



National Institute of Environmental Health Sciences
Your Environment. Your Health.

Resources and Tools to Support Integration of Geospatial and Environmental Exposure Data into Epidemiological and Clinical Health Research

Joint Annual Meeting of the International Society of Exposure Science
and the International Society for Environmental Epidemiology

Atlanta, GA
August 17, 2025





WORKSHOP AGENDA

WELCOME AND SETTING THE STAGE	8:00 AM – 8:15 AM
CHORDS DATA ECOSYSTEM	8:15 AM – 8:20 AM
CAFE DATA REPOSITORY	8:20 AM – 8:25 AM
SHOWCASE: CHORDS DATA CATALOG	8:25 AM – 8:50 AM
SHOWCASE: CAFE DATA RESOURCES	8:50 AM – 9:15 AM
TUTORIAL: AMADEUS SOFTWARE	9:15 AM – 10:05 AM
TUTORIAL: CAFE TOOLKIT	10:05 AM – 10:55 AM
FACILITATED DISCUSSION: DATA STANDARDS	10:55 AM – 11:15 AM
PANEL DISCUSSION/Q&A	11:15 AM – 11:30 AM
ADJOURN	11:30 AM

Learning Objectives

- Increase awareness of existing geospatial environmental exposure data, tools, and resources
- Improve understanding of evolving approaches for linking and harmonizing exposure and health data
- Considerations for data standards in the context of epidemiologic and clinical research

Workshop Goals

- Build a network of researchers working in this space
- Solicit feedback on tools and resources
- Disseminate resources and opportunities related to the CHORDS and CAFE programs to a wider audience
- Join our stakeholder group/community of practice

CHORDS Overview

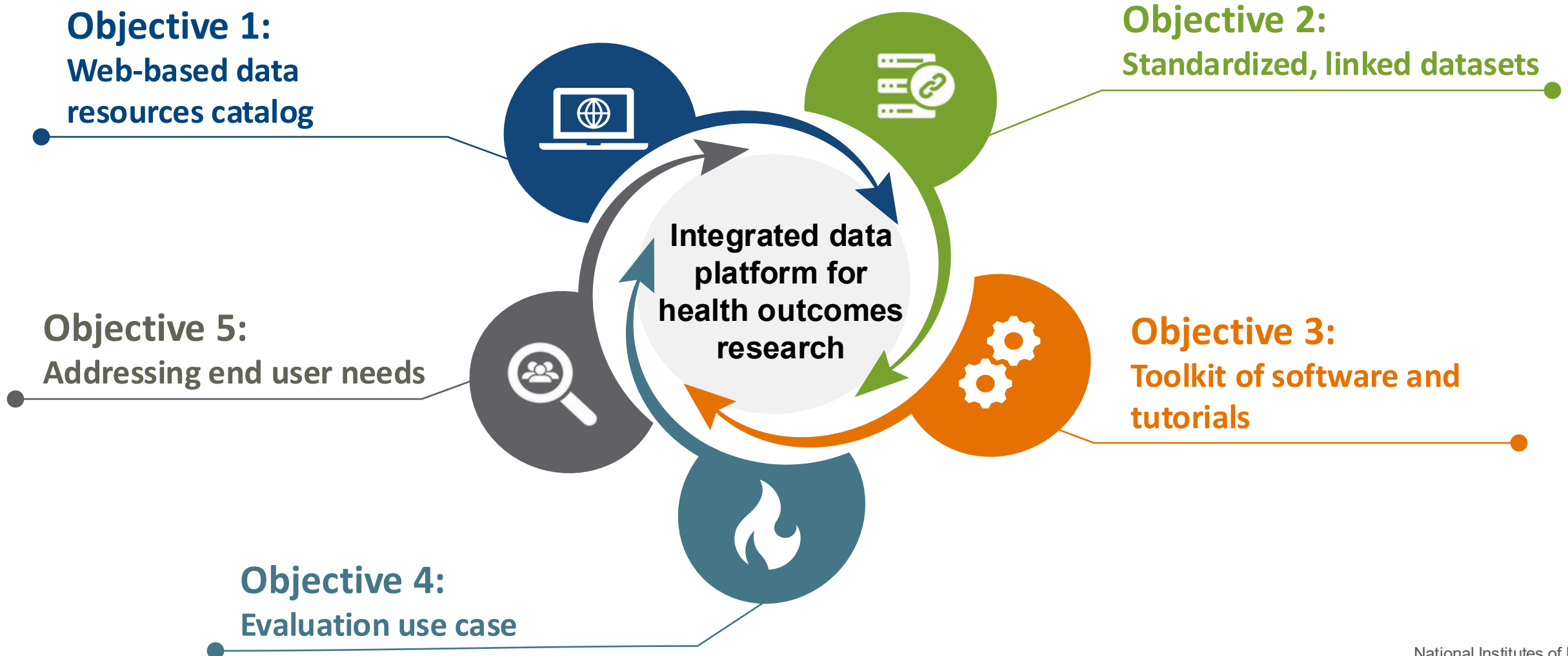
Connecting Health Outcomes Research and Data Systems (CHORDS)

Facilitating the **Linking of Environmental and Health Data** to Advance **Patient-centered Outcomes Research**



Goal: Strengthen data infrastructure to facilitate research connections between environmental exposures and health outcomes so researchers can 1) identify, analyze, and reduce the health effects associated with disaster-related events (e.g., wildfires) and 2) improve patient and population health outcomes

CHORDS Objectives and Deliverables





CHORDS Data Platform



About CHORDS

The CHORDS project aims to build and strengthen data infrastructure for patient-centered outcomes research on environment and health.



Browse Catalog [↗](#)

View a catalog of over 100 CHORDS Data Resources.



Highlighted Research

Collection of case studies intended to provide examples of research articles that examine the health effects of wildfire-related exposures and highlight the key environment and health data sets used in these studies.



Software

Find out more and download the CHORDS software.



Training and Use Cases

The CHORDS project seeks to connect researchers with guides, tutorials, and example code.

Data Catalog: Manually Curated Searchable Geospatial and Population Resources



CHORDS Data Resources

122 RESOURCES

Search studies by keyword...

Summary Statistics | Tags | Table of Records | Pagination

Show Search Filters

RESOURCE NAME	DOMAINS
OSM	Transportation, Built Environment
The OpenStreetMap (OSM) is a collection of worldwide community contributed or volunteered (crowd-sourced) global map information that includes roads, trails, cafés, railway stations, and much more. New, updated, and complete data are made available each week in both compressed XML, and a custom PBF format.	
GeoExportData Roads Streets Highways Trails Transit Stop Community-Contributed Transportation Built Environment	
NLCD-3C	Land Use And Land Cover, Greenness
The National Land Cover Database (NLCD) is a comprehensive land cover database for the contiguous United States (CONUS) that is managed by the Multi-Resolution Land Characteristics Tree Canopy (TC) Cover All Years product. It provides estimates of tree canopy density across the CONUS and represents the percentage of land area that is covered by tree canopies.	

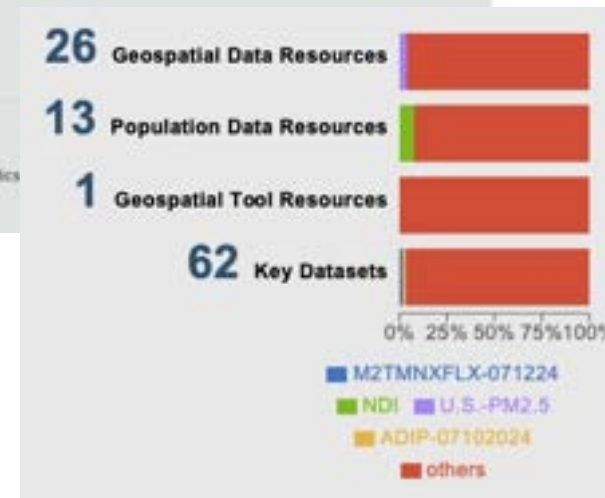
Metadata provided:

- Project information
- Measures
- Spatial characteristics
- Temporal characteristics
- Data access information
- Other:

Keywords

Citations

Publications



Open-Source Software and Tools for Simplifying Environmental Health Data Analysis

CHORDS-specific tools

- **amadeus**: Downloads large-scale environmental and weather data and gets it ready for analysis
- **beethoven**: A reproducible and extensible pipeline for air pollution exposure designed for updated and timely releases
- **chopin**: Simplifies parallelization, running many calculations or processes simultaneously, of big environmental exposure data

		
Amadeus 	Beethoven 	Chopin 
repo status: Active	repo status: WIP	repo status: WIP
Machine for Data, Environments, and User Setup for common environmental health datasets is an R package developed to improve and expedite users' access to large, publicly available geospatial datasets.	Building an Extensible, Reproducible, Test-driven, Harmonized, Open-source, Versioned, Ensemble model for air quality is an R package developed to facilitate the development of ensemble models for air quality.	Computation of Spatial Data by Hierarchical and Objective Partitioning of Inputs for Parallel Processing.

	
Contents lists available at ScienceDirect	
Environmental Modelling and Software	
journal homepage: www.elsevier.com/locate/envsoft	
Amadeus: Accessing and analyzing large scale environmental data in R	
Mitchell Manware  , Insang Song  , Eva S. Marques  , Mariana Alifa Kassien  , Lara P. Clark  , Kyle P. Messier 	



Toolkit of Code, Data, and Educational Materials

Open-source resources: Tutorials and educational materials to support different types of users (e.g., students, epidemiologists, clinicians, data managers) in accessing, processing, and integrating geospatial exposure data into health research

- **Geospatial Data Foundations**
 - Working with point, polygon, and raster data
- **Health Data Integration**
 - Linkage to Census Units
 - FHIR PIT Tutorial
- **Use Case Studies**
 - AHRQ HCUP Analysis

8 HCUP and Amadeus Smoke Plume Use Case

Clinician/Medical Professional Clinical Data Manager Community Health Worker Student

Integrating HCUP databases with Amadeus Exposure data

Date Modified: April 29, 2025

Author: Darius M. Bost

Programming Language: R

8.1 Motivation

Understanding the relationship between external environmental factors and health outcomes is critical for guiding public health strategies and policy decisions. Integrating individual patient records from the Healthcare Cost and Utilization Project (HCUP) with data from environmental datasets allows researchers to examine how elements such as air quality, wildfire emissions, and extreme temperatures impact hospital visits and healthcare utilization patterns.

Ultimately, linking HCUP and environmental exposure data enhances public health monitoring and helps researchers better quantify environmental health risks.

Use Case: Healthcare Cost and Utilization Project (HCUP) Analyses

Spatial analysis: bivariate map of asthma prevalence and heavy smoke exposure across ZIP codes in Oregon

- Map shows areas with overlapping smoke and health risks

Logistic regression model: examine relationship between asthma diagnoses and exposure to different levels of smoke density

- Exposure to medium and heavy smoke is associated with significantly increased odds of asthma

Findings can support targeted public health interventions

Legend:

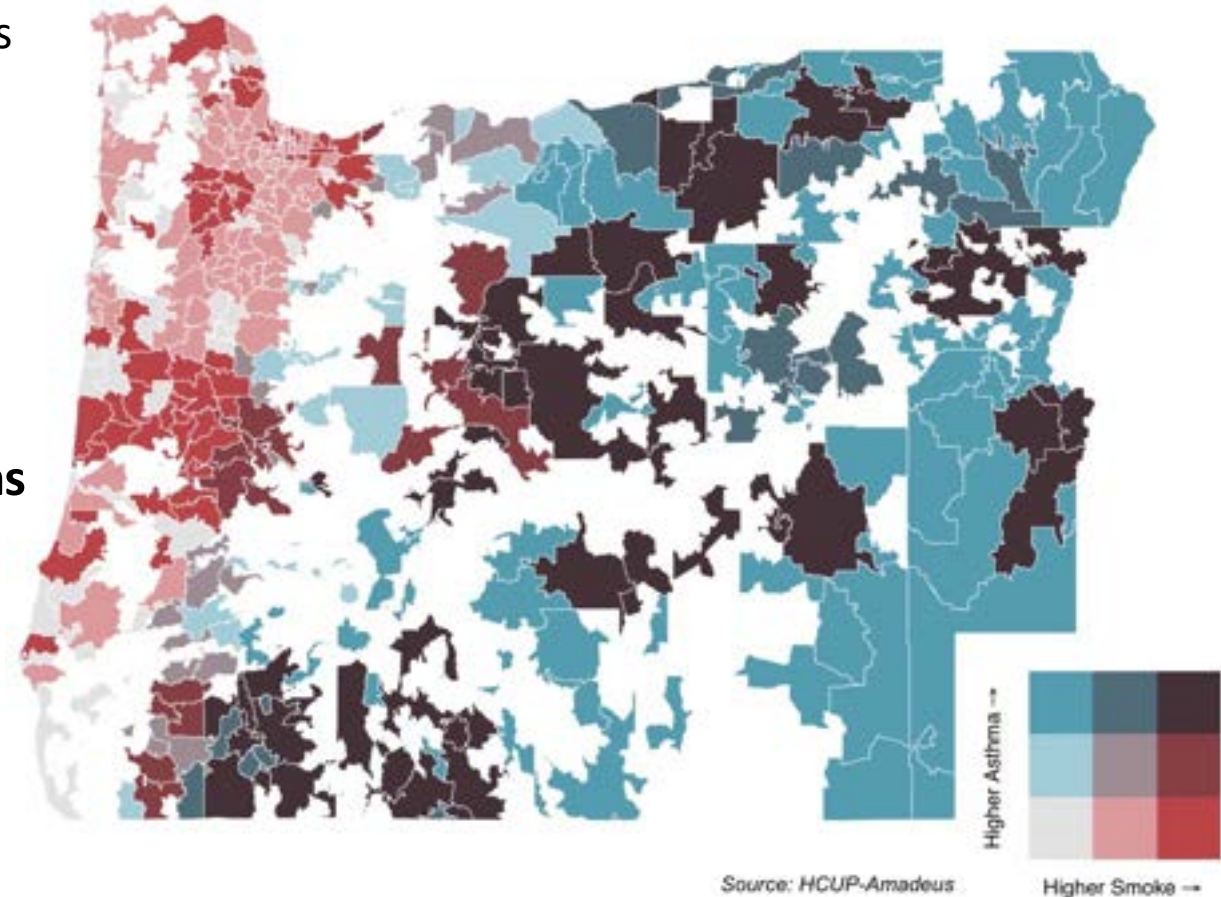
X-axis (**red**): higher smoke exposure

Y-axis (**blue**): higher asthma

Color interpretations:

- Dark red:** high smoke, low asthma
- Dark blue:** high asthma, low smoke
- Dark purple:** high smoke, high asthma
- Light gray:** low smoke, low asthma

Asthma Prevalence vs Heavy Smoke Exposure by ZIP Code
Bivariate map showing intersection of health and environmental burden



Source: HCUP-Amadeus



CHORDS Team and Contributors

Core Team

- Adam Burkholder
- Alison Motsinger-Reif
- Ann Liu
- Aubrey Miller
- Charles Schmitt
- David Fargo
- David Reif
- Kyle Messier
- Trisha Castranio

NIH

- Asif Rizwan (NCI)
- Erin Iturriaga (NHLBI)
- Regina Bures (NHLBI)
- Richard Kwok (NIA)
- Umit Tokac (NHLBI)
- Cristina Justice (AofU)

Other Agencies

- Pamela Owens (AHRQ)
- Patricia Keenan (AHRQ)
- NASA representatives

NIEHS

- Darius Bost
- Deep Patel
- Jennifer Fostel
- Lara Clark
- Marcus Jackson
- Maria Shatz
- Mariana Kassien
- Mike Conway
- Mitchell Manware
- Rupali Gupta
- Skylar Marvel
- Sue Nolte

Technical Expert Panel (TEP)

- Angela Werner (CDC)
- Caleb Dresser (Harvard)
- Cavin Ward-Caviness (EPA)
- Cole Brokamp (Univ. Cincinnati)
- Genee Smith (Johns Hopkins)
- Kevin Lane (BU/CAFÉ)
- Patrick Wall (CDC)
- Philip Awadalla (Canada Data Integ. Ctr)
- Rima Habre (USC)
- Yang Liu (Emory)



NIH-NSF Collaboration Centers: Promoting Time-Critical Extreme Weather and Natural Disaster Health Data Collection and Research

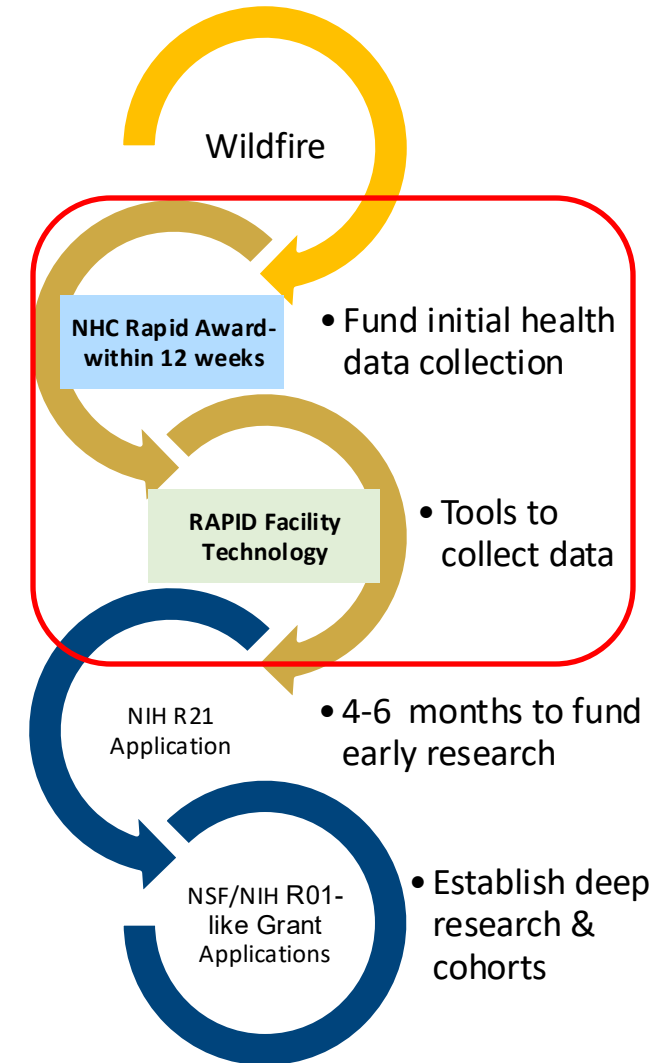


- **Natural Hazards Center (NHC) (University of Colorado-Boulder)**

- Fund awards to collect perishable health data focused on high-risk groups and time-sensitive situations (e.g., pregnancy, comorbidities, workers, at-risk populations)
- Quick response awards between \$10-50K within 12 weeks of hurricanes, wildfires, floods, tornadoes, etc.

- **RAPID Facility Program (University of Washington)**

- Acquisition and access to health-focused sensors, instrumentation, support, and training for collection of post-natural disaster exposure and health data
- Field testing resources during 2025 Los Angeles wildfires





National Institute of Environmental Health Sciences
Your Environment. Your Health.

CHORDS DATA CATALOG

Joint Annual Meeting of the International Society of Exposure Science
and the International Society for Environmental Epidemiology

Atlanta, GA
August 17, 2025



Topics



Design Principles



CHORDS Portal Phase 1



Curation Pipeline



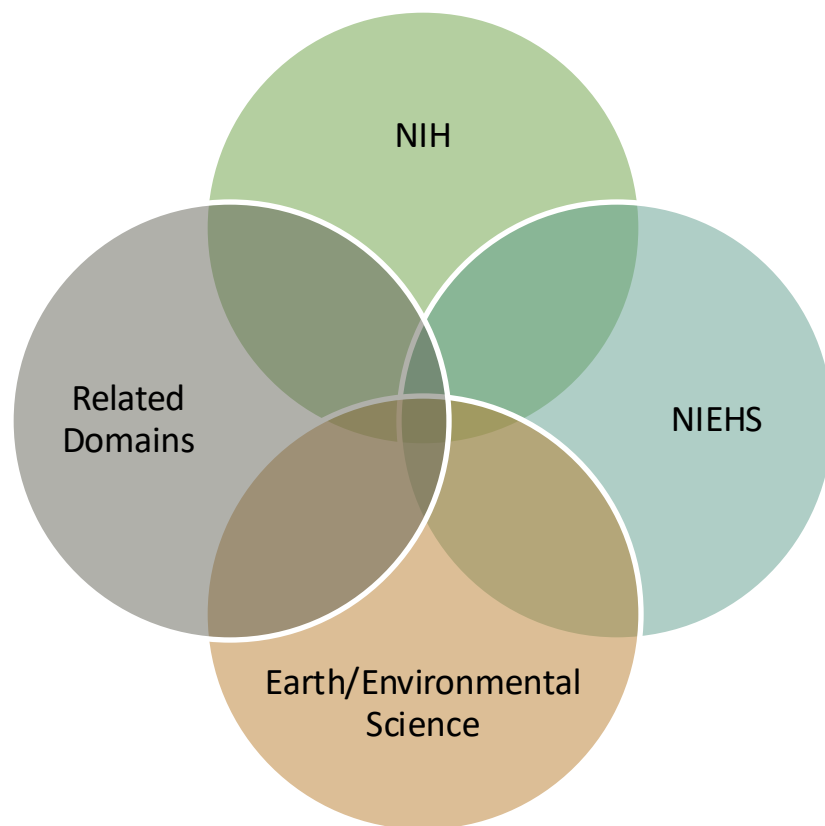
CHORDS Phase 2

Automated curation
Multivalent catalog
Formalized workflows



Lessons Learned

CHORDS in the Context of the Exposome

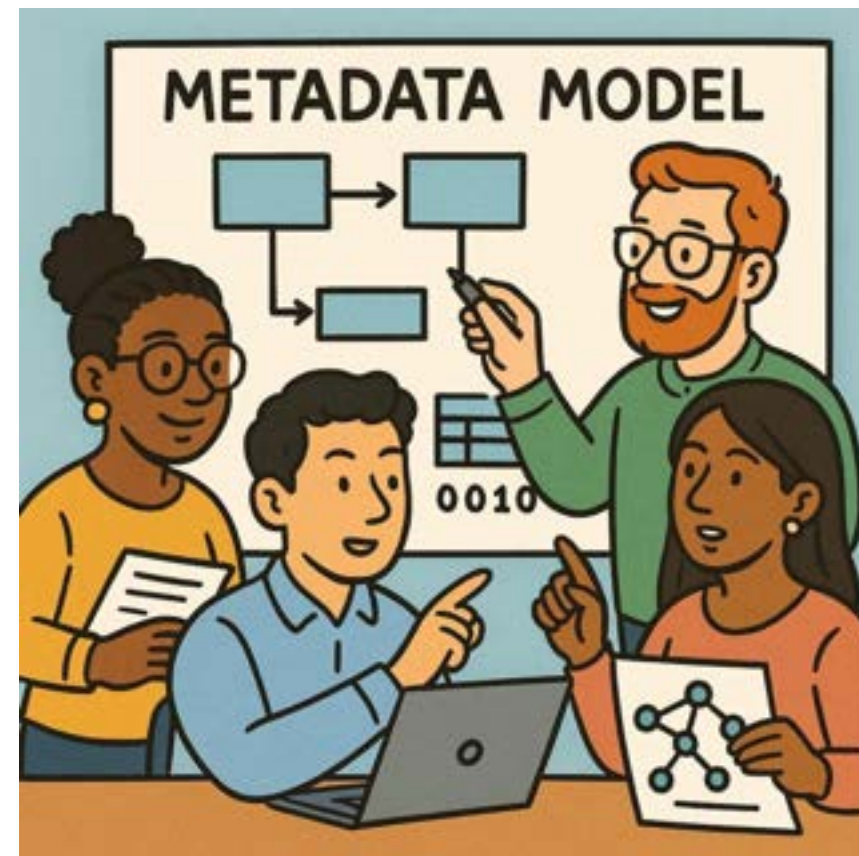


While focused on specific CHORDS use cases, this is an opportunity to think about the broader ecosystem studying the Exposome.

- Look ‘inward’ at other parts of NIH
- Ease federation with other platforms common in other domains
- Be flexible in terms of source and target schema while maintaining as much structure and validation as possible
- Think about the provenance and verification of data over time
- Prepare for increased use of AI for both metadata extraction and catalog automation as well as the use of AI in data discovery

Designing the Catalog

- This is really about the metadata model, first and foremost
- With such a broad set of domains, we would have to build out the model as we iterated
 - Tooling to allow curators to design and communicate design as well as curate the data
 - Workflows that allowed constant migration
- Community standards, ontologies for crossing environment and health not readily available
 - ECTO - <https://pmc.ncbi.nlm.nih.gov/articles/PMC9951428/>
 - EXO - <https://pmc.ncbi.nlm.nih.gov/articles/PMC3314380/>



Designing the System

- The care and feeding of the metadata model development and curation was central
- The system needed to enable the entire workflow, from design to curation to validation to presentation
- The data itself is the catalog, and the system needed flexibility to slice, dice, repurpose and federate
- The catalog needed to fit into an expanding, federated and heterogeneous cyberinfrastructure picture



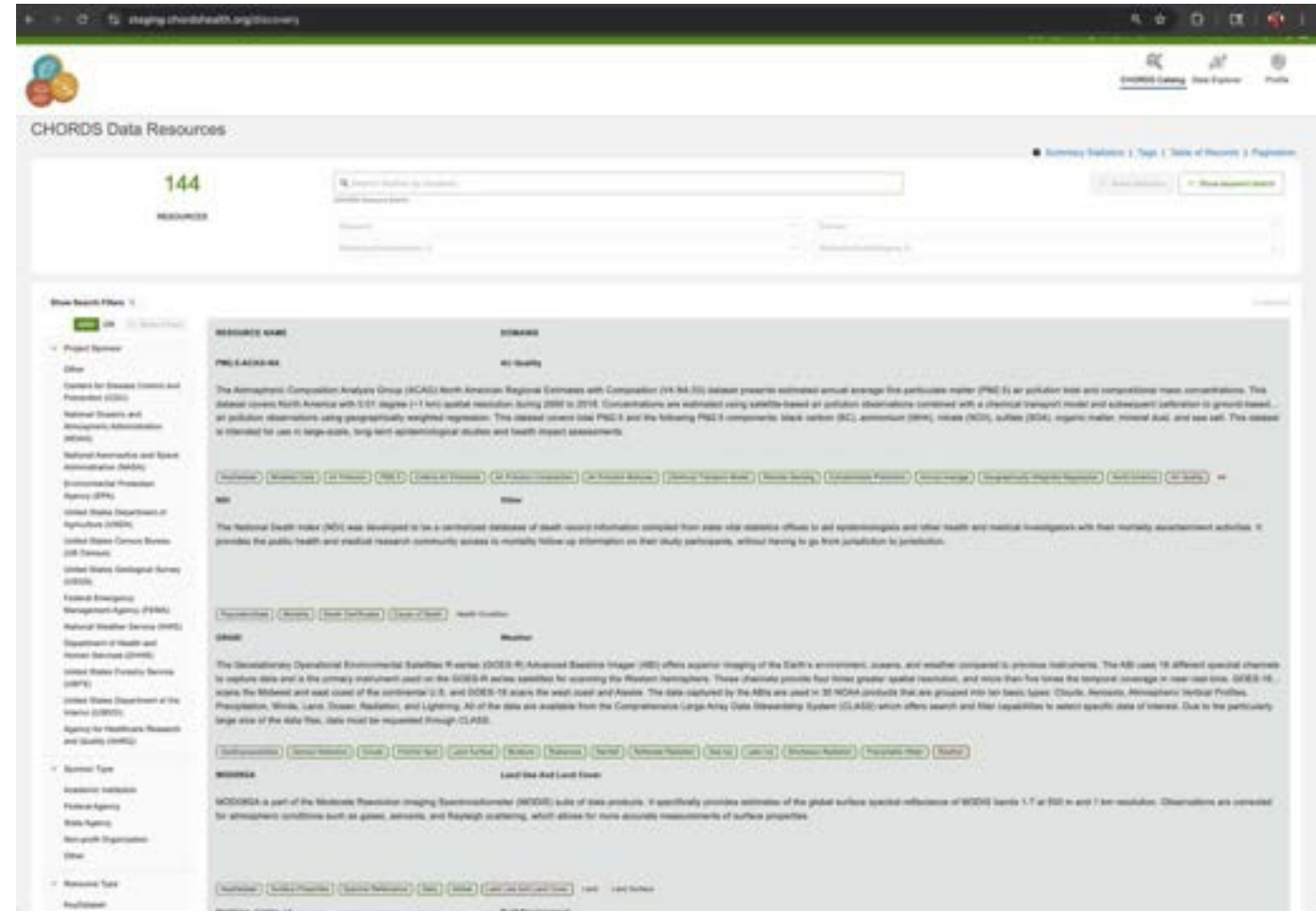
Architectural Design Principles

- Avoid building bespoke systems, use existing platforms.
- Focus on trans-NIH standards and platforms.
- Join communities (standards and open-source).
- API first (offer and utilize open API at the foundation).



The CHORDS Portal

- <https://chordshealth.org/discovery>
- Gen3 based
 - Open Source - <https://github.com/uc-cdis>
 - GA4GH Compliant
 - FedRAMP certified
- Operated by Gen3 on AWS



Data Model Centric

- Above all things, the data model was seen as the key
 - Structured
 - Flexible
 - Validated
 - Includes controlled vocabularies
 - Serializable and API Accessible

The image displays three overlapping screenshots of the CHORDS Catalog web application, illustrating its data model-centric design.

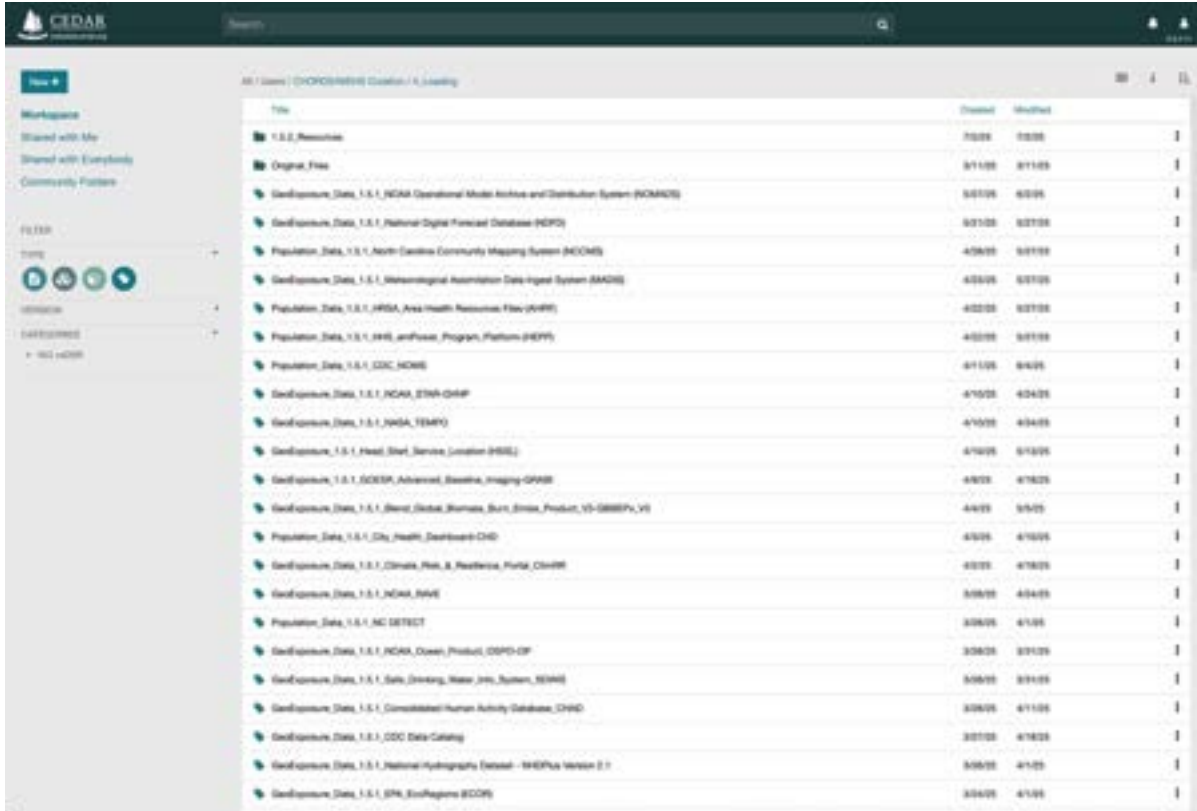
- Top Screenshot:** Shows a hierarchical data model diagram. The root node is "Program", which links to "Project". "Project" further links to "Publication" and "Core Metadata Collection". A legend on the right defines link types: Required Link (solid line), Optional Link (dashed line), Administrative (purple circle), and Data File (green circle).
- Middle Screenshot:** Shows a "Resource" form. The title is "Resource: A collection of data or resources, can have files attached or may point to an external reference". It includes a "Data File" tab and a table with fields for resource details, such as "Name", "Description", "Type", "Code", "Access", "Domain", "Keywords", and "Status".
- Bottom Screenshot:** Shows a "Query graph" interface. It features a "GraphQL" editor with tabs for "Pretty", "Merge", "Copy", "History", and "Inspect". The editor contains a GraphQL query:

```
1 {  
2   program {  
3     name  
4     id  
5   }  
6   projects {  
7     name  
8     code  
9     short_name  
10  }  
11 }
```

 The right side of the editor shows the corresponding JSON response:

```
{  
  "data": {  
    "program": {  
      "id": "00000000-0000-0000-0000-000000000000",  
      "name": "CHORDS",  
      "projects": []  
    }  
  }  
}
```

Data Curation with CEDAR



Title	Created	Modified
1.5.1_Resource	7/2/05	7/2/05
Original_File	8/11/05	8/11/05
Geospectrum_Data_1.5.1_NCAR Operational Model Archive and Distribution System (NOMADS)	5/21/05	6/2/05
Geospectrum_Data_1.5.1_National Digital Forensic Database (NDFD)	5/21/05	5/21/05
Population_Data_1.5.1_North Carolina Community Mapping System (NCCMS)	4/26/05	5/21/05
Geospectrum_Data_1.5.1_Meteorological Reanalysis Data-Ingest System (MADS)	4/26/05	5/21/05
Population_Data_1.5.1_HRSA_Area Health Resource File (AHRF)	4/22/05	5/21/05
Population_Data_1.5.1_HHS_andPower_Program_Platform (HPPF)	4/22/05	5/21/05
Population_Data_1.5.1_GDC_NOMIS	4/11/05	5/2/05
Geospectrum_Data_1.5.1_NCAR_STNA-Query	4/10/05	4/24/05
Geospectrum_Data_1.5.1_NCAR_TERRA	4/10/05	4/24/05
Geospectrum_1.5.1_Fuel_Start_Service_Location (HSL)	4/10/05	5/13/05
Geospectrum_1.5.1_SCHSP_Advanced_Essential_Imaging-SPAR	4/8/05	4/18/05
Geospectrum_Data_1.5.1_Sensor_Dataset_Biomass_Burn_Sensor_Product_VS-DBMS/Py_V2	4/6/05	5/5/05
Population_Data_1.5.1_City_Health_Dashboard-CHD	4/5/05	4/10/05
Geospectrum_Data_1.5.1_Climate_Risk_8_Resilience_Fortra_Climate	4/3/05	4/18/05
Geospectrum_Data_1.5.1_NCAR_RAVE	3/29/05	4/24/05
Population_Data_1.5.1_NC_DETECT	3/26/05	4/1/05
Geospectrum_Data_1.5.1_NCAR_Down_Product_DSDP-CP	3/26/05	3/31/05
Geospectrum_Data_1.5.1_Data_Incoming_Water_Int_System_NIWA	3/26/05	3/31/05
Geospectrum_Data_1.5.1_Consolidated Human Activity Database-CHAD	3/26/05	4/11/05
Geospectrum_Data_1.5.1_GDC_Data-Catalog	3/21/05	4/18/05
Geospectrum_Data_1.5.1_National Hydrography Dataset - NHDPlus Version 2.1	3/26/05	4/1/05
Geospectrum_Data_1.5.1_SPA_SixRegions (SCOR)	3/24/05	4/1/05

- CEDAR (<https://cedar.metadatacenter.org/>) is being used for human curation of metadata.
- CEDAR templates developed by curators, also serving as the Gen3 model specification.



Example CEDAR Template

DATA RESOURCE

Comments
Ecosystem maps provide a static classification rather than real-time environmental conditions. They do not account for seasonal changes, disturbances (e.g., wildfires), or human impacts that might alter ecosystems over time.

Intention use
Ecosystems is intended to be a spatial framework for use in assessing and monitoring ecosystems and their components.

Source name #

Update frequency #
Other

Update frequency other
Periodically

Includes other collected
False

Has doi
False

Has visualization tool
True

DESCRIPTOR DATA

Measures
Biomass, Land Cover, Land Surface Characteristics, Vegetation Cover, Vegetation Type, Vegetation Productivity, Vegetation Structure, Vegetation Greenness, Water Bodies, Other

Measures other
Terrestrial Composition, Aquatic Composition

Measurement method
Measurement method other

Time period start
2012

Time period end
2013

Time available comment

Temporal resolution #
Unaffiliated

Temporal resolution other

Spatial resolution #
State

Spatial resolution other

Spatial coverage
Continental United States, North America

Spatial coverage other

Spatial bounding box

Geometry type #
Polygon

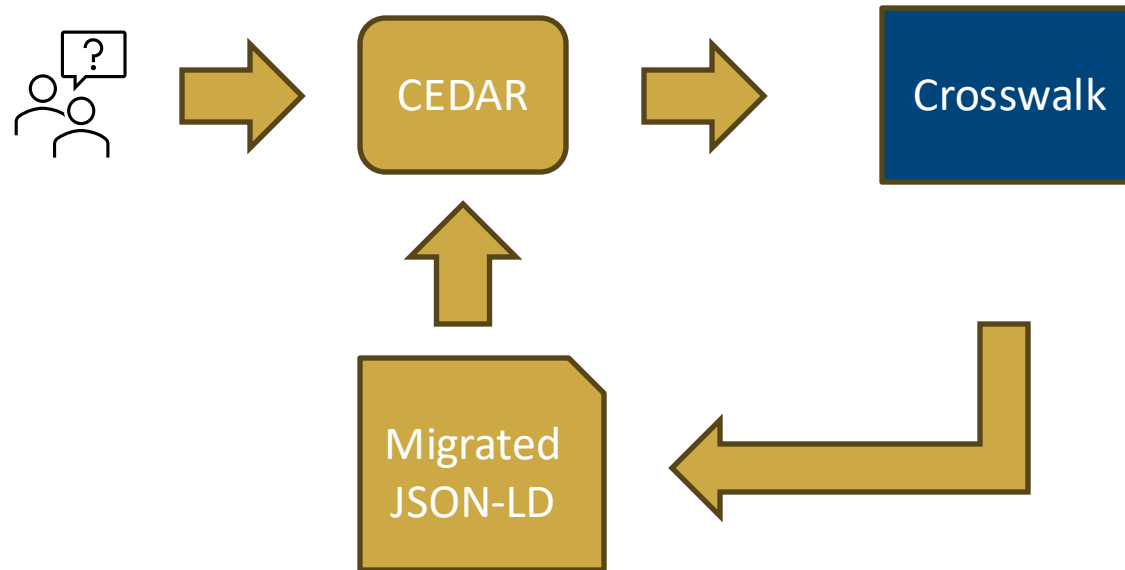
To JSON-LD

```

120 },
121 * "resource": {
122 *   "context": {
123     "resource_guid": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
124     "resource_name": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
125     "resource_type": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
126     "domain": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
127     "domain_other": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
128     "access_type": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
129     "date_added": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
130     "date_updated": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
131     "date_verified": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
132     "publication": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
133     "resource_short_name": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
134     "keywords": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
135     "resource_url": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
136     "resource_description": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
137     "payment_required": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
138     "is_static": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
139     "Resource Reference_156": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
140     "Resource Use Agreement_156": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a",
141     "resource_version": "https://schema.metadataservice.org/properties/88888888-70fa-47d7-b720-9e737e9e9a1a"
142   },
143 *   "resource_guid": {
144     "value": "88888888"
145   },
146 *   "resource_name": {
147     "value": "EPA Ecoregions"
148   },
149 *   "resource_type": {
150     "value": "Data Resource"
151   },
152 *   "domain": {
153 *     {
154       "value": "Greenness"
155     },
156 *     {
157       "value": "Other"
158     }
159   },
160 *   "domain_other": {
161 *     {
162       "value": "Greenness"
163     },
164   },
165 *   "access_type": {
166 *     {
167       "value": "Open"
168     }

```

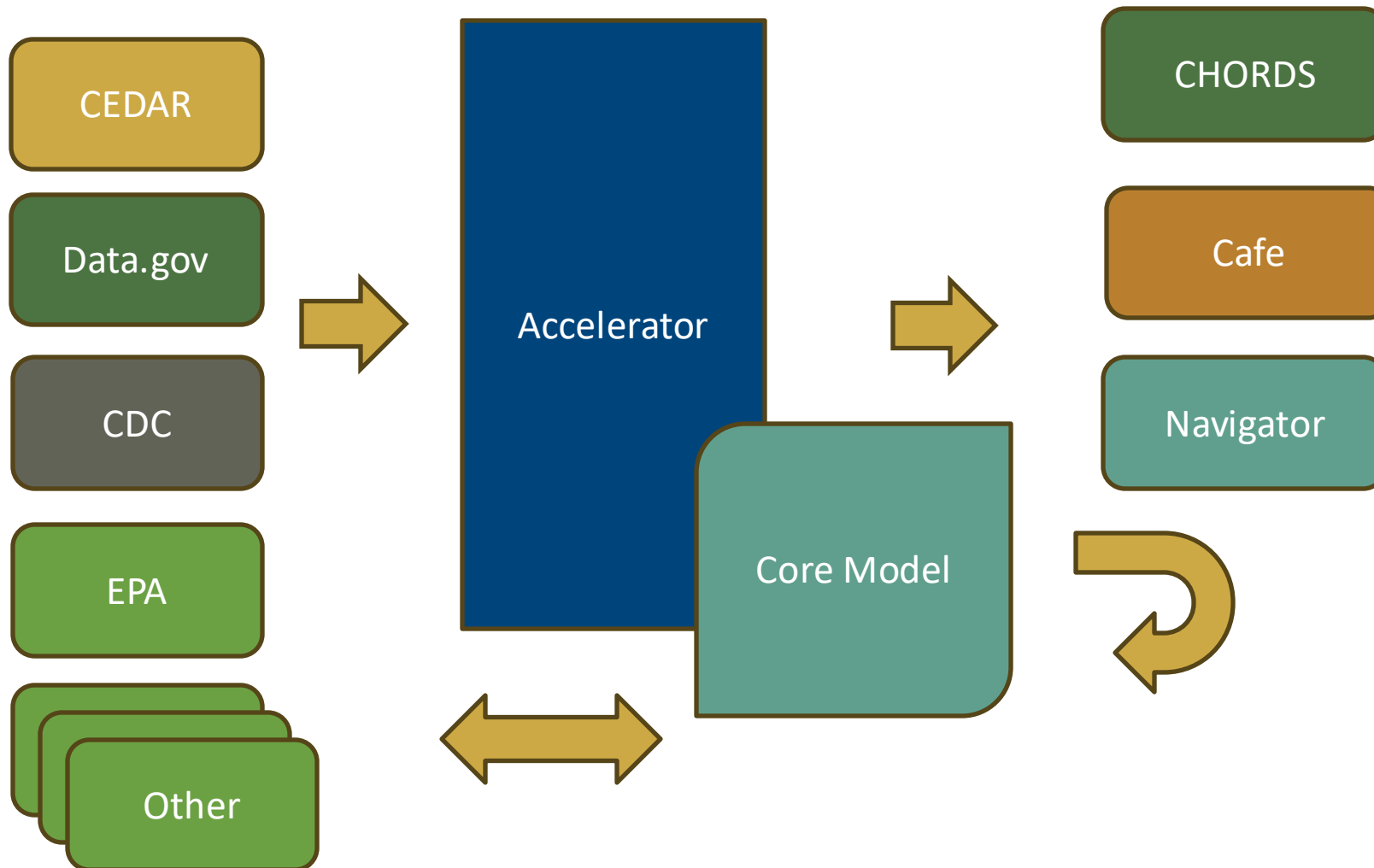
CHORDS Phase 2



Data Models Evolve

- Migration data flows have been created and must be managed
- Note the closed system, without human intervention it cannot be established which data might have disappeared or been updated
- Adding automated input from other sources and pushing data to new endpoints creates all sorts of disconnects, stale data issues

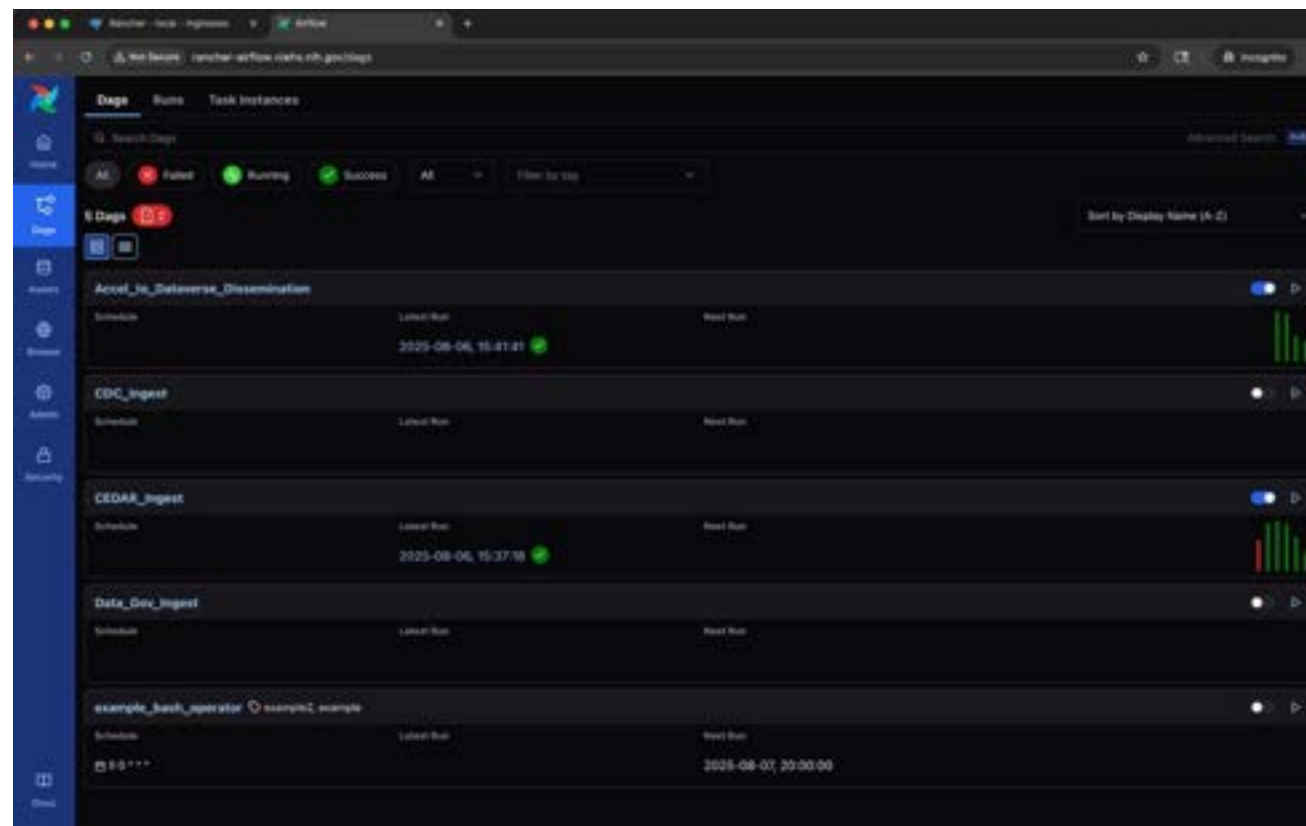
CHORDS Phase 2



- Add automated cataloging
- Create a connector architecture for adding new sources, new sinks, and migration workflows
- Automate re-validation and updates
- Project the CHORDS catalog into multiple endpoints.
- Migration workflows
- Revalidation and discovery workflows

Extending CHORDS Data Federation

- Think about catalog federation as a commodity ETL problem
- Provide plugin frameworks for adding new sources, models, crosswalks and data sinks, easing on-boarding
- Periodic, sensor-based, and curator triggered workflows



Demo – CEDAR Ingest

Trigger DAG - CEDAR_Ingest

Single Run
Trigger a single run of this DAG

Backfill
Run this DAG for a range of dates

Run Parameters

cedar_file_id *
a8b88c7-376b-4563-8a96-as/lead27118
CEDAR report ID of an individual file

cedar_api_key *
xxx
CEDAR API key

submitter_name *
Mike Conway
Name of the submitter

submitter_email *
mike.conway@nih.gov
Email of the submitter

Advanced Options

Trigger

- Common model
- Knows where it came from and when
- Knows the last time it was re-validated

MongoDB Compass - ehsk8mvt02.niehs.nih.gov:32017/accelerator.accelerator

Compass

My Queries

CONNECTIONS (2)

Search connections

- ★ accel-devel
- ★ ehsk8mvt02.niehs.nih.gov:32017
 - accelerator
 - admin
 - config
 - local

Documents 2 Aggregations Schema Indexes 1 Validation

Type a query: { field: 'value' } or [Generate Query](#)

ADD DATA EXPORT DATA UPDATE DELETE

50

```
{
  "_id": ObjectId("6893aeb5894dc1b6d33cedd"),
  "submission": {
    "submitter_name": "Mike Conway",
    "submitter_email": "mike.conway@nih.gov",
    "submitter_comment": ""
  },
  "data": {
    "program": {
      "program_name": "CEDARS",
      "preferred_label": ""
    },
    "project": {
      "project_code": "FAQSD-071324",
      "project_name": "Fused Air Quality Surface Using Downscaling (FAQSD)",
      "project_short_name": "FAQSD",
      "project_url": ""
    },
    "project_sponsor": [
      {
        "type": "",
        "name": "Environmental Protection Agency (EPA)",
        "other_type": false
      },
      {
        "type": "Federal Agency",
        "name": "",
        "other_type": false
      }
    ],
    "resource": {
      "resource_name": "NIH Environmental Protection Agency Fused Air Quality Surface Using Downscaling"
    }
  },
  "technical_metadata": {
    "created": "2025-08-06T19:36:21.777360",
    "modified": "",
    "verified": "",
    "original_source_type": "accelerator",
    "original_source_identifier": "8bb72176-1b58-46d1-b4a4-09e61a9644b9",
    "original_source_link": "cedar"
  },
  "history": [
    {
      "timestamp": "2025-08-06T19:36:21.777375",
      "msg": "accession from accelerator with identifier 8bb72176-1b58-46d1-b4a4-09e61a9644b9"
    }
  ]
}
```

Demo – CHORDS Data Dissemination to CAFE

Trigger DAG - Accel_to_Dataverse_Dissemination

Single Run **Backfill**
Trigger a single run of this DAG. Run this DAG for a range of dates.

Run Parameters

object_identifier *
Object identifier in accel database

submitter_name *
Name of the submitter

submitter_email *
Email of the submitter

Advanced Options

```
modified: ""
verified: ""
original_source_type: "accelerator"
original_source_identifier: "8bb72176-1b58-46d1-b4a4-09e61a9644b9"
original_source_link: "cedar"
- history: Array (2)
  - 0: Object
    timestamp: "2025-08-06T19:36:21.777375"
    msg: "accession from accelerator with identifier 8bb72176-1b58-46d1-b4a4-09e..."
  - 1: Object
    timestamp: "2025-08-06T19:40:25.854373"
    msg: "Disseminating document 6893aeb5084dc1bbd393cedd of type accelerator to..."
original_source: "8bb72176-1b58-46d1-b4a4-09e61a9644b9"
```

Dataverse Search User Guide

This draft version has incomplete metadata that needs to be edited before it can be published.

Requia et al. (2020) Daily 8-Hour Maximum and Ann Concentrations for the Contiguous United States, 1 v1.10 (2000–2016)

[Draft](#) [Update Metadata](#) [Incomplete Metadata](#)

2025, "Requia et al. (2020) Daily 8-Hour Maximum and Annual O3 Concentrations for the Contiguous United States, 1-km Grids, v1.10 (2000–2016)", <https://doi.org/10.5622/2525/REGOVL>. Root, DRAFT VERSION

[View Dataset](#) [Learn about Data Citation Standards](#)

Description

The Daily 8-Hour Maximum and Annual O3 Concentrations for the Contiguous United States, 1-km Grids, Version 1.10 (2000–2016) dataset contains estimated ozone (O3) air pollution concentrations. O3 concentrations are estimated using a geographically weighted ensemble modeling approach with three types of machine learning models (neural network, random forest, and gradient boosting). The approach incorporates O3 air pollution observations from ground-based monitors, land use and meteorological variables, chemical transport models, and remote sensing data, along with other sources. This dataset provides daily 8-hour maximum average and annual 8-hour maximum average O3 concentration estimates during 2000 to 2016 for the contiguous United States with 1 km spatial resolution.

Subject

Medicine, Health and Life Sciences

Keyword

Ensemble Model, Generalized Additive Model, Chemical Transport Model, Remote Sensing, Machine Learning, Air Pollution, Ozone, Concentration Prediction, Annual Average, Daily 8-Hour Maximum Average, CONUS

Related Publication

Cites: Requia, W. J., Di, Q., Silvern, R., Kelly, J. T., Koutrakis, P., Mickley, L. J., ... & Schwartz, J. (2020). An ensemble learning approach for estimating high spatiotemporal resolution of ground-level

Lessons Learned

- There are lots of good data sets out there, but **data ready to use for health applications is still a challenge**
- **Machine readable metadata is not always provided**, sometimes it must be derived from web sites or other sources
- There is **spotty availability, especially at the National level**, for things like well water quality, waterway contamination, landfill location
- At the heart of it, CHORDS is about **building a data model that can reasonably catalog a diverse corpus of data**
- As we will discuss later, human curation is difficult, we need to **increase automation**, including constantly updating and re-validating the catalog

Lessons Learned

- In large part, integrating Environmental and Health data is like comparing apples to oranges. Consistent metadata is the central goal
- To better integrate data:
 - Analyte specifics/metrics for chemicals (e.g. CAS #, levels of detection, level of sensitivity)
 - Exposure metrics – average versus peak, etc.
 - Clear time frames and methods of derivation
 - Spatial metrics, how derived, levels of confidence
 - Methods for linking time and spatial metrics to each other and to health data sets



Summary

- CHORDS is about the data model
- CHORDS is about federation
- CHORDS is about human and automated curation
- Focus on open API and frameworks so that catalog data can be shared and re-purposed
- New domains and data types are welcome, and there is a sustainable method for expansion
- Core API: <https://github.com/NIEHS/accelerator-core>
- Helm Charts: <https://github.com/NIEHS/accelerator-helm>

Thanks! Get in touch with any questions!



Mike Conway
NIEHS/NIH
mike.conway@nih.gov

Deep Patel
NIEHS/NIH
deep.patel@nih.gov

BUSPH-HSPH-CAFÉ

Research Coordinating Center (U24)



ISEE-ISES 2025



School of Public Health
Center for Climate & Health



HARVARD
T.H. CHAN

SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

H

Goals of CAFÉ



BUSPH-HSPH Climate Change and Health Research Coordinating Center



Greg Wellenius
(BU)



Francesca Dominici
(Harvard)



Amruta Nori-Sarma
(Harvard)



Danielle Braun
(Harvard)



Gaurab Basu
(Harvard)



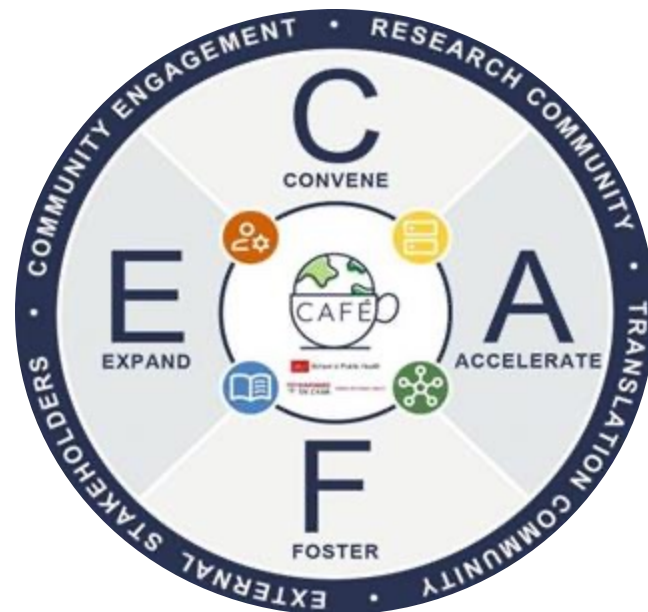
Rebecca Pearl-Martinez
(BU)



Benjamin Sovacool
(BU)



Kevin Lane
(BU)



Administrative
Function



Data
Management
Function



Resource
Function



Capacity
Building
Function

COMMITTED SUPPORTERS:

Harvard Data Science Institute, Harvard Climate Change Institute,
BUSPH Center for Climate and Health, Microsoft, AWS, ESRI, Google,
BU Events Planning, UMass Boston, Meharry Medical College, PHFI,
C40 Cities, EHRA



Pat Kinney
(BU)



Rachel Nethery
(Harvard)



Jon Levy
(BU)



Pam Templer
(BU)

What does CAFÉ provide?

Various resources, including:

- CAFÉ University webinars
- Free annual virtual meetings
- Funding for pilot studies
- Research translation lab
- Research Mentore-Mentee Connections
- Youtube tutorials
- Slack channels
- Monthly newsletter
- **Data and data management resources**

<https://climatehealthcafe.org/>

Find us on Linkedin at Climate & Health CAFÉ

Join our Community of Practice to receive our monthly newsletter and Slack channels!



CAFÉ Data Management

1. Facilitate sharing and reuse of Climate Health (CCH) data (**Dataverse**)
2. Develop and promote data science and software tools for processing, linking, and analyzing common CCH data (**Github**)
3. Provide data management and dissemination guidance
4. Identify CCH research data needs and define common data elements across the community of practice (COP)

Note: CAFE is not a data processing/analysis service center! Our focus is on curating commonly used datasets and coding pipelines.

CAFÉ: Data Repositories (Dataverse)

Goal: FAIR Data Principles

Findable

Increases visibility, citations,
and impact of research

Supports knowledge discovery
and innovation

Interoperable

Supports and promotes inter-
and cross-disciplinary data
and reuse

FAIR DATA



Accessible

Streamlines and maximizes
ability to build upon previous
research results

Attracts partnerships with
researchers and business in
allied disciplines

Reusable

Promotes use and reuse of
data allowing resources to be
allocated wisely

Improves reproducibility and
reliability of research results

Researchers

- Structure data clearly and apply good data management practices
- Document data and software
- Richly describe data using standardized metadata fields
- Apply license and/or clear terms of use

Repositories

- Assign persistent identifiers
- Structure metadata records according to a disciplinary standard or schema
- Index data as searchable resources
- Retrieve datasets according to an open protocol that supports authentication
- Preserve data files and metadata
- Track provenance and versioning

NIH Data Management Sharing Plans

- Outlines how data resulting from a research project is managed in the short and long term
- Procedures for data collection, documentation, and processing
- Plan for data sharing so others can find and access materials in perpetuity
- Living document that is frequently referred to and updated as needed
- Many funders now require data sharing or data management plans as part of the conditions of funding



Source: [Elements of an NIH Data Management and Sharing Plan](#)



NIH Data Management and Sharing Policy



Submission of a Data Management and Sharing Plan (DMSP)

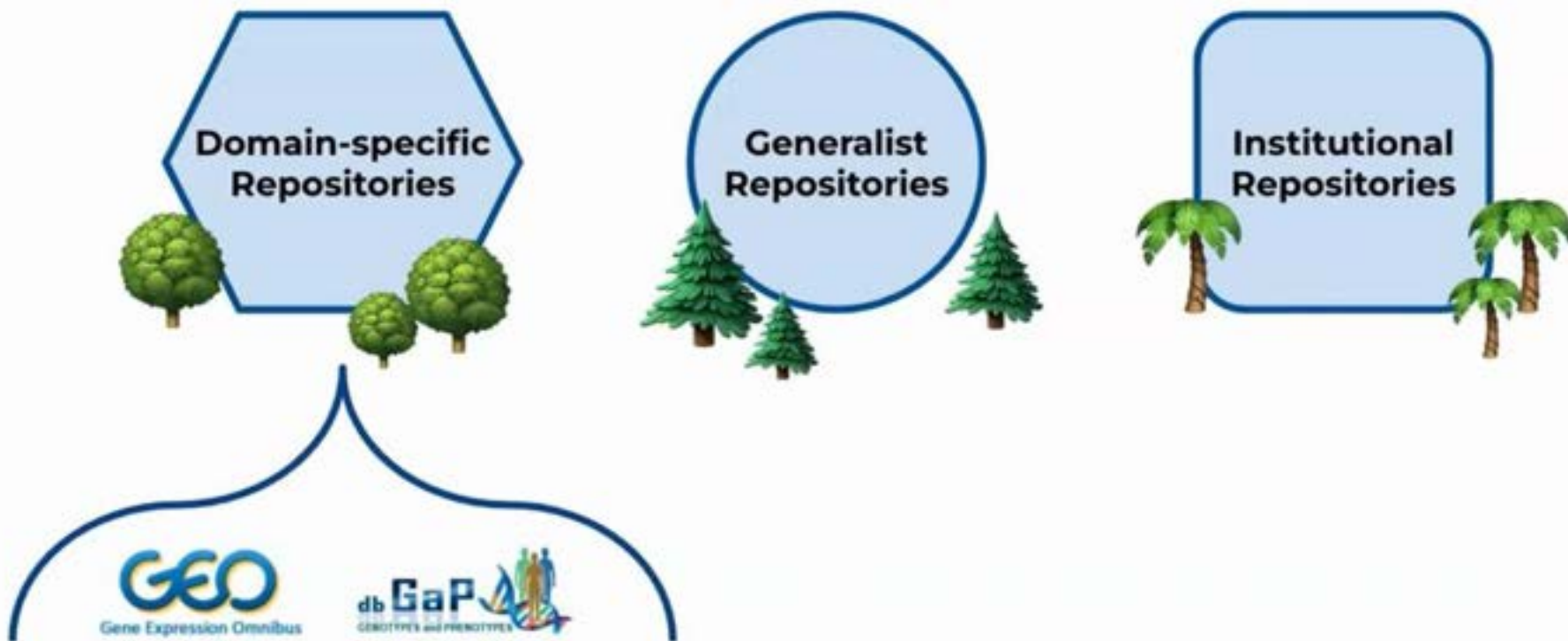


Proposed repository to be used consistent with NIH guidance

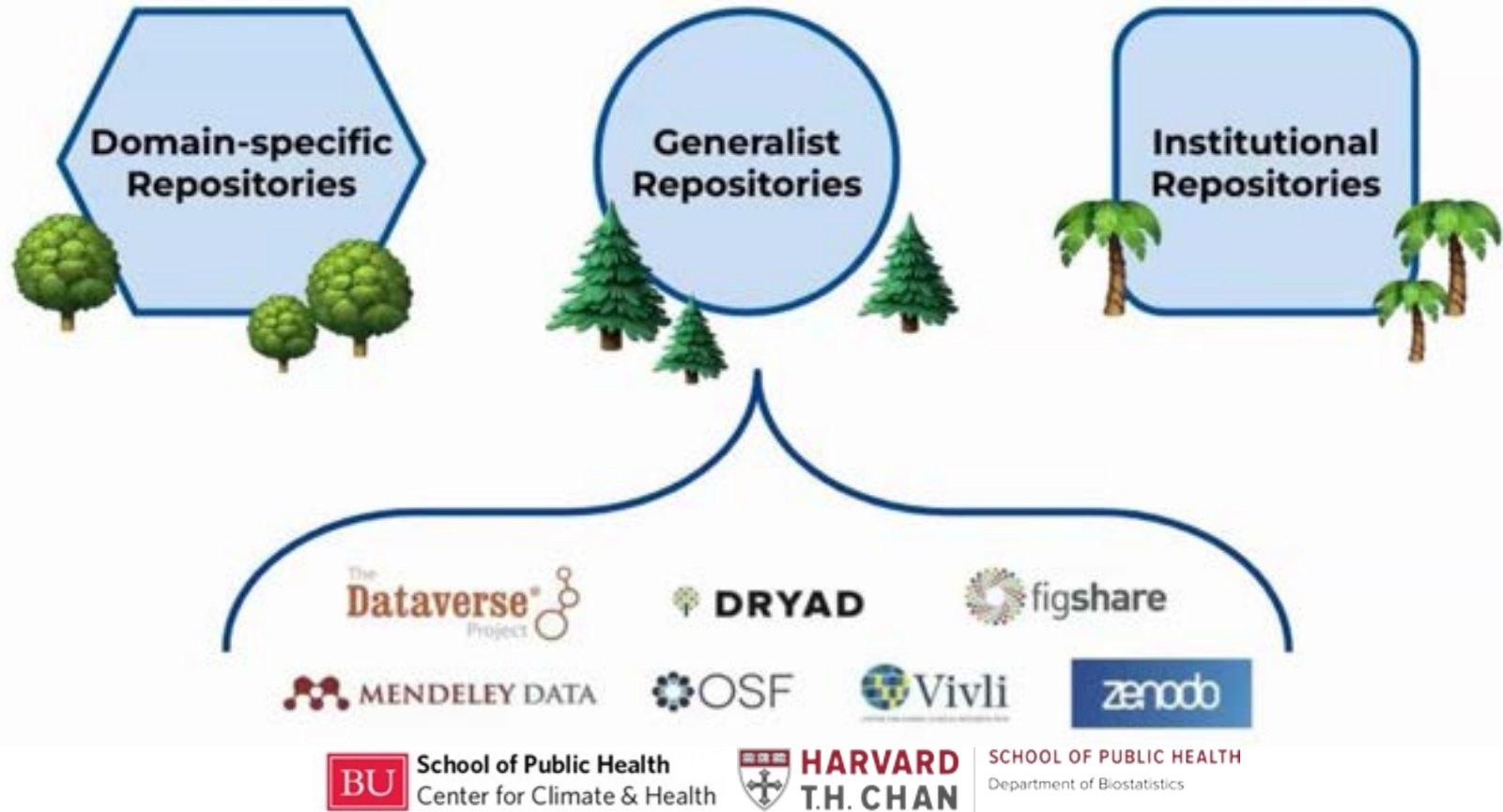


Compliance with the awardee's plan as approved by NIH ICO

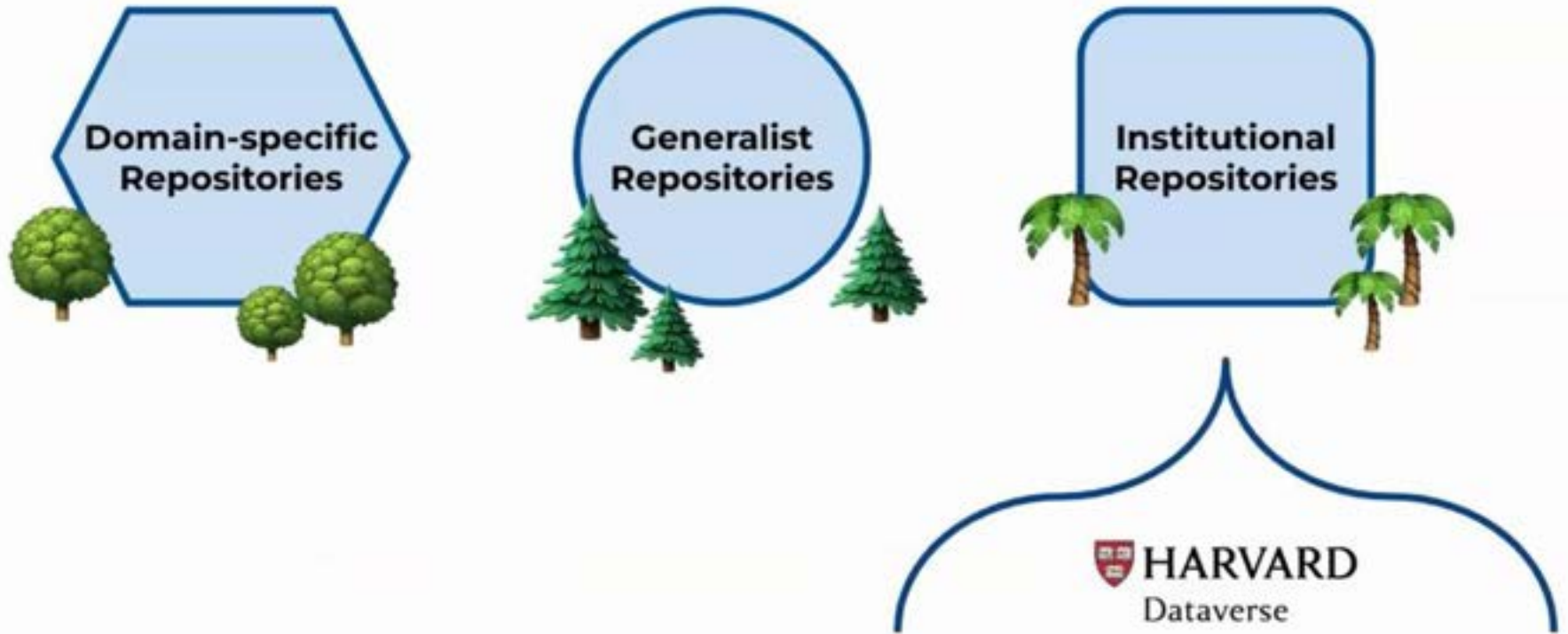
Data Repository Ecosystem



Data Repository Ecosystem



Data Repository Ecosystem

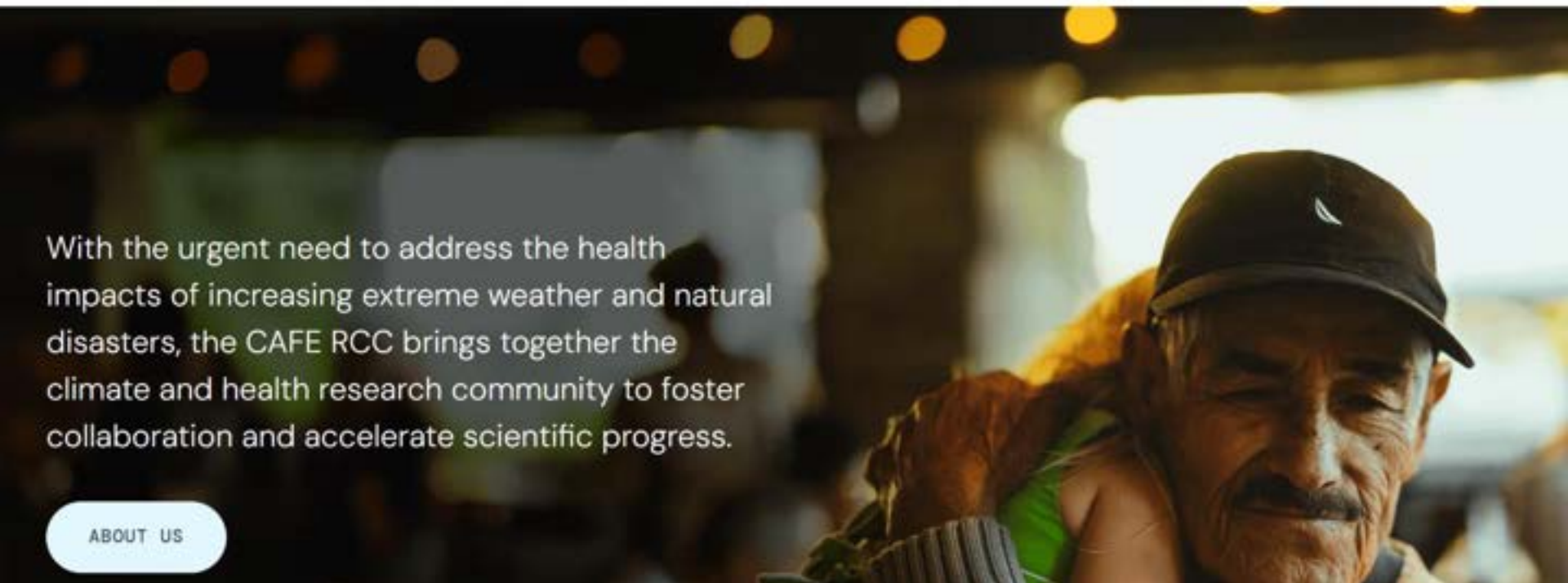


School of Public Health
Center for Climate & Health



HARVARD
T.H. CHAN

SCHOOL OF PUBLIC HEALTH
Department of Biostatistics



With the urgent need to address the health impacts of increasing extreme weather and natural disasters, the CAFE RCC brings together the climate and health research community to foster collaboration and accelerate scientific progress.

ABOUT US

<https://www.climatehealthcafe.org/>



School of Public Health
Center for Climate & Health



HARVARD
T.H. CHAN

SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

CAFÉ Dataverse Collection Highlights



- Over **700 Data submissions** deposited in <2 years.
- Most datasets have been **downloaded between 50 - 5000 times (and increasing)**.
- Most downloaded datasets include wildfire air pollution, synthetic health data, spatial aggregations and greenspace.
- New **CIESIN datasets** deposited August 2025.
- **Sub-collections**, which are a nice way for groups with multiple datasets to share data within the CAFÉ Collection in one place.

HARVARD Dataverse

Climate Change and Health Research Coordinating Center (CAFÉ) Collection
(Harvard University, Boston University)

Harvard Dataverse > Climate Change and Health Research Coordinating Center (CAFÉ) Collection >

Daily local-level estimates of ambient wildfire smoke PM2.5 for the contiguous US

Version 1.0

Childs, Marissa; Li, Jessica; Viten, Jeffrey; Heft-Neal, Sam; Driscoll, Anne; Wang, Sheme; Goulet, Carlos; Olu, Minghao; Burney, Jennifer; Burke, Marshall. 2024. "Daily local-level estimates of ambient wildfire smoke PM2.5 for the contiguous US". <https://doi.org/10.7927/Q7N71AJS>. Harvard Dataverse, V1, UNF:6:F8gqHmAWW3JmUNP5dQ== [fileUNF]

Cite Dataset > Learn about [Data Citation Standards](#)

Access Dataset >
Edit Dataset >
Link Dataset
Contact Owner > Share

Dataset Metrics >
233 Downloads >

Climate Change and Health Research Coordinating Center (CAFÉ) Collection
(Harvard University, Boston University)

Harvard Dataverse > Climate Change and Health Research Coordinating Center (CAFÉ) Collection >

Urban Park Use in the United States at Census Tract 2017 - 2021

Version 2.0

Swarni Vagstad; Michael T. Young; Stylianos Serghios; Akim Kumok; Yash Shah; Kevin J. Lane; Shreshth Bhandekar; Mansi Kansal; Tonar Shekar; Gregory A. Wallerius; Evgeniy Gashilovich. 2024. "Urban Park Use in the United States at Census Tract 2017 - 2021". <https://doi.org/10.7927/Q7N71AJS>. Harvard Dataverse, V2, UNF:6:J4zF6xM8BY1z6i2bA== [fileUNF]

Cite Dataset > Learn about [Data Citation Standards](#)

Access Dataset >
Edit Dataset >
Link Dataset
Contact Owner > Share

Dataset Metrics >
51 Downloads >

Climate Health AIR pollution study (CHAIR) in India (Ashoka University & Karolinska Institutet) **Unpublished**

Apr 9, 2024

CHAIR India (Consortium for Climate, Health & Air Pollution Research in India) research group, a global collaboration for air pollution and climate research across India. The work is led and coordinated centrally through the Institute of Environmental Medicine, Karolinska Institutet.

NSAPH (NSAPH)
Jan 30, 2024

The NSAPH Subcollection, part of the Climate-Health CAFÉ Dataverse, features data contributions from the National Studies on Air Pollution and Health (NSAPH) group based at the Harvard T.H. Chan School of Public Health. This subcollection is focused on providing datasets related...

Dataverse Extracting Data



The screenshot shows the Harvard Dataverse page for the dataset "Wildfire Smoke PM2.5 per Zipcode". A blue arrow points to the "Access Dataset" button, which is highlighted in a callout box. The callout box shows the "Access Dataset" menu with options: "Download Options", "Download ZIP (2.7 GB)", "Explore Options", and "Binder".

Dataset Information:

- Dataset:** Wildfire Smoke PM2.5 per Zipcode
- Version:** 1.0
- Access Dataset:** [Button]
- Download Options:** [Button]
- Download ZIP:** 2.7 GB
- Explore Options:** [Button]
- Binder:** [Button]

Description: This dataset contains fully aggregated measurements of daily PM2.5 from ambient wildfire smoke in the contiguous United States, spanning from 2000 to 2016. The data is sourced from a study by Chou et al. (2016), titled "Daily Local-Level Estimates of Ambient Wildfire Smoke PM2.5 for the Contiguous US" published in Environmental Science & Technology. To compute the standardized weight across different zip codes, we computed the weight calculation on a 10km grid. To compute the weight, we divided area weights that add up to 1 in each polygon of zip codes. The analysis can be replicated by the simple values per zipcode for the aforementioned period. Those interested in replicating our data processing pipeline can access it at <https://github.com/HarvardDataverse/processing>.

Subject: Earth and Environmental Sciences, Computer and Information Science

Keywords: Smoke, PM2.5, standardized weight

Related Publication: Chou, W. L., L. J. Wu, L. Hsu, S. Chou, A. Wang, S. ... & Burke, W. (2016). Daily Local-Level Estimates of Ambient Wildfire Smoke PM2.5 for the Contiguous US. Environmental Science & Technology, 50(16), 13607-13617. <https://doi.org/10.1021/acs.est.6b01000>

License: CC BY-NC 4.0

Files: 11 files

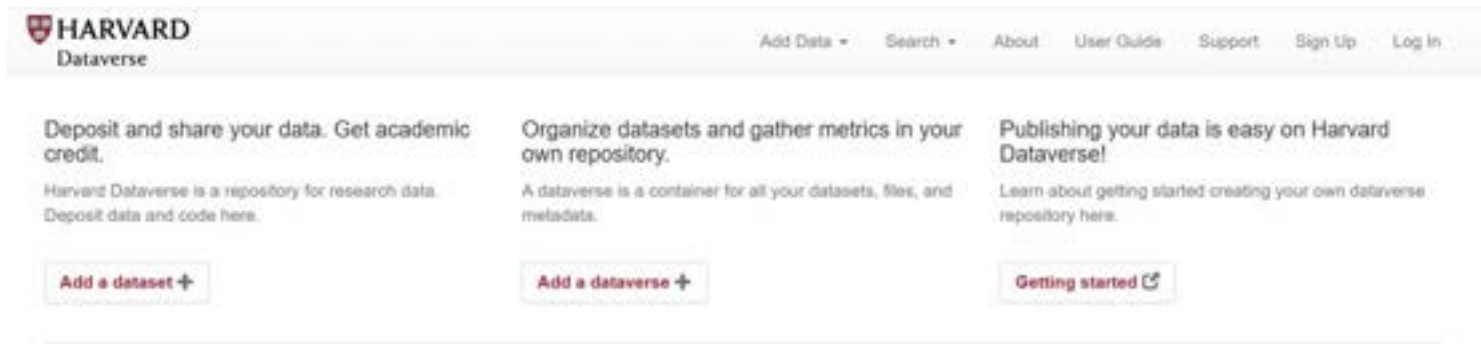
File Type: CSV

Download: [Button]

Files: 11 files

Download: [Button]

Why Create/Use a Dataverse Repository



- Can share data with collaborators or the public. Open content can be accessed directly via the UI or API, and restricted content can be requested using a "request access" feature if enabled by the data depositor.
- Assigns a DOI to every published dataset. The repository records file downloads and views and makes this information available to depositors. Depositors who create collections can also ask or require downloaders to provide information about their data re-use.
- Uses standard-compliant metadata to ensure that dataset metadata can be mapped easily to common metadata schemas to make data more preservable and interoperable.
- Provides a mechanism by which a journal's editors and reviewers can have anonymous access to a dataset or dataverse before it is made public. See [Private URLs](#)

CAFÉ Dataverse Collection: Contributing Data



Welcome to the Climate CAFÉ Data Management Page

CAFÉ Dataverse Collection

Climate-Health CAFÉ Dataverse Collection

[Contributing to the Climate-Health CAFÉ Dataverse Collection](#)

Example Datasets in the Climate-Health CAFÉ Dataverse Collection

CAFÉ GitHub Organization

Climate CAFÉ GitHub Organization

Climate CAFÉ Code-sharing Walkthrough

Getting Started with GitHub

Reproducible Data Creation Pipeline



Contributing to the Climate-Health CAFÉ Dataverse Collection

This tutorial provides step-by-step instructions on how to upload data to [Dataverse](#) and select the "Climate CAFÉ" organization as the Host Dataverse. Dataverse is a robust platform for sharing, archiving, and collaborating on research data.

To ensure the reproducibility of your data, it's essential to include a link to your GitHub repository in the dataset's description. This link should lead to the pipeline that outlines the process for creating the data.

[Contents](#)

[Prerequisites](#)

[Steps](#)

<https://climate-cafe.github.io/tutorial.html>

HARVARD Dataverse Add Data Search About User Guide Support

Harvard Dataverse >

Host Dataverse Changing the host dataverse will clear any fields you may have entered data into.

Harvard Dataverse

*Asterisks indicate required fields

Citation Metadata

Title *

Add "Replication Data for" to Title

Author *

Name * Irene, Kacia

Affiliation Harvard University

Identifier Type Select...

Identifier

Point of Contact *

Name Irene, Kacia

Affiliation Harvard University

E-mail * kirene@hsph.harvard.edu

Description * This field supports only certain HTML tags.

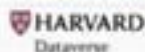


School of Public Health
Center for Climate & Health



HARVARD
T.H. CHAN

SCHOOL OF PUBLIC HEALTH
Department of Biostatistics



Convene-Accelerate-Foster-Expand

The Joint BUSPH HSPH Research Coordinating Center
of the NIH Climate Change and Health Initiative

Climate Change and Health Research Coordinating Center (CAFE) Collection

(Harvard University, Boston University)

A data repository to enable research at the intersection of climate and
human health

[Harvard Dataverse](#) >

[Contact](#) [Share](#)

Welcome to the CAFE Dataverse collection! This open collection is designed to support and enhance global research initiatives focused on understanding and mitigating the health impacts of climate change. To learn more about CAFE, please visit our [homepage](#).

Instructions

To make contributions to the CAFE collection, you'll find easy-to-follow steps in our [CAFE Dataverse upload instructions](#).

NSAPH

Your Sub-Collection Here!

Your Group's
Sub-Collection
Here!

Your Sub-Collection Here!

Search this Dataverse...



[Advanced Search](#)

[+ Add Data](#)

[Databases \(0\)](#)

[Datasets \(18\)](#)

[Files \(138\)](#)

Dataverse Category

[Organization or Institution \(1\)](#)

[Research Group \(1\)](#)

1 to 10 of 20 Results

[Sort](#)

Census data interpolated by year and zip code

Jan 27, 2022 - NSAPH



Baeth, Ben, 2022, "Census data interpolated by year and zip code", <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7927/H73K8M>, Harvard Dataverse, V1

Demographic values crosswalked from pcta to zip codes, with missing values replaced by a moving average model for each ZCTA. Data was available for the year 2000, and from 2011-2016. All other years were interpolated. Git repository: <https://github.com/NSAPH/national-causal-analy>.

Developing your CAFÉ Dataverse Collection: Custom Metadata

A screenshot of the Harvard Dataverse Custom Metadata form. The form includes fields for "Subject", "Keyword", "Term", "Controlled Vocabulary Name", and "Controlled Vocabulary URL". A red rectangle highlights the "Term", "Controlled Vocabulary Name", and "Controlled Vocabulary URL" fields. A red arrow points from the "Term" field to the "Environmental Health Sciences Glossary" page shown in the next block. The "Controlled Vocabulary Name" field contains the text "Climate Change and Human Health Glos". The "Controlled Vocabulary URL" field contains the text "https://tools.niehs.nih.gov/cchhglossary/".

CAFÉ Climate and Health Research Coordinating Center Collection



NIEHS Glossary Keyword Term

<https://www.niehs.nih.gov/health/topics/glossary>



School of Public Health
Center for Climate & Health



HARVARD
T.H. CHAN

SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

Benefits of Dataverse to Help You Meet FAIR Data Principles



Findable

- Each dataset receives a **persistent DOI**, making it uniquely identifiable and citable.
- Rich **metadata standards** enhance searchability and indexing in global registries (e.g., DataCite, Google Dataset Search, National Library of Medicine).

Accessible

- Supports **open access to data** (with options for restricted access when needed), satisfying NIH's emphasis on making data broadly available.
- Provides a stable, institutionally supported platform with **long-term data preservation**.

Interoperable

- Encourages use of **standard data formats** (e.g., CSV, JSON, Stata, SPSS) that can be reused across systems and disciplines.
- Metadata and APIs are compatible with other repositories and tools, supporting machine-readability and integration.

Reusable

- Users can attach **detailed documentation**, codebooks, and README files to promote data reuse and understanding.
- Enables application of **clear data licenses** (e.g., CC0, CC BY), fulfilling legal and ethical reuse criteria

Harvard Dataverse NIH-DMP Guidance

<https://support.dataverse.harvard.edu/harvard-dataverse-nih-dmp-guidance>

Harvard Dataverse NIH-DMP Guidance

^ NIH-GREI

Overview

NIH-GREI Initiative

Harvard Dataverse NIH-DMP Guidance

Generalist Repository Comparison Chart

Version 1, June 30th, 2023

Introduction

Researchers should consider using a combination of repository features and data sharing best practices to address the elements in the NIH Data Management and Sharing Plan (DMSP). Below are examples and general guidance for describing how Harvard Dataverse Repository addressed some of the questions in the DMSP.

Document licensed under: <https://creativecommons.org/licenses/by-nc-sa/4.0>.

DMSP Elements

- [Element 1: Data Types](#)
- [Element 2: Related Tools, Software and Code](#)
- [Element 3: Standards](#)
- [Element 4: Data Preservation, Access, and Associated Timelines](#)
- [Element 5: Access, Distribution, and Reuse Considerations](#)
- [Element 6: Oversight of Data Management](#)

Downloadable PDF of the HDV NIH-DMP Guidance information

CAFÉ Toolkit Examples: GitHub

<https://climate-cafe.github.io/intro.html>



School of Public Health
Center for Climate & Health



HARVARD
T.H. CHAN

SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

Goal: FAIR Data Principles

Findable

Increases visibility, citations,
and impact of research

Supports knowledge discovery
and innovation

Interoperable

Supports and promotes inter-
and cross-disciplinary data
and reuse

FAIR DATA



Accessible

Streamlines and maximizes
ability to build upon previous
research results

Attracts partnerships with
researchers and business in
allied disciplines

Reusable

Promotes use and reuse of
data allowing resources to be
allocated wisely

Improves reproducibility and
reliability of research results

Benefits of using GitHub for code sharing to meet FAIR Data Principles



Findable

- GitHub repositories are indexed by search engines and can be assigned **persistent identifiers (e.g., DOIs)**, making code discoverable.
- Clear project structure and metadata (e.g., README, tags, releases) improve searchability and navigation.

Accessible

- Code is **openly accessible by default** (when using public repos), satisfying NIH expectations for open data/code sharing.
- GitHub supports easy **download, cloning, and API access** to repositories.

Interoperable

- GitHub supports **standard code formats** and integrates with popular tools and languages (e.g., Python, R, Jupyter, Git).
- Enables collaboration and integration with reproducibility platforms (e.g., Binder, CodeOcean).

Reusable

- Clear **version control and documentation** support reproducibility and reuse of code by other researchers.
- Supports **open-source licenses** (e.g., MIT, Apache), making reuse legally and ethically straightforward.
- Promotes **open science** through forks, pull requests, and issues, supporting the collaborative spirit of NIH's open data initiatives.

Supports NIH DMS Plan Compliance

- Facilitates sharing of **software, scripts, and workflows** alongside datasets, as required by NIH for reproducibility.
- Documentation (e.g., README, CONTRIBUTING.md) can describe data provenance and usage instructions.

Integration with Archival Services

- GitHub integrates with **Zenodo** and other repositories to generate **citable DOIs**, ensuring long-term preservation and citation, aligning with FAIR's "Findable" and "Reusable" goals.

Enhances Collaboration and Community Engagement

- Promotes **open science** through forks, pull requests, and issues, supporting the collaborative spirit of NIH's open data initiatives.

CAFÉ Educational Resource Hub

- Crowd-sourced, searchable catalog to centralize information across experience levels, areas of interest, programming languages, etc.
- Includes coding tips, epidemiological methods, coding tutorials, and more!

The screenshot shows the CAFÉ Educational Resource Hub search results page. On the left, there is a sidebar with filters for 'Type of resource' (Book or reference text, Online course, Video or recorded webinar, Online tutorial or course, Free workshop) and 'Skill' (Communication, Data access, Data visualization, Cloud computing/computing proficiency, Data cleaning/coding strategies, Statistical analysis, Survey design/forecasting strategy, Professional development, Statistical method, Subject matter training). Below these are filters for 'Programming language' (R, SQL, Python). The main content area displays three search results, each with a title, a brief description, and the speaker's name and affiliation. The first result is 'The Relationship between Relative Extreme Temperature Events and Mortality in the United States' by Julie Solomon, M.S., from Research Data Services (Boston, Harvard University). The second result is 'CAFÉ University - How can CAFÉ help with the new NH Data Management and Sharing Plan?' by Julie Solomon, M.S., from Research Data Services (Boston, Harvard University). The third result is 'CAFÉ University: NIEHS Tools to Support Research and Communication in Climate Change and Health' by Trisha Cossentino, NIEHS. The fourth result is 'CAFÉ University: Wrangling and Processing Historic Weather Data for Climate and Health Studies' by Rachel Nethery, PhD.

<https://www.climatehealthcafe.org/educational-resource-hub>

Please fill out the form to add a resource you think might be helpful for the climate change and health community of practice.

What do you want the name or title of this resource to be?*

Name or title

Please provide the link to the resource here*

Link

What organization developed this resource?*

Organization name



Scan and Scroll
Down for
Submission
Form!



School of Public Health
Center for Climate & Health



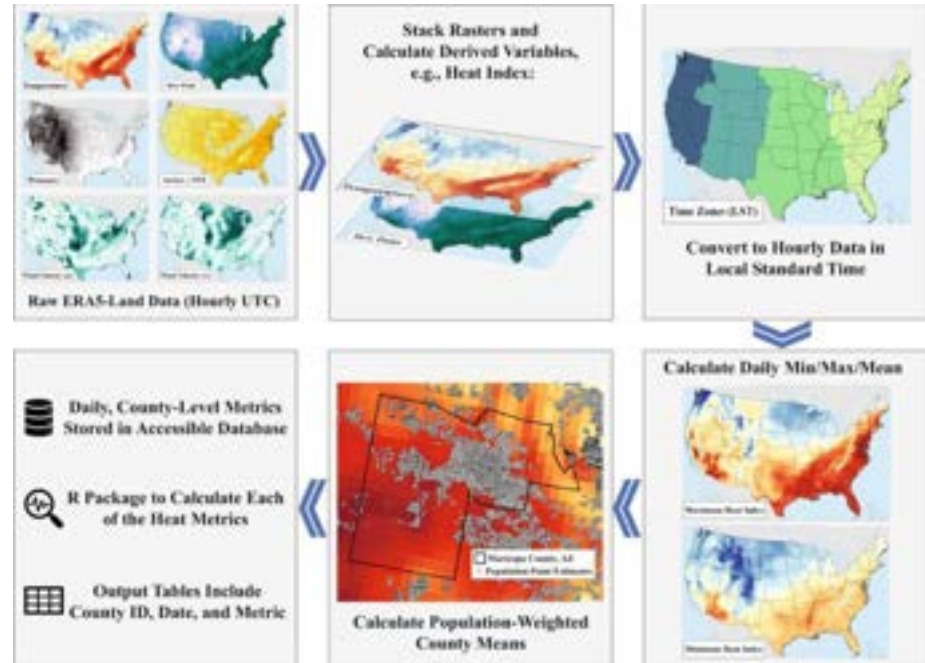
HARVARD
T.H. CHAN

SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

Wet-Bulb Globe Temperature, Universal Thermal Climate Index, and Other Heat Metrics for US Counties, 2000–2020

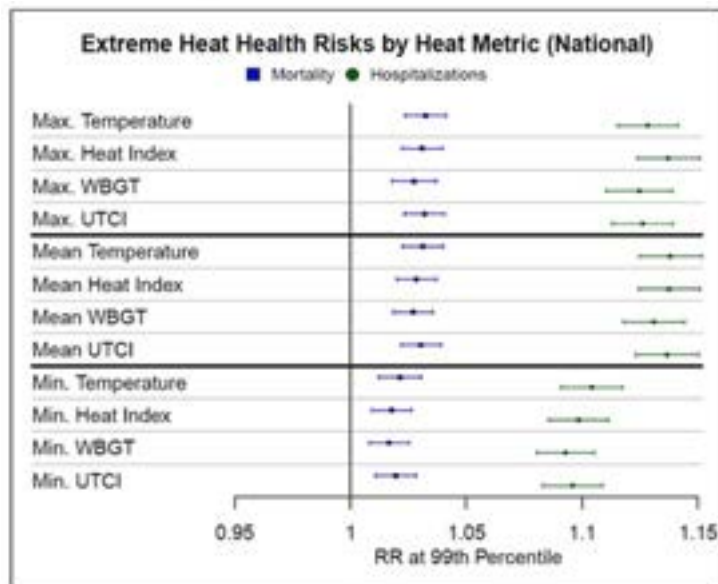
[Keith R Spangler](#)^{1,✉}, [Shixin Liang](#)^{1,2}, [Gregory A Wellenius](#)¹

- ERA-5 data is available globally with a historical record.
- CAFÉ-collection has an example for working with US and Global data examples.
- Github has R and Python versions that combine multiple existing packages for processing and merging to data.



Does choice of outdoor heat metric affect heat-related epidemiologic analyses in the US Medicare population?

Keith R. Spangler^{1,2*}, Quinn H. Adams³, Jie Kate Hu³, Danielle Braun^{3,4}, Kate R. Weinberger³, Francesca Dominici³, Gregory A. Wellenius³



- This example has been used in health research both within the United States and Globally.

Example: WBGT and UTCI ERA5 Aggregation Tutorial

- Includes template scripts for expedited batch ERA5 Download
- Applies **Spangler et al. heatmetrics R package** for step-by-step wet bulb globe and universal thermal climate index estimation
- Hourly raster to daily county aggregation, with time zone adjustment

Area-Level Estimation of Wet Bulb Globe Temperature and Universal Thermal Climate Index Pipeline

Project Overview

R code is provided for all steps from data download to county-level estimation of hourly wet bulb globe temperature

Usage

This repository provides the following tools for the query of any data from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 reanalysis dataset. The tools are designed to be used by county-level analysts. The repository includes a README file and a LICENSE file. The repository also includes a README file and a LICENSE file. The repository also includes a README file and a LICENSE file.

The download of ERA5 data in this repository adheres to the approach described in [Spangler et al. \(2023\)](#), and again uses the [heatmetrics](#) R package to efficiently query hourly ERA5 land and pressure levels data. The [heatmetrics](#) R package is then used for the processing of variables to relevant metrics of heat stress. The [heatmetrics](#) package must be installed manually. It can be downloaded from [CRAN](#).

Why Use ERA5?

When offering a dataset for analysis of the relationship between climate change and primary considerations, they include:

- **Temporal Resolution:** Depending on your question of interest hourly, daily or more appropriate. The ERA5 product we use for this analysis is hourly. That is, resolutions that can be downloaded from the [Climate Data Store](#).
- **Time Span:** Selecting a data source is also informed by the time span needed relative to extensive historical data record questions for the time span, and other

Notes on Computation

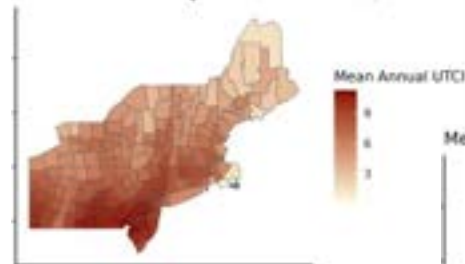
Data Sources

- ERA5 land and pressure levels data from [ECMWF](#)
- ERA5 hourly data from [Climate Data Store](#)
- [Heatmetrics](#) R package

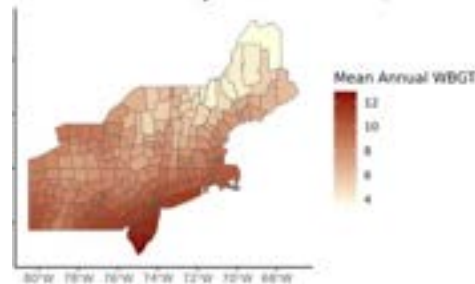
Workflow

A series of R scripts are provided in the code repository. Each script can be used to divide processing into working on ground temperatures, iteratively aggregating the time of processing. The scripts included were created for use on a university cloud computing network and may vary for your own time (learn more about high performance computing from [https://cloud.computing.org/](#)).

Mean Annual UTCI by Northeast Counties.



Mean Annual WBGT by Northeast Counties.



Required Input Data

- The GitHub will walk through how to use the Copernicus API to download ERA5-Land and ERA5
 - You will need to create a Copernicus account prior to use!
- There are several supporting datasets required for use of the pipeline as well
 - Find links and guidance directly in the GitHub scripts

• ECMWF Account Creation and API Extraction

See ecmwfr package for details!



• National Land Cover Database

Used to estimate urbanicity

National Land Cover Database (NLCD) 2011
Land Cover Conterminous United States (ver. 2.0,
May 2024)

• JRC Permanent Water

Used to mask inland water bodies

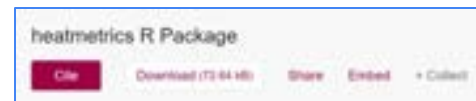
Map of permanent water bodies of the World v1
Collection: FLOODS | River Flood Hazard Maps in European and Global Scale

• Your Shapefile of Interest

```
50 # Load Shapefile. This approach involves a GIS application for the Northeast US.  
51 # downloaded using TIGER15. If you are conducting a global analysis or have  
52 # an existing shapefile, the below can be replaced to just set the shapefile path.  
53 # Import:  
54 # shapefile.path = "path/to/shapefile.shp"
```

• Heatmetrics R package from Figshare

.tar.gz file with all utci and wbgf functions

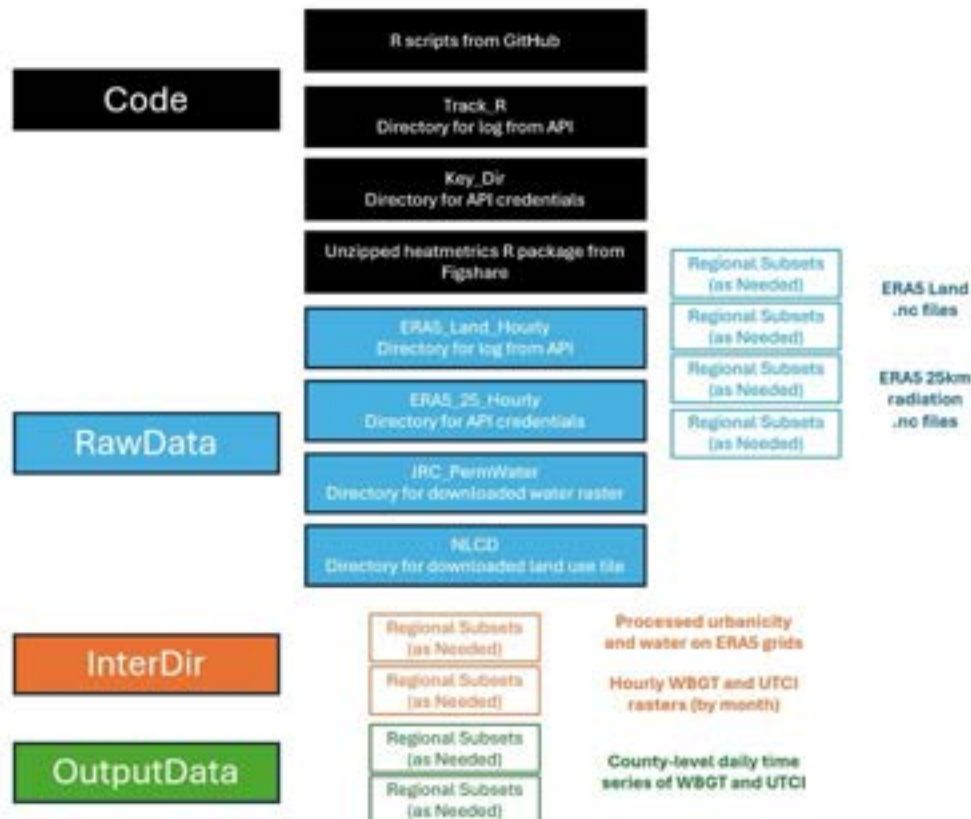


Setting Up Your Environment

- Ensuring your directory structure is created to ingest, organize, and access data is crucial at the outset
- Each R script lists where to enter your file paths at the top under 'User-Defined Parameters'

```
##### User-Defined Parameters #####  
# Set up directories to read in and output data  
#  
era_dir <- "RawData/ERAS_Hourly/" # ERAS-Land Rasters  
era_interdir <- "InterDir/" # Intermediate where static rasters output  
nlcd_dir <- "RawData/NLCD/" # where NLCD data is stored  
water_dir <- "RawData/JRC_Permanent/" # where JRC Permanent water is stored
```

Recommended Structure



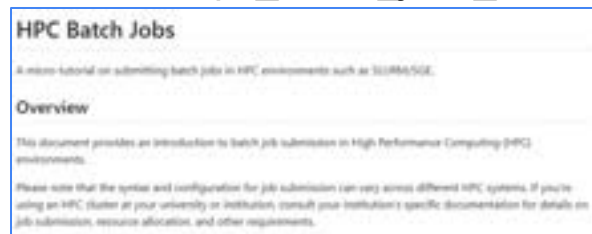
Computational Requirements and Workarounds

- WBGT and UTCI processing involves hourly data for 8+ meteorological variables and is time and compute-intensive
- The GitHub repository includes batch scripts to divide the processing and run different regions and time frames in parallel

Unfamiliar with Batch Scripting?

See the CAFÉ microtutorial!

Climate-CAFE/hpc batch jobs micro tutorial



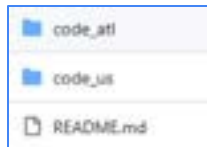
Example Script for Batch Submission to Subset Across Time and Space

[illegible]

Computational Requirements and Workarounds

- Batch scripting is an efficient way to separate large spatiotemporal processing but may not be available in all settings
- If you are processing a single county and/or a short time series (less than a year), an example processing pipeline can be found under code_atl
- Loops are used to process across months

```
# Set county. This code is developed for a single US county. These could be updated to reflect  
# any subdivision of your area of interest, as needed to divide processing into  
# more computationally efficient steps.  
#  
county_in <- 13121 # Example county is Fulton County, GA
```



- 01A_Query_ERAS_Land_CDSAPI_Batch_v2
- 01C_Download_Queue_ERAS_Land_CDSAPI_Batch_v2
- 02A_Query_ERAS_25km_CDSAPI_Batch_v1
- 02C_Download_Queue_ERAS_25km_CDSAPI_Batch_v1
- 03A_UTCI_WBGT_Intermediate_Static_v1
- 04A_UTCI_WBGT_Intermediate_Dynamic_v2
- 05A_ERAS_Extraction_Pts_v1
- 06A_Aggregate_ERAS_UTCI_UseFishnet_v2

Scripts 1+2: Downloading ERA5 Data

```
# the below is the format API request language. All of the inputs
# specified below in proper formatting can be identified by hovering a
# request using the Inspector (DB point-and-click interface for data
# discovery. https://cloud.ibm.com/console/marketplace/databases/mariadb-aws)
# select the variables, table, and result as the output format, and then
# select "Show API Request" at the bottom of the screen.
#
# Note that the target in the filename that will be imported to the path
# specified in the next part of the script. If using a link, ensure that
# that the unique features of each request are noted in the script (here
# we have each of the year, variable, and results [all our loop parameters])
# in the filename so we don't accidentally overwrite.
#
request_url = "http://
default_host_name = "mariadb-aws-us-east-1",
product_type = "mariadb",
variable = "country",
data = query_data,
time = ("2000", "2000", "2000",
        "2000", "2000", "2000",
        "2000", "2000", "2000",
        "2000", "2000", "2000",
        "2000", "2000", "2000",
        "2000", "2000", "2000",
        "2000", "2000", "2000",
        "2000", "2000", "2000"),
data_format = "text/csv",
download_path = "mariadb",
area = ("input_monthname, input_monthname, input_monthname, input_monthname",
        "target = partial('with-the-country-', yearvar, ' ', query_1, ' ', query_2, ' ', query_3, ' ', query_4, ' ', query_5, ' ', query_6, ' ', query_7, ' ', query_8, ' ', query_9, ' ', query_10, ' ', query_11, ' ', query_12, ' ', query_13, ' ', query_14, ' ', query_15, ' ', query_16, ' ', query_17, ' ', query_18, ' ', query_19, ' ', query_20, ' ', query_21, ' ', query_22, ' ', query_23, ' ', query_24, ' ', query_25, ' ', query_26, ' ', query_27, ' ', query_28, ' ', query_29, ' ', query_30, ' ', query_31, ' ', query_32, ' ', query_33, ' ', query_34, ' ', query_35, ' ', query_36, ' ', query_37, ' ', query_38, ' ', query_39, ' ', query_40, ' ', query_41, ' ', query_42, ' ', query_43, ' ', query_44, ' ', query_45, ' ', query_46, ' ', query_47, ' ', query_48, ' ', query_49, ' ', query_50, ' ', query_51, ' ', query_52, ' ', query_53, ' ', query_54, ' ', query_55, ' ', query_56, ' ', query_57, ' ', query_58, ' ', query_59, ' ', query_60, ' ', query_61, ' ', query_62, ' ', query_63, ' ', query_64, ' ', query_65, ' ', query_66, ' ', query_67, ' ', query_68, ' ', query_69, ' ', query_70, ' ', query_71, ' ', query_72, ' ', query_73, ' ', query_74, ' ', query_75, ' ', query_76, ' ', query_77, ' ', query_78, ' ', query_79, ' ', query_80, ' ', query_81, ' ', query_82, ' ', query_83, ' ', query_84, ' ', query_85, ' ', query_86, ' ', query_87, ' ', query_88, ' ', query_89, ' ', query_90, ' ', query_91, ' ', query_92, ' ', query_93, ' ', query_94, ' ', query_95, ' ', query_96, ' ', query_97, ' ', query_98, ' ', query_99, ' ', query_100, ' ', query_101, ' ', query_102, ' ', query_103, ' ', query_104, ' ', query_105, ' ', query_106, ' ', query_107, ' ', query_108, ' ', query_109, ' ', query_110, ' ', query_111, ' ', query_112, ' ', query_113, ' ', query_114, ' ', query_115, ' ', query_116, ' ', query_117, ' ', query_118, ' ', query_119, ' ', query_120, ' ', query_121, ' ', query_122, ' ', query_123, ' ', query_124, ' ', query_125, ' ', query_126, ' ', query_127, ' ', query_128, ' ', query_129, ' ', query_130, ' ', query_131, ' ', query_132, ' ', query_133, ' ', query_134, ' ', query_135, ' ', query_136, ' ', query_137, ' ', query_138, ' ', query_139, ' ', query_140, ' ', query_141, ' ', query_142, ' ', query_143, ' ', query_144, ' ', query_145, ' ', query_146, ' ', query_147, ' ', query_148, ' ', query_149, ' ', query_150, ' ', query_151, ' ', query_152, ' ', query_153, ' ', query_154, ' ', query_155, ' ', query_156, ' ', query_157, ' ', query_158, ' ', query_159, ' ', query_160, ' ', query_161, ' ', query_162, ' ', query_163, ' ', query_164, ' ', query_165, ' ', query_166, ' ', query_167, ' ', query_168, ' ', query_169, ' ', query_170, ' ', query_171, ' ', query_172, ' ', query_173, ' ', query_174, ' ', query_175, ' ', query_176, ' ', query_177, ' ', query_178, ' ', query_179, ' ', query_180, ' ', query_181, ' ', query_182, ' ', query_183, ' ', query_184, ' ', query_185, ' ', query_186, ' ', query_187, ' ', query_188, ' ', query_189, ' ', query_190, ' ', query_191, ' ', query_192, ' ', query_193, ' ', query_194, ' ', query_195, ' ', query_196, ' ', query_197, ' ', query_198, ' ', query_199, ' ', query_200, ' ', query_201, ' ', query_202, ' ', query_203, ' ', query_204, ' ', query_205, ' ', query_206, ' ', query_207, ' ', query_208, ' ', query_209, ' ', query_210, ' ', query_211, ' ', query_212, ' ', query_213, ' ', query_214, ' ', query_215, ' ', query_216, ' ', query_217, ' ', query_218, ' ', query_219, ' ', query_220, ' ', query_221, ' ', query_222, ' ', query_223, ' ', query_224, ' ', query_225, ' ', query_226, ' ', query_227, ' ', query_228, ' ', query_229, ' ', query_230, ' ', query_231, ' ', query_232, ' ', query_233, ' ', query_234, ' ', query_235, ' ', query_236, ' ', query_237, ' ', query_238, ' ', query_239, ' ', query_240, ' ', query_241, ' ', query_242, ' ', query_243, ' ', query_244, ' ', query_245, ' ', query_246, ' ', query_247, ' ', query_248, ' ', query_249, ' ', query_250, ' ', query_251, ' ', query_252, ' ', query_253, ' ', query_254, ' ', query_255, ' ', query_256, ' ', query_257, ' ', query_258, ' ', query_259, ' ', query_260, ' ', query_261, ' ', query_262, ' ', query_263, ' ', query_264, ' ', query_265, ' ', query_266, ' ', query_267, ' ', query_268, ' ', query_269, ' ', query_270, ' ', query_271, ' ', query_272, ' ', query_273, ' ', query_274, ' ', query_275, ' ', query_276, ' ', query_277, ' ', query_278, ' ', query_279, ' ', query_280, ' ', query_281, ' ', query_282, ' ', query_283, ' ', query_284, ' ', query_285, ' ', query_286, ' ', query_287, ' ', query_288, ' ', query_289, ' ', query_290, ' ', query_291, ' ', query_292, ' ', query_293, ' ', query_294, ' ', query_295, ' ', query_296, ' ', query_297, ' ', query_298, ' ', query_299, ' ', query_300, ' ', query_301, ' ', query_302, ' ', query_303, ' ', query_304, ' ', query_305, ' ', query_306, ' ', query_307, ' ', query_308, ' ', query_309, ' ', query_310, ' ', query_311, ' ', query_312, ' ', query_313, ' ', query_314, ' ', query_315, ' ', query_316, ' ', query_317, ' ', query_318, ' ', query_319, ' ', query_320, ' ', query_321, ' ', query_322, ' ', query_323, ' ', query_324, ' ', query_325, ' ', query_326, ' ', query_327, ' ', query_328, ' ', query_329, ' ', query_330, ' ', query_331, ' ', query_332, ' ', query_333, ' ', query_334, ' ', query_335, ' ', query_336, ' ', query_337, ' ', query_338, ' ', query_339, ' ', query_340, ' ', query_341, ' ', query_342, ' ', query_343, ' ', query_344, ' ', query_345, ' ', query_346, ' ', query_347, ' ', query_348, ' ', query_349, ' ', query_350, ' ', query_351, ' ', query_352, ' ', query_353, ' ', query_354, ' ', query_355, ' ', query_356, ' ', query_357, ' ', query_358, ' ', query_359, ' ', query_360, ' ', query_361, ' ', query_362, ' ', query_363, ' ', query_364, ' ', query_365, ' ', query_366, ' ', query_367, ' ', query_368, ' ', query_369, ' ', query_370, ' ', query_371, ' ', query_372, ' ', query_373, ' ', query_374, ' ', query_375, ' ', query_376, ' ', query_377, ' ', query_378, ' ', query_379, ' ', query_380, ' ', query_381, ' ', query_382, ' ', query_383, ' ', query_384, ' ', query_385, ' ', query_386, ' ', query_387, ' ', query_388, ' ', query_389, ' ', query_390, ' ', query_391, ' ', query_392, ' ', query_393, ' ', query_394, ' ', query_395, ' ', query_396, ' ', query_397, ' ', query_398, ' ', query_399, ' ', query_400, ' ', query_401, ' ', query_402, ' ', query_403, ' ', query_404, ' ', query_405, ' ', query_406, ' ', query_407, ' ', query_408, ' ', query_409, ' ', query_410, ' ', query_411, ' ', query_412, ' ', query_413, ' ', query_414, ' ', query_415, ' ', query_416, ' ', query_417, ' ', query_418, ' ', query_419, ' ', query_420, ' ', query_421, ' ', query_422, ' ', query_423, ' ', query_424, ' ', query_425, ' ', query_426, ' ', query_427, ' ', query_428, ' ', query_429, ' ', query_430, ' ', query_431, ' ', query_432, ' ', query_433, ' ', query_434, ' ', query_435, ' ', query_436, ' ', query_437, ' ', query_438, ' ', query_439, ' ', query_440, ' ', query_441, ' ', query_442, ' ', query_443, ' ', query_444, ' ', query_445, ' ', query_446, ' ', query_447, ' ', query_448, ' ', query_449, ' ', query_450, ' ', query_451, ' ', query_452, ' ', query_453, ' ', query_454, ' ', query_455, ' ', query_456, ' ', query_457, ' ', query_458, ' ', query_459, ' ', query_460, ' ', query_461, ' ', query_462, ' ', query_463, ' ', query_464, ' ', query_465, ' ', query_466, ' ', query_467, ' ', query_468, '
```

- Smaller requests move more quickly through Copernicus download queue
- Using batch submission to submit requests will expedite process
 - To further speed up this process, the API is directed to save files to a directory that has not been created on your device
 - After all requests have been processed through the queue, the directory for saving files is created.
 - Log language from the API will be saved to a file on your device so the files can be programmatically downloaded to the newly created directory.

Example Log Text which
can be used to
download files from
queue

```

Even after exiting your request is still being processed!
Retry downloading as soon as completed!
If you close your session use the following code:

wf_transfer(
url = "https://cds.climate.copernicus.eu/api/retrieve/v1/obs/
?&lang=en&format=JSON&urlp=/a/",
filenames = "era5-10m-2m-temperature-1921-01-01
"
)

```

Scripts 1+2: Downloading ERA5 Data

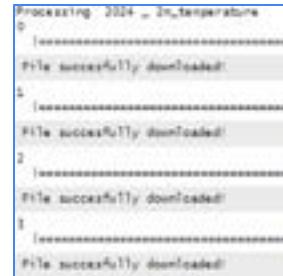
- View the status of your requests by visiting <https://cds.climate.copernicus.eu/request>
- Time from queue to completion will vary based on how many users are accessing data
- Once all requests are Complete, run script 1C to download to your device
- Downloading all variables typically takes several hours per year



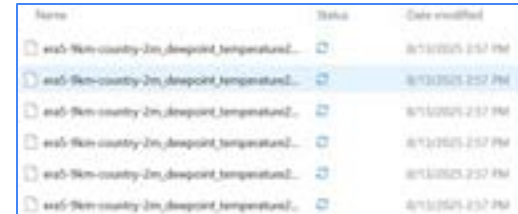
Product	Submission	Start	Status	Actions
ERA5 Land hourly data from 1979 to present	2023-08-14 10:12:00	2023-08-14 10:12:00	Pending	Cancel Download
ERA5 Land hourly data from 1979 to present	2023-08-14 10:12:00	2023-08-14 10:12:00	Pending	Cancel Download



Product	Submission	Start	Status	Actions
ERA5 Land hourly data from 1979 to present	2023-08-14 10:12:00	2023-08-14 10:12:00	Complete	Cancel Download
ERA5 Land hourly data from 1979 to present	2023-08-14 10:12:00	2023-08-14 10:12:00	Complete	Cancel Download



```
Processing 2024_2m_dewpoint
0
File successfully downloaded!
1
File successfully downloaded!
2
File successfully downloaded!
1
File successfully downloaded!
```



Name	Status	Date modified
era5-5km-country-2m_dewpoint_temperature2...	Downloaded	8/13/2025 2:57 PM
era5-5km-country-2m_dewpoint_temperature2...	Downloaded	8/13/2025 2:57 PM
era5-5km-country-2m_dewpoint_temperature2...	Downloaded	8/13/2025 2:57 PM
era5-5km-country-2m_dewpoint_temperature2...	Downloaded	8/13/2025 2:57 PM
era5-5km-country-2m_dewpoint_temperature2...	Downloaded	8/13/2025 2:57 PM

Scripts 3+4: Calculating Hourly WBGT + UTCI

- Script 3 creates 'static' intermediate raster layers for percent urbanicity, percent water, etc. that are not time-varying
- Script 4 runs for every hour you are processing. It is time and computationally intensive.
 - Set up is for batch scripting to divide by month and region

```
# For each hour in each day, loop through and recalculate solar estimate
#
for (date_in in c(dates_rast)) {

  # Track progress
  cat(date_in, "\n")

  # Extract index aligning with dates
  index_date <- which.max(date_in, time(era5_rast_srad_in))
  index_date <- c(xin(index_date) - 1, index_date[1:36])

  # Subset rasters to date specified for assessing hourly solar values
  era5_rast_srad_date <- era5_rast_srad_in[[index_date]]
  era5_rast_ssr_date <- era5_rast_ssr_in[[index_date]]
  era5_rast_strd_date <- era5_rast_strd_in[[index_date]]
  era5_rast_st_date <- era5_rast_st_in[[index_date]]

  # Initialize rasters to update
  era5_rast_srad_date2 <- era5_rast_srad_date
  era5_rast_ssr_date2 <- era5_rast_ssr_date
  era5_rast_strd_date2 <- era5_rast_strd_date
  era5_rast_st_date2 <- era5_rast_st_date

  # Loop through times and revise to be subtract previous hour
  #
```

Solar Radiation
Input Conversion
to Independent
Hourly Estimate

```
# Use wrapper function to have numeric and raster inputs
wbgt_wrapper <- function(x) {

  wbgt(
    year = as.numeric(x[1]), # your fixed year
    month = as.numeric(x[2]), # your fixed month
    day = day_d, # your fixed day
    lat = x[3],
    lon = x[4],
    solar = x[5],
    cte = x[6],
    fair = x[7],
    pres = x[8],
    fair = x[9],
    relhum = x[10],
    speed = x[11],
    speed2 = 18,
    dt = -0.051,
    urban = x[12]
  )
}

# Apply wrapper across raster values for the layer
result <- apply(rstack, wbgt_wrapper)

# Add WBGT layer
names(result) <- "wbgt"

# Add result to base
output_wbgt[[1]] <- result
```

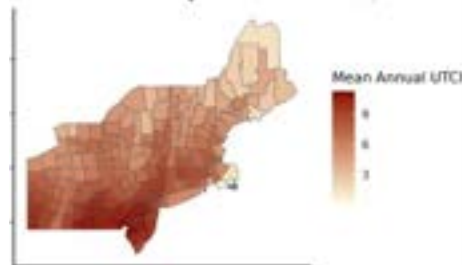
Use heatmetrics
package for WBGT
and UTCI formula

Scripts 5+6: Temporal and Spatial Aggregation

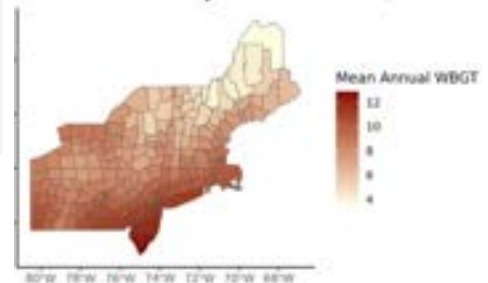
- Script 5 combines the ERA5 raster grid with the input polygon. This can then be used to extract the hourly raster values to the county
- Script 6 includes the daily and county-level aggregation, to create the final outputs for use

```
93 # ~~~~~ CREATE A FISHNET GRID OF THE RASTER EXTENT ~~~~~ #
94 #
95 # Here, we are making a shapefile that is a fishnet grid of the raster extent.
96 # It will essentially be a polygon of lines surrounding each ERA5 cell.
97 #
98 # Reference/credit: https://gis.stackexchange.com/a/243585
99 #
100 era_raster <- era_stack[[1]]
101 era_extent <- ext(era_raster)
102 xmin <- era_extent[1]
103 xmax <- era_extent[2]
104 ymin <- era_extent[3]
105 ymax <- era_extent[4]
```

Mean Annual UTCI by Northeast Counties.



Mean Annual WBGT by Northeast Counties.



WBGT and UTCI Next Steps

- Nationwide county products to be posted to CAFÉ Dataverse
- Update package with population weighting to county aggregation
- Enhance quality control procedures to flag values that may be more prone to error due to presence of water bodies, unexpected input values, etc. in line with Spangler et al.

	StCoFIPS	Date	Tmin_C	Tmax_C	Tmean_C	TDmin_C	TDmax_C	TDmean_C	NETmin_C	NETmax_C	NETmean_C	Himin_C	HImax_C	HImean_C	HXmin_C	HXmax_C
1	01001	2000-01-02	14.62	21.63	17.28	14.45	16.52	15.38	8.10	15.61	10.88	14.72	21.54	17.40	18.21	25.38
2	01003	2000-01-02	16.26	22.61	18.80	15.82	18.43	17.18	10.10	15.88	12.35	16.51	22.75	19.12	20.76	27.66
3	01005	2000-01-02	13.88	20.09	16.23	13.80	16.28	14.97	7.29	14.83	10.17	13.92	20.19	16.33	17.08	24.69
4	01007	2000-01-02	13.57	21.02	16.62	13.46	16.68	15.04	6.52	15.05	10.09	13.58	21.02	16.72	16.58	25.22
5	01009	2000-01-02	13.49	18.49	15.29	13.35	15.80	14.52	5.84	12.15	8.19	13.48	18.57	15.36	16.44	22.79
6	01011	2000-01-02	14.25	20.18	16.41	14.09	16.11	15.07	7.80	14.44	10.17	14.31	20.21	16.51	17.62	24.50
	HXmean_C	WBGTmin_C	WBGTmax_C	WBGTmean_C	UTCImin_C	UTCImax_C	UTCImean_C	Flag_T	Flag_TD	Flag_NET	Flag_HI	Flag_HX	Flag_WBGT	Flag_UTCI		
1	21.41	14.21	20.49	16.73	11.98	24.88	16.57	0	0	0	0	0	0	0		
2	24.13	15.78	22.33	18.65	13.32	25.46	17.92	0	0	0	0	0	0	0		
3	20.12	13.40	19.67	15.88	11.59	23.10	15.76	0	0	0	0	0	0	0		
4	20.56	13.17	20.20	16.15	9.71	24.62	15.55	0	0	0	0	0	0	0		
5	18.91	13.10	18.63	15.88	9.77	21.11	13.55	0	0	0	0	0	0	0		
6	20.36	13.78	19.51	16.02	11.84	22.45	15.75	0	0	0	0	0	0	0		

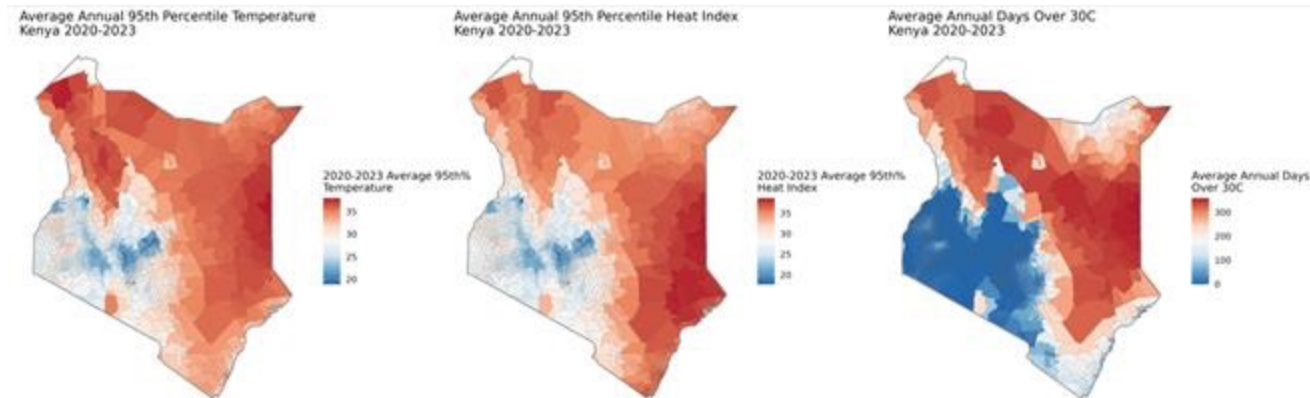


ERA5 Daily Heat Aggregation

<https://github.com/Climate-CAFE/era5-daily-heat-aggregation>



- Use existing R package to demonstrate process for query of ERA5 data with variable, time, and geographic extent inputs
- Across Kenya administrative boundaries, compute daily summary heat measures, including daily max temperature, heat index, and more



OpenStreetMap Road Metric Calculations

- ❖ **Objective:** Calculate total road lengths by type (e.g., highways, local streets) for municipalities in Mexico, adaptable to other countries
- ❖ **Data:** Uses OpenStreetMap (OSM) roads from Geofabrik, GADM shapefiles for boundaries, and R scripts to calculate and aggregate road sums.
 - Bash scripts included for fast processing.



CAFÉ GitHub: US Census Query

- Example census tutorial on CAFÉ Youtube

<https://www.youtube.com/watch?v=zC5nA3AVMpE>



CAFÉ GitHub Organization: R & Python Code Examples



Data Source	Product	Coverage
ERA5 Temperature	Example using Kenya Daily Heat Measure Statistics 2000-2023 (Example script, Dataverse entry)	Historic Weather Data, to 1950 including broader atmospheric levels
NASA FLDAS Temperature and Water Availability	Example using Iraq global administrative wards for 2000-2020 (Example script, Dataverse entry)	Measure monthly and annual temperature and drought metrics
Global Open StreetMap Road Exposure Metrics	Example using Mexico Administrative areas (Example script, Dataverse entry)	Calculate commonly used road measures into different geographical areas
PRISM Temperature	US ZCTA and County Population-Weighted 2020 Daily Max Temperature (Example script)	Historic Weather Data, to 1980 with high spatial resolution
Climate Data Online Weather Stations	US City-Level Time Series of Extreme Heat Days, 1950-Present (Example script, Dataverse entry)	Historic Weather Data, to 1940 for specific locations
CMIP6 Projected Temperature	US Tract-Level Daily Heat Threshold Measures at 2050 (Example script, Dataverse entry)	Projected Weather Data, through 2099 across SSPs
Global climate types and pm25 end-to-end pipelines	Aggregations for up to five embedded levels of geo boundaries for all countries in the globe (Pipelines, Dataverse entry)	Historical estimates, from 1998-2022



School of Public Health
Center for Climate & Health



HARVARD
T.H. CHAN

SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

CAFÉ Educational Resource Hub

- Crowd-sourced, searchable catalog to centralize information across experience levels, areas of interest, programming languages, etc.
- Includes coding tips, epidemiological methods, coding tutorials, and more!

Showing 111 of 128 items

Search:

Type of resource: **A**

- ☐ BOOK OR EBOOK/NOI 1047
- ☐ ONLINE COURSE
- ☐ VIDEO OR RECORDED WEBINAR
- ☐ ONLINE SYMPOSIUM OR VIDEO/FILE
- ☐ CODE/REPOSITORY

Only **A**

- ☐ COMMUNICATION
- ☐ DATA ACCESS
- ☐ DATA ORGANIZATION
- ☐ CLOUD COMPUTING/COMPUTING ENVIRONMENT
- ☐ DATA CLEANING/ORGANIZATION STRATEGIES
- ☐ GEOGRAPHICAL ANALYSIS
- ☐ GRAPH/PROSPECT/ANALYSIS/TOOL/FILE
- ☐ PROFESSIONAL DEVELOPMENT
- ☐ STATISTICAL METHOD
- ☐ SUBJECT MATTER TRAINING

Programming language: **A**

- ☐ R
- ☐ SAS
- ☐ SPSS/STATA/SQL

A The Relationship between Relative Extreme Temperature Events and Mortality in the United States

A video lecture discussing the relationship between relative extreme temperature events and mortality in the United States.

DATA ACCESS | SUBJECT MATTER TRAINING | VIDEO OR RECORDED WEBINAR | INTERMEDIATE | FREE

A CAFE University - How can CAFE help with the new NIH Data Management and Sharing Plan?

Speaker:
Julie Goldman, M.S.
Research Data Services (RDS), Harvard University

COMMUNICATION | SUBJECT MATTER TRAINING | VIDEO OR RECORDED WEBINAR | INTERMEDIATE | FREE

A CAFE University: NEHS Tools to Support Research and Communication in Climate Change and Health

Speaker: Trisha Casmiano, NEHS

DATA CLEANING/ORGANIZATION STRATEGIES | SUBJECT MATTER TRAINING | VIDEO OR RECORDED WEBINAR | INTERMEDIATE | FREE | PYTHON | R

A CAFE University: Wrangling and Processing Historic Weather Data for Climate and Health Studies

Speaker: Rachel Nethery, PhD
Co-lead for CAFE Capacity Building Function
Assistant Professor of Biostatistics, Harvard T.H. Chan School of Public Health

Please fill out the form to add a resource you think might be helpful for the climate change and health community of practice.

What do you want the name or title of this resource to be?*

Name or title

Please provide the link to the resource here*

Link

What organization developed this resource?*

Organization name

Scan and Scroll
Down for
Submission
Form!



CACHE

Center for Aging, Climate & Health

<https://agingclimatehealth.org/>



School of Public Health
Center for Climate & Health



HARVARD
T.H. CHAN

SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

An NIA-funded initiative to advance life-course research on the AD/ADRD **exposome** (U24AG088894)

What is the exposome?

People experience a variety of exposures from their homes and communities. These exposures, **broadly defined to capture social and behavioral exposures** beyond physical and chemical, are the **exposome**. Independently and combined, they can impact people's risk for developing Alzheimer's Disease and Alzheimer's Disease Related Dementias (AD/ADRD).

GECC Aims

1. **Identify** research priorities via consensus-building among a broad range of stakeholders
2. **Develop** guidance for measuring, harmonizing, and using exposome data
3. **Create** novel exposome measures and data
4. **Disseminate** open-access resources and products to the broader research community

Domains



Climate



Physical Environment



Social Environment



Policy

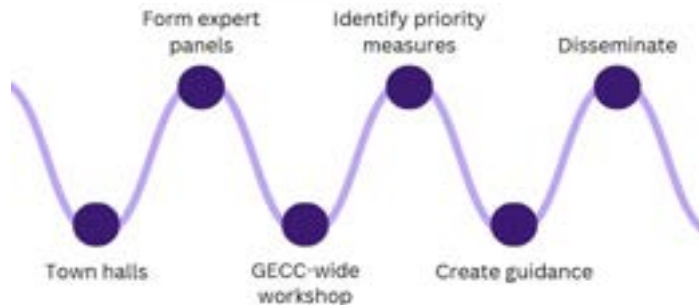


Community Services



Life Experiences

Process



Learn more at
www.gatewayexposome.org



Thank You!!

Connect with CAFÉ

Visit us at our website

<https://climatehealthcafe.org/>

Find us on LinkedIn at Climate & Health CAFÉ

Join our Community of Practice to receive our monthly newsletter and Slack channels!





Integration of Geospatial Data into Epidemiological and Clinical Health Research

Kyle P Messier

ISES/ISEE 2025 Pre-Meeting Workshop 06

Aug 17, 2025

Outline

Geospatial Exposure Data

Health Data

Integration Challenges

Tools to Support Geospatial Calculations and Integrations

Amadeus Examples

Open-Source Community

Questions



Geospatial Data Sources

Environmental and Exposure Variables



- NASA EARTHDATA
- Multi-Resolution Land Characteristics Consortium – NLCD
- Zenodo – EU Open Research Repository
- CAFÉ Dataverse (<https://dataverse.harvard.edu/dataverse.xhtml?alias=CAFE>)
- CHORDS Catalog (<https://chordshealth.org/discovery>)



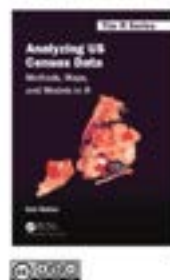
Health Data

Population and Person Level Data

- US Census
- CDC Places
- CDC Wonder



This is the online home of *Analyzing US Census Data: Methods, Maps, and Models in R*, a book published with CRC Press in 2023.



Address or Individual Level Data

- Point location of individual with health outcomes
- Epidemiological cohorts
- Electronic Health Records
- Minimum need is 1 spatial location
- Better is addresses over life course
- Best is detailed tracking
- Privacy concerns with address or time-course data



Data Integration

Spatial and Temporal Challenges

REVIEW ARTICLE OPEN

A review of geospatial exposure models and approaches for health data integration

Lara P. Clark¹, Daniel Zilber², Charles Schmitt¹, David C. Fargo³, David M. Reif², Alison A. Motsinger-Reif⁴ and Kyle P. Messier^{2,4,5*}

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2024

BACKGROUND: Geospatial methods are common in environmental exposure assessments and increasingly integrated with health data to generate comprehensive models of environmental impacts on public health.

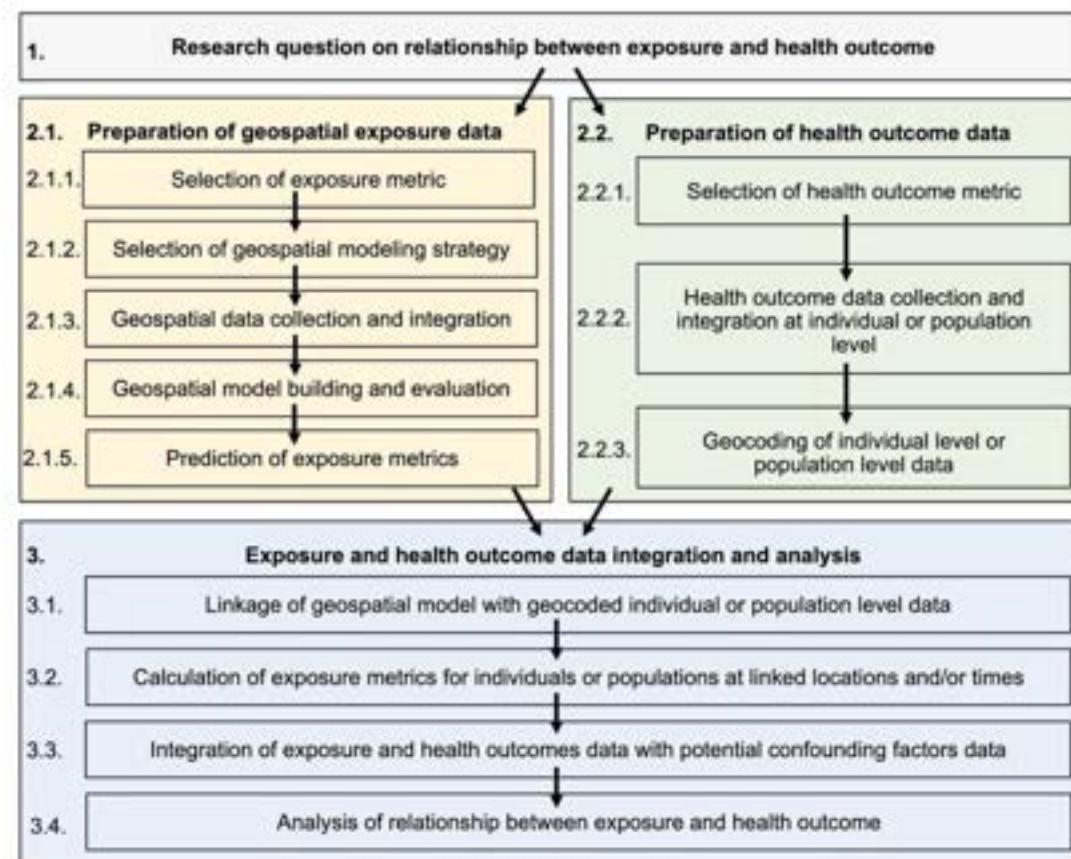
OBJECTIVE: Our objective is to review geospatial exposure models and approaches for health data integration in environmental health applications.

METHODS: We conduct a literature review and synthesis.

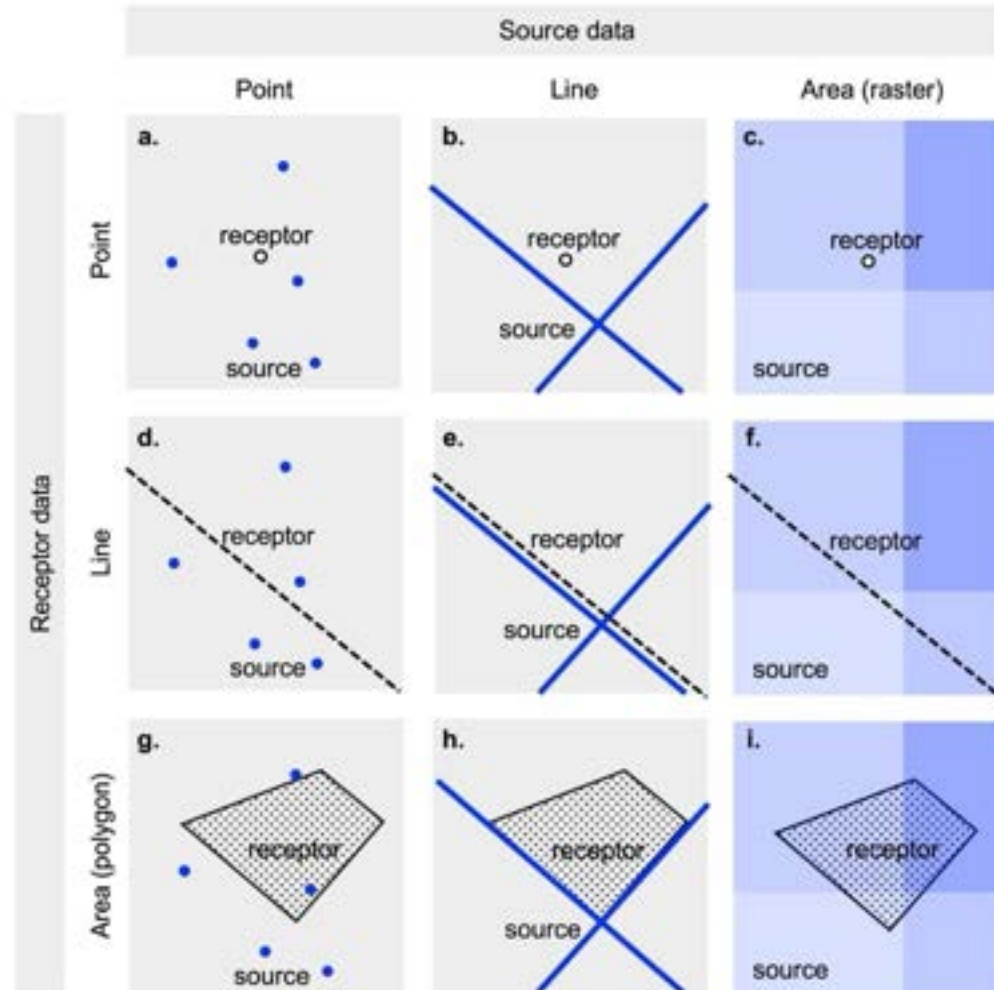
RESULTS: First, we discuss key concepts and terminology for geospatial exposure data and models. Second, we provide an overview of workflows in geospatial exposure model development and health data integration. Third, we review modeling approaches, including proximity-based, statistical, and mechanistic approaches, across diverse exposure types, such as air quality, water quality, climate, and socioeconomic factors. For each model type, we provide descriptions, general equations, and example applications for environmental exposure assessment. Fourth, we discuss the approaches used to integrate geospatial exposure data and health data, such as methods to link data sources with disparate spatial and temporal scales. Fifth, we describe the landscape of open-source tools supporting these workflows.

Keywords: Exposome; Environmental public health; Exposure modeling; Spatiotemporal; Toxicology; Linkage

Journal of Exposure Science & Environmental Epidemiology (2025) 35:131–148; <https://doi.org/10.1038/s41370-024-00712-8>

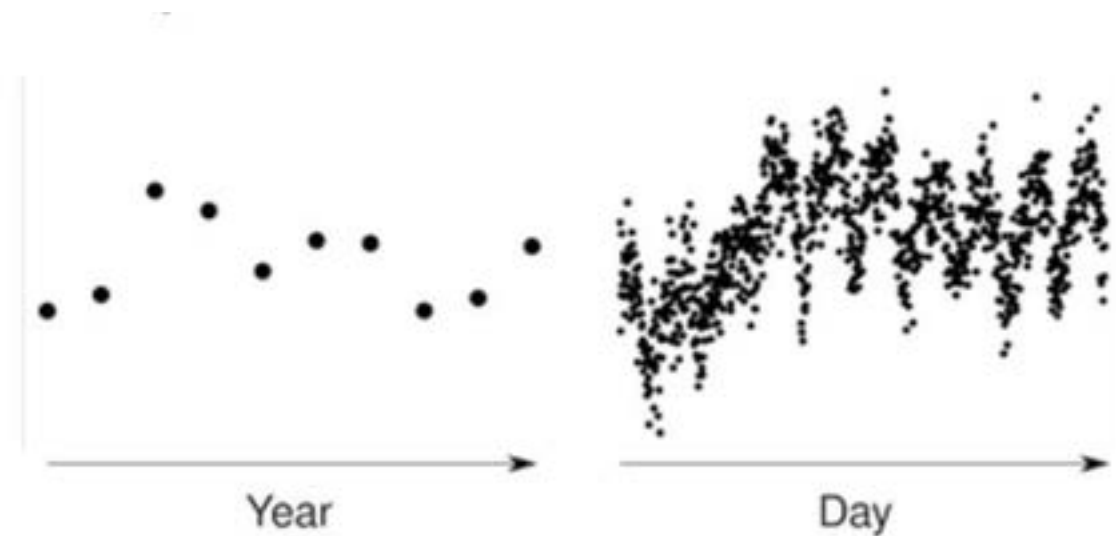
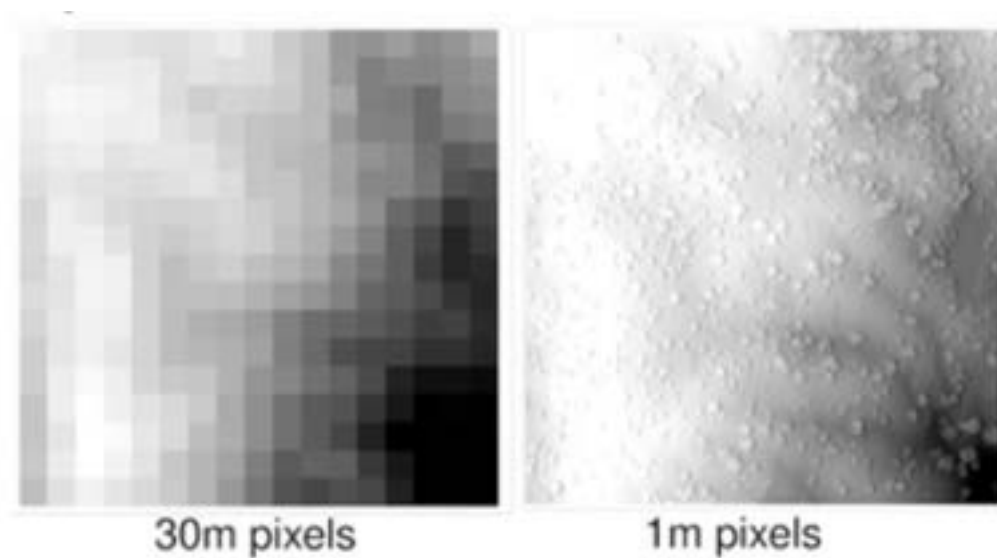


Spatial Geometry Relationships



- Source data → Geospatial and exposure data sources
- Receptor data → residential addresses, census tracts, etc.
- Integrating geospatial data to residential address → Row 1
- Integrating geospatial data to census tracts → Row 3

Spatial and Temporal Resolution



Tools to Help

- Amadeus
- Chopin
- Heat Exposure from Eva Marques
 - Samba
 - Mercury
 - Brassens
- Beethoven (in development)





National Institute of Environmental Health Sciences
Your Environment. Your Health.



amadeus: Accessing and analyzing large-scale environmental data in R

Mitchell Manware, Insang Song, Eva Marques, Mariana Alifa Kassien, Lara P Clark, Kyle P Messier

ISES/ISEE 2025 Pre-Meeting Workshop 06
Aug 17, 2025



“The development of innovative data science and data-driven approaches, such as integration of increasingly large and diverse data types and data sources, holds enormous potential for advancing key research needs across the environmental health sciences.”

- NIEHS, Strategic Plan 2025 – 2029, Area 4: Data Science and Computational Biology

F A I R



Findable



Accessible



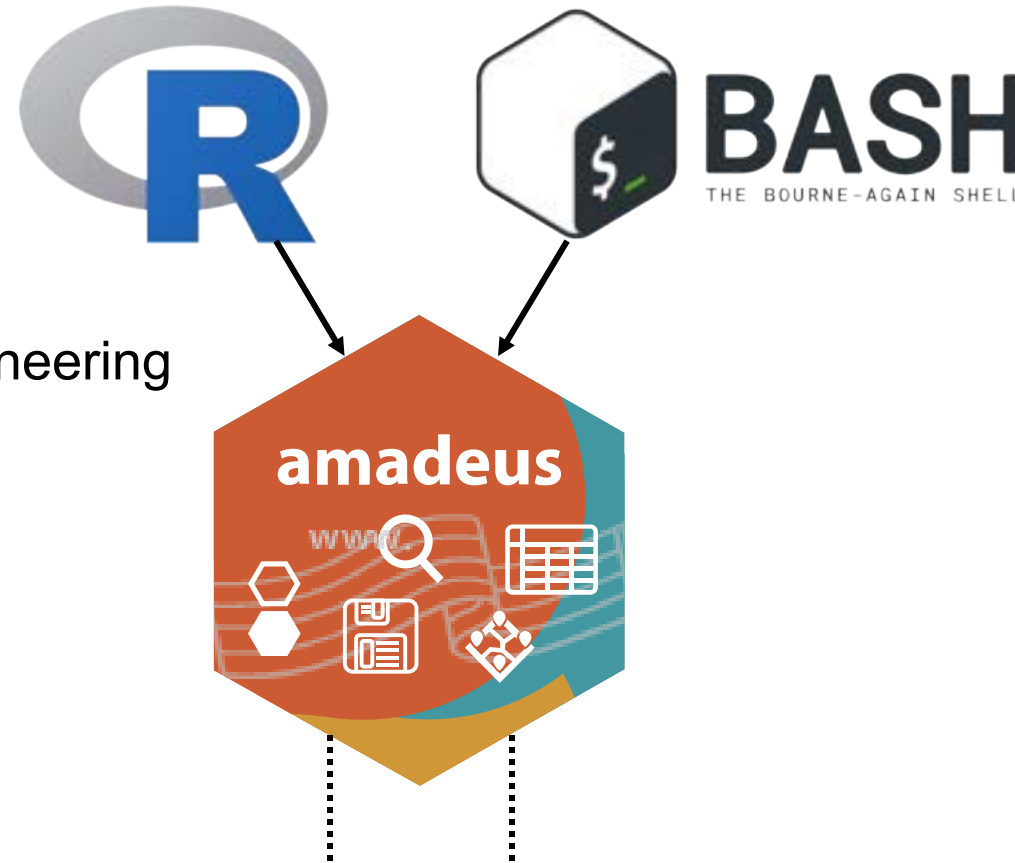
Interoperable

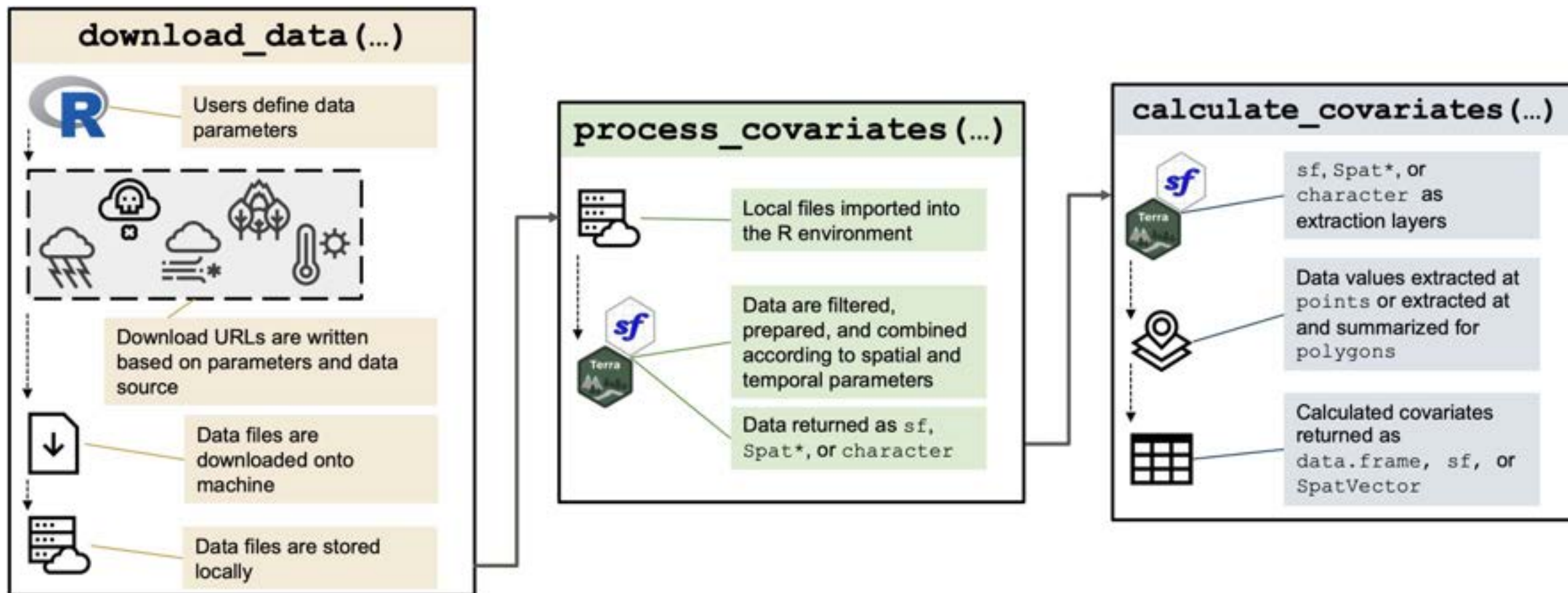


Reusable

amadeus

- Data access, preparation, and covariate calculation
 - Features, model features, feature engineering
- 16 data sources; >1,000 variables
- Interoperable
- Extensible
- Test-driven development
- Open source

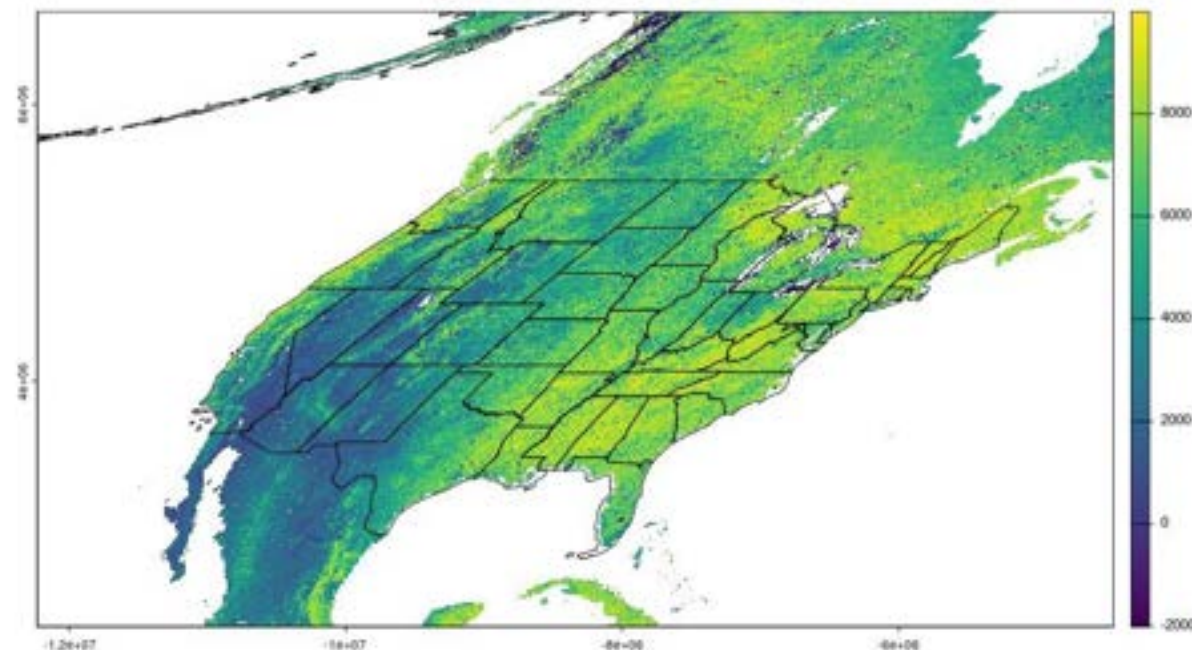




NASA Moderate Resolution Imaging Spectroradiometer (MODIS)

```
amadeus::download_modis(  
  product = "MOD13Q1",  
  version = "61",  
  horizontal_tiles = c(7, 13),  
  vertical_tiles = c(3, 6),  
  nasa_earth_data_token = readLines("~/nasa_token.txt"),  
  date = c("2023-06-01", "2023-06-10"),  
  directory_to_save = directory_mod13q1,  
  acknowledgement = TRUE,  
  download = TRUE,  
  remove_command = TRUE,  
  hash = TRUE  
)  
ndvi_states <- amadeus::calculate_modis(  
  from = list.files(  
    directory_mod13q1, full.names = TRUE, recursive = TRUE  
  ),  
  locs = tigris::states(year = 2023),  
  locs_id = "NAME",  
  radius = 0,  
  name_covariates = "NDVI_0_",  
  subdataset = "(NDVI)",  
  geom = "sf"  
)
```

- Normalized Difference Vegetation Index (NDVI)
- Mean zonal statistics computed for U.S. states and territories
- 24 lines



Auxiliary functions

- Internal functions manage complexities
- Various file formats
- Native coordinate reference systems
- Extraction location geometries
 - Point or polygon

```
calc_prepare_locs <-
...
#### project extraction locations to processed data
sites_e <- process_locs_vector(
  locs,
  terra::crs(from),
  radius
)
...
```

```
process_flatten_sds <- function(
  path = NULL,
  subdataset = NULL,
  fun_agg = "mean",
  ...
) {
  status_curv <- suppressWarnings(terra::is.rotated(terra::rast(path)))
  if (any(status_curv)) {
    stop(
      "The raster is curvilinear. Please rectify or warp
      the input then flatten it manually."
    )
  }
  ...
}
```

```
calc_worker <-
...
#### extract layer data at sites
if (terra::geomtype(locs_vector) == "polygons") {
  ## apply exactextractr::exact_extract for polygons
  sites_extracted_layer <- exactextractr::exact_extract(
    data_layer,
    sf::st_as_sf(locs_vector),
    progress = FALSE,
    force_df = TRUE,
    fun = fun,
    max_cells_in_memory = max_cells
  )
} else if (terra::geomtype(locs_vector) == "points") {
  #### apply terra::extract for points
  sites_extracted_layer <- terra::extract(
    data_layer,
    locs_vector,
    method = "simple",
    ID = FALSE,
    bind = FALSE,
    na.rm = TRUE
  )
}
...
```


Interoperability

- Output classes are from popular spatial packages
- Linking to U.S. Census data
- Data analysis pipelines
- Beyond R

```
# R
amadeus::download_data(
  dataset_name = "narr",
  variable = "shum",
  year = 2021,
  directory_to_save = "./data/",
  acknowledgement = TRUE,
  download = TRUE,
  remove_command = TRUE
)
```

```
# Python
import netCDF4
shum_python = netCDF4.Dataset("data/shum/shum.202101.nc")
shum_python.variables["shum"]
```

```
## <class 'netCDF4._netCDF4.Variable'>
## float32 shum(time, level, y, x)
##   GRIB_id: 51
##   GRIB_name: SPFH
##   _FillValue: 9.96921e+36
##   coordinates: lat lon
##   grid_mapping: Lambert_Conformal
##   level_desc: Pressure Levels
##   standard_name: specific humidity
```

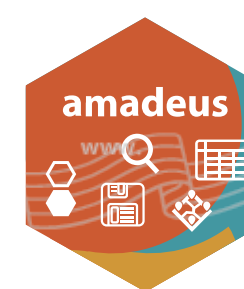


Test-driven development coverage 99.68%

- Reliability and stability
- Error catching during development
- Interpretable error messages
- Unit tests
 - URLs return successful response status (`download_*`)
 - Correct object classes (`process_*`; `calculate_*`)
 - Identify missing or improper parameters (all functions)
- Integration tests
 - No errors when `process_*` run with live `download_*` files



+



CRAN 1.2.3

repo status Active

amadeus: Accessing and Analyzing Large-Scale Environmental Data

Functions are designed to facilitate access to and utility with large scale, publicly available environmental data in R. The package contains functions for downloading raw data files from web URLs (`download_data()`), processing the raw data files into clean spatial objects (`process_covariates()`), and extracting values from the spatial data objects at point and polygon locations (`calculate_covariates()`). These functions call a series of source-specific functions which are tailored to each data source/dataset's particular URL structure, data format, and spatial/temporal resolution. The functions are tested, versioned, and open source and open access. For `sum_rdc()` method details, see Messier, Akita, and Serre (2012) <[doi:10.1021/es201152a](https://doi.org/10.1021/es201152a)>.

Version: 1.2.3
Depends: R (≥ 4.1.0)
Imports: dplyr, sf, sfutils, stats, rvest, methods, data.table, httr, rvest, rvestextract, utils, stringr, testthat (≥ 3.0.0), stars, tidyr, rlang, rhdgltools, archive, collapse, Rdpack
Suggests: covr, scater, knitr, rmarkdown, byronom, FSN, doRNG, dcrtools, stringr, tidy, spelling
Published: 2025-01-16
DOI: [10.32614/CRAN.package.amadeus](https://doi.org/10.32614/CRAN.package.amadeus)
Author: Mitchell Manware [aut, ctb], Insang Song [aut, ctb], Eva Marques [aut, ctb], Mariana Alifa Kassien [aut, ctb], Elizabeth Scholl [ctb], Kyle Messier [aut, cre], Spatiotemporal Exposures and Toxicology Group [cph]
Maintainer: Kyle Messier <kyle.messier@nih.gov>
BugReports: <https://github.com/NIHSE/amadeus/issues>
License: MIT + file LICENSE
URL: <https://nichs.github.io/amadeus/>
NeedsCompilation: no
Language: en-US
Materials: README, NEWS
CRAN checks: [amadeus results](#)

Documentation:

Reference manual: [amadeus.html](#), [amadeus.pdf](#)
Vignettes: [Climatology Lab gridMET \(source, R code\)](#), [North American Regional Reanalysis \(NARR\) \(source, R code\)](#), [Climatology Lab TerraClimate \(source, R code\)](#)

Downloads:

Package source: [amadeus_1.2.3.tar.gz](#)
Windows binaries: r-devel: [amadeus_1.2.3.zip](#), r-release: [amadeus_1.2.3.zip](#), r-oldrel: [amadeus_1.2.3.zip](#)
macOS binaries: r-release (arm64): [amadeus_1.2.3.rgz](#), r-oldrel (arm64): [amadeus_1.2.3.rgz](#), r-release (x86_64): [amadeus_1.2.3.rgz](#), r-oldrel (x86_64): [amadeus_1.2.3.rgz](#)
Old sources: [amadeus.archive](#)

Linking:

Please use the canonical form <https://CRAN.R-project.org/package=amadeus> to link to this page.

amadeus



R CMD check passing coverage 98.72% test passing pkgdown passing repo status Active
CRAN 1.2.3.4 Downloads: 5007

amadeus is a mechanism for data, environments, and user setup for common environmental and weather datasets in R. **amadeus** has been developed to improve access to and utility with large scale, publicly available environmental data in R.

See the peer-reviewed publication, [Amadeus: Accessing and analyzing large scale environmental data in R](#), for full description and details.

Cite **amadeus** as: > Manware, M., Song, I., Marques, E. S., Kassien, M. A., Clark, L. P., & Messier, K. P. (2025). Amadeus: Accessing and analyzing large scale environmental data in R. Environmental Modelling & Software, 186, 106352.

Installation

amadeus can be installed from CRAN with `install.packages()` or from GitHub with `pak`.

```
install.packages("amadeus")
```

```
pak::pak("NIHSE/amadeus")
```

Download

download_data accesses and downloads raw geospatial data from a variety of open source data repositories. The function is a wrapper that calls source-specific download functions, each of which account for the source's unique combination of URL, file naming conventions, and data types. Download functions cover the following sources:

Links

[View on CRAN](#)
[Browse source code](#)
[Report a bug](#)

License

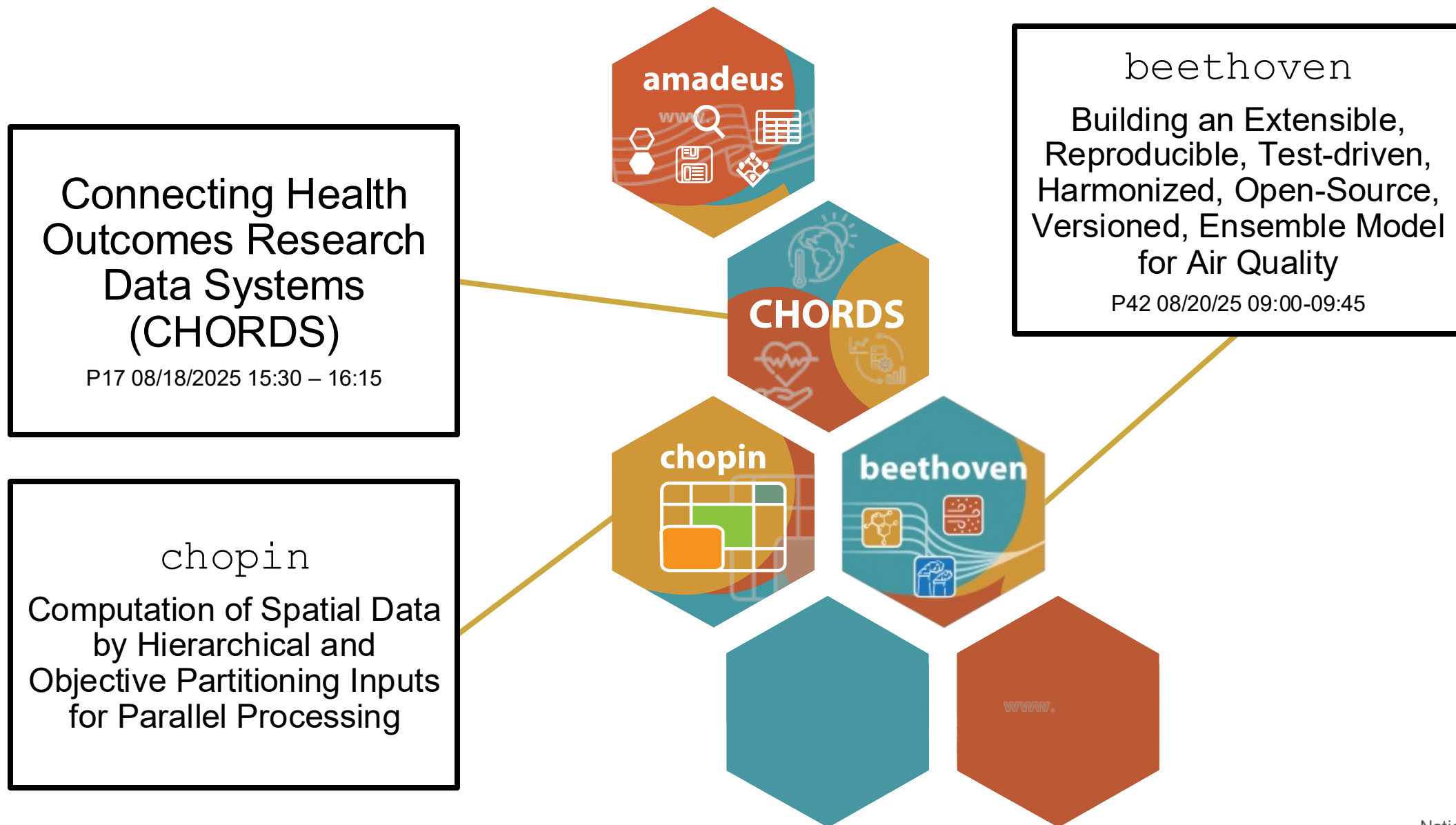
[Full license](#)
[MIT + file LICENSE](#)

Citation

[Citing amadeus](#)

Developers

Mitchell Manware
Author, contributor
Insang Song
Author, contributor
Eva Marques
Author, contributor
Mariana Alifa Kassien
Author, contributor
Kyle Messier
Author, maintainer
[More about authors...](#)



chopin



Convenient parallel geospatial operations
via data partitioning and automatic
deployment

- Regular grid, hierarchical, balanced

Custom and common gridding (Uber H3,
DGGRID) options

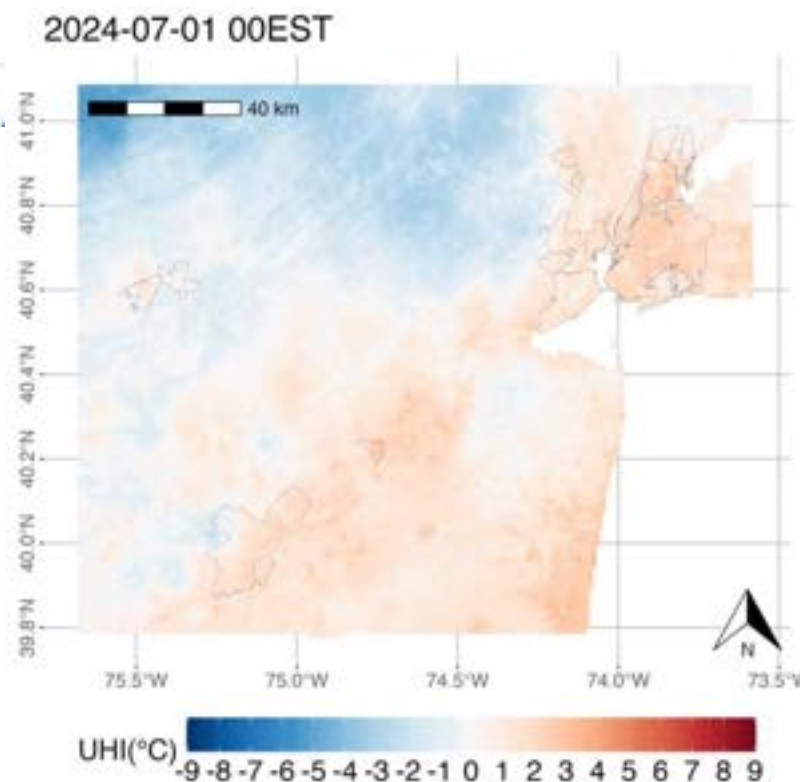
Available in CRAN

Published in *SoftwareX*



lint passing DOI 10.5281/zenodo.15595978

samba: Spatiotemporal Air-temperature Model with Bayesian Approach



https://github.com/NIEHS/urbanheat_pipeline

A factory to generate air temperature data at hourly resolution on US cities!

This project is an assembly line using [brassens](#), [samba](#) and [mercury](#) libraries to automatically generate 2-meters air temperature rasters on a list of US cities and observed months.

This pipeline has generate casestudies on the top largest Urban Census areas on different climatic events (heatwave, blizzard, typical weather...) and seasons. This dataset has been designed for environmental epidemiology and public health studies on (extreme) temperature exposure in US cities. It can also be leveraged by urban climatologists to improve our understanding of the spatiotemporal evolution of the urban heat island with regard to the variety of city layout and climatic region in the US.

To cite the dataset: Marques, E., & Messier, K. (2025). Urban air temperature at high spatiotemporal resolution on major US cities (soon available) [Dataset]. Harvard Dataverse. <https://doi.org/10.7910/DVN/HNVCBR756>

Urban heat island movie generation

If you want to create a movie from the 2m-air temperature raster:

- go on our [urbanheat_pipeline Github repository](#).
- download `container/container_movies.def`, `create_movie.sh` and `create_maps_for_movie.R`
- open a terminal and build the container (you need to have `apptainer` installed). It might take a few minutes.

```
apptainer build --fakeroot container_movies.sif container_movies.def
```

[Urban air temperature at high spatiotemporal resolution on major US cities](#)



Jun 24, 2025

Marques, Eva; Messier, Kyle, 2025, "Urban air temperature at high spatiotemporal resolution on major US cities", <https://doi.org/10.7910/DVN/HNVCBR>, Harvard Dataverse, V2

... This dataset has been designed for environmental epidemiology and public health studies on (extreme) temperature exposure in US cities. It can also be leveraged by **urban** climatologists to improve our understanding of the spatiotemporal evolution of the **urban heat** island with regard to the variety of city layout and climatic region in the US. ...

Related Publication Citation: @article{marques2025improved, title={Improved high resolution heat exposure assessment with personal weather stations and spatiotemporal Bayesian models}, author={Marques, Eva and Messier, Kyle P}, journal={Authorea Preprints}, year={2025}, publisher={Authorea} }

F



Findable

- CRAN
- GitHub
- Publication[†]
- CHORDS

A



Accessible

- Open source
 - No cost

I



Interoperable

- `sf` and `terra`
- Python and other programming languages

R



Reusable

- Local files
- Re-executable code

[†]Manuscript published at *Environmental Modelling & Software*



Amadeus Examples

Spatial and Temporal Integration



Amadeus & Open-Source Community

That's you!

Participating in Open-Source Projects

LINKS

- [View on CRAN](#)
- [Browse source code](#)
- [Report a bug](#)

LICENSE

- [Full license](#)
- [MIT + file LICENSE](#)

COMMUNITY

- [Contributing guide](#)
- [Code of conduct](#)

CITATION

- [Citing parsnip](#)

13 labels

Label	Description
new dataset	Add a dataset to spatiotemporal rating
enhancement	New feature or request
documentation	Improvements or additions to documentation
checkboxlist	
bug	Something isn't working
enhancement	How can "amadeus" be improved?
help wanted	Extra attention is needed
wontfix	This will not be worked on
question	Further information is requested

1. Open Issues
2. Ask Questions
3. Report Bugs
4. Help with Documentation
5. Contribute examples

Acknowledgements

- Mitchell Manware, M.S.
- Insang Song, Ph.D.
- Connecting Health Outcomes Research and Data Systems Team
 - Aubrey K Miller, M.D., M.P.H.
 - Charles P Schmitt, Ph.D.
 - David M. Reif, Ph.D.
 - Adam B. Burkholder, M.S.
 - Lara P Clark, Ph.D.
 - Alison A. Motsinger-Reif, Ph.D.
 - Ann Liu, Ph.D., M.P.H.
 - Trisha M. Castranio

Questions?



Scan for the GitHub codebase.

Dowle, M., & Srinivasan, A. (2024). data.table: Extension of 'data.frame'. R package version 1.14.2. <https://cran.r-project.org/package=data.table>

GitHub, Inc. (n.d.). *GitHub logo*. <https://github.com/logos>

GitHub, Inc. (n.d.). *GitHub Invertocat logo*. <https://github.com/logos>

GNU Project. (2024). *Bash (Bourne Again Shell) [Computer software]*. Free Software Foundation. <https://www.gnu.org/software/bash/>

Hijmans, R. J. (2024). terra: Spatial data analysis package for R. R package version 1.7-19. <https://cran.r-project.org/package=terra>

Maccherone, B., and Frazier, S. (n.d.). *Moderate Resolution Imaging Spectroradiometer – Data*. <https://modis.gsfc.nasa.gov/data/>

Marsner. (n.d.) *Why Test-Driven Development (TDD)*. <https://marsner.com/blog/why-test-driven-development-tdd/>

Miquel, M., & Dorman, M. (2024). targets: Make dynamic and reproducible workflows. R package version 1.0.0. <https://cran.r-project.org/package=targets>

NASA Earth Observing System Data and Information System. (n.d.). *MODIS Terra Vegetation Indices (MOD13A1) [Data set]*. NASA Goddard Space Flight Center. <https://modis.gsfc.nasa.gov/data/>

National Aeronautics and Space Administration. (n.d.). *NASA logo*. <https://www.nasa.gov/>

National Institute of Environmental Health Sciences. (2024, August 14). Climate and Health Outcomes Research Data Systems (CHORDS). <https://www.niehs.nih.gov/research/programs/chords>

National Institute of Environmental Health Sciences. (2024, September 23). *Area 4: Data Science and Computational Biology*. <https://www.niehs.nih.gov/about/strategicplan/research/data-science>

National Library of Medicine. (n.d.). *FAIR Data*. <https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-200.html>

National Oceanic and Atmospheric Administration. (n.d.). *NOAA logo*. <https://www.noaa.gov>

Pebesma, E. J. (2024). sf: Simple features for R. R package version 1.0-13. <https://cran.r-project.org/package=sf>

Python Software Foundation. (2024). Python programming language. Version 3.11.5. <https://www.python.org/>

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>

Song, I. & Messier, K. (2024). chopin: Computation of Spatial Data by Hierarchical and Objective Partitioning of Inputs for Parallel Processing. R package version 0.9.0. <https://docs.ropensci.org/chopin>

Song, I., Marques, E., Manware, M., et al. (2024). beethoven: Building an Extensible, rEproducibile, Test-driven, Harmonized, Open-source, Versioned, ENsemble model for air quality. R package version 0.4.3. <https://github.com/NIEHS/beethoven>

U.S. Department of the Interior. (n.d.). *DOI logo*. <https://www.doi.gov/>

U.S. Environmental Protection Agency. (n.d.). *EPA logo*. <https://www.epa.gov/>

U.S. Geological Survey. (n.d.). *USGS logo*. <https://www.usgs.gov/>

University of California, Merced. (n.d.). *SNRI logo*. <https://snri.ucmerced.edu/>

Walker, K. (2024). tidycensus: Load US census data from the US Census Bureau's API. R package version 1.0.0. <https://cran.r-project.org/package=tidycensus>

Walker, K. (2024). tigris: R package for loading TIGER/Line shapefiles. R package version 1.0.0. <https://cran.r-project.org/package=tigris>

Wickham, H. (2024). *testthat: Get started with testing*. R package version 3.1.0. <https://cran.r-project.org/package=testthat>

Wickham, H., François, R., Henry, L., & Müller, K. (2024). dplyr: A grammar of data manipulation. R package version 1.1.2. <https://cran.r-project.org/package=dplyr>

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>



National Institute of Environmental Health Sciences
Your Environment. Your Health.

Efforts to standardize geospatial data elements and to create geospatial CDEs

Maria Shatz, PhD
Office of Data Science
National Institute of Environmental Health Sciences (NIEHS)
Research Triangle Park, NC, USA
e-mail: maria.shatz@nih.gov



Joint Annual Meeting
of the International Society of
Exposure Science and the
International Society for
Environmental Epidemiology



Conflict of Interest Statement

The author declares no conflict of interest.



Examples of geospatial measures for health

Geospatial Measure	Description	Example Data Sources	Public Health Use
Air Quality Index (AQI), PM _{2.5} , NO ₂ , O ₃	Measures air pollution levels by location	EPA AirNow, NASA satellite remote sensing	Identify high-risk areas for respiratory/cardiovascular illness; issue health alerts
Proximity to Pollution Sources	Distance to highways, animal feed facilities, oil wells, waste sites	National Land Cover Database, EPA Toxic Release Inventory	Assess exposure risk and plan zoning/regulatory actions
Noise Pollution Maps	Sound intensity across neighborhoods	Environmental monitoring stations noise sensors	Study effects on sleep, cardiovascular health, stress
Green Space Availability	Access to parks, tree canopy coverage	Satellite imagery, OpenStreetMap	Promote physical activity, mental health benefits
Wildfire and smoke exposure	Location based exposure to smoke and particle pollution	AirNow Fire and Smoke Map, US EPA and Forest Service	Enables the public to take protective actions against wildfire smoke.
Population Density & Demographics	Distribution of population groups (age, income, etc.)	Census data	Plan resource allocation and emergency response
Disease Incidence Heat Maps	Spatial distribution of reported cases	Public health registries, GIS geocoding	Detect clusters/outbreaks for rapid response
Flood/Heatwave/Wildfire Risk Maps	Areas prone to climate hazards	FEMA flood maps, NOAA climate data, satellite fire tracking	Prepare evacuation plans, target public health messaging
Water Quality Mapping	Location of contamination in rivers, wells, and systems	EPA Safe Drinking Water Information System	Prevent and mitigate waterborne disease outbreaks

Value of standardized data elements

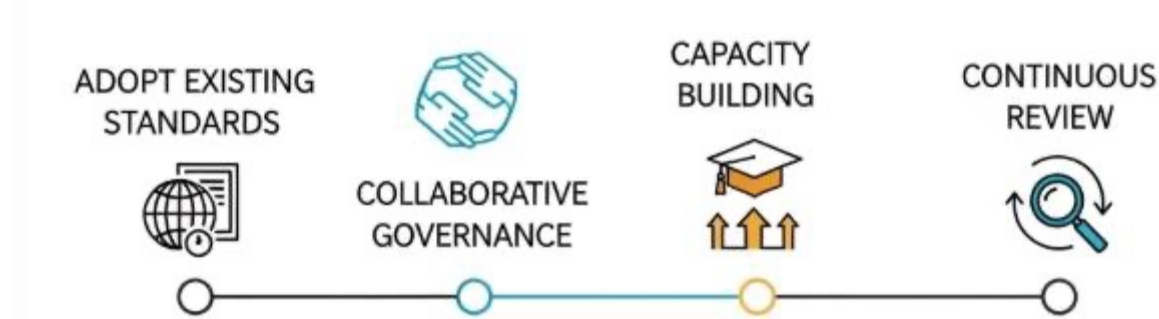
Reason	Impact if standardized	Impact if Not standardized
Consistency	PM2.5 in one state means the same as PM2.5 in another	Apples-to-oranges comparisons between regions and cohorts
Data Integration	Easily merge environmental, health, and demographic datasets	Incompatible formats, loss of detail
Reproducibility	Other researchers can verify and replicate findings	Conflicting results due to method differences
Scalability	GIS tools and dashboards work in multiple jurisdictions	Costly customizations for each location
Public Trust	Clear, consistent public health communication	Confusion, misinformation, reduced engagement

Core Principles for Standardization

- **Clear Definitions** – Agree on exactly what is measured (e.g., “walkability” defined by X metrics).
- **Common Units** – Use the same measurement units (e.g., micrograms/m³ for PM2.5).
- **Temporal Alignment** – Match timeframes (e.g., daily averages vs. annual means).
- **Spatial Resolution** – Define the geographic scale (e.g., census tract, zip code, 1 km² grid).
- **Metadata Documentation** – Record data sources, collection methods, and processing steps.
- **Interoperable Formats** – Use open, standardized GIS data formats (GeoJSON, shapefiles).



Implementation Path



- **Adopt Existing Standards** – Use established frameworks (CDC’s SVI, WHO air quality guidelines).
- **Collaborative Governance** – Engage diverse stakeholders in setting norms.
- **Capacity Building** – Train staff in GIS, data cleaning, and metadata creation.
- **Continuous Review** – Update standards as technology and health priorities evolve.

Challenges

- While there are standards for health data (e.g. ICD-10, LOINC, SNOMED) or for geospatial data (Open Geospatial Consortium, OGD) there are no standard on intersection of the two fields
- Multiple existing measures with no community consensus
- Researchers are not aware of the existing standards
- Variety of data sources and computational methods
- Due to the temporal element the values are dynamic;
- The presentation should be not only machine-readable but easily reused for computation with different parameters

Approaches to standardization of data elements

How do we move forward as a community to improve this situation?

NIH proposed creating and endorsing CDEs and publishing them in the CDE repository moving forward

What is a Common Data Element (CDE)?

- A CDE is a fixed representation of a variable comprising one defined question
- paired with a specified set of allowable responses including value range and units if applicable,
- mapped to semantic concepts and codes like UMLS, SNOMED AND
- that are **used in common** across multiple: studies, research sites, initiatives, datasets, etc.

Question (variable): *What is the participant's height?*

Allowable answers: *Feet and inches in whole numbers*

- *For example, if all clinical studies supported by an IC were required to collect height data with this question and answer pair,*
- *this data element would be common to all of those studies - it would be a **Common Data Element***

NIH CDE Governance Committee

- The CDE Governance Committee of the NIH CDE Task Force will determine whether CDEs meet **criteria that merit their designation as “NIH-endorsed CDEs.”**
- The criteria:
 - Submission from a **recognized NIH body** (e.g., NIH ICO, trans-NIH committee, etc.)
 - **Clear definition** of the variable
 - For assessment instruments, documented evidence of **reliability and validity**
 - Both human- and machine-readable **formats** preferred
 - Clear **licensing** and IP status for re-use (prefer CC-by or open source)
- NIH-endorsed CDEs will be available through the NIH CDE Repository

Proposed structure for geospatial CDEs

Is change in health condition/outcome linked to an environmental factor?

Q: What change in health condition has the patient experienced?

A: List: Hospitalization/worsening of symptoms/new symptoms

Q: Which health condition/outcome was affected?

A: A standardized list of health outcomes (asthma, cardiovascular event, fatigue, etc.)

Q: What was patient's main exposure location during period of interest

A: Location of residential or work or recreation, address or zip code. How do we justify and describe our choice of "main" location?

Q: What environmental factor has the patient been exposed to?

A: A standardized list of environmental factors (direct measurement of allergens, smoke, pesticides, air pollution, weather extremes, or modeling based on proximity to animal feed facilities, agricultural fields). Do we use individual factors or some aggregated indices?

Q: For how long the patient was exposed during period of interest

A: Duration of exposure standard during period of interest (period of interest can be anything from last week to second trimester of pregnancy, duration can be a single number or a range - >1h, 2-5h, 6-12h etc.)

Proposed structure for geospatial CDEs

Q: What was the intensity of the exposure?

A: Numeric value

Units

Value type: Computed or directly measured

If computed

Source of data

Model used (algorithm, meteorological data source if used)

Calculated value is: range, mean, median, max, min etc. during indicated above ***period of interest***

If directly measured

Method,

Instrument

Beyond metadata schema

- Can we propose a core set of environmental geospatial CDEs, perhaps bundled by disease? Reminder: CDEs are **used in common** across multiple: studies, research sites, initiatives, datasets, etc.
- Ongoing effort – conversion PEGS (Personalized Environment and Genes Study) surveys developed in coordination with NHANES studies into CDEs – funded by ODSS and in coordination with NCI team

Existing Efforts on EHS Standards

- **ESIP Geo4Health cluster** – recent session ***Communicating through Bio Geo divide*** during the annual meeting
- **Gravity Project** is part of HL7 initiative that seeks to identify data elements and associated value sets to represent SDOH information documented in electronic health records (EHRs) across clinical activities. They recently started an effort on identifying **Community Level Domains: Food Access, Neighborhood Safety, Access to Green Space** data elements.
- **GA4GH** recently created **Human Exposome Data Standards Group** and are exploring ways to support geolocation and exposure data in **Phenopackets**, will be hosting a Human Exposome session titled Bridging Location and Phenotype
- **American Medical Informatics Association (AMIA) Climate, Health and Informatics WG**
- **Network for Exposomics in the United States (Nexus)** is building a **Geospatial Sciences Hub** to promote Understanding of Place-Based Exposures
- **Gateway Exposome Coordinating Center (GECC)** aims to develop guidance for measuring, harmonizing, and using exposome data