Database Integration Workshop: Building the Data Capacity for Food-Energy-Water Research

Workshop Report

September 11, 2018
Raleigh, North Carolina

Yuan Yao[1], Runze Huang[2], Richard Venditti[1], Kai Lan[1], Zhenzhen Zhang[3]
[1]Department of Forest Biomaterials, North Carolina State University, Raleigh, NC
[2]ExLattice, Inc. Raleigh, NC
[3]Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC

Funded by USDA-NIFA, Award Number 2018-67021-27696

Table of Contents

Executive Summary 2
Participants 3
Workshop Agenda 4
Session 1 Report – Governmental Data Capacity & Vision 6
Session 2 Report – Data Capacity Mapping & Vision 9
Current Data Capacity and Sources for FEW Research 9
Gaps and Data Needs 16
The Vision of Database Integration and Data Sharing 18
Session 3 Report – Data Integration and Synthesis 20
Session 4 Report – Challenges and Barriers 22
Session 5 Report – Action Plan and Roadmapping 23

Please cite this report as:
Yuan Yao, Runze Huang, Richard Venditti, Kai Lan, Zhenzhen Zhang, Database Integration Workshop: Building the Data Capacity for Food-Energy-Water Research, North Carolina State University, Raleigh, NC, 2019. URL https://faculty.cnr.ncsu.edu/yuanyao/database-integration-workshop-building-the-data-capacity-for-food-energy-water-research/.

Executive Summary

Energy and water are two critical resources for food production. Developing sustainable agricultural systems requires wise balanced management of food, energy, and water systems (FEW). Intensive efforts have been made by the research community, government agencies, and industry to generate data to meet the needs of various stakeholders, but such data are highly scattered and often used separately. To improve the system-wide sustainability of agricultural systems along with their interactions with water and energy supplies, answers to the following questions are critical:

- What are the frontiers of data from both public and private sources related to food, energy, and water systems?
- How can we leverage and integrate existing U.S. government-wide databases for new insights?
- Who should be involved and how can we encourage data generating, sharing, and engagement from a broad range of stakeholders in government, academia, and industry?

The Database Integration Workshop: Building the Data Capacity for Food-Energy-Water Research was held on September 11, 2018 at North Carolina State University. The objective of the workshop was to provide a clear vision for better measuring, understanding, and promoting system-wide sustainability of FEW systems through large-scale data sharing and engagement. The workshop gathered researchers, analysts, and program leaders from universities, research institutes, national labs, and government agencies with expertise in FEW systems and/or data management. The workshop included group discussions and keynote presentations (Sessions 1 and 3). The outcomes of the workshop include:

- A comprehensive list of data resources available and related to FEW, top gaps and data needs the participants identified and voted for, and the vision of database integration and data sharing (Session 2).
- Top challenges and knowledge gaps workshop participants identified and ranked (Session 4).
- Action plans with short-term and long-term goals to address the top challenges (Session 5).

Participants

Organizing Committee

Yuan Yao (Chair)	North Carolina State University
Runze Huang	ExLattice, Inc.
Richard Venditti	North Carolina State University

Keynote Speakers

James P. Dobrowolski	USDA National Institute for Food and Agriculture
Patrick Canning	USDA, Economic Research Service
Dwane Young	U.S. Environmental Protection Agency
Casey Burleyson	Pacific Northwest National Laboratory
Eric Masanet	Northwestern University
Chandra Krintz	University of California, Santa Barbara
Gale Boyd	Duke University, SSRI
Yuan Yao	North Carolina State University

Participants

Alberta Carpenter	National Renewable Energy Laboratory
Alicia Lindauer	U.S. Department of Energy
Antony Williams	U.S. Environmental Protection Agency
Deepti Salvi	North Carolina State University
Elizabeth Nichols	North Carolina State University
Erika Mack	USDA Forest Service
Hongmei Gu	U.S. Forest Service, Forest Products Laboratory
Inna Kouper	Indiana University
Jennifer Dunn	Northwestern-Argonne Institute for Science and Engineering
Jordan Kern	North Carolina State University
Kemafor Ogan	North Carolina State University
KP Sandeep	North Carolina State University
Lingjuan Wang Li	North Carolina State University
May Wu	Argonne National Laboratory
Natalie Nelson	North Carolina State University
Melanie Hedgespeth	North Carolina State University
Michael Wang	Argonne National Laboratory
Rachel Emerson	Idaho National Laboratory
Raju Vatsavai	North Carolina State University
Robert Beach	RTI International
Sakineh Tavakkoli	Johns Hopkins University
Sarah Rehkamp	USDA, Economic Research Service
Vincent Tidwell	Sandia National Laboratories
Kai Lan	North Carolina State University
Zhenzhen Zhang	North Carolina State University

Workshop Agenda

The workshop agenda was adjusted and condensed from 1.5 days to 1 day because of Hurricane Florence.

Tuesday, Sept 11th, Room 3210, Talley Student Union, NCSU Campus

8:00 am	**Registration and Networking**
8:30 am	**Workshop Overview** Dr. Yuan Yao, NC State University
9:00 am	**Session 1**: **Governmental Data Capacity & Vision** Food and Agriculture Cyberinformatics and Tools NIFA's Initiative for Data Science in Agriculture *Dr. James Dobrowolski, National Program Leader, USDA, National Institute of Food and Agriculture* An Overview of FEDS: The Food Environment Data System *Dr. Patrick Canning, Senior Economist, USDA, Economic Research Service* Water Data Integration *Dwane Young, Chief Water Data Integration Branch, EPA Office of Water* Bridge Building at 70 mph: Data Management in an Active DOE Office of Science Project *Dr. Casey Burleyson, Data Sciences Scientist, Pacific Northwest National Laboratory*
10:45 am	Break
11:00 am	**Session 2: Breakout Group Discussion** Key Questions: • What is the current data capacity? • What are overlapping/gap areas among different databases and data sources? • What is the vision of future integrated database and data sharing?
12:00 pm	**Working Lunch & Networking**
1:00 pm	**Session 3: Data Integration and Synthesis** National Energy Statistics: Opportunities and Challenges for Food-Energy-Water (FEW) Data Integration *Dr. Eric Masanet, Associate Professor, Northwestern University, previous Head of the Energy Demand Technology Unit, International Energy Agency*

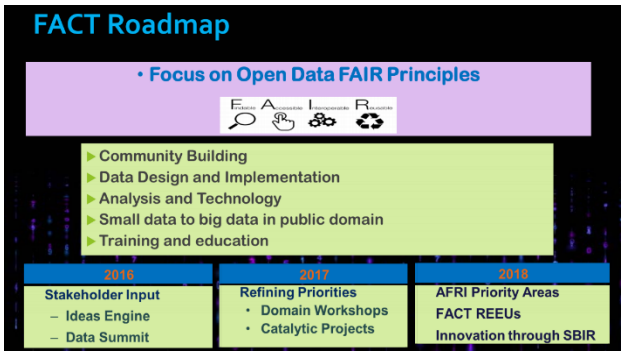
	SmartFarm: Data Integration & Systems for Agriculture-based, FEW Research *Dr. Chandra Krintz, Professor, Computer Science, UC Santa Barbara; SmartFarm and RACELab Director* Federal Statistical Research Data Centers *Dr. Gale Boyd, Director of the Triangle Research Data Center, Duke University* Promoting Bioeconomy for Sustainable Food-Energy-Water Systems: The Need of Interdisciplinary Research from a Data Point of View *Dr. Yuan Yao, Assistant Professor of Sustainability Science and Engineering, NC State University*
2:45 pm	**Break**
3:00 pm	**Session 4: Challenges and Barriers** Key Questions: • What are the challenges for database integration and data synthesis across different agencies or databases (e.g., technical, educational, and political/governmental challenges)?
3:45 pm	**Break and Assessable Participant-Identified Challenges**
4:00 pm	**Session 5: Action Plan and Road Mapping** Key questions: • What are short-term and long-term goals for government agencies/researchers/communities to pursue? • What are short-term and long-term efforts that you would like to propose to the entire FEW related communities?
5:30 pm	**Group Presentations**
6:00 pm	Dinner and Conclusion Remarks at Room 3222

Session 1 Report – Governmental Data Capacity & Vision

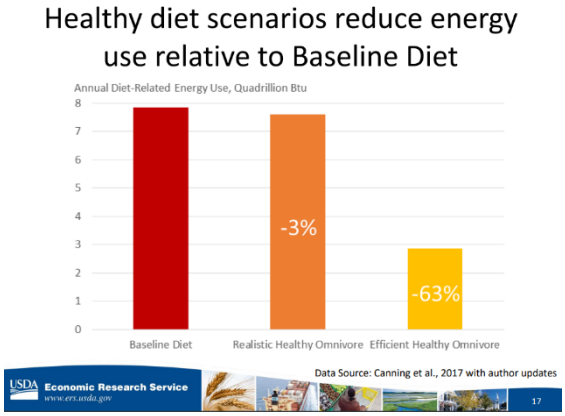
Four keynote speakers from different agencies gave presentations in this session. The slides of the presentations can be found on the project website:

https://faculty.cnr.ncsu.edu/yuanyao/database-integration-workshop-building-the-data-capacity-for-food-energy-water-research/

Dr. James Dobrowolski is the National Program Leader from the USDA National Institute of Food and Agriculture (NIFA). He presented NIFA's Initiative for Data Science in Agriculture – FACT (Food and Agriculture Cyberinformatics and Tools). Dr. Dobrowolski highlighted major challenges in managing agriculture data that are evolving and increasing across the food supply chains, such as non-digital or fractionated data and uneven accessibility. Based on stakeholders' inputs in previous NIFA workshops, the top four critical areas are: (1) data infrastructure and management, (2) applications and use of data, entities affected by data, (3) creation, collection, provenance, and characteristics of data, (4) training, programs, student, and knowledge needs around data. Dr. Dobrowolski discussed NIFA's data opportunities and showed the FACT Roadmap that focuses on Open Data FAIR principles (Findable, Accessible, Interoperable, and Reusable). He specifically pointed out that stakeholders are encouraged to provide input through the FACT summit, workshops, and listening sessions to develop science priorities.

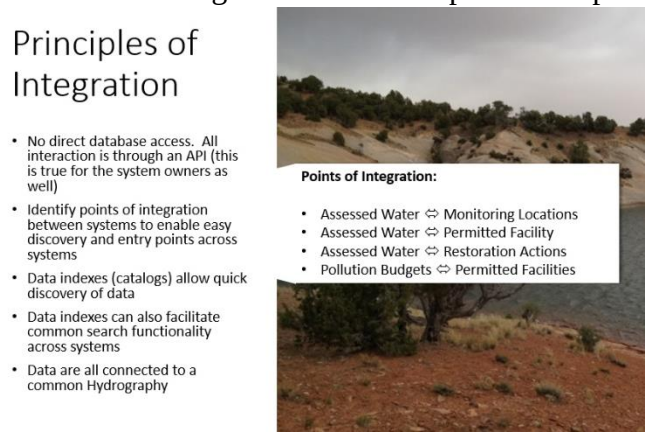


Dr. Patrick Canning is the Senior Economist from the USDA Economic Research Service. He presented an overview of FEDS: The Food Environment Data System. FEDS is an Environmental Input-Output (EIO) model that can quantify monetary and environmental flows throughout the U.S. economy. Dr. Canning highlighted that the unique attributes of FEDS include (1) data development to look at resource use over time, (2) the use of the methodology adopted by the United Nations Statistical Commission, and (3) adaptability to other countries although the current focus is U.S. food systems. Dr. Canning showed the results of resource use of current American diets across the food life-cycle and pointed out that electricity was the most used energy commodity by the U.S. food system in 2012. He also noted in 2007 U.S. diets meat

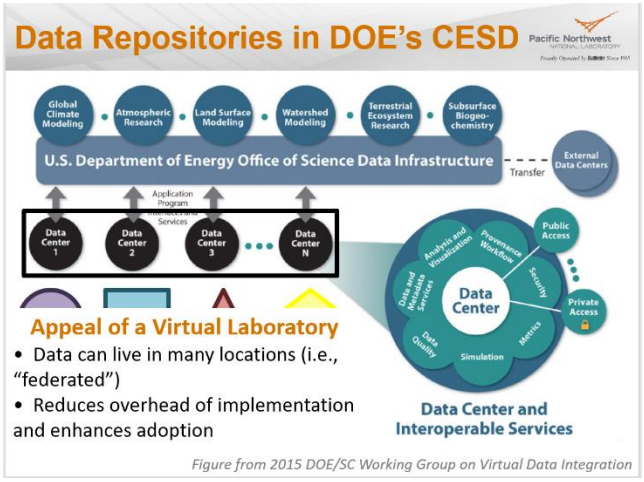


consumed the most water. Then Dr. Canning presented four interesting case studies when Americans began to adopt healthier diets. Interestingly, healthier diets could lead to energy and Greenhouse Gas (GHG) emissions reductions but not necessarily water use reduction. For FEDS, Dr. Canning emphasized that the future directions of data and modeling will include providing regularly updated data products and building models to introduce consumer and producer feedback for policy analysis.

Mr. Dwane Young is the Chief Water Data Integration Branch in the EPA Office of Water. He presented water data integration efforts by the EPA. Mr. Young introduced four pillars of open water data – standards, metadata, common hydrography, and discoverability. Mr. Young pointed out that not all water data contains these four elements. Based on the characteristics of current open water data, Mr. Young discussed the principles of water data integration that are standards-based and API supported. They have defined outputs to allow points integration between systems and data indexed for easy discovery. He used Water Quality Exchange (WQX) as a demonstration of the standards-based approach. Mr. Young highlighted that WQX is effective in data sharing as it does not depend on a particular technology but allows partners to map their systems to WQX. He also introduced another case study of data sharing with a focus on sensor data – Interoperable Watersheds Network. Mr. Young mentioned that a few current challenges and problems still exist in data standards, metadata, and architecture. He then showed several projects addressing those challenges, such as IWN's Open Architecture, Hydrologic Networks, and Catchment-based indexing approach. Mr. Young concluded that integrated data allows a broader capacity in providing service and enhances public engagement.



Dr. Casey Burleyson is the Data Sciences Scientist at Pacific Northwest National Laboratory. He presented data management strategies in an active DOE Office of Science Project – Integrated Multi-sector, Multi-scale Modeling (IM3). The project involves the collaboration of nine institutes with a goal to improve the understanding of the responses of the complex human-earth system to different stresses. Dr. Burleyson pointed out that the fundamental challenge recognized by the team is the balance between short-term deliverables



and the long-term goal of creating best practices, reproducibility, and reuse in data management. Dr. Burleyson then presented approaches and strategies used in the IM3 project, including the platforms for the data and code repositories, and the public websites where data and codes associated with specific publications can be shared. He also presented data repositories in DOE's Climate and Environmental Sciences Division, where unique data centers are mapped to individual programs and data are integrated and connected via a "virtual laboratory." Dr. Burleyson discussed changing technology and shifting platforms and recommended a focus on standardization on core elements in data management such as metadata, user credentials, and use metrics.

Session 2 Report – Data Capacity Mapping & Vision

Current Data Capacity and Sources for FEW Research

This session was divided into three parts:

(1) First 20 minutes – group discussion. Participants were asked to use flipchart paper to list databases they knew related to FEW. An example table was given to guide participants to list information regarding each database, including the name, the aspects it covers (food/energy/water), agency/sources (e.g., USDA, EPA, DOE), scale and resolution (e.g., national/county level, year/real-time, process/facility level), and whether it is publicly available.

(2) Second 20 minutes – group discussion. Participants were asked to discuss the following three key questions:

- Are there any overlapping/gap areas among different databases/sources?
- What do you need for your research?
- What is your vision of future integrated database and data sharing?

(3) Last 20 minutes – group presentation. Each group presented and summarized their findings and discussions.

Table 1 summarizes the databases identified by workshop participants. The databases were categorized based on their relevance to one or multiple FEW systems. The project team added a URL link and short descriptions for each database.

Table 1 Data Sources Related to FEW Research

Database	**F/E/W**	**Agency/Source**	**Scale/Resolution**	**Publicly available**	**URL**	**Brief Description**
Chicago Data Portal - urban	Food	Chicago	Chicago based	Y	https://data.cityofchicago.org/	Chicago's Data Portal is developed to grant accesses to government data, including datasets of departments, services, facilities, and performance.
Food Availability Data System		USDA ERS	USA/Annual	Y	https://www.ers.usda.gov/data-products/food-availability-per-capita-data-system/	This database includes three data series on food and nutrient availability for consumption at the national level, including food availability, loss-adjusted food availability, and nutrient availability.
Household Surveys		World Bank	Global	Y	http://microdata.worldbank.org/index.php/home	The database provides the data collected through sample surveys of households by the World Bank, including the food consumption and living standards of households surveys.
National Land Cover Database (NLCD)		USGS	USA/30-meter/5-year	Y	https://www.mrlc.gov/nlcd2011.php	The NLCD database offers the wall-to-wall, spatially explicit, national land cover changes and changing trends.
Cropland Data Layer (CDL)		USDA NASS	USA/30-meter /Annual	Y	https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php	CDL provides a crop-specific land cover classification data of more than 100 crop categories grown in the U.S.
Food Databases (FDA Database)		FDA/CDC	USA/State Level	Y	https://www.fda.gov/Food/default.htm	The food databases by the FDA provide the information of food substances consumed in the U.S.
Substance Registration System (FDA SRS)		FDA	USA/3-6 monthly	Y	https://fdasis.nlm.nih.gov/srs/	The database provides the unique ingredient identifiers for substances in drugs, biologics, foods, and devices.
Food Commodity Intake Database		USDA/EPA	USA/Regional/ Annual	Y	https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/food-commodity-intake-database-fcid/	This database provides the intake data regarding food commodities rather than food consumption (e.g., wheat flour & eggs vs noodles).
State Energy Data System (SEDS)	Energy	EIA	USA/State Level/Annual	Y	https://www.eia.gov/state/seds/	This database contains the data of energy production, consumption, prices, and expenditures at the state level in the USA.

Bioenergy Feedstock Library		INL	USA/Annual-daily	Y	https://bioenergylibrary.inl.gov/Home/Home.aspx	This database managed by Idaho National Laboratory has the data for physical, chemical and conversion characteristics of biomass feedstock.
Emissions & Generation Resource Integrated Database (eGrid)		EPA	USA/Unit/ Annual	Y	https://www.epa.gov/energy/emissions-generation-resource-integrated-database-egrid	This database provides the environmental characteristics of electric power generated in the United States, including emissions, net generations, resource mix, and other attributes.
Commodity Flow Survey (CFS)		BTS	USA/State Level/Annual	Y	https://www.bts.gov/content/commodity-flow-survey-overview	This database is part of the Economic Census and is updated every 5 years. It contains approximately 100,000 establishments from the industries of mining, manufacturing, wholesale trade, and auxiliaries.
Energy Information Administration (EIA)		DOE	USA/Plant Level	Y	https://www.eia.gov/electricity/data/eia923/	The database covers monthly and annual electric power data on electricity generation, fuel consumption, fossil fuel stocks, and receipts at the power plant and prime mover level.
PA Department of Environmental Protection		PA State Department	PA State	Y	https://www.dep.pa.gov/Pages/default.aspx	The PA DEP provides state-wide energy data in five main categories: comprehensive energy data, energy efficiency, renewables, production and pricing, and fuels.
EIA Annual Energy Outlook (AEO)		DOE/EIA	USA/annual-daily	Y	https://www.eia.gov/outlooks/aeo/	The report provides modeled projections of domestic U.S. energy markets through 2050.
Greenhouse Gas Reporting Program dataset (GHG Reporting)		EPA	USA/Unit/ Annual	Y	https://www.epa.gov/ghgreporting/ghg-reporting-program-data-sets	This program provides GHG emissions data from large emitting facilities, suppliers of fossil fuels and industrial gases, and facilities that inject CO_2 underground.
Motor Vehicle Emission Simulator (MOVES)		EPA	USA/County & Project Level/ 4-5 year	Y	https://www.epa.gov/moves/latest-version-motor-vehicle-emission-simulator-moves	MOVES is an emission modeling system assessing emissions from mobile sources (e.g., mobile fuel combustion) for criteria air pollutants, GHG, and air toxics at different levels.
HydroClient	Water	CUAHSI	Global/daily	Y	https://data.cuahsi.org/	This is a platform that provides water data from diverse sources including U.S. federal agencies, international governments, non-profit organizations, and academic projects. Four types of data are provided: meteorology, ground water, surface water, and water quality.

Safe Drinking Water Information System (SDWIS)		EPA	USA/Point Level	Y (Local N)	https://www3.epa.gov/enviro/facts/sdwis/search.html	This database contains information about public water systems and their violations against drinking water regulations.
National Water Model (NWM)		NOAA	USA/Local Level/Daily	Y	http://water.noaa.gov/about/nwm	A hydrologic model that simulates streamflow over the entire United States.
Water Quality Portal		EPA/USGS/NWQMC	USA/Local Level	Y	https://www.waterqualitydata.us/	The database contains the water quality monitoring data including physical, chemical, and biological quality data collected by federal, state, tribal, and local agencies.
Water Quality Watch		USGS	USA/Local Level/Real Time	Y	https://waterwatch.usgs.gov/wqwatch/	This database provides real-time and historic water quality data in surface waters.
Water Data Exchange (WaDE)		Western States Water Council (WSWC)	USA/Local & Watershed Level	Y	http://www.westernstateswater.org/wade/	The database includes physical water supply data, water use data, institutional and regulatory constraints, water allocation information, consumptive uses, return flows, and any other targeted data.
National Pollutant Discharge Elimination System (NPDES) Permits		EPA	USA/Local Level	Y	https://www.epa.gov/npdes	This dataset provides the data of regulating point sources that discharge pollutants to waters.
Underground Injection Control (UIC)		EPA	USA/Local Level	Y	https://www.epa.gov/uic	This program provides data of injection wells (generally for storing CO_2, disposing waste, enhancing oil production, mining, and preventing saltwater intrusion) developed in the U.S. by six different well categories.
National Ground-Water Monitoring Network (NGWMN)		USGS	USA/Local Level	Y	https://cida.usgs.gov/ngwmn/	The database is a collection of groundwater monitoring network data across the nation, including water levels, quality, and well construction.
Streamflow Data		USGS	USA/Point Level/15-min-daily	Y	https://waterwatch.usgs.gov/?id=ww_current	The site provides the real-time data of streamflow in the U.S.
USGS Surface-Water Data		USGS	USA/County Level/Daily	Y	https://waterdata.usgs.gov/nwis/sw	The dataset includes time-series data of stream levels, discharging streamflow, reservoir and lake levels, surface-water quality, and rainfall.
National Hydrography Database (NHD)		USGS	USA/State/Hydrologic Unit (HU8, HU4)	Y	https://www.usgs.gov/core-science-systems/ngp/national-hydrography	The main datasets are the National Hydrography Dataset, Watershed Boundary Dataset, and NHDPlus High Resolution.

Soil Moisture Data		NASA	USA/Local Level	Y	https://smap.jpl.nasa.gov/data/	This database includes data products of soil moisture at different levels.
Gravity Recovery and Climate Experiment (GRACE) Groundwater (GW) Data		NASA	USA/Local Level	Y	https://grace.jpl.nasa.gov/applications/groundwater/	This program contains measured data of groundwater changes by observing changes in the Earth's gravity field.
Impaired Waters and Total Maximum Daily Loads (TMDLs) Dataset and Tools		EPA	USA/Local Level	Y	https://www.epa.gov/tmdl/resources-tools-and-databases-about-impaired-waters-and-tmdls	Tools and datasets in the EPA related to TMDLs and water assessment.
Manufacturing Energy Consumption Survey (MECS)	Food-Energy	EIA	USA/Unit Level/ 4-year	Y	https://www.eia.gov/consumption/manufacturing/	MECS provides energy consumption data by industry and regions in the United States, including the energy consumption of the food sector.
Landsat Data	Energy-Water	NASA	USA/Local Level	Y	https://landsat.gsfc.nasa.gov/data/	The dataset records reflected and emitted energy in various wavelengths of the electromagnetic spectrum. Landsat data have been used to monitor water quality, glacier recession, sea ice movement, coral reef health, land use change, etc.
National Environmental Methods Index (NEMI)		EPA/USGS/NWQMC	USA	Y	https://www.nemi.gov/home/	The NEMI provides the methods and procedures at multiple stages of the monitoring process (e.g., monitor chemicals and metals in water).
AWARE-US		ANL	USA/County Level	Y	https://greet.es.anl.gov/publication-aware_us	This work used the AWARE method for applications in the U.S. (AWARE-US) by incorporating measured runoff and human water use data at U.S. county-level resolution.
Soil Survey Geographic Database (SSURGO)	Food-Water	USDA	USA/Regional (1:12,000 to 1:63,360)	Y	https://www.nrcs.usda.gov/wps/portal/nrcs/detail/soils/survey/?cid=nrcs142p2_053627	This database contains the soil-related information over the U.S. including available water capacity, soil reaction, electrical conductivity, frequency of flooding, land type, and other information categories.
National Water Information System Web (NWISWeb)		USGS	USA/County Level/Annual	Y	https://www.usgs.gov/nwis-national-water-information-system	The integrated water database contains surface water data (e.g., gage height and streamflow), groundwater data (e.g., water level), and water quality data (e.g., temperature, pH, nutrients) in the U.S. The database has the water-use data for agriculture and industrial sectors.

GREET	Food-Energy-Water	ANL	USA/Process Level	Y	https://greet.es.anl.gov/	GREET is a life cycle modeling tool that simulates emissions and energy consumption of different vehicles and fuel combinations.
U.S. Life Cycle Inventory Database (USLCI Database)		NREL	USA/Process Level	Y	https://www.nrel.gov/lci/	The USLCI database contains life cycle inventory data of materials production and product manufacturing, including food, energy, and water-related systems.
Census of Agriculture Quick Stats		USDA (NASS)	USA/Multi-scale/Multi-level	Y	https://quickstats.nass.usda.gov/	A comprehensive tool for accessing agricultural data published by NASS, including land use, ownership, operator characteristics, production practices (e.g., irrigation), income and expenditures (e.g., fuel input expenditure).
Toxics Release Inventory (TRI)		EPA	USA/Plant Level	Y	https://www.epa.gov/toxics-release-inventory-tri-program	This dataset provides information on toxic chemical release, pollution prevention, and waste management activities reported by industrial and federal facilities, including sectors of food, utilities, and petroleum products.
Permit Limits and Discharge Monitoring Report (DMR)		EPA	USA/Plant Level	Y	https://echo.epa.gov/tools/data-downloads/icis-npdes-dmr-and-limit-data-set	The DMR dataset provides the information and data of the permitted dischargers in the national file.
Regional and Global Climate Database		USGS	USA/County Level	Y	http://regclim.coas.oregonstate.edu/	The database offers global and regional climate data and access to the publications' data, figures, and information.
National Centers for Environmental Information (NCEI)		NOAA	USA/Hourly	Y	https://www.ncdc.noaa.gov/	The organization hosts data and models of archives on earth including comprehensive oceanic, atmospheric, and geophysical data.
USDA LCA Commons		USDA	USA/County	Y	https://www.lcacommons.gov/	This database provides LCA data of agricultural unit processes for crop plantation, swine, poultry, and beef.
EARTHDATA (Earth Observing System Data and Information System (EOSDIS))		NASA	USA/Local Level/Near Real Time	Y	https://earthdata.nasa.gov/	EARTHDATA is a comprehensive collection of datasets about the physical, chemical, and biological systems of the earth.
New and Transparent United States Environmentally Extended Input-Output Model (USEEIO)		EPA	USA/National Level	Y	https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NRMRL&dirEntryId=336332	USEEIO is an environmentally extended input-output model of the United States. It used the data on economic transactions between 389 industry sectors and related environmental data to build a life cycle model of 385 US goods and services. The environmental indicators include land; water; energy and mineral usage and emissions

						of greenhouse gases; criteria air pollutants; nutrients and toxics.
Crop Budget		State Agriculture Office	State Level/Annual	Y	Varied according to states in the U.S.	The Crop Budget contains projection data for farm revenue, variable cost, fixed cost, and net income, commonly including costs of fuel and water applications.
Census Population Estimates	Others	US Census Bureau	USA/County Level/Annual	Y	https://www.census.gov/programs-surveys/popest/about.html	Demographic data that can be used for FEW related analysis

Gaps and Data Needs

Although workshop participants identified a large number of databases, they agreed that large data gaps still exist and hinder current and future interdisciplinary FEW research. Specifically, workshop participants identified the following gaps.

(1) Overlapping while mismatching

Workshop participants agreed that overlaps exist among different databases (e.g., water data collected by different government agencies), and at the same time, mismatches exist for datasets collected and used to characterize the "same" systems. Participants identified and discussed different examples. Two groups mentioned mismatches between the datasets collected at different scales (e.g., agriculture irrigation data collected at farm scale then aggregated to state level matches the water data monitored at the state level). One group mentioned mismatches due to different sampling methods (e.g., datasets collected by government agencies versus collected by private sectors). Standardization on data sampling, documentation, and reporting, as well as harmonization are needed.

(2) High-resolution data

High-resolution data at both temporal and geospatial scales are needed. For example, from the food perspective, field-level data for agriculture yields, practices, related costs, and rescaled changes are currently lacking. For water data, gaps exist in real-time water conditions, groundwater quality, consumption use, withdrawal, and reuse. Regarding energy data, some participants mentioned the needs for more recent and frequent data (e.g., yearly data versus data reported every 5 years).

Another aspect of high-resolution data discussed by participants refers to data aggregation. Many participants mentioned national, regional, or sectoral data that can quantify resource exchanges and interactions among FEW systems. However, those data are highly aggregated, and it is challenging to use them to understand the complex interactions within FEW systems at process-, product-, and individual- or entity- level. A typical example is food manufacturing, given the different food products, manufacturing process configurations, and raw materials used.The energy and water consumption data of individual food manufacturing plants could have large variations. Such variations can only be understood if such data are available at plant- or process- level.

Despite the needs of high-resolution data, a few participants raised the question regarding whether higher-resolution data are always needed. High-resolution data commonly need larger footprints and more effort to collect, store, process, and reuse. A middle ground may need to be determined on a case-by-case basis before intensive investment in generating high-resolution data is made.

(3) Data accessibility and usability

Although many databases identified by workshop participants are publicly available, participants agreed that the availability does not necessarily mean high usability. Two groups mentioned issues in public engagement. They pointed out that it is difficult for the public to access and use many databases because of the technology barrier. Participants concluded government databases should not only be useful for researchers but also for the public, especially for diverse groups of citizens. Many application programming interface (API) packages available could be helpful to develop user-friendly applications to enhance public engagement in data collection and use. In addition, some databases are available with "restricted use" due to confidentiality (e.g., confidential business information (CBI), confidential survey, and identity information). How to enhance the usability of this type of data without breaking confidentiality is an open question and needs to be explored.

(4) Data discoverability

Data usability and discoverability are two different concepts raised by workshop participants. The former focuses on the access and use of data, while the latter refers to the pathways and means to discover the datasets. A few participants shared their experiences and strategies of data searching and knowledge discovery. They commented that the current data discovery process (e.g., using Google, or Web of Science) highly depends on the experience of data users and keywords they select. Thus, they do not know what datasets or aspects they may miss (referred to as "we do not know what we do not know"). This is a challenge faced by many data users, especially for those who conduct interdisciplinary research such as FEW. The Digital Object Identifier system (DOI) for data, similar to DOI for scholar publications, could be a solution to enhance both discoverability and re-usability of data, especially for those embedded in journal publications.

(5) Data needs for inter-, multi-, and trans-disciplinary researchers

Most databases are generated and managed for specific purposes and disciplines. Given that most FEW research involves inter-, multi-, and trans-disciplinary collaboration, one question raised by workshop participants is how current databases could better support the research relevant to FEW. Many participants shared their experiences collecting data across different disciplines and sources, then performing integration and synthesis by themselves. This process is not only time-consuming and challenging, but also brings in uncertainty. However, connecting different databases without clear purposes and strategies may also be difficult and unnecessary. Some participants suggested that it would be helpful to survey and understand the general data needs of researchers in FEW relevant areas, in order to gather and integrate datasets that can meet their needs. For example, different data formats, access interfaces, and data processing languages commonly discourage researchers across disciplines to use and

synthesize data. Developing packages that can effectively transform different data formats and languages would be very helpful.

(6) Data uncertainty and missing data

Data uncertainty is a challenge faced by almost all data users. It is critical to understand the sources of uncertainty (e.g., uncertainty due to sampling, measurement methods, aggregation/disaggregation methods) and the quantity of uncertainty. This information should be reported or at least mentioned and discussed when data are presented. Another gap workshop participants mentioned is missing data in existing datasets. Simply disregarding missing data may introduce bias. Imputation is a potential solution, but more data or more advanced computational technology is needed for meaningful and effective imputation.

The Vision of Database Integration and Data Sharing

Future database integration and data sharing should address the needs and gaps highlighted in the previous discussion. Participants provided the following visions:

- Standardization on data collection and documentation, metadata, data presentation, and data access
- Free and easy discovery, access, and reuse. The integrated databases or interfaces should allow researchers and data users to identify and reuse datasets they need more efficiently. For example, DOI may be established and linked to individual databases and datasets in publications. Datasets are either presented in a similar format or they are such that they could be efficiently transformed to meet the needs of different researchers.
- High quality and consistency across datasets collected and presented
- Data from the public sector and private sector well supplemented
- Programming codes, another form of data, well used to generate data
- Technology development and support to fill missing data and data harmonization
- Convenient mechanism to allow the use of restricted data
- Delegated responsibility on data harmonization and integration
- Cultural changes to encourage and stimulate data sharing

One interesting discussion centralized on distributed database management. Some participants argued that FEW related databases should be centrally managed for easy discovery and access. Others argued that the integration across all databases related to FEW might not be feasible or necessary given that many datasets are too large to be moved, processed, and downloaded by users. A vision proposed and agreed upon by many participants is to have an integrated interface allowing clustered datasets (see Figure 1). Datasets would be clustered as individual “Hubs” in a way that can be used by researchers interested in a specific topic (e.g., water quality, water use, crop yields, or energy consumption). A centralized interface would manage different data hubs. This could be a potential infrastructure to enhance the discoverability, usability, and consistency of databases.

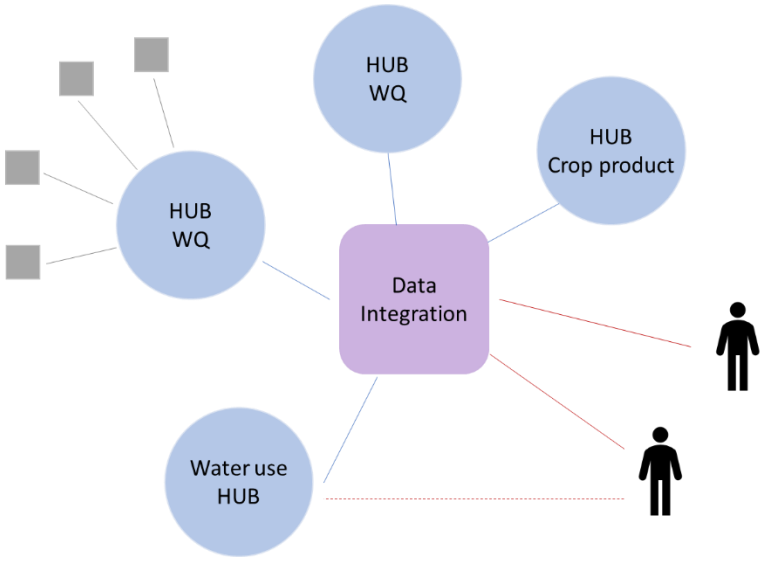


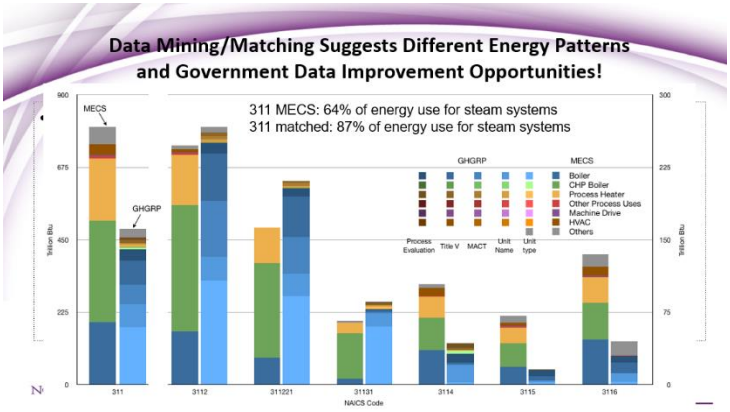
Figure 1. A schematic of integrated interface allowing clustered datasets

Session 3 Report – Data Integration and Synthesis

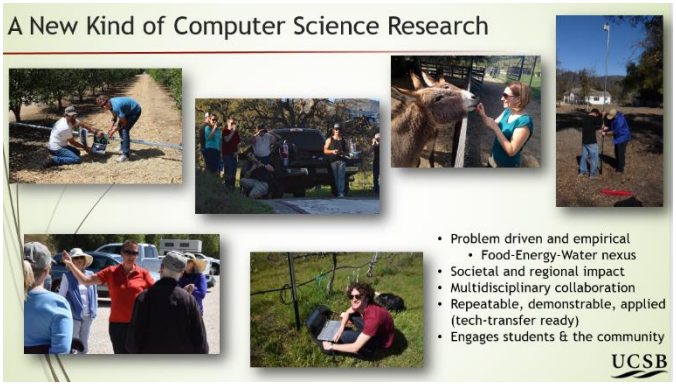
Four keynote speakers from different institutions gave presentations on data integration and synthesis. The slides of the presentations can be found on the project website:

https://faculty.cnr.ncsu.edu/yuanyao/database-integration-workshop-building-the-data-capacity-for-food-energy-water-research/

Dr. Eric Masanet is an Associate Professor at Northwestern University. He was previously the head of the Energy Demand Technology Unit in the International Energy Agency (IEA). Dr. Masanet presented "National Energy Statistics: Opportunities and Challenges for Food-Energy-Water (FEW) Data Integration." He first used three case studies on international, national, and process-level FEW decisions to introduce gaps and needs in bridging available data and FEW models. Dr. Masanet then discussed the IEA's and U.S. approaches in collecting energy efficiency data at different resolutions to better understand energy progress at different industrial sectors and sub-sectors. He highlighted the strengths and weaknesses of both approaches and introduced another two projects using process-level data and modeling for FEW decision making. Then Dr. Masanet discussed the mismatch between databases that are commonly used in FEW modeling. He concluded that data mining and matching may provide improvement opportunities for government data.



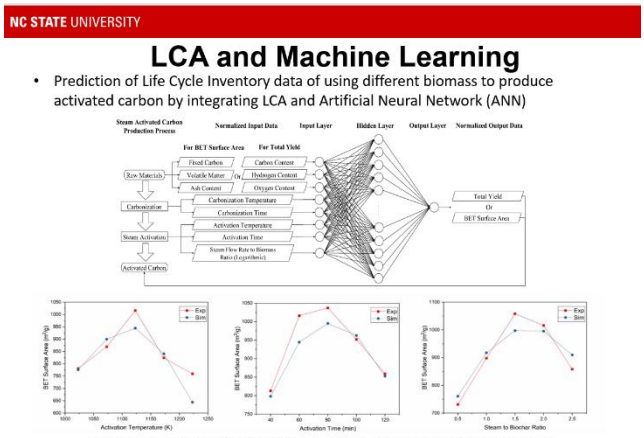
Dr. Chandra Krintz is a Professor in Computer Science at the University of California Santa Barbara. She is also the Director of SmartFarm and RACELab. Dr. Krintze presented: "SmartFarm: Data Integration & Systems for Agriculture-based, FEW Research." Dr. Krintz started by discussing recent developments in cloud computing and data analytics. She then gave a few examples of the need to tailor cloud and data analytics to address the critical needs and complex challenges of food production, such as irrigation scheduling, disease/pest management, and farm-to-fork tracking. Dr. Krintz emphasized that interface and prediction are critical to connect data and decision



making. Dr. Krintz introduced the concept of edge clouds that can be connected with a public cloud to community or university cloud services. She discussed the SmartFarm research project that focuses on a self-managed edge cloud system to provide farmers with secure data analysis for problems such as precision applications of water and pesticides, forest prediction, and damage mitigation. Dr. Krintz mentioned several challenges of integrating agriculture data. She also highlighted that this is a new area of computer science research that is highly problem driven (FEW focused) and needs multidisciplinary collaborations.

Dr. Gale Boyd is the Director of the Triangle Research Data Center at Duke University. Dr. Boyd introduced the Triangle Research Data Center (TRDC) that is a partnership between the U.S. Bureau of Census and Duke University, in cooperation with the University of North Carolina at Chapel Hill, North Carolina State University (NCSU), and RTI International (RTI). TRDC is one of the Federal Statistical Research Data Centers that provides authorized access to restricted use microdata for statistical purposes only. It could be a mechanism to allow and encourage the use of data containing confidential information. Dr. Boyd also discussed the challenges of matching data from different databases and highlighted the potential use of administration data collected by different government agencies.

Dr. Yuan Yao is an Assistant Professor of Sustainability Science and Engineering at North Carolina State University. She presented: "Promoting Bioeconomy for Sustainable Food-Energy-Water Systems: The Need of Interdisciplinary Research from a Data Point of View." Dr. Yao introduced the critical role of biomass in FEW sustainability and highlighted major challenges in applying Life Cycle Assessment (LCA) to support biomass decision making. Dr. Yao pointed out that LCA needs to be integrated with data and modeling techniques in other disciplines to better address temporal and geospatial dynamics of biomass utilization, which is hard to quantify and understand through traditional LCA approaches. Moreover, data are a major challenge for such integration as in many other interdisciplinary research areas. Dr. Yao presented three case studies from both modeling and data perspectives, including integrating LCA with agent-based modeling, geographic information systems (GIS), and machine learning. Dr. Yao also discussed image data, which could provide beneficial land use information for bioeconomy development. In the end, Dr. Yao called for more data sharing and integration across disciplines. She also highlighted the needs for capacity building and infrastructure development to support increasing interdisciplinary research projects in FEW areas.



Preliminary results of the prediction of surface area of activated carbon versus experimental results

Liao, M., Kelley, S. and Yao, Y.* (2018) Artificial neural network based modeling for the prediction of yield and surface area of activated carbon from steam activation, Journal of Bioresource Technology (in preparation)

Session 4 Report – Challenges and Barriers

In this session, participants first discussed in groups the challenges and barriers for integrating, sharing, and synthesizing databases across different agencies and sources. Each group needed to identify the top five challenges that they believe to be critical. After the discussion, each participant voted for the top challenges (limited to 4 votes/each). Similar challenges were grouped to avoid duplicated voting. Table 2 shows the identified challenges and voting results. The top four topics/challenges guided the discussion in the following session.

Table 2 Top Challenges Identified by Workshop Participants

Grouped Topics	Challenges Identified	Vote
Metadata and database discoverability	Metadata	25
	Database discoverability	
	Better ways to discover database	
	Universe of databases is not fully characterized	
	Data standards and metadata standards are no small tasks.	
High-level coordination	Lack of coordination at the federal level (can be improved by state-level cooperation)	19
Database access	Access and agencies' firewalls	17
	Easy access to the database	
	Legacy data	
	Separate interface from implementation	
Politics in database governance	Data governance	17
	Ownership decision-making legacy	
	Rules and policy of data access in different agencies	
	Politics of data providers	
Data gaps	Gaps between metadata and the needs of interdisciplinary researchers	10
	The ways data are represented to make sure they can be used by others in an effective and meaningful way	
	Activity data gaps	
Mismatch among databases	Temporal and spatial mismatch	9
New generation of data	Emerging new data from conventional data	7
	Next generation data	
Public and private interaction	Public and private interaction	6
Data integration	Data integration requires curation	4
Data sharing	The environment of data sharing is evolving	1
Data generality	Tradeoffs between generally and specifically restricted needs with composition	1

Session 5 Report – Action Plan and Roadmapping

In this session, participants discussed short-term and long-term goals and efforts needed to address the top four challenges identified in the previous session. Workshop participants proposed different action plans, and a few aspects mentioned by most are highlighted below:

- There is a need to identify the best practice in other domains regarding database integration and data sharing (e.g., biology and healthcare). Reinventing the wheel could be avoided if effective strategies were developed and could be adapted to FEW research.
- Pilot projects on FEW database integrations are needed to test and identify feasible action plans for large-scale implementation across the U.S. Lessons could be learned by monitoring challenges encountered and solutions developed by the pilot projects. It will be useful to investigate the efforts related to data standards and structures, coordination with government agencies, accessibility and discoverability, and governance. It is possible to identify generalized solutions and those unique to the specific research topics through the pilot projects. The results and lessons learned from those projects could shed light on potential pathways to success.
- There is a need to develop necessary incentives and infrastructure to promote data sharing among all stakeholders (e.g., government, academia, the private sector, and the public). For academics who are both the data users and data providers, many actions proposed by workshop participants are related to either academic publication or funding mechanisms for academic research. Participants agreed that the current accessibility of data in academic publications and research projects is very limited.

The short-term and long-term goals identified by workshop participants to address each of the four top-ranked challenges are listed as follows:

Challenge 1 Data Standards and Structure Management

Short-Term

- Reuse but simplify existing standards, such as Data Documentation Initiative (https://www.ddialliance.org/) and Open Geospatial Consortium (http://www.opengeospatial.org/)
- Use open structure markup (e.g., schema.org)
- Identify successes and failures in other domains (e.g., data sharing and integration in healthcare, earth sciences)
- Investigate and explore existing infrastructure (e.g., Earth System Grid Federation, https://esgf.llnl.gov/)

Long-Term

- Develop standardized and inter-operable metadata standards for FEW related data
- Implement and incorporate the best practices from other domains.

Challenge 2 Lack of Coordination at the Federal Level for Data

Short-Term

- Enhance inter-agency interactions and communications (e.g., working groups)
- Establish coordination organizations (e.g., HydroShare, https://www.hydroshare.org/).
- Support staff communication and collaboration
- Identify data mismatches among databases
- Enhance data management that has a better tolerance of changes in data custodians

Long-Term

- Delegated groups who oversee federal statistics
- Government support for data integration
- Expand use of inter-agency working groups (e.g., Federal Interagency Council on Statistical Policy, https://nces.ed.gov/FCSM/index.asp) through mandated participation, empowerment, and accountability
- Develop MOUs (memorandum of understanding) between agencies

Challenge 3 Accessibility and Discoverability of Data and Databases

Short-Term

- Enhance the awareness and use of various data sources (e.g., education, workshops, and social media)
- Compile existing available FEW data
- Use fitness for use framework (description, formats, and other aspects to match data uses)
- Create indices at the variable/measurement level (e.g., a question bank for surveys or measurement database for resources)

Long-Term

- All data findable and accessible through well-known federated portals
- Engagement with the private sector (e.g., Google data search) and the public on FEW data
- Improvement in persistent identification infrastructure for data (e.g., DOI, citation standards, data artifact)
- Regularized/standard data flow practice in publications

Challenge 4 Data Governance

Short-Term

- Clear and consistent policies for current data availability, ownership and use
- Identify best practice for data sharing (e.g., data archiving)
- Understand current rules and policies for government and academic data; develop consistent language and terminology for data across different communities and domains
- Provide training and compliance for restricted data
- Promote open access policies at universities
- Encourage data sharing and reporting across federally funded research projects

Long-Term

- Establish federated portals where FEW data are findable and accessible
- Engage with private sectors on necessary policies and protections to make more private data available
- Engage with the public on data collection and reuse
- Improve persistent identification infrastructure for data (e.g., DOI, citation standards, data artifact)
- Implement best practice of data sharing and archiving
- Develop transparent rules for government and academic data
- Promote and build consensus across government, the private sector, and academia
- Create a mandatory requirement of data sharing for funded research