High Performance Computing for Address Level Climate Data Extraction

Presented at the 2022 American Association of Geographers Conference

AAG
AMERICAN ASSOCIATION
of GEOGRAPHERS

Weihe (Wendy) Guan, Jeff Blossom, Devika Kakkar,
Center for Geographic Analysis, Harvard University

Center for Geographic Analysis
Harvard University

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

NSF

Spatiotemporal Thinking, Computing and Applications
STC

NIH

IQSS
The Institute for Quantitative Social Science

Project Introduction

Project Viva[1] - A Boston area based study of a cohort of some 2,000 mothers and children.

Source: https://www.hms.harvard.edu/viva/

Key objective: Examining the effects of climate related environmental exposures (temperature, precipitation, humidity) at cohort address locations over time.

The Environmental influences on Child Health Outcomes (ECHO) Program is a national effort to enhance the health of children and adolescents through research that may help inform healthcare practices, programs, and policies. Project Viva is 1 of over 60 cohorts across the US, shown on the map below, that together form the ECHO Program.

Source: The Viva ECHO fact sheet.

This work is sponsored by Dr. Diane Gold of the Harvard T.H. Chan School of Public Health (HSPH) within the NIH-ECHO program, grant UH3OD023286. Data analysis assistance with the Viva cohort provided by Heike Gibson of HSPH. This work is also partially sponsored by NSF Award #1841403.

Project Overview (continued)

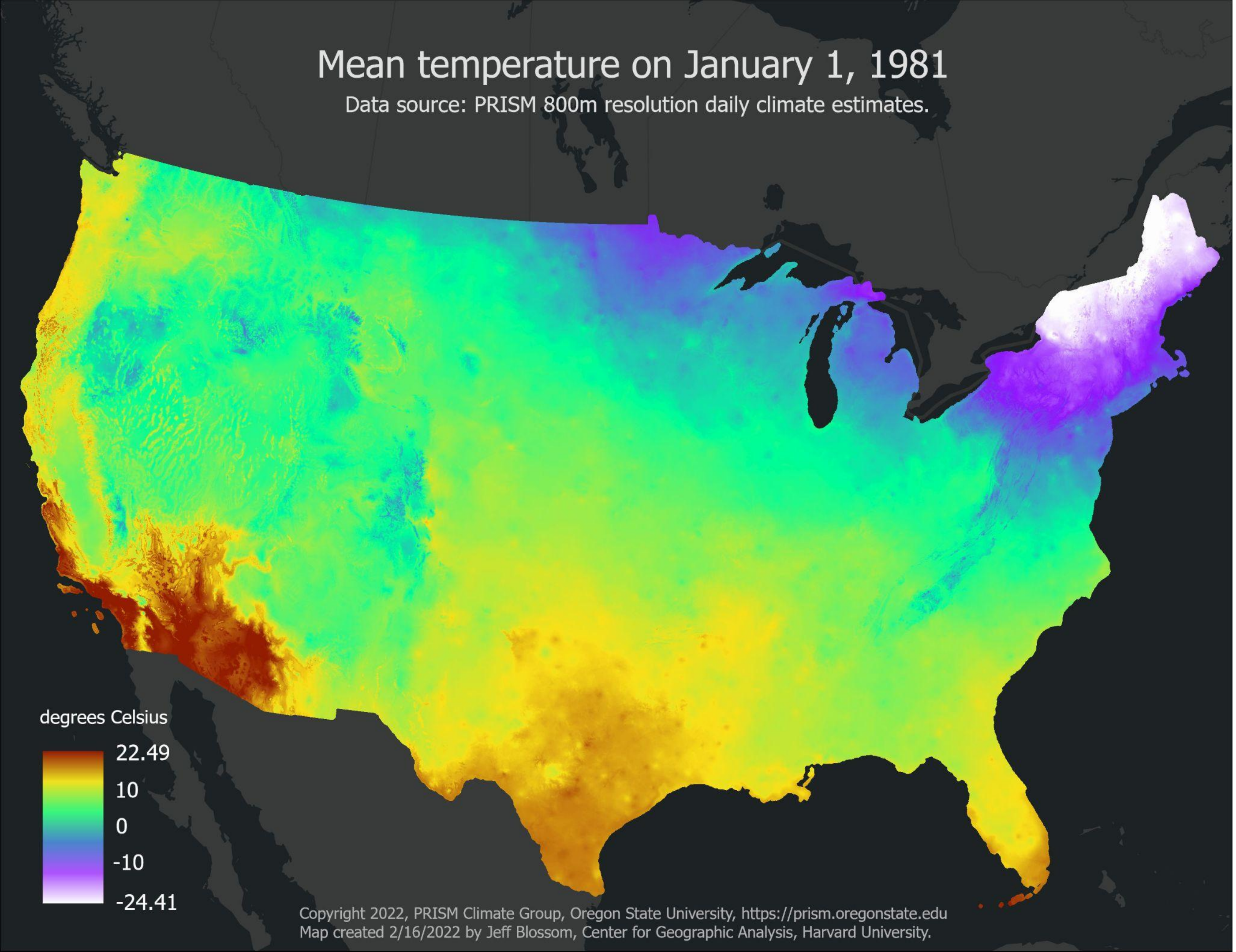
Project Viva[1] - Boston area based study of a cohort of some 2,000 mothers and children

Key objective: Calculating climate related environmental exposures (temperature, precipitation, humidity) at cohort address locations over time.

Input Climate Data: 800-meter resolution PRISM[2] Spatial Climate Dataset in .BIL raster format

- **Spatial Extent**: 48 contiguous United States
- **Number of Variables**: 7 daily climate variables: precipitation, temperature (min, mean, max), vapor pressure deficit (min,max) dew point
- **Temporal Extent**: 40 years from 1981 – 2020, one raster per variable per day
- **Total size**: 8 TB, over 100K rasters, each 85 MB

Project Overview (continued)



Mean temperature on January 1, 1981

Data source: PRISM 800m resolution daily climate estimates.

Copyright 2022, PRISM Climate Group, Oregon State University, https://prism.oregonstate.edu
Map created 2/16/2022 by Jeff Blossom, Center for Geographic Analysis, Harvard University.

Map example using PRISM 800m climate data.

PRISM CLIMATE GROUP

Oregon State University
https://prism.oregonstate.edu/

Project Overview (continued)

Project Viva[1] - Boston area based study of a cohort of some 2,000 mothers and children

Key objective: Calculating climate related environmental exposures (temperature, precipitation, humidity) at cohort address locations over time.

Input Cohort Data: 4,796 cohort address locations over a period of 19 years (1999 - 2017)

- **Input data format*:**

```
address_id,Longitude,Latitude,Start_date,End_date
001_1,-88.8896,30.8862,19991128,20021226
001_2,-89.5246,34.6690,20021227,20110104
002_1,-72.2499,42.4215,19991227,20030221
002_2,-70.7325,-41.9593,20030222,20100103
002_3,-69.6060,46.1955,20100104,20160105
```

- **Total Number of “patient-days” for climate data extraction:** 10.3 million

*Longitude, Latitude locations listed here are randomly determined, they are not actual patient locations.

To be completed by Jeff

Traditional Methods

- **R based Processing of PRISM data**
  - Using *“exactextractr”* and *exact_extract()* command to process PRISM rasters
  - Takes **2-3 weeks** to process **1 climate variable** for **14 years** for **4*4 km** resolution PRISM for **3-digit zcta** level (~1000 shapefiles)

Solution: Available Computing Tools and Resources

PostGIS

Spatial PostgreSQL

+

FAS RC

FAS RESEARCH COMPUTING
HARVARD UNIVERSITY
FACULTY OF ARTS & SCIENCES

- Powerful, open source object-relational database system
- CPU based processing; just a few GPU based functions
- Numerous spatial data processing capabilities (over 500 in total)
- Powerful ***Raster*** data processing capabilities available

- Compute: 100,000 compute nodes, 8-64 cores/node, 12Gb to 512Gb memory/node
- Software: CentOS 7, Slurm job manager, Singularity, 1000+ scientific tools and programs
- Storage: 100 GB (Home dir), 4TB+ (Lab storage), 2.4PB (Global scratch),
- **#144** in TOP500 Supercomputers in the world

Challenges: GIS Big Data Processing (Raster Data)

Major processing steps

- Creation of database and loading of Climate Rasters and patient addresses
- Extraction of 7 climate variables for all persons/days;calculation of additional climate variables
- Automation of processes and scaling of solution on Cluster Computing environment

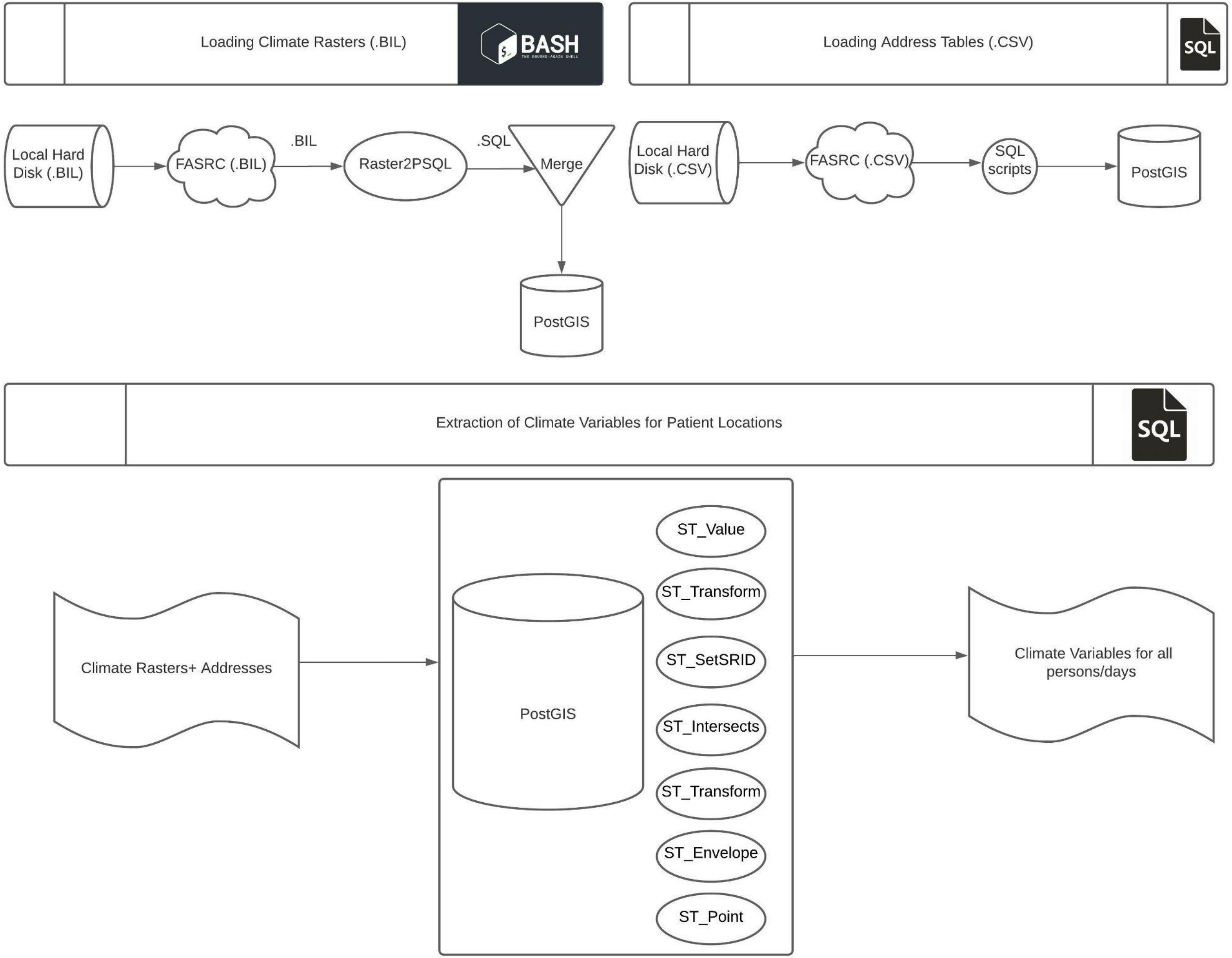
Obstacles

- For this large scale of raster data:
  - Traditional methods are slow, costly and inefficient
  - Local resources are insufficient; Cloud/ Cluster resources are needed for scaling
  - Single operator is insufficient; combination of powerful spatial operators needed
  - Manual operations are not possible; automation is needed

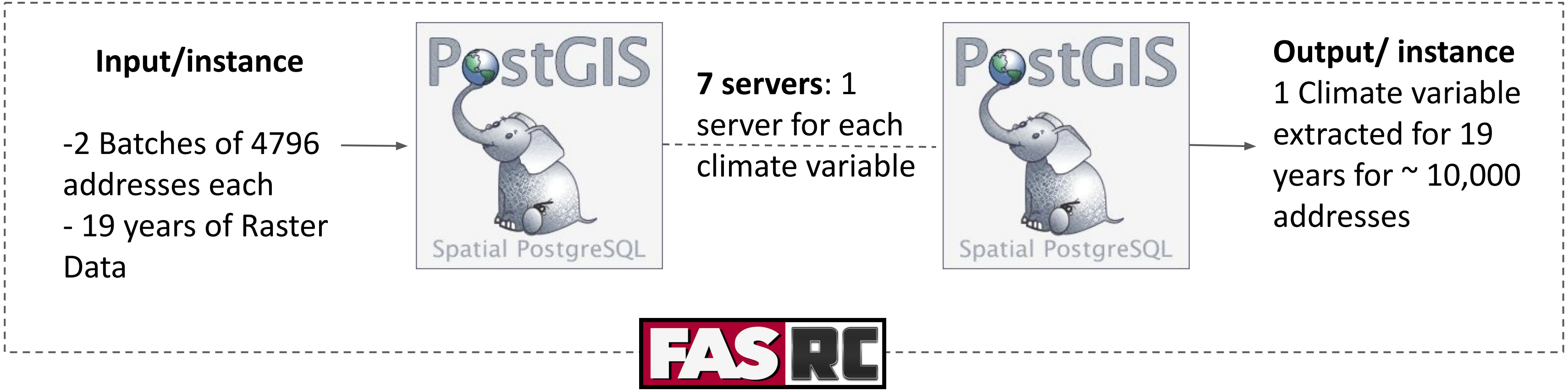
Solution

- Development of proof-of-concept using combination of powerful spatial operators
- Automation of this solution using combination of bash and SQL scripting
- Optimal scaling of this solution on Harvard’s High-Performance Computing Cluster

Solution: Architecture Diagram



Solution: Scaling on Harvard’s Computing Cluster



Input/instance

-2 Batches of 4796 addresses each
- 19 years of Raster Data

7 servers: 1 server for each climate variable

Output/ instance
1 Climate variable extracted for 19 years for ~ 10,000 addresses

- **Scaling** can be applied in two ways due to independence of attributes:
  - **Across years (# years = 20):** More resources, slower processing
  - **Across climate variables (# variables = 7):** Less resources, faster processing
- Optimal Performance was obtained by scaling across the variables:
  - **4 days/variable** for processing 20 years of data for 5000 locations
  - **24 hours** for loading 20 years of raster data

Results

- 7 Climate Variable for all person/days, Total Number of “patient-days” for climate data extraction: **10.3 Million**
- Absolute and relative humidity were calculated using the existing mean temperature and dewpoint variables for all person/days
- Unified solution for 9 climate variables for all person/days were extracted as shown below

```
address_id,day,ppt,tmean,tmin,tmax,tdmean,vpdmin,vpdmax,rh,ah
001_1,19991128,3.125,12.500,11.0,15.5,7.810,0.126,9.864,73.095,8.033
001_1,19991129,4.646,6.300,4.43,10.54,0.710,0.245,6.525,67.436,4.992
001_1,19991130,0.000,9.070,7.2,14.56,-4.740,3.493,12.423,37.357,3.307
001_1,19991201,0.000,12.760,5.34,17.45,-4.090,5.817,15.749,30.701,3.429
001_1,19991201,0.647,13.420,8.65,19.34,2.250,1.930,17.131,46.738,5.438
```

Solution: Novelty of our approach

Scalable
Can be easily scaled to bigger datasets

Leverages Existing Resources
Implemented on Harvard's Computing Cluster

Replicable
Can be replicated on other raster datasets and clusters

High-Performance
Computing of Raster Big data

Time-efficient
- **4 days/variable** for 19 years of data for 4,796 locations
- **24 hours** for loading 19 years of raster data

Open Source
PostGIS based processing

Challenges Solved

- Automated solution using bash and SQL scripting
- Optimal scaling of the solution using Harvard’s High-Performance Computing Cluster.

Ongoing Work

- Two Harvard medical researchers are both currently working with the Viva climate data extracts in health outcome related studies.
- Preparing to use our solution for additional cohorts.

Future Applications

- Distribute the solution to work on non-Harvard computing environments. The Viva cohort is a Harvard study, allowing for processing data in a secure Harvard controlled environment. Other cohorts are spread out among many Universities, with most Institutional Review Board restricting cohort data to residing on local environments, handled by IRB approved personnel.
- Our approach can be applied to the free PRISM 4km resolution climate data, or any geospatial study involving extracting values from temporal raster data such as NDVI, night lights, etc.

Thank you!

Center for
Geographic Analysis
Harvard University

Wendy Guan - wguan@cga.harvard.edu

Jeff Blossom - jblossom@cga.harvard.edu

http://gis.harvard.edu

Devika Kakkar - kakkar@fas.harvard.edu

References

[1] Project Viva: https://www.hms.harvard.edu/viva/

[2] PRISM Climate Data: https://prism.oregonstate.edu/

[3] Introduction to Cluster Computing on FASRC:
https://www.rc.fas.harvard.edu/wp-content/uploads/2019/12/Intro-to-Cannon.pdf

[4] About Postgis: https://postgis.net/

[5] OmniSci: https://www.omnisci.com/

[6] Children’s Respiratory and Environment Workgroup (CREW):
https://www.rhoworld.com/federal-project-pages/childrens-respiratory-and-environment-workgroup-crew/

[7] National Health Institute ECHO project:
https://www.nih.gov/research-training/environmental-influences-child-health-outcomes-echo-program