USGS
science for a changing world

Science Analytics and Synthesis Program

Community for Data Integration 2018 Funded Project Report

Open-File Report 2020–1062

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

Community for Data Integration 2018 Funded Project Report

By Leslie Hsu, Caitlin M. Andrews, John B. Bradford, Daniel D. Buscombe, Katherine J. Chase, Wesley M. Daniel, Jeanne M. Jones, Pam Fuller, Benjamin B. Mirus, Matthew E. Neilson, Hans W. Vraga, Jessica J. Walker, Dennis H. Walworth, Jonathan Warrick, Jake Weltzin, Daniel J. Wieferich, and Nathan J. Wood

Science Analytics and Synthesis Program

Open-File Report 2020–1062

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

U.S. Department of the Interior
DAVID BERNHARDT, Secretary

U.S. Geological Survey
James F. Reilly II, Director

U.S. Geological Survey, Reston, Virginia: 2020

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment—visit https://www.usgs.gov or call 1–888–ASK–USGS.

For an overview of USGS information products, including maps, imagery, and publications, visit https://store.usgs.gov/.

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Although this information product, for the most part, is in the public domain, it also may contain copyrighted materials as noted in the text. Permission to reproduce copyrighted items must be secured from the copyright owner.

Suggested citation:
Hsu, L., Andrews, C.M., Bradford, J.B., Buscombe, D.D., Chase, K.J., Daniel, W.M., Jones, J.M., Fuller, P., Mirus, B.B., Neilson, M.E., Vraga, H.W., Walker, J.J., Walworth, D.H., Warrick, J., Weltzin, J., Wieferich, D.J., and Wood, N.J., 2020, Community for Data Integration 2018 funded project report: U.S. Geological Survey Open-File Report 1062, 9 p., https://doi.org/10.3133/ofr20201062.

ISSN 2331-1258 (online)

Contents

Figures

Abbreviations

API	Application Programming Interface
3DEP	3D Elevation Program
CDI	Community for Data Integration
GCM	Global Circulation Model
GIS	Geographic Information Systems
ISO	International Organization for Standards
lidar	light detection and ranging
NAS	Nonindigenous Aquatic Species
USGS	U.S. Geological Survey

Community for Data Integration 2018 Funded Project Report

By Leslie Hsu,[1] Caitlin M. Andrews,[1] John B. Bradford,[1] Daniel D. Buscombe,[2] Katherine J. Chase,[1] Wesley M. Daniel,[1] Jeanne M. Jones,[1] Pam Fuller,[1] Benjamin B. Mirus,[1] Matthew E. Neilson,[1] Hans W. Vraga,[1] Jessica J. Walker,[1] Dennis H. Walworth,[1] Jonathan Warrick,[1] Jake Weltzin,[1] Daniel J. Wieferich,[1] and Nathan J. Wood[1]

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 10 projects funded in fiscal year 2018, 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 80 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 improve access to Earth and biological science data, models, and other outputs; and
- develop, organize, and share knowledge and best practices in data integration.

This report provides a summary of the 10 projects funded in 2018, outlining their goals, activities, and accomplishments.

[1]U.S. Geological Survey.

[2]Northern Arizona University.

Community for Data Integration Funded Projects, Fiscal Year 2018

The 10 projects funded in fiscal year 2018 illustrate the breadth of USGS research. Topics covered include drought risk, ice jams, landslides, light detection and ranging (lidar) processing and analysis, natural language processing, deep neural networks, non-native species, and metadata standards.

Interactive Web-Based Tool for Anticipating Long-Term Drought Risk

Principal Investigator—John B. Bradford (USGS)
Coinvestigator and partner—Caitlin M. Andrews (USGS)

Droughts are becoming more frequent and severe, and this trend is expected to continue in coming decades (Parry and others, 2007). Drought effects on natural resources include reduced water availability for plants and humans; increased insect, disease, and fire outbreaks; and increased vegetation mortality (Ostroff and others, 2017). The project team developed a publicly available online interactive tool that allows natural resource managers to access and visualize site-specific, historical, and future water availability. Users can set site qualities that affect the ability of a place to retain and use water, including soil texture and vegetation composition. With these inputs, as well as site-specific historical and future climate data for 11 different global circulation models (GCMs) and two representative concentration pathways, ecohydrology simulations are performed on the fly by the SOILWAT2 model on an Amazon Web Services EC2 (Elastic Compute Cloud). Users are presented with raw and synthesized data in the form of figures that allow for the exploration of long-term historical perspectives to recognize their site's natural range of variability and future projections of 21st century drought, so they can quantify their site's risk of novel drought exposure (fig. 1).

Related links—

- *See project page at* https://www.sciencebase.gov/catalog/item/5acd21aae4b0e2c2dd155dea
- *GitHub repository for the Dryland Ecology Research Group:* https://github.com/DrylandEcology

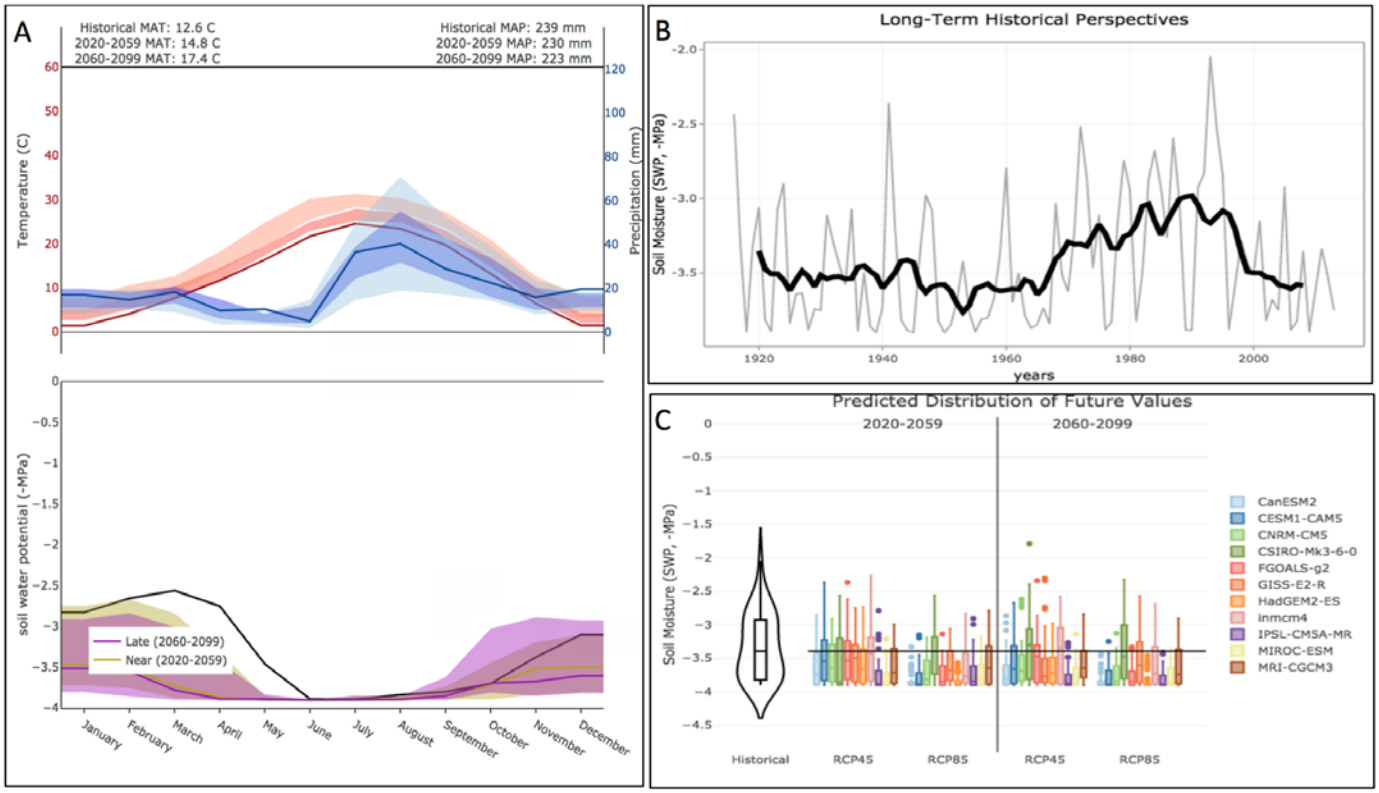

Figure 1. Example simulation output figures from the Long-Term Drought Visualizer representing long-term historical and future patterns of climate and water availability. (*A*) Annual mean patterns for three time periods (historical (1974–2013) (line), near future (2020-2059) (narrow band), long-term future (2060–2099) (wide band), with futures represented as a range (min and max) of variability across 11 global circulation models (GCMs) for representative concentration pathway 85 (RCP85). Variables shown are temperature (red) and precipitation (blue) in the upper panel and soil moisture (purple, yellow) in the lower panel. (*B*) Long-term annual historical (1916–2013) time series of water availability (annual points in light grey and a 10-year rolling mean in black). (*C*) Distributions of annual future water availability for each GCM (colored data) and RCP, compared to the historical values (black).

ICE! Ice Jam Hazard Mobile-Friendly Website

Principal Investigator—Katherine J. Chase (USGS)

Coinvestigators and partners—Hans W. Vraga, Lauren M. Privette, and Jeremy K. Newson (USGS)

Ice jams are formed when floating ice accumulates on rivers behind natural or human-made features and are a major hazard because of their potential to significantly reduce the flow of the river and cause flooding. The project team worked with the U.S. Army Corps of Engineers, National Weather Service, Silver Jackets (an interagency team working toward reducing flood risk), and USGS stakeholders to develop a mobile-friendly prototype of an ice jam hazard website and reporting system. The prototype (fig. 2) shows how ice jam conditions can be recorded nationwide. The public can view and download ice jam information. Historic ice jam locations and frequencies, as well as information on potentially hazardous developing ice jams, are all valuable data. Given the science, modeling, and hazard warning potential provided by these data, continued development of this system is widely supported by stakeholders and partners.

The prototype system consists of: (1) an Angular-Material framework JavaScript client hosted on Amazon Web Services (AWS) Simple Storage Service, (2) .Net Core web services deployed using containerization on AWS Fargate, and (3) an AWS Relational Database Service PostgreSQL instance. This cloud-first system enriches the collective USGS experience in working with the latest cloud technologies.

Related links—

- *See project page at:* https://www.sciencebase.gov/catalog/item/5b9198e5e4b0702d0e808b76
- *Website*: https://test.wim.usgs.gov/icejams/
- *Web Services Documentation*: https://test.wim.usgs.gov/icejamsservices/#/
- *Client code repository*: https://github.com/USGS-WiM/ice-jams
- *Services code repository*: https://github.com/USGS-WiM/IceJamsservices

National Alert Risk Mapper for Nonindigenous Aquatic Species

Principal Investigator—Pam Fuller (USGS)

Co-investigators and partners—Ian A. Pfingsten (USGS), Wesley M. Daniel (USGS), Matthew E. Neilson (USGS), Dana Infante (Michigan State University), James Ballard (Gulf States Marine Fisheries Commission), and Arthur Cooper (Michigan State University)

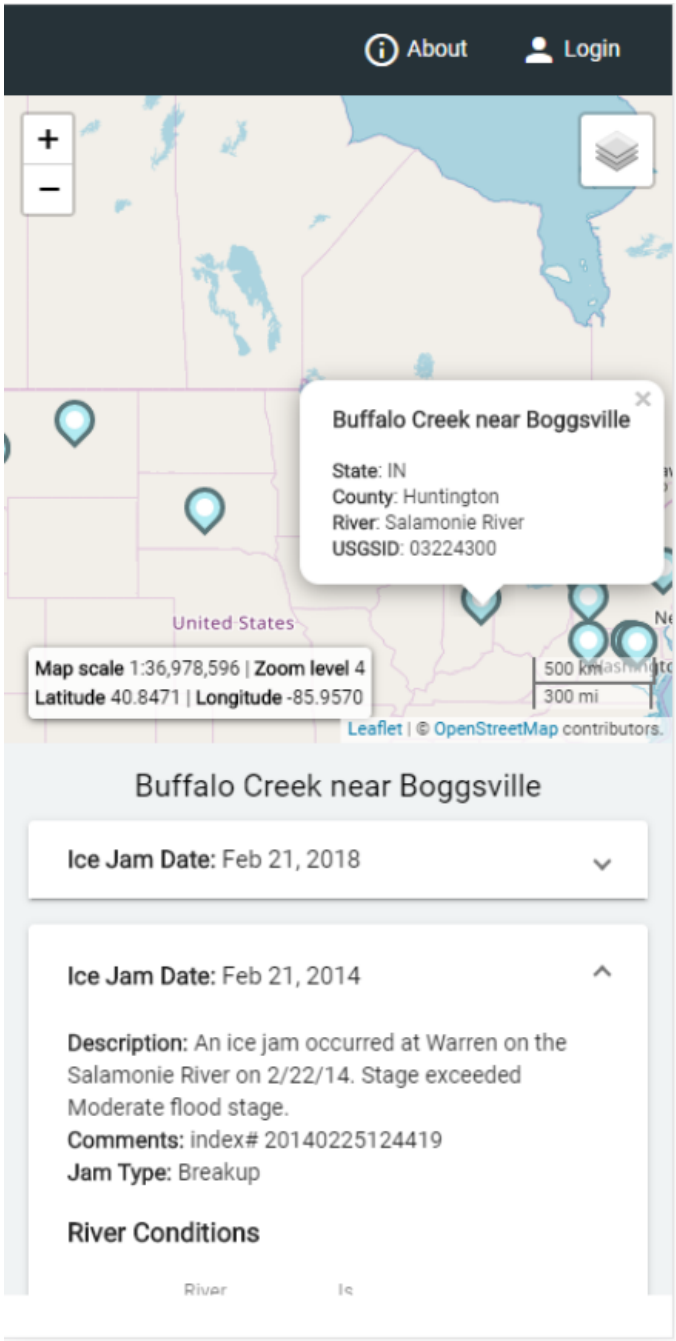

Figure 2. Screen capture of the mobile user interface for the Ice Jams system for the northern area of the United States, showing information for a site in Indiana.

The Nonindigenous Aquatic Species (NAS) Database and Alert System provide a framework for the rapid dissemination of new invasions as they are incorporated into the NAS Database. The system notifies registered users of new sightings of non-native species as part of national-scale early detection and rapid response systems. Originally NAS alerts included only a simple text description of the sighting location with no indication of new areas at risk of invasion. The NAS group requested funding from the FY18 Community for Data Integration group to improve a pilot project, the Alert Risk Mapper (ARM) tool. This tool was used to develop maps which characterize the potential water bodies at risk from a new non-native species sighting. The Alert Risk Mapper tool (fig. 3) was enhanced by expanding its geographical extent to the entire contiguous United States and Hawaii, along with automating portions of the mapmaking process to improve its effectiveness as a tool for early detection and rapid response.

Related links—

- *See project page at*: https://www.sciencebase.gov/catalog/item/5acd257ae4b0e2c2dd155df5
- *Nondindigenous Aquatic Species Alert Risk Mapper*: https://nas.er.usgs.gov/AlertSystem/default.aspx

Integrating Disparate Spatial Datasets from Local to National Scale for Web-Based Visualization and Analysis: A Case Study Compiling U.S. Landslide Inventories

Principal Investigator—Benjamin Mirus (USGS)

Co-investigators and partners—Rex L. Baum (USGS), Robert G. Schmitt (USGS), Eric S. Jones (USGS), Jeremy Lancaster (California Geological Survey), Stephen Slaughter (Washington State Department of Natural Resources), Dalia Kirschbaum (NASA), Matthew Crawford (Kentucky Geological Survey), and William Burns (Oregon Department of Geology and Mineral Industries)

Spatial data on landslide occurrence across the United States varies greatly in quality, accessibility, and extent. This problem of data variability is common across USGS Mission Areas; it presents an obstacle to developing national-scale products and to identifying areas with relatively good or bad data coverage. The project team compiled available data of known landslides into a national-scale, searchable online map (fig. 4), which greatly increases public access to landslide hazard information. Additionally, the project held a workshop with landslide researchers and practitioners and sought broader input on the topic from the CDI community. Based on recommendations, the project investigators identified a limited subset of essential attributes for inclusion in its product. The investigators also defined a quantitative metric for confidence in data quality and developed simple scripts to assign confidence to each landslide in the inventory. This approach can be used to inform other national-scale products

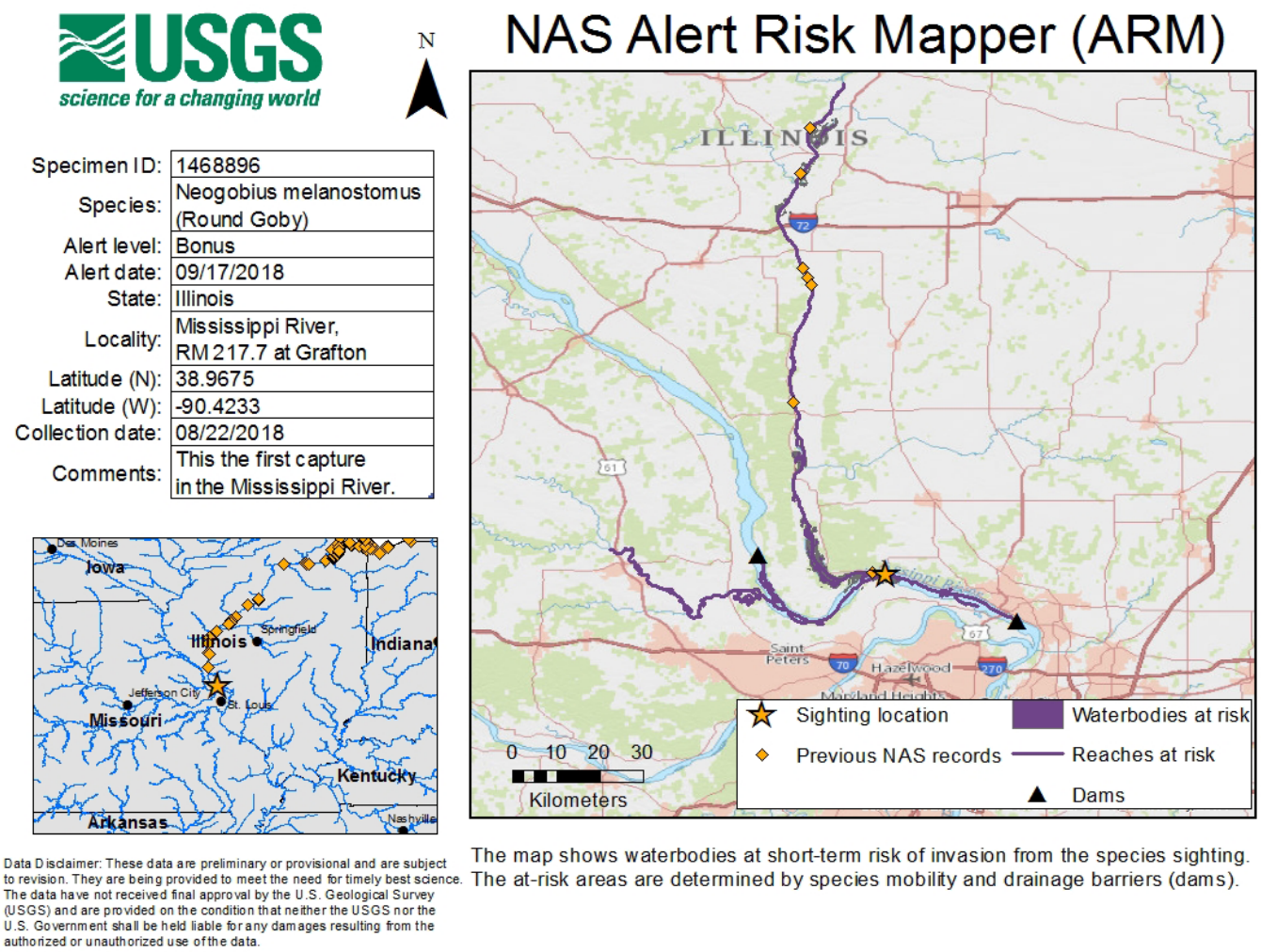

Figure 3. Waterbodies at short-term risk of invasion after a new non-native species sighting of Neogobius melanostomas (Round Goby) on the Mississippi River, north of St. Louis, MO. The at-risk areas are determined by species mobility and drainage barriers (dams).

to be compiled from disparate spatial datasets, particularly for projects with limited resources allocated to data management and maintenance.

Related links—

- *See project page at*: https://www.sciencebase.gov/catalog/item/5acd2600e4b0e2c2dd155dfa
- *Web application*: https://www.usgs.gov/maps/national-landslides-map-and-data
- *Publication*: https://doi.org/10.1007/s10346-020-01424-4 (Mirus and others, 2020)
- *Data Release*: https://doi.org/10.5066/P9E2A37P (Jones and others, 2019)

Knowledge Extraction Algorithms (KEA): Turning Literature into Data

Principal Investigator—Matthew E. Neilson (USGS)

Co-investigators and partners—Daniel J. Wieferich (USGS), Shanan Peters (University of Wisconsin Madison), and Wesley M. Daniel (USGS)

Identifying, extracting, and mobilizing information from recent and historical literature is a time-consuming part of organizing and collating synthetic data productions. This project explored the use of algorithm-based methods to identify and extract occurrence information from the GeoDeepDive literature database to support upkeep of the Nonindigenous Aquatic Species (NAS) database (fig. 5). The GeoDeepDive API (application programming interface) was extended to include query capabilities on terms from the Integrated Taxonomic Information System (ITIS). This functionality helped support identification of

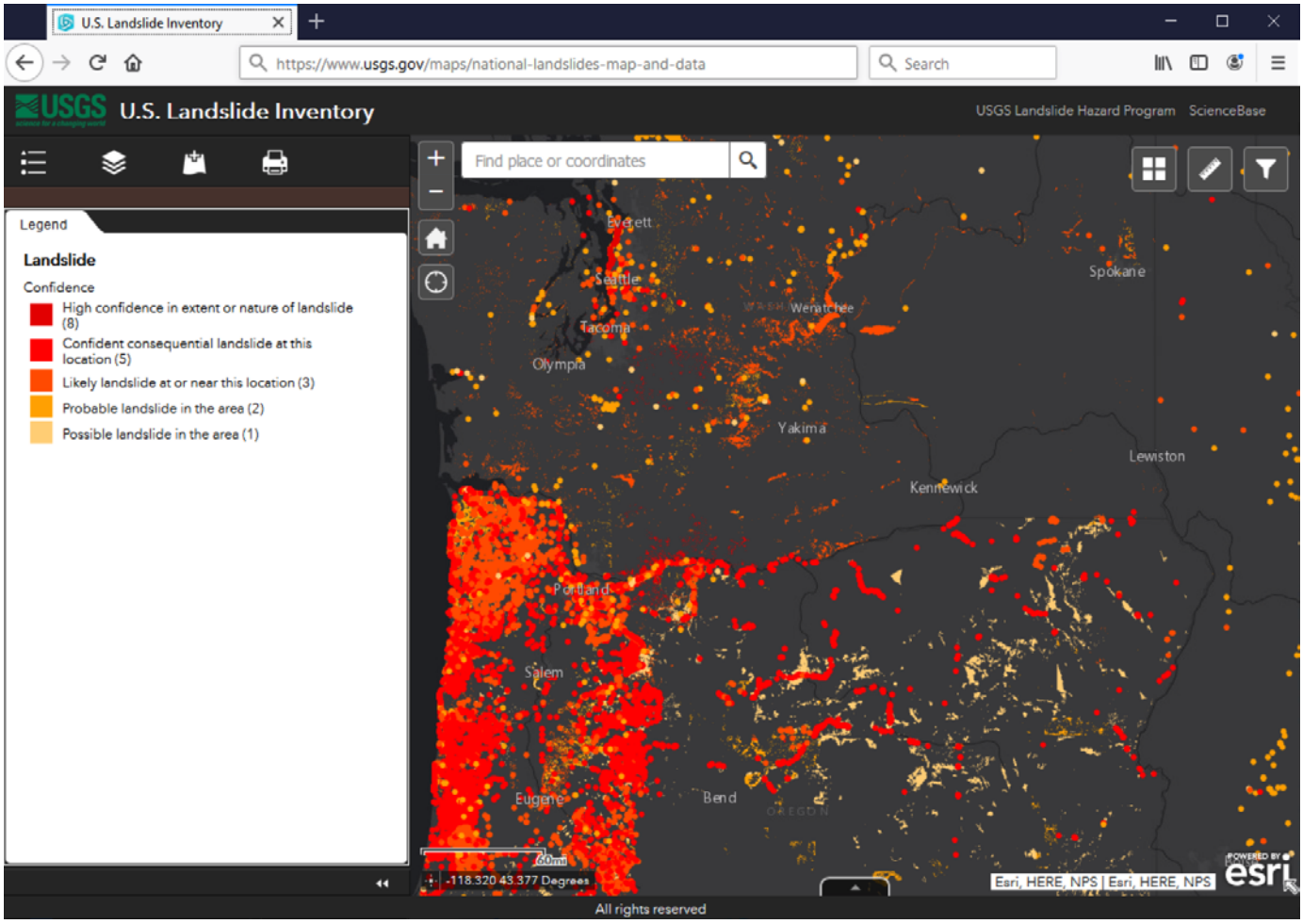

Figure 4. Screenshot of the U.S. Landslide Inventory web application showing landslides in Washington and Oregon.

literature mentioning and (or) focusing on species that are tracked by the Nonindigenous Aquatic Species Database. These methods were paired with algorithms to extract location information associated with mentions of specific terms. Efforts are underway to continue improving these algorithms and workflow.

Related links—

- *See project page at*: https://www.sciencebase.gov/catalog/item/5acd2680e4b0e2c2dd155dfd
- *GeoDeepDive public API endpoint describing the dictionary of terms from ITIS*: https://geodeepdive.org/api/dictionaries?dict=ITIS
- *GeoDeepDive public API endpoint for searching terms, including from ITIS*: https://geodeepdive.org/api/terms

Investigation of Lidar Data Processing and Analysis in the Cloud

Principal Investigator—Jessica J. Walker (USGS)

Co-investigators and partners—Harold House and Dionne Zoanni (USGS)

Lower technical and financial barriers have led to a proliferation of lidar point-cloud datasets acquired to support diverse USGS projects (fig. 6). The objective of this effort was to implement an open-source, cloud-based solution through USGS Cloud Hosting Solutions (CHS) that would address the needs of the growing USGS lidar community. The project team proposed to allow users to upload point-cloud datasets to CHS-administered Amazon Web Services storage where open-source packages Entwine and Potree would provide visualization and manipulation via a local web browser. This functionality for individual datasets would mirror services available for USGS 3D Elevation Program (3DEP) data at the time of this project. After the project team determined that the software packages Entwine and Potree could not satisfy internal technical and security requirements, it explored alternative possibilities to securely and readily transfer, process, share,

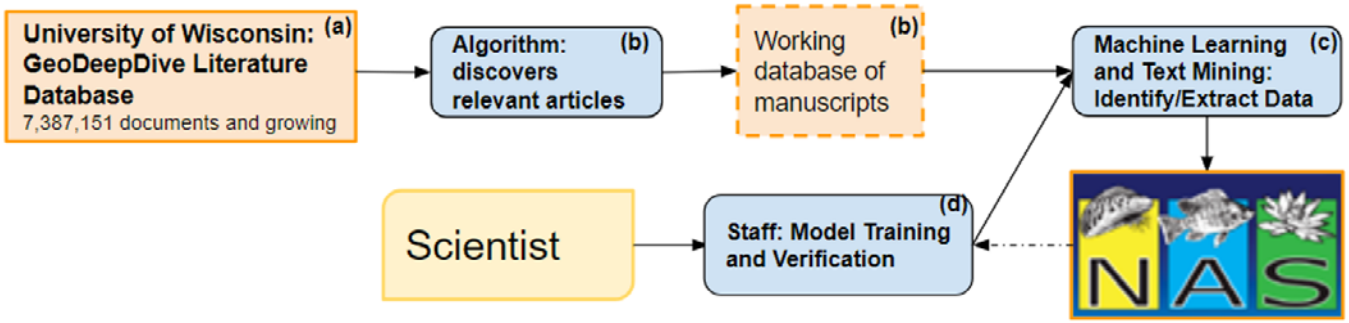

Figure 5. Workflow for Knowledge Extraction Algorithms (KEA). The orange-outlined rectangles are databases, the blue rectangles are action steps, and the yellow rectangle represents the scientist-user. First, relevant articles that potentially have species occurrence data are extracted from the large GeoDeepDive literature database into a smaller database that can be acted on. Next, machine learning and text mining identify and extract the specific passages with data that could be incorporated into the Nonindigenous Aquatic Species (NAS) database. An iterative process of model training and verification by a human scientist refines the algorithms before data is extracted to the NAS database.

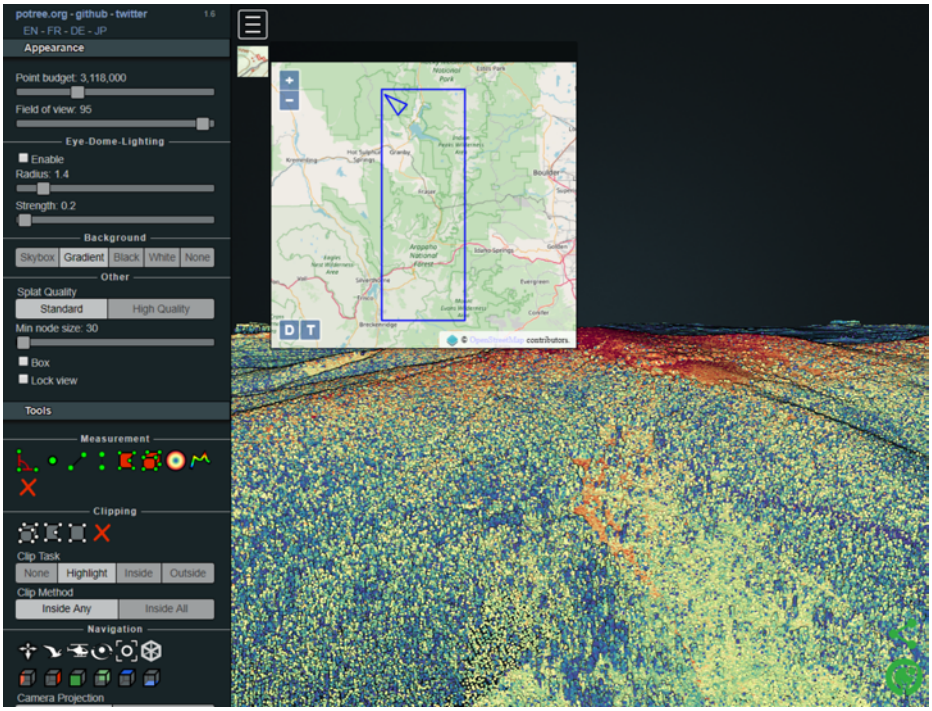

Figure 6. Web browser displaying USGS 3DEP (3D Elevation Program) point-cloud data acquired over Grand County, Colo. The triangle in the 2D overview map indicates the viewer's perspective – the viewer is looking outward from the vertex shared by the two equal sides of the triangle toward the remaining side. The colors represent the elevation of each point, from low (blue) to high (red). The lidar data are stored as a public dataset on Amazon Web Services and piped through open-source packages Potree and Entwine (accessible through https://usgs.entwine.io). The combination demonstrates the potential of cloud-based and open-source solutions for the efficient and rapid display of large-scale lidar datasets.

and visualize large-volume point-cloud datasets via the cloud. The project team summarized the lessons learned throughout the course of the project for lidar practitioners in the USGS.

Related links—

- *See project page at*: https://www.sciencebase.gov/catalog/item/5b919e87e4b0702d0e808b9d

Content Specifications to Enable USGS Transition to ISO Metadata Standard

Principal Investigator—Dennis Walworth and Frances Lightsom (USGS)

Co-investigators and partners—Josh Bradley (Fish and Wildlife Service), Tara Bell (USGS), Andrew LaMotte (USGS), Jennifer Walter (USGS), Lisa Zolly (USGS), Sophie Hou (National Center for Atmospheric Research), Anna Milan (National Oceanographic and Atmospheric Administration)

International metadata standards known collectively as ISO 19115, where ISO stands for the International Organization for Standardization, are an option for describing and documenting USGS data. The open-ended nature of the ISO standard provides great flexibility and vocabulary to describe research products. However, that flexibility means that there are few constraints to guide authors to standardized, robust documentation across the bureau. This project brought together members of the USGS data community to develop diverse, modular, content specifications to define standard USGS ISO metadata content requirements (fig. 7). Using these content specifications, metadata authors will be guided to include appropriate metadata fields for the type of data being described, producing complete, high-quality metadata while also using the flexibility and modern features of the ISO standard. As a step toward this goal, this project convened a workshop of data specialists from across the bureau and proposed initial content specification modules to aid in authoring metadata records in the international standard. These reusable modules are Basic (required by all USGS metadata), Geospatial, Biological, and Lineage.

Related links—

- *See project page at*: https://www.sciencebase.gov/catalog/item/5acd27a0e4b0e2c2dd155e01
- *Content specification modules repository*: https://doi.org/10.5066/P99A3JNQ

Mapping Land-use, Hazard Vulnerability, and Habitat Suitability Using Deep Neural Networks

Principal Investigator—Jonathan Warrick (USGS)

Co-investigators and partners—Daniel D. Buscombe (Northern Arizona University), Paul E. Grams (USGS), Jenna A. Brown (USGS), and Christopher R. Sherwood (USGS)

Deep learning is a computer analysis technique inspired by the human brain's ability to learn. The computer analysis involves using several layers of artificial neural networks to learn and subsequently recognize patterns in data, forming the basis of many state-of-the-art applications from self-driving cars to drug discovery and cancer detection. Deep neural networks can learn many levels of abstraction, and outperform many other types of automated classification algorithms. This project developed software tools, resources, and two training workshops that allow USGS scientists to apply deep learning to remotely sensed imagery and better understand natural hazards and habitats across the Nation (fig. 8). The tools and training resources are available from GitHub, and their use has been the subject of several publications.

Related links—

- *See project page at*: https://www.sciencebase.gov/catalog/item/5acd2923e4b0e2c2dd155e09
- *For Deep Learning tools and example data*: https://github.com/dbuscombe-usgs/dl_tools
- *Workshop background information*: https://github.com/dbuscombe-usgs/cdi_dl_workshop

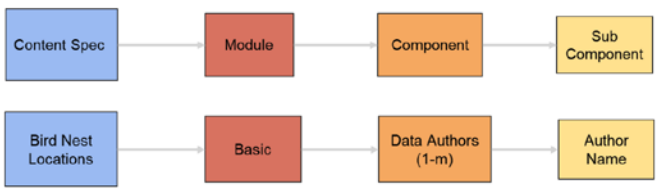

Figure 7. Conceptual architecture of a content specification. The top row shows the conceptual architecture of a content specification. The bottom row shows an example of a content specification defined for bird nest observational data. The bird nest location specification includes a basic module which requires at least one author component. The author component requires an author name as a sub-component.

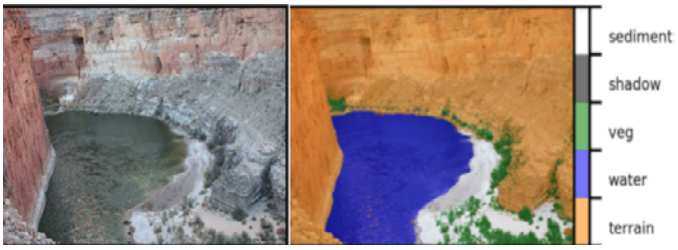

Figure 8. Example of Deep Neural Network classification using the techniques provided in the Community for Data Integration-funded workshop. The image on the left is an oblique aerial photograph of a section of the Colorado River in Grand Canyon at South Canyon (51 km downstream of Lees Ferry, AZ), the image on the right is a segmentation of that photo into areas of sediment, shadow, vegetation, water, and terrain.

Workflows to Support Integrated Predictive Science Capacity: Forecasting Invasive Species for Natural Resource Planning and Risk Assessment

Principal Investigator—Jake Weltzin (USGS)

Co-investigators and partners—Theresa M. Crimmins (University of Arizona), Katharine L. Gerst (University of Arizona), R. Lee Marsh (University of Arizona), Erin E. Posthumus (University of Arizona), Alyssa H. Rosemartin (University of Arizona), Jeff Switzer (University of Arizona), and Toni Lyn Morelli (USGS)

Insect pests cost billions of dollars per year globally, negatively impacting food crops and infrastructure and contributing to the spread of disease (Crimmins and others, 2020). Timely information regarding developmental stages of pests can facilitate early detection and control, increasing efficiency and effectiveness of management interventions. To address this need, the USA National Phenology Network created a suite of Pheno Forecast map products relevant to science and management. Pheno Forecasts indicate, for a specified day, the status of the insect's target life cycle stage in real time across the contiguous United States (fig. 9). These risk maps enhance decision-making and short-term planning by both natural resource managers and members of the public. The approach for stakeholder engagement and the technical workflows developed to generate the Pheno Forecast maps are replicable and readily transferrable within USGS. The National Phenology Network staff sought input and critique of the maps and delivery from end users over the course of product and tool development and used this input to shape the products.

Related links—

- *See project page at:* https://www.sciencebase.gov/catalog/item/5acd27b3e4b0e2c2dd155e03
- *Information sheet:* https://pubs.er.usgs.gov/publication/70204926
- *Publication:* https://doi.org/10.1093/aesa/saz026 (Crimmins and others, 2020)
- *Pheno Forecasts Web page:* https://www.usanpn.org/data/forecasts
- *Web service documentation:* https://github.com/usa-npn/npn-geo-services
- *Workflow to generate, cache, and manage raster Pheno Forecast maps*: https://github.com/usa-npn/gridded_models

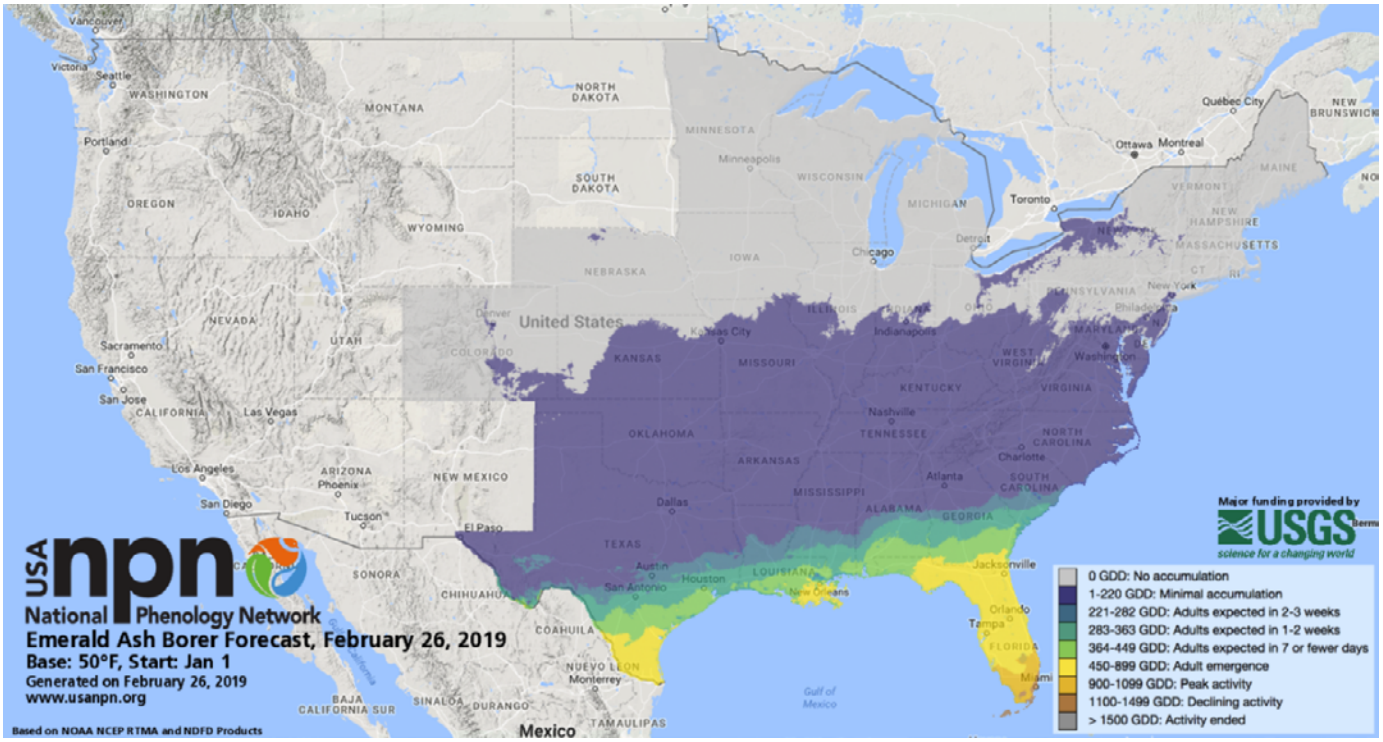

Figure 9. Pheno Forecast map for emerald ash borer for February 26, 2019. Colors indicate the status of adult emergence. The status of a location is determined by comparing the local Growing Degree Day (GDD) accumulation to a published heat accumulation threshold for the life cycle stage.

Risk Map

Principal Investigator—Nathan J. Wood (USGS)

Co-investigators and partners—Jeanne M. Jones, Kevin D. Henry, Jason T. Sherba, and Peter Ng (USGS)

The CDI Risk Map Project developed modular tools and services to benefit a wide group of scientists and managers that deal with various aspects of risk research and planning. Risk is the potential that exposure to a hazard will lead to a negative consequence to an asset such as human or natural resources. This project builds on a Department of the Interior project that is developing geospatial layers and other analytical results that visualize multihazard exposure to various DOI assets (Wood and others, 2019). The CDI Risk Map team developed the following tools: a spatial database of hazards and assets, an API (application programming interface) to query the data, web services with Geoserver (an open-source geospatial server), and a modular map viewer and related infographics using the open source visualization framework TerriaJS (fig. 10).

Related links—

- *See project page at*: https://www.sciencebase.gov/catalog/item/5b91a0c2e4b0702d0e808bb2

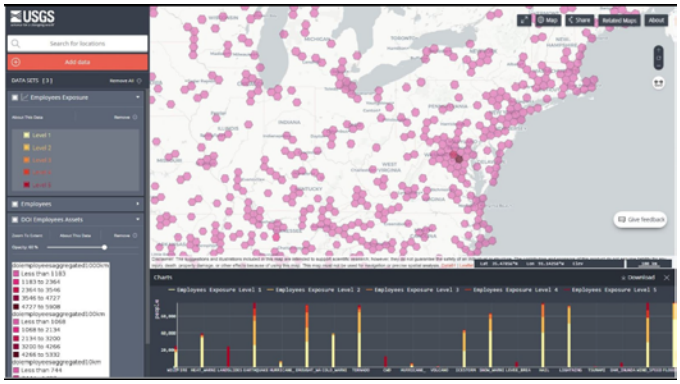

Figure 10. A screenshot of the risk map and infographics built on TerriaJS. The large map area is a heatmap of Department of the Interior employee assets. The bar chart in the lower part of the figure shows exposure of the asset to different hazards such as earthquakes, droughts, and tornadoes. The left-most area is an explanation of the data including the color scale for the different data displayed.

References Cited

Crimmins, T.M., Gerst, K.L., Huerta, D.G., Marsh, R.L., Posthumus, E.E., Rosemartin, A.H., Switzer, J., Weltzin, J.F., Coop, L., Dietschler, N., Herms, D.A., Limbu, S., Trotter, R.T., and Whitmore, M., 2020, Short-term forecasts of insect phenology inform pest management: Annals of the Entomological Society of America, v. 113, no. 2, 10 p., accessed March 2020, at https://doi.org/10.1093/aesa/saz026.

Jones, E.S., Mirus, B.B., Schmitt, R.G., Baum, R.L., Burns, W.J., Crawford, M., Godt, J.W., Kirschbaum, D.B., Lancaster, J.T., Lindsey, K.O., McCoy, K.E., Slaughter, S., and Stanley, T.A., 2019, Map of landslide inventories across the United States: U.S. Geological Survey data release, accessed March 2020, at https://doi.org/10.5066/P9E2A37P.

Mirus, B.B., Jones, E., Baum, R.L., Godt, J.W., Schmitt, R., Burns, W.J., Crawford, M., Kirschbaum, D., Lancaster, J., Lindsey, K.O., McCoy, K., Slaughter, S., and Stanley, T., 2020, Landslides across the United States—Occurrence, Susceptibility, and Data Limitations: Landslides, accessed May 29, 2020 at https://doi.org/10.1007/s10346-020-01424-4.

Ostroff, A.C., Muhlfeld, C.C., Lambert, P.M., Booth, N.L., Carter, S.L., Stoker, J.M., and Focazio, M.J., 2017, USGS integrated drought science: U.S. Geological Survey Circular 1430, 24 p., accessed April 16, 2020, at https://doi.org/10.3133/cir1430.

Parry, M.L., Canziani, O.F., Palutikof, J.P., van der Linden, P.J., and Hanson, C.E., eds., 2007, Climate Change 2007—Impacts, adaptation and vulnerability—Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change: Cambridge, UK, Cambridge University Press, 982 p.

Wood, N., Pennaz, A., Ludwig, K., Jones, J., Henry, K., Sherba, J., Ng, P., Marineau, J., and Juskie, J., 2019, Assessing hazards and risks at the Department of the Interior—A workshop report: U.S. Geological Survey Circular 1453, 42 p., accessed March 2020, at https://doi.org/10.3133/cir1453.

ISSN 2331-1258 (online)
https://doi.org/10.3133/ofr20201062