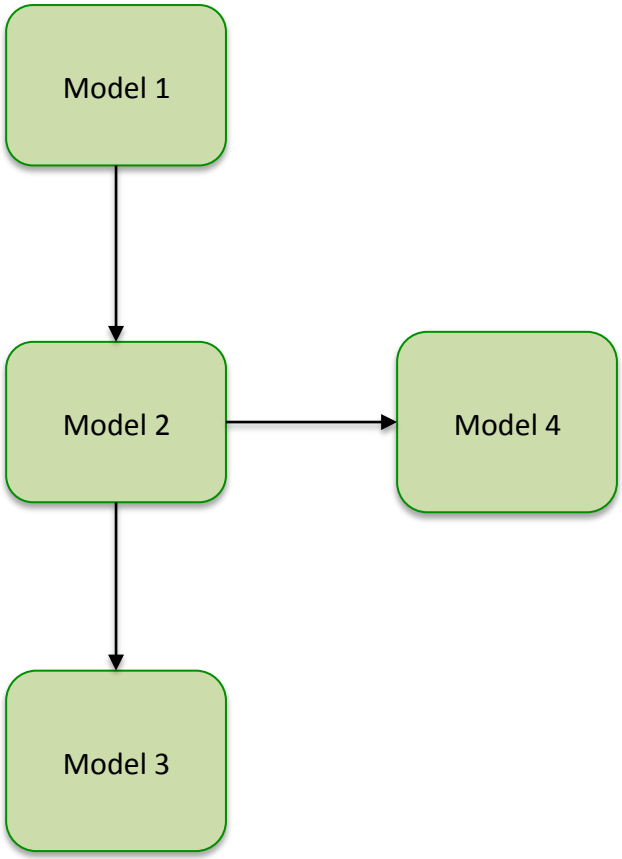
http://repository.iutahepscor.org

- Web-based system for iUTAH researchers to submit and publish data and models
- System supports curation of datasets
- Integrates the submission and presentation of data and metadata
- Supports discovery and access of datasets to a wide audience
- Supports storage and archival
- Datasets are private until approved by a moderator



Cyberinfrastructure for Coupled Modeling Design Considerations

- **Link independent models by their inputs and outputs**
  - Leverages OpenMI concepts
  - Feed-forward simulation
  - Feed-back simulation
- **Database centric**
  - Tight integration with database storage systems
  - Direct ingestion of observation data (GAMUT)
  - Simulation results storage and archival
- **Extended to use additional data sources**
  - Web services, e.g. USGS, NOAA
  - OpenDAP
- **Platform independent design**
  - Written in Python
  - Open source technology
  - Numerical libraries and solvers
- **Goals**
  - Quickly develop scientific model components
  - Seamlessly couple them with others
  - Archive results in an easily accessible and shareable manner

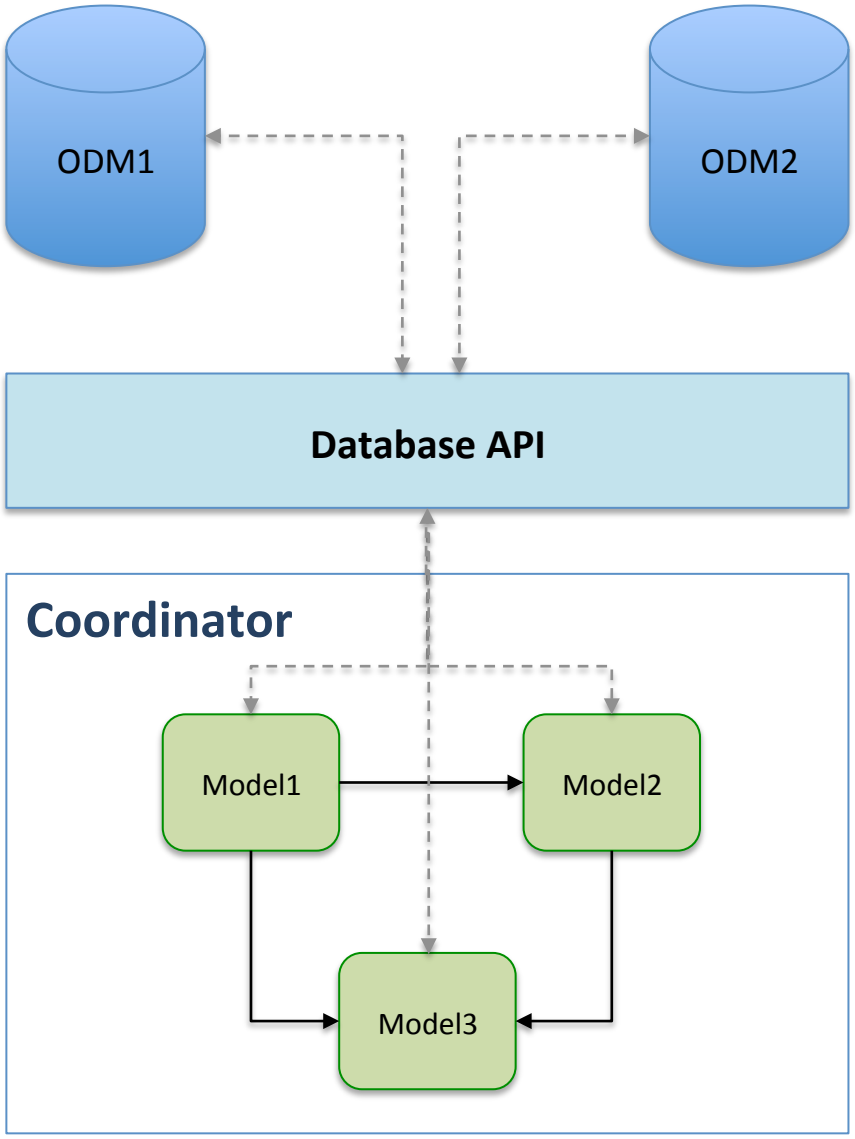


Generalized Software Architecture

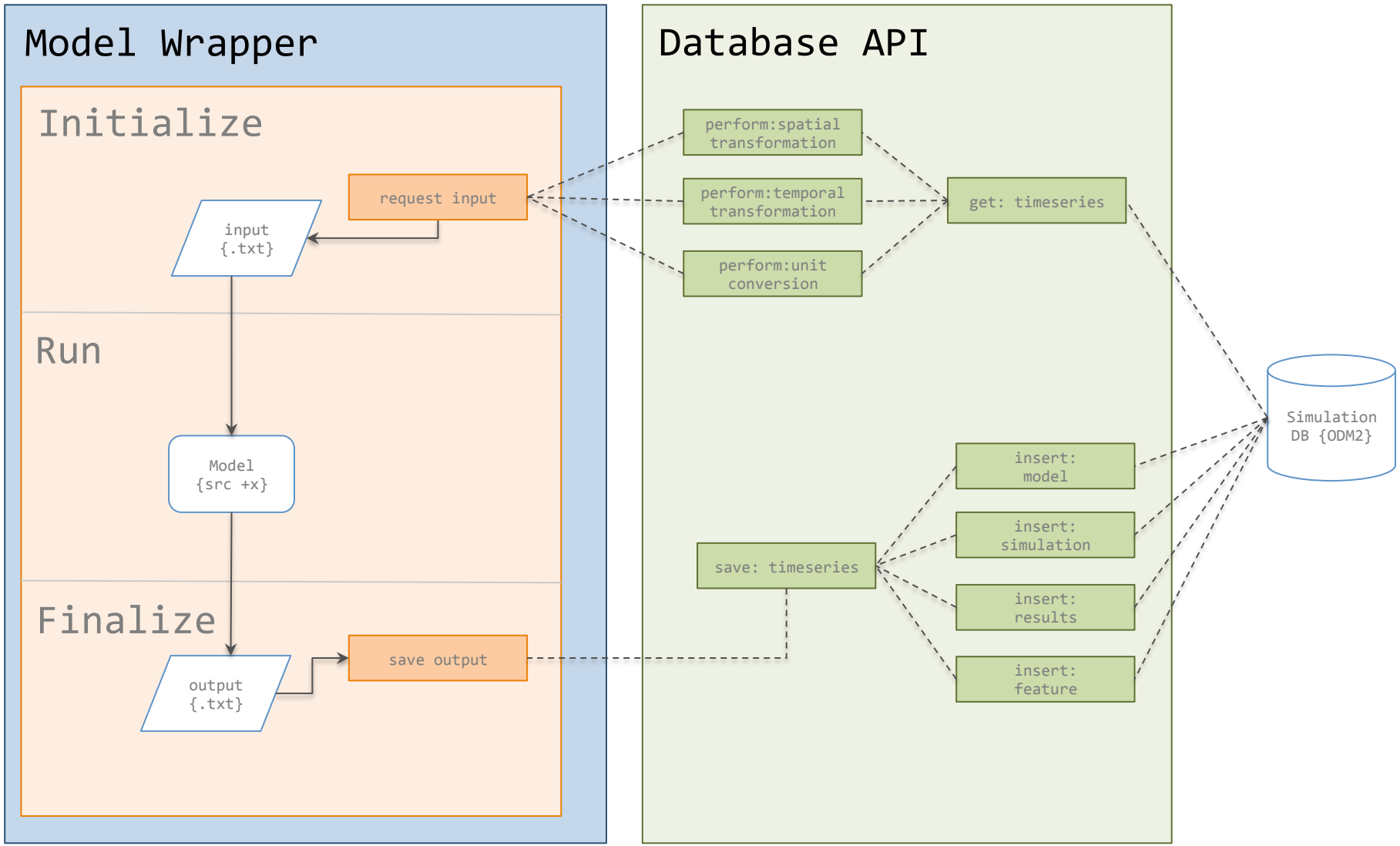
Database Storage
- GIS enabled database
- Local and/or remote connections
- Input forcings
- Output results

Application Programming Interface
- Database read/write abstraction
- Easy database interaction

Model Linking and Execution
- Model outputs are connected to inputs
- Models communicate with databases through API
- Models are simplified using a software wrapper
- All data transfer is handled by the coordinator
- Spatial, temporal, and unit conversions are handled by API
- Tracking of simulation result provenance



Basic Model Workflow



Cyberinfrastructure Team Immediate Next Steps

- **RFA1**:
  - Interpretive/graphical presentation of GAMUT infrastructure
  - Continued work on data QA/QC software.
- **RFA2**:
  - Interactive visualization of household survey results
- **RFA3**:
  - Continued development of coupled modeling infrastructure
  - Interactive visualization of the “Framework”
- **Hardware**:
  - Increasing capacity of storage infrastructure for virtualization
- **Hydroinformatics**:
  - We are offering the course again this fall!

Questions?

Jeff Horsburgh
jeff.horsburgh@usu.edu

Amber Jones
amber.jones@usu.edu