Science Synthesis, Analysis, and Research Program

Community for Data Integration 2020 Project Report

COMMUNITY for DATA INTEGRATION

Open-File Report 2024–1027

U.S. Department of the Interior
U.S. Geological Survey

Community for Data Integration 2020 Project Report

By Leslie Hsu, Emily G. Chapin, Theodore B. Barnhart, Amanda E. Cravens, Richard A. Erickson, Jason Ferrante, Aaron Fox, Nathaniel P. Hitt, Margaret Hunter, Katharine Kolb, Jared R. Peacock, Matthew D. Petkewich, Sasha C. Reed, Terry L. Sohl, and Tanja N. Williamson

Science Synthesis, Analysis, and Research Program

Open–File Report 2024–1027

U.S. Geological Survey, Reston, Virginia: 2024

Contents

Figures

Abbreviations

CDI	Community for Data Integration
CSI	Coastal Salinity Index
eDNA	environmental DNA
FAIR	findable, accessible, interoperable, reusable
FY	fiscal year
IRIS	Incorporated Research Institutions for Seismology
MT	magnetotelluric
NAS	Nonindigenous Aquatic Species
USGS	U.S. Geological Survey

Community for Data Integration 2020 Project Report

By Leslie Hsu, Emily G. Chapin, Theodore B. Barnhart, Amanda E. Cravens, Richard A. Erickson, Jason Ferrante, Aaron Fox, Nathaniel P. Hitt, Margaret Hunter, Katharine Kolb, Jared R. Peacock, Matthew D. Petkewich, Sasha C. Reed, Terry L. Sohl, and Tanja N. Williamson

Abstract

The U.S. Geological Survey Community for Data Integration annually funds small projects focusing on data integration for interdisciplinary research, innovative data management, and demonstration of new technologies. This report provides a summary of the 12 projects funded in fiscal year 2020, outlining their goals, activities, and accomplishments.

Introduction

The U.S. Geological Survey's (USGS) Community for Data Integration (CDI) annually funds projects focusing on data integration for interdisciplinary research, innovative data management, and demonstration of new technologies. Since 2010, the CDI has funded more than 110 projects. The CDI supports projects that

- focus on targeted efforts that yield near-term benefits to Earth and biological science;

- leverage existing capabilities and data;

- implement and demonstrate innovative solutions (for example, methodologies, tools, or integration concepts) that could be used or replicated by others at scales from project to enterprise;

- preserve, expose, and increase access to Earth and bio-logical science data, models, and other outputs; and

- develop, organize, and share knowledge and best prac-tices in data integration.

This report provides a summary of the 12 projects funded in fiscal year 2020 (FY 2020), outlining their goals, activities, and accomplishments. Proposals in FY 2020 were encouraged to address one of the following optional themes:

1. Projects that address one of the major components of the USGS Director's vision.

    a. Integrate USGS data into a comprehensive data lake.

    b. Develop, test, and apply an integrated predictive science capability that incorporates data, interpreta-tions, and knowledge that spans discipline boundar-ies, geographies, and sectors.

    c. Provide actionable intelligence that can be used via dashboards and applications to enhance situational awareness, provide new operational capabilities, and inform decision making using technologies such as artificial intelligence, machine learning, and high-performance computing.

2. Tools and methods supporting wildland fire and water prediction, aligned with the Earth monitoring, analyses, and projections (EarthMAP) vision (Jenni and oth-ers, 2017).

3. Producing findable, accessible, interoperable, reusable (FAIR) data and tools for integrated predictive science capacity.

4. Reusing or repurposing modular tools such as those that were developed from previous CDI projects, including the CDI Risk Map.

The twelve projects funded in FY 2020 illustrate the scope of USGS research. Topics that are covered include artificial intelligence and machine learning, citizen science, FAIR practices, remote sensing, streamgage networks, cloud environments, Jupyter Notebooks, and standardized data formats. The projects in this report are organized by primary Science Support Framework elements (fig. 1) as specified by the project leads.

The CDI Science Support Framework is a conceptual structure that illustrates how research data flow through the processes and products upon which the CDI operates. CDI communities of practice contribute to scientific advancement, information management and integration through computa-tional tools and services; management, policy, and standards; and data and information assets. Categorizing projects funded by the CDI in the Science Support Framework elements helps to track how CDI activities are advancing the CDI commu-nity's knowledge of Earth systems.

Science Support Framework elements include data man-agement, knowledge management, the stages of the science data lifecycle (planning, acquisition, processing, analysis,

preservation, publishing and sharing), applications, web services, semantics, information, data assets, science-project support, and communities of practice (Faundeen and others, 2013).

In figure 1, the Science Support Framework breaks down the "what" and the "how" of the CDI. The vertical elements represent products and tools that contribute to the knowledge and understanding of the Earth's systems, and the horizontal elements represent the processes, implementation, and human-data-technology interactions used to achieve data integration.
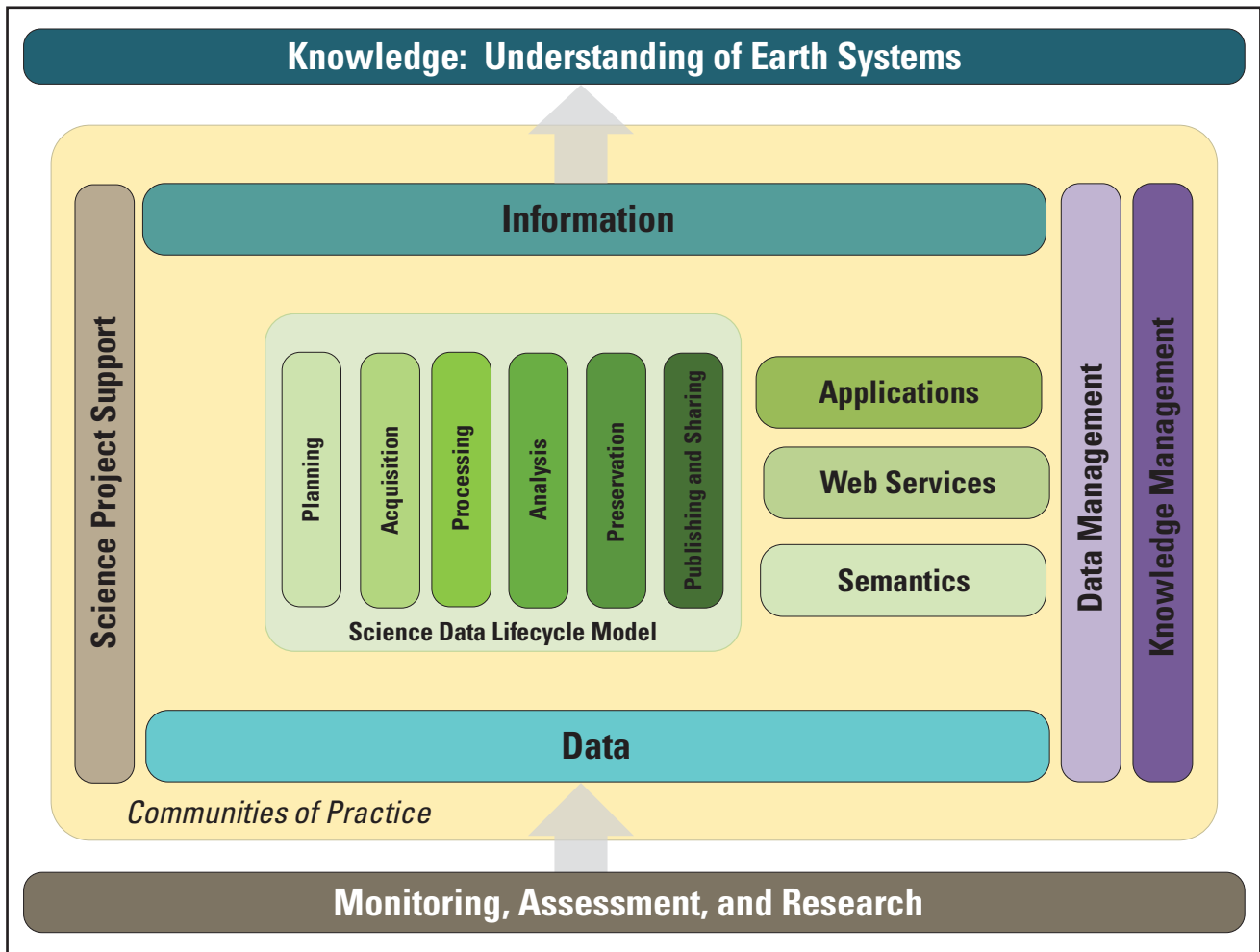
Figure 1.   Diagram showing the Community for Data Integration (CDI) Science Support Framework (Chang and others, 2013).

Data Management

"Developing Metadata Standards and an Archivable and Exchangeable File Format for Magnetotelluric Data"

Jared R. Peacock (USGS), Anna Kelbert (USGS), Andy Frassetto (Incorporated Research Institutions for Seismology; IRIS)

Science Support Framework elements: Data management, publishing and sharing, preservation

Product types: Publication, data release

Significance: Increases accessibility and interoperability of magnetotelluric (MT) data (spatiotemporal and geophysical data; fig. 2)

What need does the project address? MT is an electromagnetic method used in geophysics that is sensitive to variations in subsurface electrical resistivity. Metadata standards, archivable and exchangeable file formats, and open-source software tools supports reading and writing of MT data files. These capabilities also support the USGS and IRIS because the USGS has an obligation to U.S. Congress to collect MT data across the southwestern United States and archive those data to IRIS.

How do project outcomes address the need? Outcomes of this project include a metadata standard for MT time-series data, an open-source Python package to read and write standard metadata (mt-metadata), and an open-source Python package (MTH5) that reads and writes a standard HDF5 file with standard metadata. These outcomes allow MT data to be archived in a standard way, regardless of the data repository, while propelling MT time-series data to follow FAIR principles.

How can users take advantage of project outcomes? Users may read the publication "MTH5—An Archive and Exchangeable Data Format for Magnetotelluric Time Series Data" (Peacock and others, 2022), view the metadata standards (Peacock and others, 2021), and view the open-source software tools mt-metadata (https://code.usgs.gov/gmeg/mt-metadata) and MTH5 (https://code.usgs.gov/gmeg/mth5). An example workflow is demonstrated in figure 2.

In figure 2, magnetotelluric (MT) data collected by different users and instruments are formatted into a Hierarchical Data Format 5 (HDF5) standard using the open-source Python package, MTH5. The associated package, mt-metadata, provides tools to read and write standardized MT metadata. Data can be archived at the Incorporated Research Institutions for Seismology (IRIS) Data Management Center (DMC), U.S. Geological Survey (USGS) ScienceBase, or other data repositories. These data can be input into data-processing software.

Related links

- MT metadata standards: https://doi.org/10.5066/P9AXGKEV

- Mt-metadata software package: https://doi.org/10.5066/P13JBD4V

- MTH5 software package: https://doi.org/10.5066/P13YMLX9

Figure 2. Diagram showing an example workflow to make magnetotelluric (MT) time-series data findable, accessible, interoperable, and reusable (FAIR).

Science Data Lifecycle—Processing

"Building a Framework to Compute Continuous Grids of Basin Characteristics for the Contiguous United States"

Theodore B. Barnhart (USGS), August R. Schultz (USGS), Florence E. Thompson (USGS), Toby L. Welborn (USGS), T. Roy Sando (USGS), Seth A. Siefken (USGS), Alan H. Rea (USGS), Roland Viger (USGS), and Peter M. McCarthy (USGS)

Science Support Framework elements: Processing, web services, publishing and sharing

Product types: Software release, data release, web service

Significance: Creates actionable data that fill a gap in USGS basin-characteristic products (fig. 3).

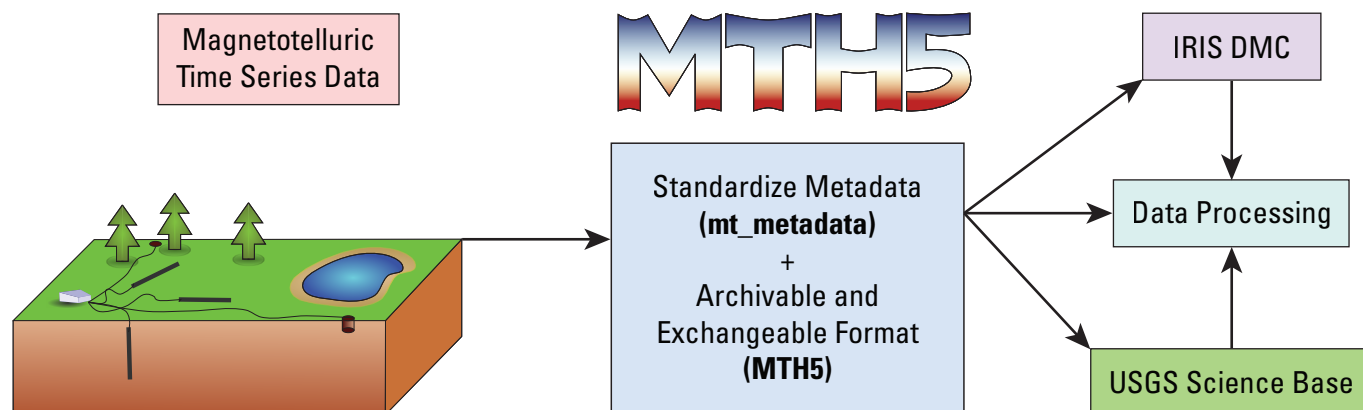**What need does the project address?** The project addresses consistent and seamless basin characteristics, describing the defining features of watersheds that are not available nationally at a 30-meter resolution. These data are useful for machine learning, statistical modeling, and process-based hydrologic modeling. Traditional methods for generating basin characteristics are time and computationally intensive, which limits their use for widespread applications without significant investment.

How do project outcomes address the need? We formalized a method to pre-compute basin characteristics for the contiguous United States using previously available flow-direction grids. We then used this method to produce a pilot dataset of nationally consistent, pre-computed basin characteristics at a 30-meter resolution, called flow-conditioned parameter grids (fig. 3).

How can users take advantage of project outcomes? The software to pre-compute basin characteristics is available as a USGS software release (Barnhart and others, 2020) and the pilot dataset of pre-computed basin characteristics may be accessed through a USGS data release (Barnhart and others, 2021). Additionally, queries against the pilot, pre-computed basin-characteristics grid may be made through a USGS StreamStats API service, found at this link: https://streamstats.usgs.gov/docs/streamstatsservices/#/.

Related links

- Tools to create flow-conditioned parameter grids: https://doi.org/10.5066/P9W8UZ47

- Flow-Conditioned Parameter Grids for the Contiguous United States: https://doi.org/10.5066/P9HUWM6Q

- StreamStats API service: https://streamstats.usgs.gov/docs/streamstatsservices/#/

"Enabling Artificial Intelligence for Citizen Science in Fish Ecology"

Nathaniel P. Hitt (USGS), Benjamin H. Letcher (USGS), Mona Arami (USGS), Natalya Rapstine (USGS), Karmann Kessler (USGS), Sheng Li (University of Georgia), Zhongliang Zhou (University of Georgia), Sean Simmons (Anglers Atlas), Nicholas Polys (Virginia Polytechnic and State University)

Science Support Framework elements: Data, processing, communities of practice

Product types: Data release, publication

Significance: Enables fish-population tracking for fish management and conservation through artificial-intelligence analysis of fish imagery (fig. 4).

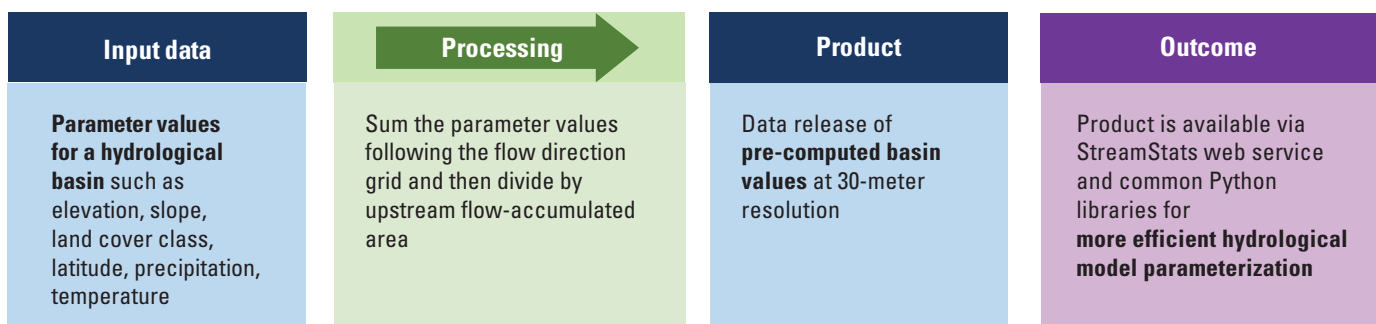| Input data	Processing	Product	Outcome
Parameter values for a hydrological basin such as elevation, slope, land cover class, latitude, precipitation, temperature	Sum the parameter values following the flow direction grid and then divide by upstream flow-accumulated area	Data release of **pre-computed basin values** at 30-meter resolution	Product is available via StreamStats web service and common Python libraries for **more efficient hydrological model parameterization**

Figure 3.   Flow chart explaining the input, processing, product, and outcome of the project to compute continuous grids of basin characteristics for the conterminous United States.

Figure 4. Image showing the pattern, color, and texture of marks and spots on a *Salvelinus fontinalis* (Brook trout). Modified from Zhou and others (2022).

What need does the project address? Artificial intelligence (AI) is revolutionizing ecology and conservation by enabling the identification and recognition of individual animals in nature from photos and videos (fig. 4). Our project develops training data and code to advance this technology for freshwater-fish management and conservation. Specifically, individual-fish recognition with AI would enable mark-recapture models to assess population size, growth, and movement dynamics.

How do project outcomes address the need? We developed training data and source code for individual fish recognition using a deep-learning framework (convolutional neural networks). Through the application of transfer-learning methods, we achieved predictive accuracy of nearly 90 percent for individual Brook trout classification from images.

How can users take advantage of project outcomes? Training data are available in the data releases "Annotated Fish Imagery Data for Individual and Species Recognition with Deep Learning" (Hitt and others, 2021a) and "Brook Trout Imagery Data for Individual Recognition with Deep Learning" (Hitt and others, 2022). The work is described in the publications "Pigmentation-Based Visual Learning for Salvelinus Fontinalis Individual Recognition" (Zhou and others, 2022) and "Comparison of Underwater Video with Electrofishing and Dive-Counts for Stream Fish Abundance Estimation" (Hitt and others, 2021b).

Related links

- Annotated fish imagery data for individual and species recognition with deep learning: https://doi.org/10.5066/P9NMVL2Q

- Brook trout imagery data for individual recognition with deep learning: https://doi.org/10.5066/P94UL1Z1

Science Data Lifecycle—Analysis

"Establishing Linkages Among USGS Land Use, Water Use, Runoff, and Recharge Models"

Terry L. Sohl (USGS), Ward E. Sanford (USGS), Gabriel Senay (USGS)

Science Support Framework elements: Analysis, applications, processing

Product types: Data release, publication

Significance: Informs the projection of water use and irrigation demand in the Delaware River Basin (fig. 5).

What need does the project address? The intersection of water-resource use and availability with human land use and land management strongly affects economic vitality, environmental health, and human health and well-being. Integrated modeling can address the feedbacks between water, land, and climate processes, producing actionable information that can be used by land managers and decision-makers to anticipate and plan for future scenarios of natural and human Earth-system change (Dornbierer and others, 2021b).

How do project outcomes address the need? The linkage of three previously disparate efforts—remote-sensing-based research on evapotranspiration and water use, land-use modeling (fig. 5), and runoff and recharge modeling—connects the vital elements of the water budget and increases our understanding of the land-water-climate system. This project serves as a platform for the continued development of an interdisciplinary predictive-modeling capability within USGS and breaks down the conceptual, practical, and technical challenges in modeling the integrated Earth system.

EXPLANATION

- ▢ Hay and alfalfa
- ▢ Corn
- ▢ Soybeans
- ▢ Fallow and idle
- ▢ Wheat
- ▢ Fruits and vegetables
- ▢ Other crops
- ▢ Open water
- ▢ Developed low density
- ▢ Developed high density
- ▢ Barren
- ▢ Deciduous forest
- ▢ Evergreen forest
- ▢ Mixed forest
- ▢ Shrubland
- ▢ Grassland and pasture
- ▢ Woody wetlands
- ▢ Herbaceous wetlands
- ▢ Perennial grass
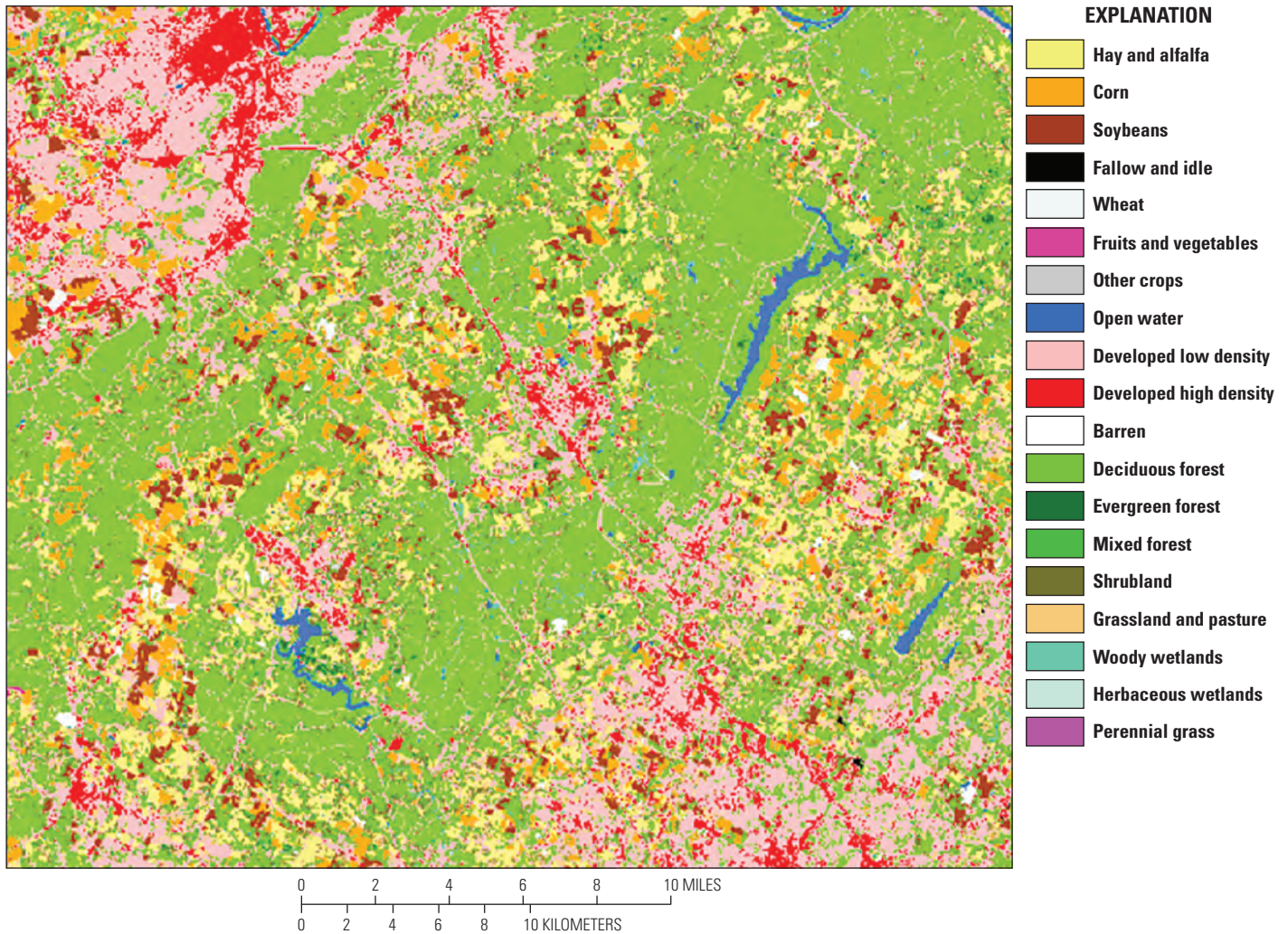
Figure 5. Map showing land-use and land-cover models, such as the above region from the Delaware River Basin, are used in integrated models to inform crop water-use estimates and irrigation requirements. Modified from Dornbierer and others (2021b).

How can users take advantage of project outcomes?
A complete suite of historical, current, and future land-use and land-cover data for the Delaware River Basin are available to download, with maps at 10-year intervals from 1680–2100 and multiple scenarios available from 2018–2100; similarly, crop water-use estimates and irrigation requirements have been produced for the concurrent land-cover distributions from 1950–2100. These data are distributed alongside the companion land-use and land-cover data (Dornbierer and others, 2021a) and described in Dornbierer and others (2021b).

- Related link—Long-term database of historical, current, and future land cover for the Delaware River Basin (1680 through 2100): https://doi.org/10.5066/P93J4Z2W

"Using Machine Learning to Map Topographic-Soil and Densely Patterned Sub-Surface Agricultural Drainage (Tile Drains) from Satellite Imagery"

Tanja N. Williamson (USGS), Alexander O. Headman (USGS), Fleford S. Redoloza (USGS), Michael E. Wieczorek (USGS), Barry Allred (U.S. Department of Agriculture Agricultural Research Service [ARS])

Science Support Framework elements: Analysis, information, publishing and sharing

Product types: Data release, publication

Significance: Improves hydrologic modeling with a national map layer of tile drainage (fig. 6).

Figure 6.　Image showing a 0.5 kilometers squared (km²) area from a 2008 satellite image showing tile-drains in Michigan farm fields (left) that were used to train the model and model output of the delineated tile locations using the U-Net model (right). Image modified from Redoloza and others (2023).

What need does the project address? Tile drainage is a type of subsurface drainage system that removes excess soil water from agricultural fields. Every state in the U.S. has agricultural fields with tile drainage and they are most common in the poorly drained fields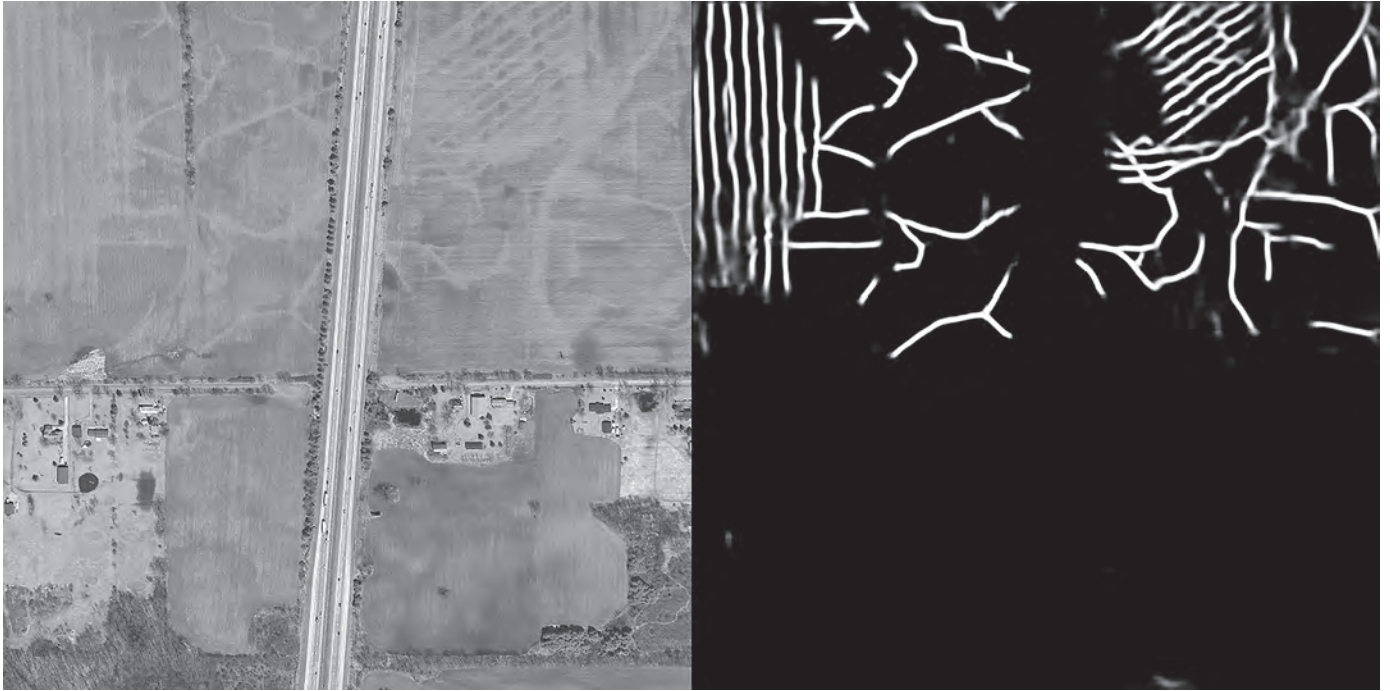 of the Midwest (U.S. Department of Agriculture, 2024). Without knowing how tile-drain extent has changed over time, it is difficult to differentiate how streamflow and water quality have changed because of spatial extent and characteristics of tile-drain networks. Our method delineates tile drains in satellite imagery, providing a way to look at historical imagery and use satellite data to maintain an up-to-date geospatial layer of tile-drain extents in basins of interest. In figure 6, on the right panel, the white lines are the delineated tile-drain locations, while the black areas denote the spacing between individual tile-drain lines. The spacing is one of the ways in which tile drainage has changed in recent decades.

How do project outcomes address the need? A fully convolutional neural network (U-Net) was trained to delineate tile drains in satellite imagery by incorporating methods from Ronneberger and others (2015) and Zhang and others (2018). A Jupyter Notebook (which can be run locally or in the cloud) is used to pre-process panchromatic satellite images (approximately 0.5 meters ground resolution) to create the format for processing (0.5 km² patches). The model is then run for the group of pre-processed data and produces a binary raster of "tile" and "not tile" for each patch (fig. 6). These patches are then reintegrated to the original spatial extent of the satellite imagery.

How can users take advantage of project outcomes? Training data are available in an image library (Williamson and Hoefling, 2023), which includes tile-drained landscapes, traced tile-drained landscapes, and examples of tile-drain types. This image library documents the initial data used to train the model and provides an example of how additional images could be integrated to improve the model. The work is described in "Machine-Learning Model to Delineate Sub-Surface Agricultural Drainage from Satellite Imagery" (Redoloza and others, 2023).

- Related link—Image library: https://doi.org/ 10.5066/P9KSZ382

Science Data Lifecycle—Publication/ Sharing

"So, You Want to Build a Decision Support Tool? Successes, Pitfalls, and Lessons Learned for Tool Design and Development"

Amanda E. Cravens (USGS), Nicole M. Herman-Mercer (USGS), Amanda D. Stoltz (USGS)

Science Support Framework elements: Publishing and sharing, planning, applications

Product types: Publication

Significance: Increases efficiency and streamlines production of decision-support tools at the USGS (fig. 7).

What need does the project address? There is a need across the USGS to streamline and implement best practices in decision-support system development. Decision-support system development may require substantial investments of time and money throughout a project's lifecycle, as well as a host of decisions in design, science output, and user and stakeholder engagement. Across the USGS, significant expertise and experience exists in developing decision support systems, but there is a need to facilitate the sharing of the knowledge learned.

How do project outcomes address the need? By use of an online survey (54 responses) and semi-structured interviews (23 interviews), the project team collected empirical qualitative, and quantitative data

Use an adaptive, iterative
design process

Collaborate across disciplines
and organizations

Engage with the target
users of the tool

Develop an empirical
understanding of use and usability

Plan for the tool's
full life span

Figure 7.   Diagram showing a summary of key points to consider when developing U.S. Geological Survey decision-support tools. This project compiled recommendations, best practices, and lessons learned from decision support system developers across the U.S. Geological Survey. Image from Stoltz and others (2023).

about USGS scientists' experiences designing and implementing decision-support systems. This dataset provides an opportunity to identify what is working well and what could be improved upon with the USGS decision-support tool efficiency.

How can users take advantage of project outcomes? Stoltz and others (2023) summarizes the recommendations, best practices, and lessons learned from decision support system developers across the USGS (fig. 7).

"Implementing FAIR Practices—Storing and Displaying eDNA Data in the USGS Nonindigenous Aquatic Species Database"

Margaret Hunter (USGS), Jason Ferrante (USGS), Matthew E. Neilson (USGS), and Wesley M. Daniel (USGS)

Science Support Framework elements: Publishing and sharing, data, semantics

Product types: Publication

Significance: Makes eDNA occurrence data more accessible and interoperable (fig. 8).

What need does the project address? Environmental DNA (eDNA) studies have allowed for the identification and biosurveillance of numerous invasive and threatened aquatic species. This information supports managers in their decision-making efforts which requires that data and metadata be produced and reported in an accurate and standardized fashion to improve confidence in the results.

How do project outcomes address the need? The project team developed robust standards for accurate data, while allowing flexibility in the protocols. The project team worked to gain the Department of the Interior and community consensus on eDNA standards for public display via seven town halls, four invited reviews, and developed a process for submitting eDNA data to the USGS Nonindigenous Aquatic Species (NAS) database (U.S. Geological Survey, 2023a) for visualization through an online map viewer (fig. 8). NAS currently houses and displays visual accounts of nonindigenous species and allows for predictions of species spread and movement corridors. The expansion of the database ensures that eDNA data are FAIR, providing a findable source of data that is accessible to the public and managers alike, interoperable for use with existing and future management tools, and ensures reusability for downstream-secondary analyses, such as the development of species distribution models.
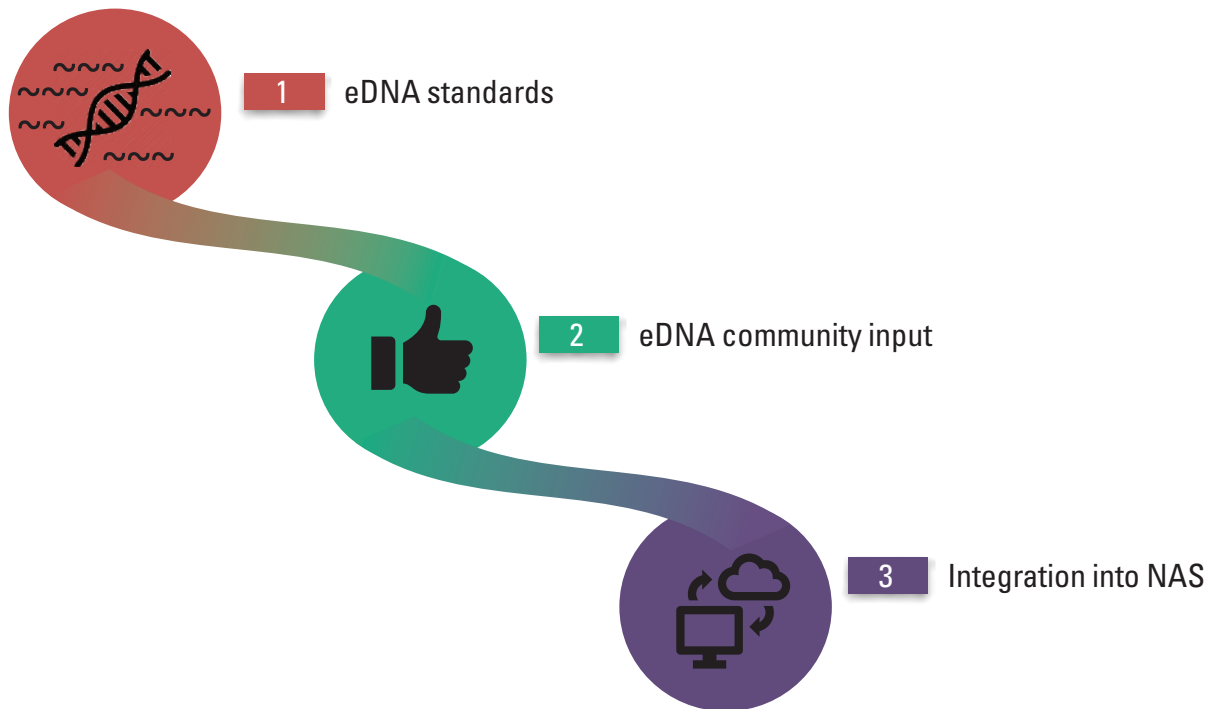

Figure 8.   Chart showing the process of illustrating the development of community consensus-derived standards for submission and public display of environmental DNA data on the U.S. Geological Survey Nonindigenous Aquatic Species database (U.S. Geological Survey, 2023a).

How can users take advantage of project outcomes?
These data can be used to inform decision-makers regarding the initiation of rapid-response efforts to invasive species, to improve the estimation of cryptic species occurrence rates, to map invasion pathways, and to improve monitoring of eradication efforts. Researchers can also use the data to inform the design of future studies. An explanation about eDNA in the NAS database is available at U.S. Geological Survey (2023a). A description of the process of developing the standard for NAS is documented in "Gaining Decision-Maker Confidence Through Community Consensus—Developing Environmental DNA Standards for Data Display on the USGS Nonindigenous Aquatic Species Database" (Ferrante and others, 2022).

• Related link—Website for eDNA in the NAS database: https://nas.er.usgs.gov/eDNA/

Applications

"Developing a 'Fire-Aware' Streamgage Network by Integrating USGS Enterprise Databases"

Katharine R. Kolb (USGS), Brian A. Ebel (USGS), Todd J. Hawbaker (USGS), Peter M. McCarthy (USGS), Sheila F. Murphy (USGS), Paul F. Steblein (USGS)

Science Support Framework elements: Applications, processing, science-project support

Product types: Data release, code repository

Significance: Provides actionable data for wildland fire-recovery efforts (fig. 9).

What need does the project address? Wildfires affect streams and rivers when they burn vegetation and scorch the ground, making floods more likely to happen and reducing water quality (Murphy and others, 2023). Timely information before and after a fire can assist public managers, first responders, fire scientists, and hydrologists with mitigating floods and planning proper water-treatment strategies.

How do project outcomes address the need? The National StreamStats Beta Application (StreamStats Team, 2022) provides a platform for users to delineate basins, compute percent-burned area in a basin, and trace an imaginary raindrop's path from the edge of a fire to the nearest water-course and then downstream (fig. 9). This project added the equations from Moody (2012). Users can now calculate additional

basin characteristics and estimate post-wildfire peak flows in streams and rivers for the Upper Colorado and Gunnison River Basins.

How can users take advantage of project outcomes?
The information is available through the National StreamStats Beta Application at StreamStats Team (2022). Underlying geographic information systems (GIS) layers are available through the data release "Basin Characteristic Layers for the Upper Colorado & Gunnison Rivers Pilot Project for StreamStats 2020" (Kolb and others, 2021), and project code is available at https://code.usgs.gov/StreamStats/clients/StreamStats-National. More information about the National Streamflow Statistics Program (NSS) is found at McCarthy (2019).

Related links

• National StreamStats Beta Application: https://streamstats.usgs.gov/national-beta/

• Basin Characteristic Layers for the Upper Colorado & Gunnison Rivers Pilot Project for StreamStats 2020: https://doi.org/10.5066/P9M46B9M

• Project code repository: https://code.usgs.gov/StreamStats/clients/StreamStats-National

• StreamStats webpage: https://www.usgs.gov/software/national-streamflow-statistics-nss-application-formerly-nss-program

"Real-Time Coastal Salinity Index for Monitoring Coastal Drought and Ecological Response to Changing Salinity Values"

Matthew D. Petkewich (USGS), Kirsten Lackstrom (Carolinas Integrated Sciences and Assessments), Bryan J. McCloskey (USGS), Andrea S. Medenblik (USGS), and Simeon Yurek (USGS)

Science Support Framework elements: Applications, analysis, data

Product types: Web link, code repository

Significance: Delivers real-time coastal drought conditions to climatologists and coastal-resource managers (fig. 10).

What need does the project address? Many coastal ecosystems are experiencing departures from historical salinity conditions due to changing land use (such as channelization and urbanization) and climate patterns (including increased frequency, severity, or duration of floods and droughts). Coastal habitats, biota, and water resources are negatively impacted by increased frequency and severity of extreme salinity-disturbance event.
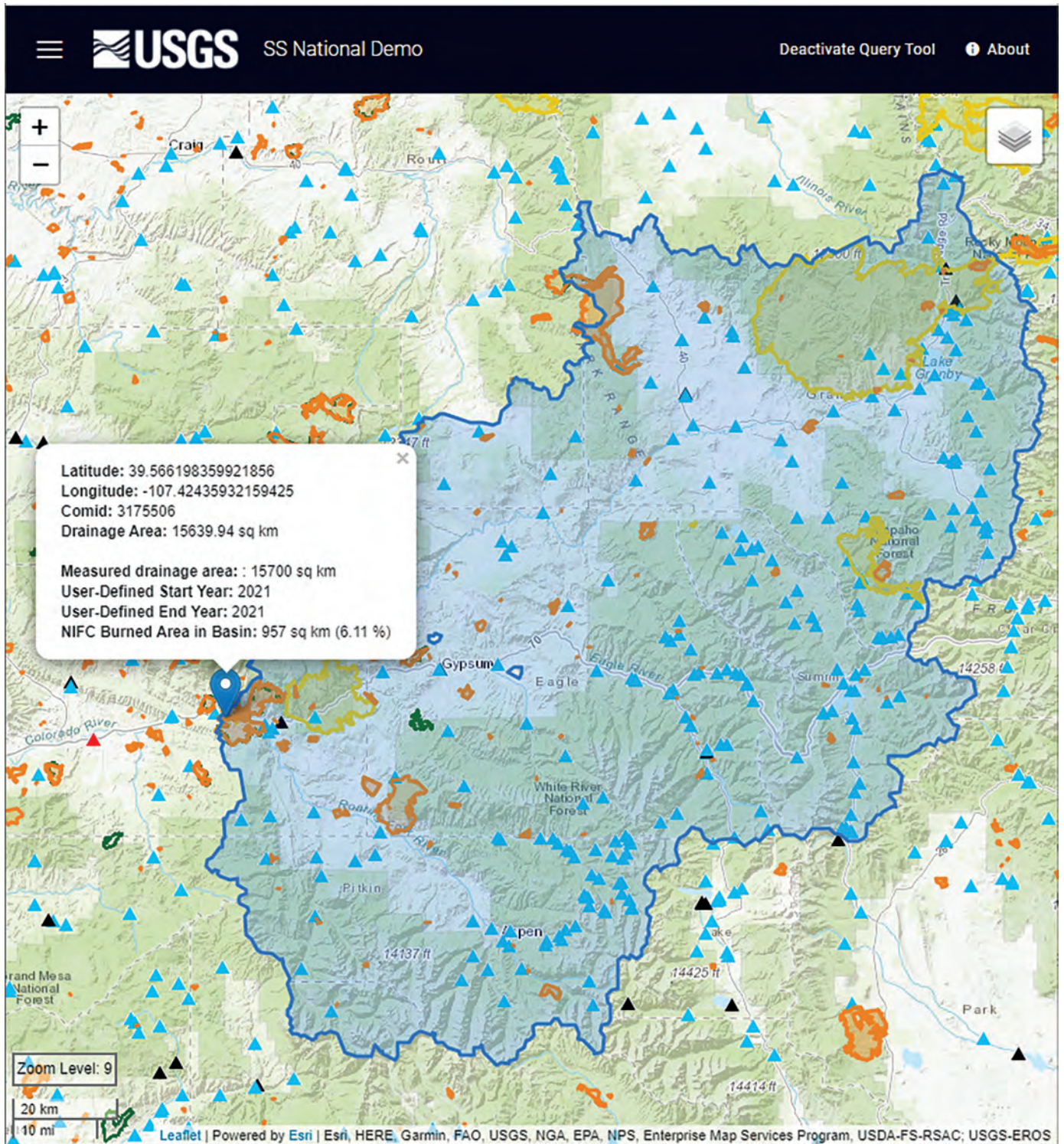
Figure 9.   Screenshot showing a sample output from the StreamStats national application fire-hydrology demo website (U.S. Geological Survey, 2019).

How do project outcomes address the need? The U.S. Geological Survey developed the Coastal Salinity Index (CSI) to identify and communicate salinity anomalies (disturbance events) through quantitative analyses of long-term salinity records. This project makes the CSI useful as a monitoring, forecasting, and decision-making tool, extending the existing web platform to enable real-time reporting of disturbance events as they unfold throughout the Eastern Atlantic and Gulf of Mexico coastlines of the United States (fig. 10). The project added the new ability to analyze spatial dynamics in salinity anomalies, which can be compared with concurrent terrestrial drought monitoring maps. Data from the USGS, the National Park Service (NPS), and the National Estuarine Research Reserve System (NERRS) gages are integrated into the website.

How can users take advantage of project outcomes? An online data portal allows real-time dissemination and graphical presentation of CSI calculations, depicting the effects of changing salinity that impact coastal estuarine ecosystems, fish habitat, and freshwater availability for municipal and industrial use. The CSI R package and associated scripts are available at McCloskey (2023).

In figure 10, various icons represent the U.S. Geological Survey (circle), the National Estuarine Research Reserves (square), and the National Park Service (triangle) salinity gages. Colors represent CSI classification, such as higher-than-normal conditions (shown in yellow, orange, red) in the Northeast United States and normal (shown in white) to lower-than-normal conditions (shown in gray and blue) in the Southeast United States and southern Florida. When selected, website pop-up boxes list the station information, CSI value, and links to the originating agency-station page, input salinity, output CSI data, and CSI graphs.

Related links

- Index web application: https://apps.usgs.gov/sawsc/csi/index.html

- Project code and scripts: https://code.usgs.gov/water/eden/CSI

"Grass-Cast: A Multi-Agency Tool Using Remote Sensing, Modeling, and On-the-Ground Science to Forecast Grassland Productivity in the Southwest"

Sasha C. Reed (USGS), William Smith (University of Arizona)

Science Support Framework elements: Applications, planning, communities of practice

Product types: Web link, publication

Significance: Provides predictions about grassland productivity to decision makers (fig. 11).

What need does the project address? Rangeland ecosystems are one of the largest single providers of agroecological services in the United States. The plant growth of these rangelands helps determine the amount of forage available for livestock and wildlife and determines the potential for fire likelihood and plant restoration success. Every spring, rangeland managers face the same difficult challenge—trying to approximate how much and where grass will be available during the upcoming growing season. Accordingly, a predictive understanding of the upcoming growing season's rangeland production could greatly support their decisions about where and at what stocking rates to allow livestock on public and private lands and for prioritizing limited resources for wildlife health, restoration efforts, and (or) fire planning.

How do project outcomes address the need? We have created a rangeland productivity forecasting tool for the southwestern United States. This tool gives intuitive visual representations of forecasts for the upcoming growing seasons' productivity (fig. 11). Data in the tool are updated every two weeks so the forecasts can gain accuracy as the growing season progresses. This tool allows public land managers, ranchers, and the public to easily access expected grassland growth to monitor conditions in other regions in the southwestern United States. The tool is being used by the Bureau of Land Management (BLM), the Bureau of Indian Affairs (BIA), the U.S. Fish and Wildlife Service (FWS), the National Resource Conservation Service (NRCS), numerous state agencies, and insurance companies.

How can users take advantage of project outcomes? The tool is free and publicly available at https://grasscast.unl.edu. The science behind the calculation of gross primary productivity is described in the paper "Satellite Solar-Induced Chlorophyll Fluorescence and Near-Infrared Reflectance Capture Complementary Aspects of Dryland Vegetation Productivity Dynamics" (Wang and others, 2022).

- Related link—GrassCast web tool: https://grasscast.unl.edu

Real-time Coastal Salinity Index Map

Figure 10.   Screenshot of the Coastal Salinity Index (CSI) website showing real-time salinity gages across the Eastern Atlantic Coast and the Gulf of Mexico of the United States (U.S. Geological Survey, 2023b).

**Percent difference in predicted summer 2023 (June–September)
ANNP compared to 36-year mean annual precipitation—
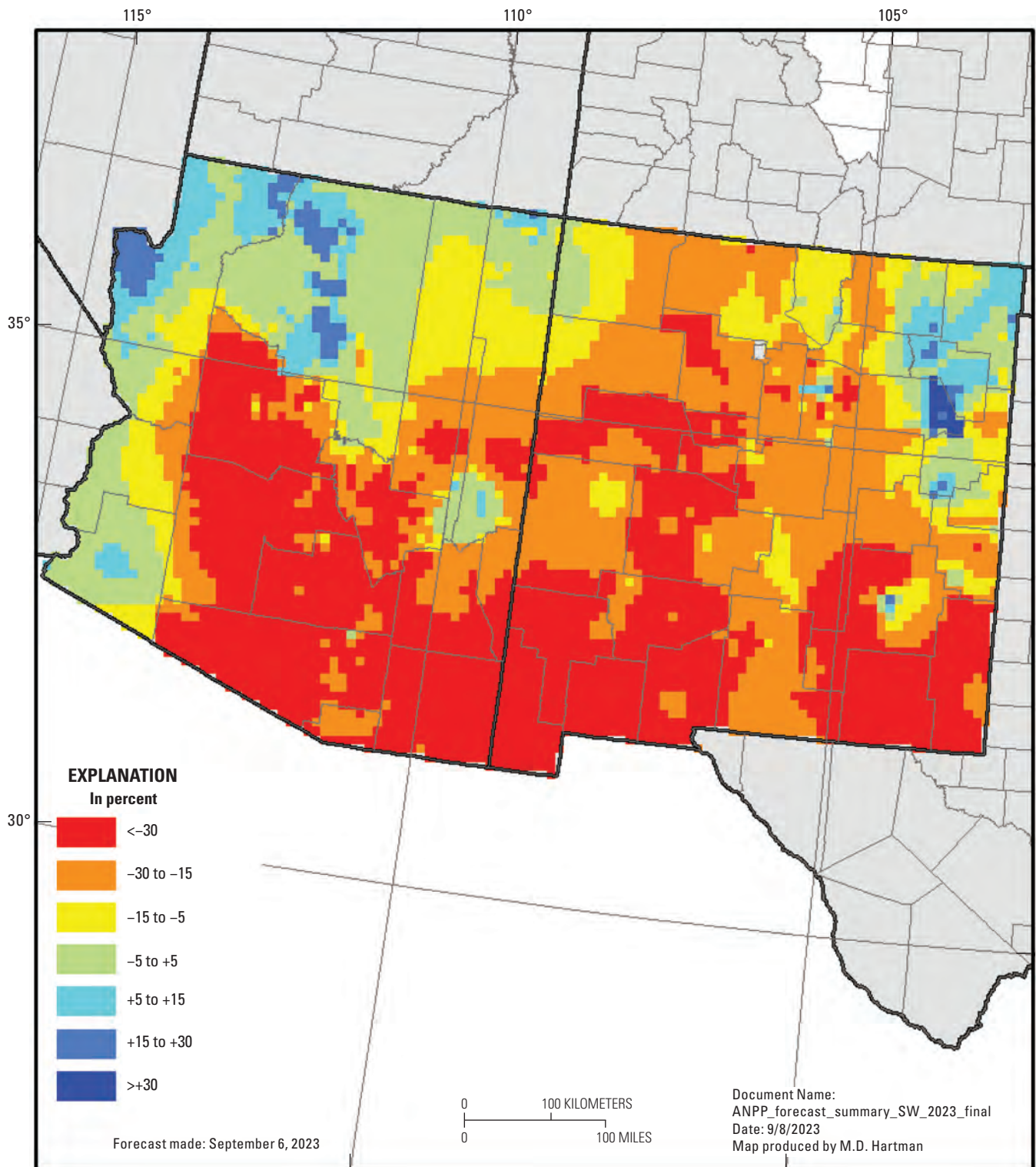Final summer 2023 forecast**

Figure 11.   Map of a GrassCast forecast of the upcoming season's productivity for Arizona and New Mexico (forecast made in September 2023). Data suggest a much below average plant-growth year compared to the mean annual productivity from 1984 to 2019.

Knowledge Management

"Using Jupyter Notebooks to Tell Data Stories and Create Reusable Workflows"

Richard A. Erickson (USGS), Edward Bulliner (USGS)

Science Support Framework elements: Knowledge management, science-project support, communities of practice

Product types: Software release, publication

Significance: Increases understanding and use of USGS data and of Pangeo (fig. 12).

What need does the project address? USGS scientists increasingly seek to share and collaborate with one another while working on data and code. Jupyter Notebooks are a tool that allow people to share data and code, and Pangeo is a platform that allows people to use Jupyter Notebooks in the USGS cloud; however, few examples exist on how to use Pangeo and no formal documentation exists for USGS scientists on how to use Pangeo.

How do project outcomes address the need? We created curated examples of the use of Jupyter Notebooks, showcasing interdisciplinary examples across the USGS (fig. 12). We also created tutorials on how to use USGS Pangeo and Jupyter Notebooks.

How can users take advantage of project outcomes? Users can utilize the Jupyter Notebook examples and tutorials to create unique data stories and document reusable workflows for their own work (Erickson and others, 2020). Erickson and others (2021) summarizes lessons from the project.

• Related link—Data stories examples and tutorials: https://doi.org/10.5066/P9NDQRX6
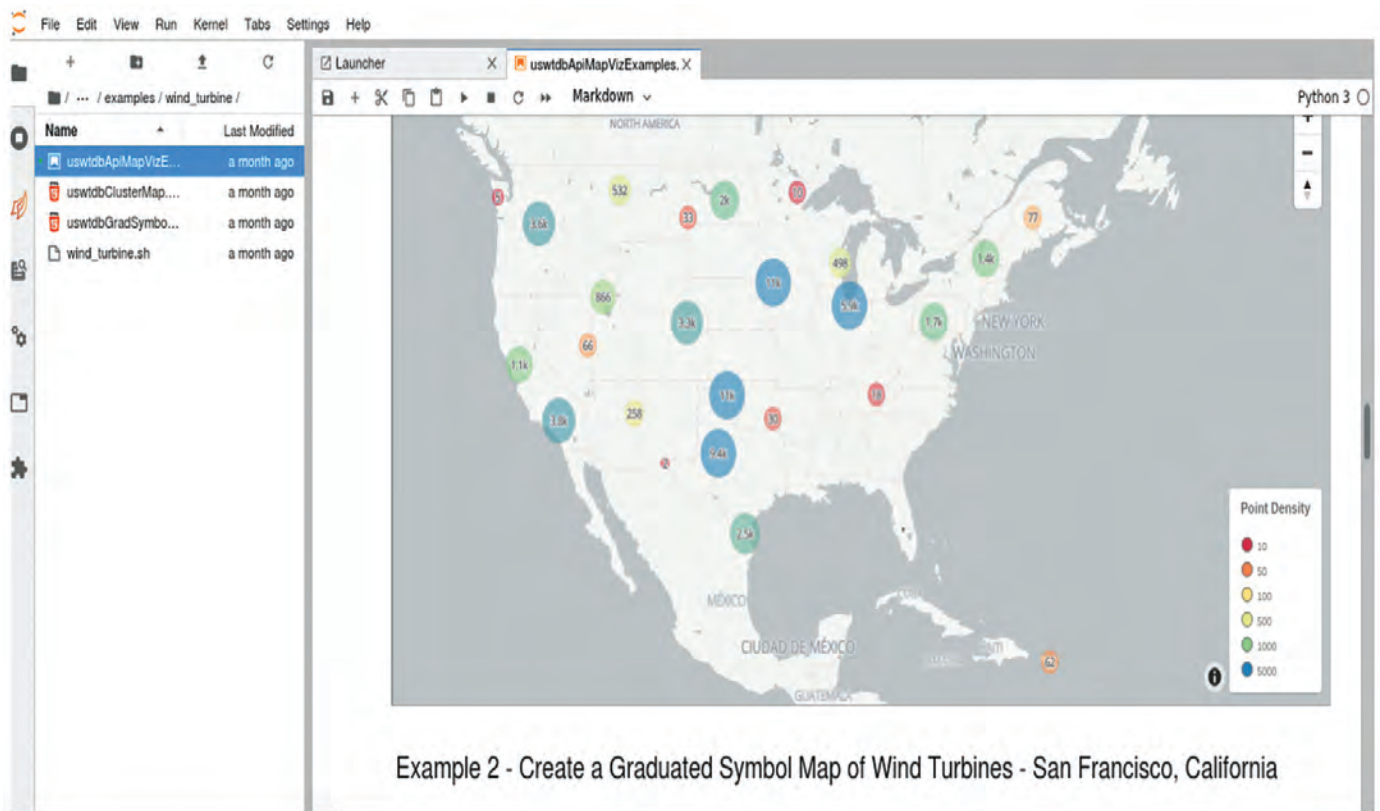

Figure 12. Example using Python code from Jupyter Notebooks to build a cluster map visualization of wind turbine locations in the conterminous United States. The Jupyter Notebook was created by Chris Garrity and released as part of a USGS software release (Erickson and others, 2020). The Notebook uses wind turbine location data from U.S. Wind Turbine Database (Hoen and others, 2018).

"USGS Cloud Environment Cookbook"

Aaron J. Fox (USGS), Caitlin M. Andrews (USGS), Travis J. Harrison (USGS)

Science Support Framework elements: Knowledge management, information, applications

Significance: Helps compile information for USGS employees to utilize the cloud environment (fig. 13)

What need does the project address? The USGS Cloud Environment Cookbook addressed the need for USGS employees to have a web space to share their experience working with a new technical environment. The goal was to provide a centralized location where it was easy for users to contribute cloud examples and use cases (fig. 13).

How do project outcomes address the need? The Cookbook provided a forum to share knowledge about working in the USGS cloud environment by using a straightforward "recipe" template and a common lightweight markup language "Markdown". The recipe for examples included guidance, such as starting with the simplest possible example, then introducing complexities one at a time, linking to other documents rather than re-explaining concepts, describing the problem and not assuming familiarity, and explaining the process rather than just the end result. It also includes guidance explaining the advantages and disadvantages of the strategy, including when it is most appropriate; mentioning alternative solutions; and providing code-bases or links necessary for reproducing the example.

How can users take advantage of project outcomes? Users could access the Cookbook to view recipes that had been submitted and find instructions for contributing their own recipes in the project's code repository. The project has been superseded by USGS Cloud Hosting Solutions site content and is no longer accessible.

Figure 13.   A schematic of the process envisioned by the Cloud Environment Cookbook to crowdsource community knowledge.

Conclusion

Community for Data Integration projects in fiscal year 2020 covered a wide variety of disciplinary and technical topics. Outputs from these projects ranged from code repositories and web applications to data releases and publications. CDI releases a request for proposals each year. CDI projects from every fiscal year beginning in 2010 can be viewed at https://www.usgs.gov/centers/community-for-data-integration-cdi/science/all-funded-projects.

Acknowledgments

References Cited

Barnhart, T.B., Sando, R., Siefken, S.A., McCarthy, P.M., and Rea, A.H., 2020, Flow-conditioned parameter grid tools: U.S. Geological Survey Software Release, version 1.0, accessed March 19, 2024, at https://doi.org/10.5066/P9W8UZ47.

Barnhart, T.B., Schultz, A.R., Siefken, S.A., Thompson, F., Welborn, T., Sando, T.R., Rea, A.H., and McCarthy, P.M., 2021, Flow-conditioned parameter grids for the contiguous United States—A pilot, seamless basin characteristic dataset: U.S. Geological Survey data release, accessed March 19, 2024, at https://doi.org/10.5066/P9HUWM6Q.

Chang, M.Y., Carlino, J., Barnes, C., Blodgett, D.L., Bock, A.R., Everette, A.L., Fernette, G.L., Flint, L.E., Gordon, J.M., Govoni, D.L., Hay, L.E., Henkel, H.S., Hines, M.K., Holl, S.L., Homer, C.G., Hutchison, V.B., Ignizio, D.A., Kern, T.J., Lightsom, F.L., Markstrom, S.L., O'Donnell, M.S., Schei, J.L., Schmid, L.A., Schoephoester, K.M., Schweitzer, P.N., Skagen, S.K., Sullivan, D.J., Talbert, C., and Warren, M.P., 2015, Community for data integration 2013 annual report: U.S. Geological Survey Open-File Report 2015–1005, 36 p., accessed March 19, 2024 at http://doi.org/10.3133/ofr20151005.

Dornbierer, J.M., Wika, S., Robison, C.J., Rouze, G.S., and Sohl, T.L., 2021a, Long-term database of historical, current, and future land cover for the Delaware River Basin (1680 through 2100): U.S. Geological Survey data release, accessed March 19, 2024, at https://doi.org/10.5066/P93J4Z2W.

Dornbierer, J., Wika, S., Robison, C., Rouze, G., and Sohl, T., 2021b, Prototyping a methodology for long-term (1680–2100) historical-to-future landscape modeling for the conterminous United States: Land, v. 10, no. 536, 31 p., accessed March 19, 2024, at https://doi.org/10.3390/land10050536.

Erickson, R.A., Bulliner, E.A., Bristol, S., Fienen, M.N., Garrity, C., Kline, K.L., Nowacki, D.J., Roberts, N.J., Burnett, J.L., and Hebert, J.L., 2020. Jupyter Data Stories: U.S. Geological Survey software release, accessed March 19, 2024, at https://doi.org/10.5066/P9NDQRX6.

Erickson, R.A., Burnett, J.L., Wiltermuth, M.T., Bulliner, E.A., and Hsu, L., 2021, Paths to computational fluency for natural resource educators, researchers, and managers: Natural Resource Modeling, v. 34, no. 3, 21 p., accessed March 19, 2024, at https://doi.org/10.1111/nrm.12318.

Faundeen, J., Burley, T.E., Carlino, J.A., Govoni, D.L., Henkel, H.S., Holl, S.L., Hutchison, V.B., Martín, E., Montgomery, E.T., Ladino, C., Tessler, S., and Zolly, L.S., 2013, The United States Geological Survey Science Data Lifecycle Model: U.S. Geological Survey Open-File Report 2013–1265, 4 p., accessed March 19, 2024, at https://doi.org/10.3133/ofr20131265.

Ferrante, J.A., Daniel, W.M., Freedman, J.A., Klymus, K.E., Neilson, M.E., Passamaneck, Y., Rees, C.B., Sepulveda, A., and Hunter, M.E., 2022, Gaining decision-maker confidence through community consensus—Developing environmental DNA standards for data display on the USGS nonindigenous aquatic species database: Management of Biological Invasions, v. 13, no. 4, p. 809–832, accessed March 19, 2024, at https://doi.org/10.3391/mbi.2022.13.4.15.

Hitt, N.P., Kessler, K.G., and Letcher, B.H., 2021a, Annotated fish imagery data for individual and species recognition with deep learning: U.S. Geological Survey data release, accessed March 19, 2024, at https://doi.org/10.5066/P9NMVL2Q.

Hitt, N.P., Rogers, K.M., Snyder, C.D., and Dolloff, C.A., 2021b, Comparison of underwater video with electrofishing and dive counts for stream fish abundance estimation: Transactions of the American Fisheries Society, v. 150, no. 1, p. 24–37, accessed March 19, 2024, at https://doi.org/10.1002/tafs.10245.

Hitt, N.P., Kessler, K.G., and Letcher, B.H., 2022. Brook trout imagery data for individual recognition with deep learning: U.S. Geological Survey data release, accessed March 19, 2024, at https://doi.org/10.5066/P94UL1Z1.

Hoen, B.D., Diffendorfer, J.E., Rand, J.T., Kramer, L.A., Garrity, C.P., and Hunt, H.E., 2018, United States wind turbine database (version 6.1, November 28, 2023): U.S. Geological Survey, American Clean Power Association, and Lawrence Berkeley National Laboratory data release, accessed March 19, 2024, at https://doi.org/10.5066/F7TX3DN0.

Jackson, P.R., 2017a, Continuous monitoring and synoptic mapping of nearshore water quality, currents, and bathymetry in Lake Michigan at 63rd Street Beach at Hyde Park, Illinois: U.S. Geological Survey data release, accessed May 20, 2024, at https://doi.org/10.5066/F75Q4T9W.

Jackson, P.R., 2017b, Continuous monitoring and synoptic mapping of nearshore water quality, currents, and bathymetry in Lake Michigan at Jeorse Park Beach near Gary, Indiana: U.S. Geological Survey data release, accessed May 20, 2024, at https://doi.org/10.5066/F7PN93V7.

Jackson, P.R., and Dupre, D.H., 2016, Three-dimensional point measurements of basic water-quality parameters in Hoover Reservoir near Westerville, Ohio, August 25 and 27, 2015: U.S. Geological Survey data release, accessed May 20, 2024, at http://dx.doi.org/10.5066/F70863D8.

Jenni, K.E., Goldhaber, M.B., Betancourt, J.L., Baron, J.S., Bristol, S., Cantrill, M., Exter, P.E., Focazio, M.J., Haines, J.W., Hay, L.E., Hsu, L., Labson, V.F., Lafferty, K.D., Ludwig, K.A., Milly, P.C.D, Morelli, T.L., Morman, S.A., Nassar, N.T., Newman, T.R., Ostroff, A.C., Read, J.S., Reed, S.C., Shapiro, C.D., Smith, R.A., Sanford, W.E., Sohl, T.L., Stets, E.G., Terando, A.J., Tillitt, D.E., Tischler, M.A., Toccalino, P.L., Wald, D.J., Waldrop, M.P., Wein, A., Weltzin, J.F., and Zimmerman, C.E., 2017, Grand challenges for integrated USGS science—A workshop report: U.S. Geological Survey Open-File Report 2017–1076, 94 p., accessed March 19, 2024, at https://doi.org/10.3133/ofr20171076.

Kolb, K.R., Rowley, T.H., and Barnhart, T.B., 2021, Basin characteristic layers for the Upper Colorado & Gunnison Rivers pilot project for StreamStats 2020: U.S. Geological Survey data release, accessed March 19, 2024, at https://doi.org/10.5066/P9M46B9M.

McCarthy, P.M., 2019, National streamflow statistics (NSS) application—formerly NSS program: U.S. Geological Survey software release, accessed March 19, 2024, at https://www.usgs.gov/software/national-streamflow-statistics-nss-application-formerly-nss-program.

McCloskey, B., 2023, Coastal salinity index R package: U.S. Geological Survey Coastal Salinity Index website, accessed March 19, 2024, at https://code.usgs.gov/water/eden/CSI.

Moody, J.A., 2012, An analytical method for predicting post-wildfire peak discharges: U.S. Geological Survey Scientific Investigations Report 2011–5236, 36 p., accessed March 19, 2024, at https://doi.org/10.3133/sir20115236.

Murphy, S.F., Alpers, C.N., Anderson, C.W., Banta, J.R., Blake, J.M., Carpenter, K.D., Clark, G.D., Clow, D.W., Hempel, L.A., Martin, D.A., Meador, M.R., Mendez, G.O., Mueller-Solger, A.B., Stewart, M.A., Payne, S.E., Peterman, C.L., Ebel, B.A., 2023, A call for strategic water-quality monitoring to advance assessment and prediction of wildfire impacts on water supplies: Frontiers in Water, 9 p., accessed March 19, 2024, at https://doi.org/10.3389/frwa.2023.1144225.

National Drought Mitigation Center, 2024, Grassland productivity forecast: Lincoln, Nebraska, National Drought Mitigation Center website, accessed February 29, 2024, at https://grasscast.unl.edu.

Peacock, J.R., Kelbert, A., Kappler, K, 2024, MT-metadata—Open-source Python package to work with magnetotelluric metadata: U.S. Geological Survey software release, accessed May 11, 2024, at https://doi.org/10.5066/P13JBD4V.

Peacock, J.R., Kappler, K., 2024, MTH5—An archivable and exchangeable HDF5 format for magnetotelluric data: U.S. Geological Survey software release, accessed May 11, 2024, at https://doi.org/10.5066/P13YMLX9.

Peacock, J.R., Frassetto, A., Kelbert, A., Egbert, G., Smirnov, M., Schultz, A.C., Kappler, K.N., Ronan, T., and Trabant, C., 2021, Metadata standards for magnetotelluric time series data: U.S. Geological Survey data release, accessed March 19, 2024, at https://doi.org/10.5066/P9AXGKEV.

Peacock, J., Kappler, K., Heagy, L., Ronan, T., Kelbert, A., and Frassetto, A., 2022, MTH5—An archive and exchangeable data format for magnetotelluric time series data: Computers & Geosciences, v. 162, 14 p., accessed March 19, 2024, at https://doi.org/10.1016/j.cageo.2022.105102.

Redoloza, F.S., Williamson, T.N., Headman, A.O., and Allred, B.J., 2023, Machine-learning model to delineate sub-surface agricultural drainage from satellite imagery: Journal of Environmental Quality, v. 52, no. 4, p. 907–921, accessed March 19, 2024, at https://doi.org/10.1002/jeq2.20493.

Ronneberger, O., Fischer, P., and Brox, T., 2015, U-Net—Convolutional networks for biomedical image segmentation *in* Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., 2015, Medical image computing and computer-assisted intervention—MICCAI 2015: Munich, Germany, October 5–9. 2015, Springer International Publishing, p. 234–241, accessed March 19, 2024, at https://doi.org/10.1007/978-3-319-24574-4_28.

Stoltz, A.D., Cravens, A.E., Herman-Mercer, N.M, and Hou, C.Y., 2023. So, you want to build a decision support tool? Assessing successes, barriers, and lessons learned for tool design and development: U.S. Geological Survey Scientific Investigations Report 2023–5076, 32 p., 1 app., accessed March 19, 2024, at https://doi.org/10.3133/sir20235076.

StreamStats Team, 2022, National streamstats beta application: U.S. Geological Survey web page, accessed March 19, 2024, at https://www.usgs.gov/tools/national-streamstats-beta-application.

U.S. Department of Agriculture, 2024, National agricultural statistics service, U.S. Department of Agriculture website, accessed May 11, 2024, at https://www.nass.usda.gov.

U.S. Geological Survey, 2023a, NAS—Nonindigenous aquatic species: U.S. Geological Survey web page, accessed March 19, 2024, at https://nas.er.usgs.gov/eDNA/.

U.S. Geological Survey, 2023b, Coastal salinity index: U.S. Geological survey web page, accessed February 24, 2021, at https://apps.usgs.gov/sawsc/csi/index.html.

Wang, X., Biederman, J.A., Knowles, J.F., Scott, R.L., Turner, A.J., Dannenberg, M.P., Köhler, P., Frankenberg, C., Litvak, M.E., Flerchinger, G.N., Law, B.E., Kwon, H., Reed, S.C., Parton, W.J., Barron-Gafford, G.A., and Smith, W.K., 2022, Satellite solar-induced chlorophyll fluorescence and near-infrared reflectance capture complementary aspects of dryland vegetation productivity dynamics: Remote Sensing of Environment, v. 270, 36 p., accessed March 19, 2024, at https://doi.org/10.1016/j.rse.2021.112858.

Williamson, T.N., and Hoefling, D.J., 2023, Machine learning with satellite imagery to document the historical transition from topographic to dense sub-surface agricultural drainage networks (tile drains): U.S. Geological Survey data release, accessed March 19, 2024, at https://doi.org/10.5066/P9KSZ382.

Zhang, Z., Liu, Q., and Wang, Y., 2018, Road extraction by deep residual U-Net: IEEE Geoscience and Remote Sensing Letters, v. 15, no. 5, p. 749–753, accessed March 19, 2024, at https://doi.org/10.1109/LGRS.2018.2802944.

Zhou, Z., Hitt, N.P., Letcher, B., Shi, W., and Li, S., 2022. Pigmentation-based visual learning for Salvelinus fontinalis individual reidentification, Proceedings of the IEEE International Conference on Big Data 2022, Osaka, Japan, December 17–20, 2022: IEEE, New York City, N.Y., p. 6850–6852, accessed March 19, 2024, at https://doi.org/10.1109/BigData55660.2022.10020966.

Glossary

Data release   A formal USGS data release that has gone through FSP review and approval.

Publication   Peer-reviewed publication (USGS series or external journal publication).

Software   Executable or compiled code that can be downloaded to your own machine.

Source code   A code repository for the project's source code.

Web application   An interactive application that runs on a web browser.

Web link   A project webpage, Wikipedia page, white paper, or online resources that do not fit other categories.

Web service   A service endpoint URL where your service can be accessed by a client application.