Open Access Datasets from Federal Government Agencies

Suzanne Reinman, Government Documents suzanne.reinman@okstate.edu

Oklahoma State University Library

Finding U.S. Government Open Access Data Sets

- Data and data sets published by U.S. government agencies are open access and are available to researchers to use
 - Learn the key access points for finding federal data

Open Data

Open Data

- Open data is data that can be freely used, re-used and redistributed by anyone subject only, at most, to the requirement to attribute and sharealike.
 Cite as any other resource
 - Available and easily accessed
 - Re-use and redistribution permitted
 - Universal participation
 - Interoperability (between computer systems)
 - Structured data that is machine-readable (CSV, XML, JSON, etc.)
 - Produced by governments
 - Select research data is also open

Why Open Data

- Open data, especially open government data, is a tremendous resource that is largely underutilized.
- Governments produce a great amount of data and government data is public by law making it open and made available for others to use.

FAIR Principles

FAIR principles

- To improve the Findability, Accessibility, Interoperability, and Reuse of digital assets
 - o In 2016, the 'FAIR Guiding Principles for scientific data management and stewardship' were published in Scientific Data.
 - O The ultimate goal of FAIR is to optimise the reuse of data
 - o www.go-fair.org/fair-principles/

FAIR principles for data are increasingly being applied to government data sources

U.S. Government Data and Statistics

- The U.S. Government is the largest publisher in the world and produces primary source data and statistics
 - Federal information is open and free and without copyright
- Researchers/information professionals: identify sets that may be useful for research – reuse instead of creating data sets on your own

Open Federal Data

- Data results from U.S. agency missions, research within the agency, or research that they fund through grants.
 - The 2013 Federal Open Data Policy and 2018 Open Government Data Act require federal agencies to publish their information online as open data, using standardized, machinereadable data formats, with their metadata included in the Data.gov catalog
 - Federally funded research has specific data publishing requirements

Making government more open and supporting scientific discovery

https://www.digitalgov.gov/open-data-policy-m-13-13/ https://data.gov/open-gov/

Directories for open government (global) data

Open Data Repositories

- There is not one comprehensive directory for all open data
 - *U.S. government agency sources included but not comprehensive
- *Open Access Directory Data Repositories

OAD

- oad.simmons.edu/oadwiki/Data repositories
- List of repositories and databases for open data by subject
- *Registry of Research Data Repositories
 - re3data.org
 - Global registry of research data repositories



List of Open Government Data Sites (Global) - Wikipedia

Azerbaijan

Bangladesh

Bahrain

Belarus

Belgium

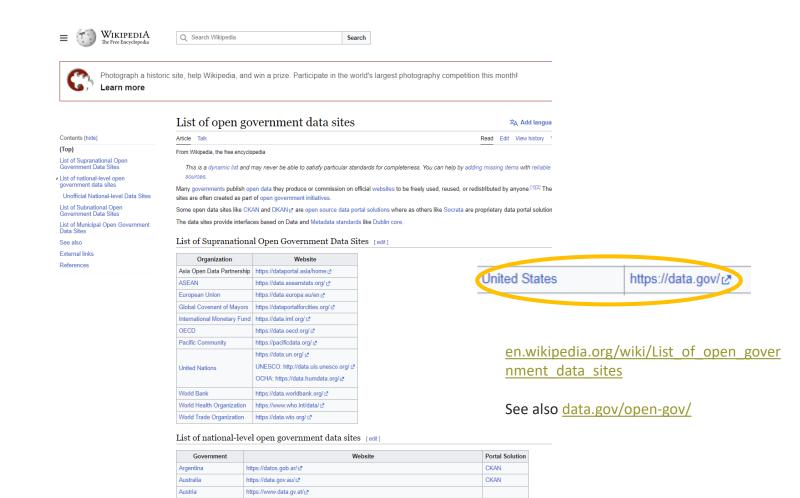
https://www.opendata.az/₽

https://www.data.gov.bh/ ≥

http://data.gov.bd/ 28

https://opendata.by/₽

http://data.gov.be/ 2 Bosnia and Herzegovina https://cbbh.ba/Content/Read/1133#₽ https://dados.gov.br/ r₹ https://www.data.gov.bn/r2



Custom

Directories for open *university* research and data

Research Data Repositories

- Researchers upload research/data, DOI assigned
- Limited content available for reuse, dependent on reuse license
- Mostly for universities/academic community
- Do not include government
- Figshare, Dryad, Open Science Framework, institutional repositories etc.
- Federal agency repositories for data/research from federally sponsored or internal research

Finding U.S. Government Open Data Sets

Locating Federal Open Access Data Sets

- Data.gov (Start here, will not include all sets)
- Agency websites
 - USA.gov
 - Data catalogs
 - Statistical agencies
- Sites by Subject
 - Science.gov
- Agency repositories
- Google Dataset Search

Data.gov



DATA



Data.gov is the United States government's open data website.

For details on Data.gov's new home page redesign, please click here.

Open Government

Data.gov is the federal government's open data site, and aims to make government more open and accountable.

Opening government data increases citizen participation in government, creates opportunities for economic development, and informs decision making in both the private and public sectors.

Data.gov implements The OPEN Government Data Act (Title II of the Foundations for Evidence-Based Policymaking Act of 2018, Public Law 115-435). The OPEN Government Data Act makes Data.gov a requirement in statute, rather than a policy. It requires federal agencies to publish their information online as open data, using standardized, machine-readable data formats, with their metadata included in the Data.gov catalog.

Data.gov is working with an expanded group of federal agencies to include their datasets in Data.gov as they implement the law. In addition, the law requires that GSA work with the Office of Management and Budget and the Office of Government Information Services to establish an "online repository of tools, best practices, and schema standards to facilitate the adoption of open data practices across the Federal Government." That repository can be found at resources.data.gov.

Data.gov was built with open source software. Anyone, especially local, state, and foreign governments are welcome to borrow the code behind Data.gov.

Open Data in the United States

Numerous states, cities, and counties have launched open data sites by collaborating with these non-federal data sources, Data.gov is able to include this data in the catalog. Searches on the Data.gov catalog will return relevant datasets from both federal and non-federal sources.



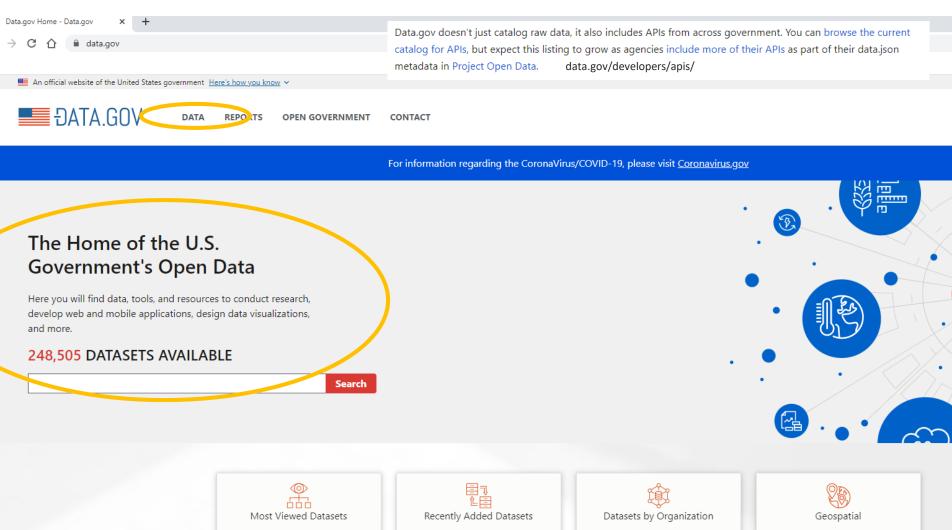
Download the full list of current international open data sites: CSV.

Download the full list of current open data sites in the United States: CSV.

data.ok.gov

Data.gov

Data.gov is the United States government's open data website. It provides access to datasets published by agencies across the federal government. Data.gov is intended to provide access to government open data to the public, achieve agency missions, drive innovation, fuel economic activity, and uphold the ideals of an open and transparent government.

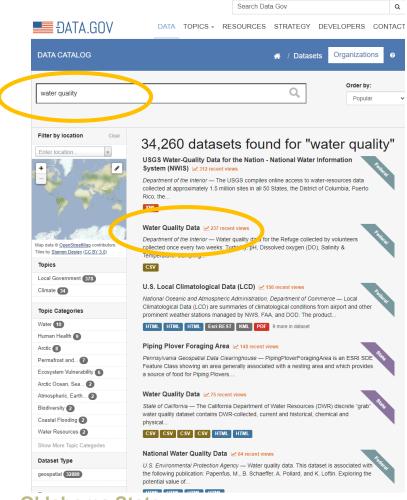


Oklahoma State University

Data.gov

Water quality

Topic Categories, Organizations



Search on main page or Data tab in menu

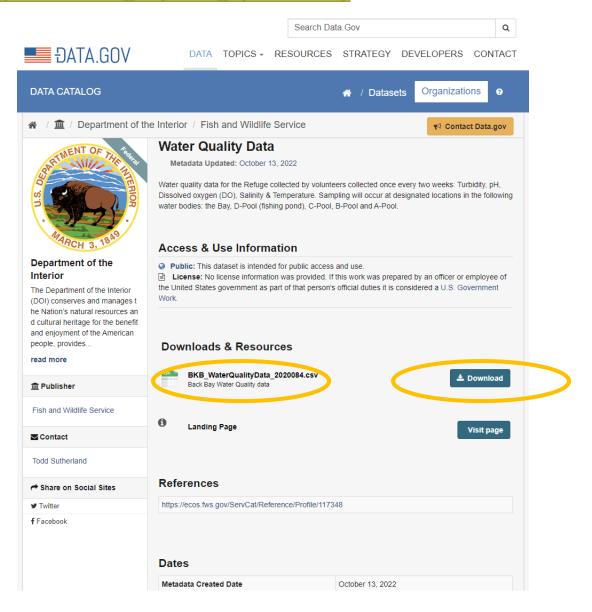
Filter by:

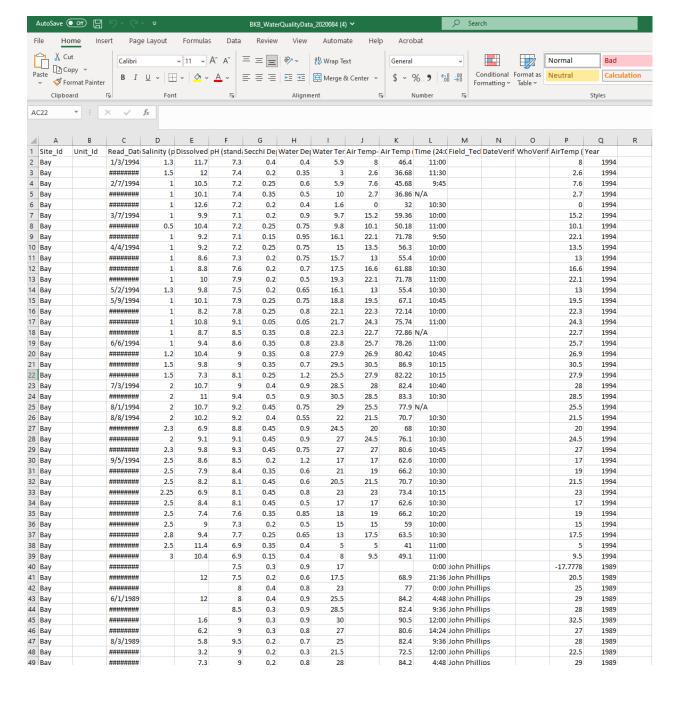
- Topic Categories
- File formats
- Organization (Agency)

Oklahoma State University

Back Bay National Wildlife Refuge

https://www.fws.gov/refuge/back-bay





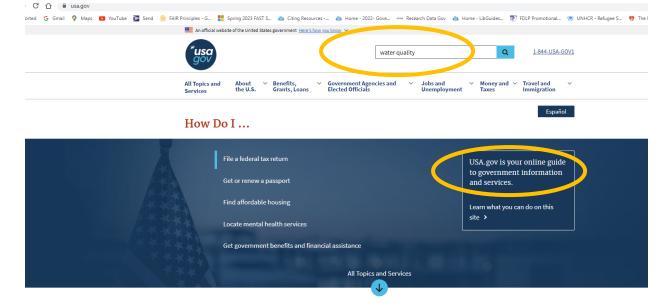
Agency websites

USA.gov

- Identify federal agencies related to your research area
 - Review their websites, data catalog, statistical agency

Agencies are including a Data section on their websites

USA.gov



U.S. Government Services and Information



Groundwater Quality | U.S. Geological Survey

of observation wells exists to ...

www.usgs.gov/special-topics/water-science-school/science/groundwater-quality

The U.S. Geological Survey (USGS) is involved in monitoring the Nation's groundwater supplies. A national network

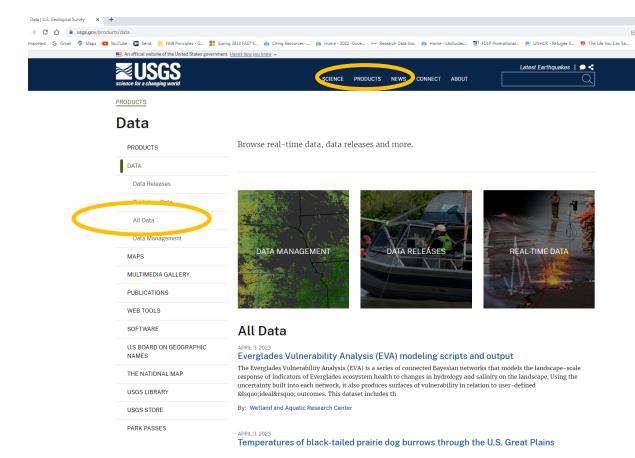
U.S. Geologial Survey (USGS) usgs.gov

Science agency for the U.S. Department of Interior

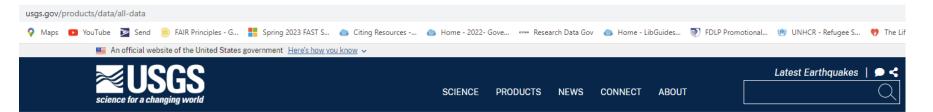
- The Nation's largest water, earth, and biological science and civilian mapping agency.
 Collects, monitor, analyze, and provide science about natural resource conditions, issues, and problems
- Products -> All Data -> Topics
- Top page -> Real-time Data Information

Also includes:

- Water Resources of the United States
 - https://water.usgs.gov/index.html
- Water Data for the Nation
 - https://waterdata.usgs.gov/nwis/
- Water Quality Data portal (USGS and EPA)
 - o https://www.waterqualitydata.us/

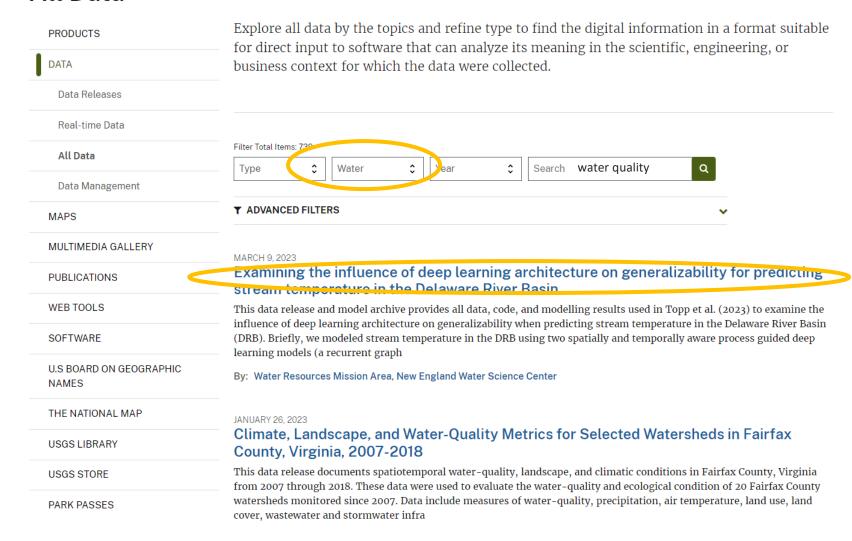


See Products -> All data



PRODUCTS DATA

All Data



Examining the influence of deep learning architecture on generalizability for predicting stream temperature in the Delaware River Basin



Dates

Publication Date : 2023-03-07 Start Date : 1980-04-01 End Date : 2019-12-31

Citation

Topp, S.N., Barclay, J.R., Diaz, J.A., Sun, A.Y., Jia, X., Lu, D., Sadler, J.M., and Appling, A.P., 2023, Examining the influence of deep learning architecture on generalizability for predicting stream temperature in the Delaware River Basin: U.S. Geological Survey data release, https://doi.org/10.5066/P9HU7BLR.

Summary

This data release and model archive provides all data, code, and modelling results used in Topp et al. (2023) to examine the influence of deep learning architecture on generalizability when predicting stream temperature in the Delaware River Basin (DRB). Briefly, we modeled stream temperature in the DRB using two spatially and temporally aware process guided deep learning models (a recurrent graph convolution network - RGCN, and a temporal convolution graph model - Graph WaveNet). The associated manuscript explores how the architectural differences between the two models influence how they learn spatial and temporal relationships, and how those learned relationships influence a model's ability to accurately predict stream temperature as domains shift towards out-of-bounds conditions. This data release and model archive contains three zipped folders for 1) Data Preparation, 2) Modelling Code, and 3) Model Predictions. Instructions for running data preparation code and modelling code can be found in the README.md files in 01 Data Prep and 02 Model Code respectively.

Contacts

Point of Contact: Simon N Topp

Originator: Simon N Topp, Janet R Barclay, Jeremy A Diaz, Alexander Y. Sun, Xiaowei Jia, Dan Lu,

Jeffrey M Sadler, Alison P Appling

Metadata Contact: Simon N Topp

Publisher: U.S. Geological Survey

Distributor: U.S. Geological Survey - ScienceBase

SDC Data Owner: Integrated Information Dissemination Division

USGS Mission Area: Water Resources

Attached Files 🛶

Click on title to download individual files attached to this item or download all files listed below as a compressed file.





application/fgdc+xml

Map »



Spatial Services

ScienceBase WMS:

https://www.sciencebase.gov/catal

Communities

USGS Data Release Products **

Tags

Categories : Data

Harvest Set: USGS Science Data Catalog (SDC)

Theme: climate change, domain shift, environment, explainable AI (XAI), generalization, hybrid modeling, inlandWaters, machine learning, modeling, process guided deep learning, stream temperature, temperature, water, water resources

Place: DE, Delaware, MD, Maryland, NJ, NY, New Jersey, New York, PA, Pennsylvania, US, United States

USGS Scientific Topic Keyword : Hydrology, Water Quality, Water Resources

Types: Map Service, OGC WFS Layer, OGC WMS Layer, OGC WMS Service

Provenance

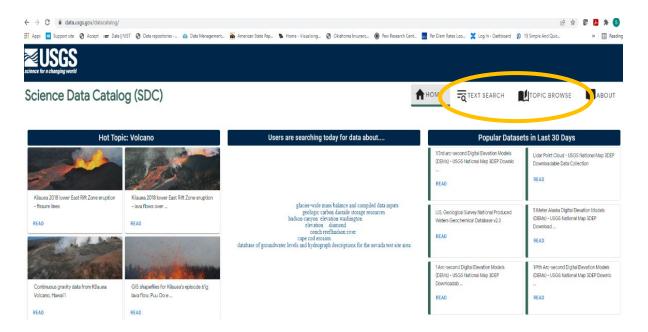
DATA CATALOGS

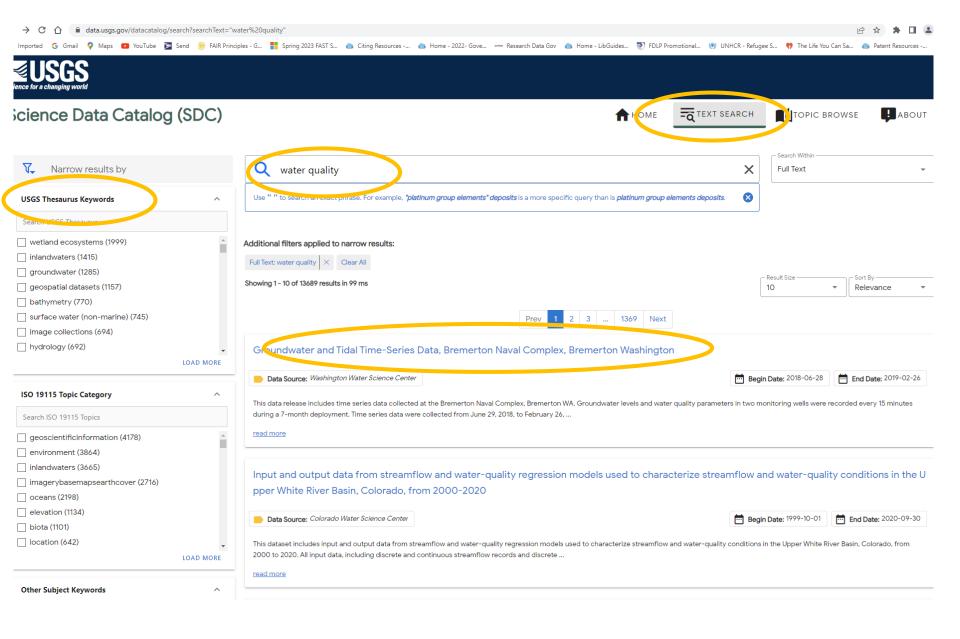
Agency data catalogs in addition to the data on their websites

Data Catalog URL Stem	Agency
data.boem.gov	Bureau of Ocean Energy Management (Interior)
data.bsee.gov	Bureau of Safety and Environmental Enforcement (Interior)
data.cdc.gov	Centers for Disease Control and Prevention (Health and Human Services)
data.census.gov	Census Bureau. This is the replacement for American FactFinder (Commerce)
data.cms.gov	Centers for Medicare & Medicaid Services (Health and Human Services)
data.commerce.gov	Commerce Department
data.defense.gov	Defense Department
data.doi.gov	Interior Department
data.faa.gov	Federal Aviation Administration (Transportation)
data.fdic.gov	Federal Deposit Insurance Corporation (Independent)
data.globalchange.gov	U.S. Global Change Research Program (13 Government Agencies)
data.healthcare.gov	Managed by Centers for Disease Control and Prevention (Health and Human Services)
data.hrsa.gov	Health Resources & Services Administration (Health and Human Services)
data.hud.gov	Housing and Urban Development
data.medicaid.gov	Centers for Disease Control and Prevention (Health and Human Services)
data.medicare.gov	Centers for Disease Control and Prevention (Health and Human Services)
data.nasa.gov	National Aeronautics and Space Administration (Independent)
data.nationalservice.gov	Corporation for National & Community Service (Independent)
data.nist.gov	National Institute of Standards and Technology (Commerce)
data.noaa.gov	National Oceanic and Atmospheric Administration (Commerce)
data.nrel.gov	National Renewable Energy Laboratory (Energy)
data.usgs.gov	United States Geological Survey (Interior)

Data Catalog - USGS Science Data Catalog

- Data.usgs.gov
- Provides seamless access to USGS research and monitoring data from across the nation.
 - Text Search
 - Topic Browse -> Water Resources -> Water Quality

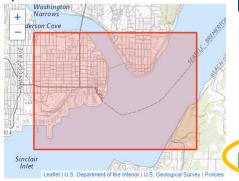




Science Data Catalog (SDC)



Spatial Extent of Data



USGS Data Source

Washington Water Science Center

ISO 19115 Topic Category

<u>health</u>

Place Keywords

water quality, tidal

Groundwater and Tidal Time-Series Data, Bremerton Naval Complex, Bremerton Washington

This data release includes time series data collected at the Bremerton Naval Complex, Bremerton WA. Groundwater levels and water quality parameters in two monitoring wells were recorded every 15 minutes during a 7-month deployment. Time series data were collected from June 29, 2018, to February 26, 2019. Field deployment details and quality assurance methods are included in the following paragraphs. Groundwater monitoring well MW-709 was monitored using two data loggers. The first data logger is a non-vented pressure transducer (In-Situ Rugged TROLL 100) that was deployed resting on the well bottom (BOT). The well measurement point (MP) elevation is 16.86 feet above NAVD88 with a total depth from MP of 30.00 feet and the transducer rested at -14.14 feet below NAVD88. Data collected at this location is referenced by National Water Information System (NWIS) site ID 473324122382202. Temperature and depth were recorded every 15-mintes for the entire deployment. Time-series data from this...

read more

⊈ GET DATA AND METADATA

Author(s)	Opatz, Chad C D Peterson, Norman T Conn, Kathleen E G Huffman, Raegan L
Publication Date	2019-08-30
Beginning Date of Data	2018-06-28
End Date of Data	2019-02-26
Data Contact	CHAD C. OPATZ
DOI	https://doi.org/10.5066/P9IKD5FW
Citation	Opatz, C. C., Peterson, N. T., Conn, K. E., & Huffman, R. L. (2019). Groundwater and Tidal Time-Series Data, Bremerton Naval Complex, Bremerton Washington [Data set]. U.S. Geological Survey. https://doi.org/10.5066/P9IKD5FW
Metadata Contact	CHAD C. OPATZ
Metadata Date	2022-08-29
Related Publication	
Views of this Record	2
Citations of these data	
Rights	public

VIEW FULL METADATA AS ORIGINAL FGDC CSDGM

Sites by Subject

Sites by Subject

- Sites that combine agency data and research in a <u>subject.gov</u> format
- Determine and review subject sites:
 - Science.gov
 - EarthData
 - https://www.earthdata.nasa.gov/
 - Climate.gov

Science.gov



Science.gov searches over 60 databases and over 2,200 scientific websites to provide users with access to more than 200 million pages of authoritative federal science information including research and lovelopment results.

What's New:

View latest announcements from federal agencies that support open science initiatives, programs, and events



U.S. Federal Science Agencies' Public Access Plans

How To Submit Research Papers to Funding Agencies

Inventory of Agency APIs and Other Services for Public Access Collections

Science.gov is governed by the interagency Science.gov Alliance. Participating agencies are:















About

STEM Opportunities





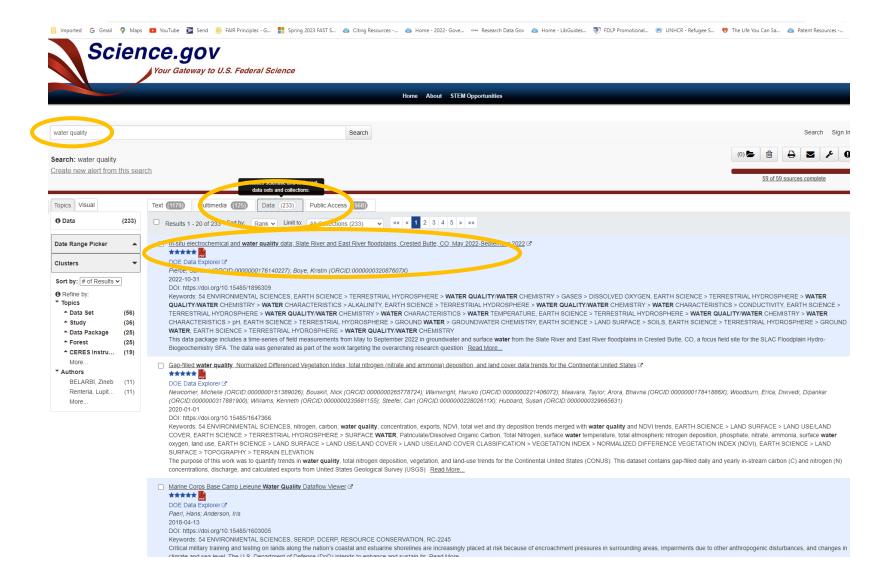








- Keyword search
- Refine by:
 - Topic
 - Category(Advanced Search)
 - Data in results



Agency Repositories

Repositories

- Investigators/researchers funded by federal agencies are required by public access policies to deposit data and research in agency repositories
- Article and Data Sharing Requirements by Federal Agencies
 - See U.S. federal funder requirements for sharing research articles and research data
 - datasharing.sparcopen.org/
 - See also <u>science.gov/publicAccess.html</u> for public access plans for research funded by the Federal Government
 - Ongoing development
 - Repositories are included in:

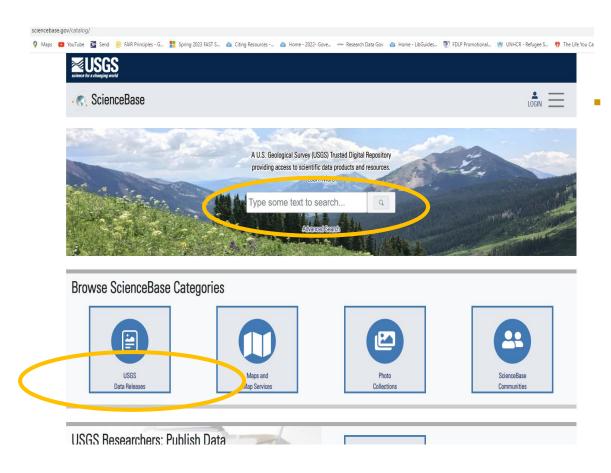


Data repositories for **Federally Funded Research**

- Select Repositories
 - USGS ScienceBase USGS
 - Ag Data Commons USDA
 - DOE Data Explorer DOE

Create account for access

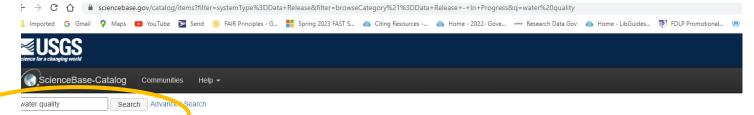
USGS ScienceBase



Digital repository and collaborative data management platform providing public access to machine readable data and metadata. ScienceBase also serves as a cross-program institutional repository, ensuring long-term stewardship of the results of federally-funded research, when a more specific Program-level data repository is not available.

sciencebase.gov/catalog/

See Data Releases



filters: System Type: Data Release (X) > Categories. NOT Data Release - In Progress (X)

Filters

Date Range

- Past year (605)
- · Past month (87) Past week (16)
- Past day (0)
- · Custom range...

Extensions

- · Shapefile (68)
- · ArcGIS Service Definition
- Raster (12)
- · Citation (3)
- · NetCDF OPeNDAP Service
- OGC Web Service (2)
- · ArcGIS REST Service (1)

Types

- Map Service (529)
- OGC WMS Layer (503)
- OGC WFS Laver (502)
- · OGC WMS Service (443)
- Citation (167)
- More...

Contacts

- . U.S. Geological Survey -ScienceBase
- Water Resources
- . U.S. Geological Survey
- · U.S. Geological Survey
- Ecosystems
- More...

Categories

- Data (1191)
- · Data Release Revised (74)
- · Publication (3)
- pre-SM502.8 (3)
- Data Release Under

Quality-Assurance and Quality-Control Data for Discrete Water-Quality Samples Collected in McHenry County, Illinois, 2020

In June and July of 2020, 45 groundwater wells in McHenry County, Illinois, were sampled for water quality (field properties, major ions, nutrients, and trace metals) and 12 wells were sampled for contaminants of emerging concern (pharmaceuticals, pesticides, and wastewater indicator compounds). Quality-assurance and quality-control samples were collected during the June and July 2020 sampling that included equipment blanks, field blanks, and replicates. The results of these samples were used to understand the sources of bias and variability associated with sample collection, processing, storage, and shipping. This data release contains one comma separated values files containing the results of the quality-control.

Categories: Data; Tags: Illinois, McHenry County, USGS Science Data Catalog (SDC), United States, Water Quality, All tags...

Uinta Basin Ground and Surface Water Quality Records

This database includes water quality records compiled at the request of the Bureau of Land Management for the Uinta Basin in Utah and Colorado. All data contained within this database are publically available via the United States Geological Survey's National Water Information System (NWIS) database (available at https://waterdata.usgs.gov/nwis). Data collected and provided by the states of Utah and Colorado were also analyzed, however, this data had also been submitted to, and was available in, NWIS by the time this database was compiled. This database includes all available water-quality data for the Uinta Basin. For groundwater samples this includes data up to October 24, 2014; for surface water samples this... Types: Citation; Tags: USGS Science Data Catalog (SDC), Uinta Basin, water quality

Water-quality data collected for Kewaunee County, Wisconsin

We deployed a sampler to characterize water quality from a household well tapping a shallow fractured dolomite aquifer in northeast Wisconsin. The sampler was deployed from January to May 2017, and monitored temperature, nitrate, chloride, specific conductance, and fluorescent dissolved organic matter on a minute time step; water was directed to sequential microbial filters during three recharge periods that ranged from 5 to 20 days. Results from the automated sampler demonstrate the dynamic nature of the household water quality, especially with regard to microbial targets, which were shown to vary 1 to 2 orders of magnitude during a single sampling event. We believe assessments of pathogen occurrence and concentration,.. Categories: Data; Tags: Door County, Geochemistry, USGS Science Data Catalog (SDC), Wisconsin, automated sampling, All tags...

Water Quality data from the Providence Water Supply Board for tributary streams to the Scituate Reservoir, water year 2017

Water-quality data were collected by the Providence Water Supply Board (PWSB) from tributaries in the Scituate Reservoir drainage area, October 2016 through September 2017 (water year 2017). Water-quality samples were collected by the PWSB either monthly or quarterly at fixed stations on 36 tributaries in the Scituate Reservoir drainage area. These data were used to calculate instantaneous loads and yields of constituents in a report by the U.S. Geological Survey. Water-quality samples were collected following a strict sampling schedule so that water-quality samples would be representative of various weather conditions. Samples were analyzed at the PWSB water-quality laboratory at the P.J. Holton Water Purification.. Categories: Data; Tags: Bacteria, Chloride, Cranston, Drinking water use, Foster, All tags...

Water Quality data from the Providence Water Supply Board for tributary streams to the Scituate Reservoir, water year 2015

Water-quality data were collected by the Providence Water Supply Board (PWSB) from tributaries in the Scituate Reservoir drainage area, October 2014 through September 2015 (water year 2015). These data were used to calculate instantaneous loads



Searching water quality in Data Releases

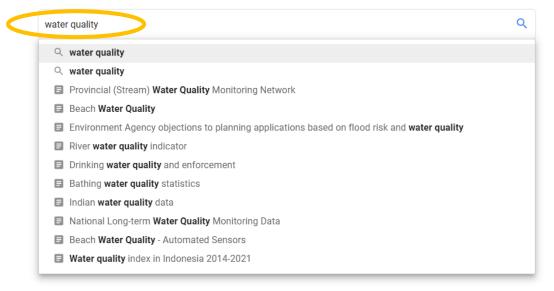
Google Dataset Search

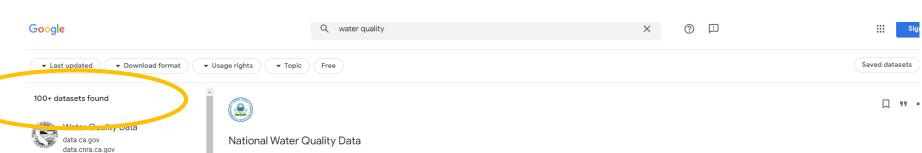
Google Dataset Search

- Search engine for datasets
 - https://datasetsearch.research.google.com/

Google

Dataset Search





kaggle Water quality

E csv Updated Mar 22, 2023

kaggle.com

≡ zip

Updated Jun 30, 2021



application/rdfxml +5
Updated Mar 7, 2023



National Long-term Water Quality Monitoring Data

open.canada.ca catalogue.arctic-sdi.org +3more

csv, esri rest, html +2
Updated May 20, 2022

Explore at: catalog.data.gov

353 scholarly articles cite this dataset (View in Google Scholar)

Dataset updated

Dec 14, 2020

Dataset provided by

United States Environmental Protection Agency

United States Description

Area covered

Water quality data. This dataset is associated with the following publication: Papenfus, M., B. Schaeffer, A. Pollard, and K. Loftin. Exploring the potential value of satellite remote sensing to monitor chorophyll-a for U.S. lakes and reservoirs.. ENVIRONMENTAL MONITORING AND ASSESSMENT. Springer, New York, NY, USA, 192: 808, (2020).

Open Refine — working with datasets

- OpenRefine is a powerful free, <u>open-source tool for working with messy data</u>: cleaning it; transforming it from one format into another; and extending it with web services and external data.
 - o <u>openrefine.org/</u>
 - o <u>info.library.okstate.edu/data</u>
 - Carpentries at OSU OpenRefine workshop with UCO November 16 & 17 info.library.okstate.edu/c.php?g=902101&p=6491976

