RINX 2.0: A Containerized Climate Raster Information Extraction System on OpenShift Cloud Environment

Devika Jain[1], Jeff Blossom[1], Jack Hayes[1], Heike Gibson[2], Sheryl Rifas-Shimann[3], Diane R. Gold[2]

[1] Center for Geographic Analysis, Harvard University, Cambridge MA
[2] Harvard T.H. Chan School of Public Health, Harvard University, Boston MA
[3] Department of Population Medicine, Harvard Medical School, Boston, MA

Keywords: Climate data, Big data, Cloud Computing, Raster Data, Geospatial, OpenShift, Containers

Abstract

RINX (Raster INformation eXtraction) 2.0 is an advanced solution for efficiently extracting climate data from large raster datasets in a cloud computing environment. Building upon the original RINX 1.0, which utilized high-performance computing clusters, RINX 2.0 leverages cloud technologies such as OpenShift and PostGIS to handle massive datasets and automate the extraction process. The system supports large-scale spatiotemporal raster extractions, processing over 158 million data points from the 15TB PRISM climate dataset. Here, we describe the architecture, methods, and tools used in RINX 2.0, including containerized environments, automated data pipelines, and integration with the New England Research Cloud. The system was deployed for the Environmental influences on Child Health Outcomes (ECHO) project, providing valuable insights into environmental health research. We present performance statistics, data management strategies, and the development of a user interface for real-time querying and visualization of results.

1. Introduction

RINX (Raster INformation eXtraction) is an end-to-end solution developed by the authors for automatic extraction of information from large raster datasets. RINX 1.0 utilized PostGIS in a high-performance compute cluster environment, (Kakkar et al., 2022) and now we present RINX 2.0 which utilizes open-source technologies in a cloud computing environment. The input for RINX is a set of geo-referenced raster datasets and a set of point locations from which the information is to be extracted. The output for RINX is a structured representation of extracted information from the raster datasets for each data point in CSV text format. The loading and processing of the input datasets to RINX 2.0 is accomplished using a combination of Bash and SQL scripts deployed in a containerized environment optimized for efficient automation. This environment uses the open technologies OpenShift and Crunchy Data to feed the input raster and point location data into the spatial database PostGIS for extraction. RINX 2.0 was created to aid the study of environmental conditions and how they affect the health of people over their lifespans for the Environmental influences on Child Health Outcomes (ECHO) (National Institute of Health. n.d.) project.

2. Use-Case

2.1 Overview of Existing Methods

Given the sheer size of the available swaths of remotely sensed data, raster value extraction using a vector point location input is becoming increasingly common among Geographic Information System software applications for performing analysis on areal subsets to improve computational efficiency (Spangler et al., 2019, Jung, 2013, Lee et al., 2021, Wang et al., 1954, Reddy, 2018, Goodchild et al., 1997, Laney, 2001). Even with the recent surge of niche spatio-temporal data analytics systems, there are limited case studies of architectures specific to raster value extraction using vector point location input despite the plethora of relevant software packages (Alam et al., 2022). The RINX 1.0 solution for extraction of spatio-temporal big raster data is the first large-scale, cluster-based solution for locally stored raster information extraction (Kakkar et al., 2022).

Even with the recent surge of niche spatio-temporal data analytics systems, there are limited case studies of architectures specific to raster value extraction using vector point location input despite the plethora of relevant software packages (Alam et al., 2022). The RINX 1.0 solution for extraction of spatio-temporal big raster data is the first large-scale, cluster-based solution for locally stored raster information extraction (Kakkar et al., 2022). Its architecture utilizes BASH scripting, SQL scripting, and PostGIS functionality to extract climate raster information scaled on High Performance Compute Clusters and proved to perform 70,160,615 extractions on 4.5TB of disk-stored raster data in five days. RINX 2.0 was developed to improve the efficiency of RINX 1.0 by moving it to a cloud-based environment.

Cloud environments prove to be incredibly beneficial when it comes to increased scalability, flexibility, collaboration, and data processing (Kumar et al., 2015). Various case studies have shown the increased efficiency of raster handling and analysis through cloud-based solutions as well (Li et al., 2020, Xiao et al., 2023, Enescu et al., 2021, Xie, 2016). Examples of using a cloud-based architecture to improve upon raster value extraction using vector point location input specifically has shown significant advancements in processing speed in this realm as well (Crego et al., 2021). Despite these advancements in cloud architecture development, there is limited literature on cloud solutions for raster value extraction using point location input that are compatible with disk-based data rather than server-based, open-source data catalogs.

2.2 Data

The Environmental influences on Child Health Outcomes (ECHO) (National Institute of Health. n.d.) program is a nation-wide project in the United States funded by the National

Institutes of Health. ECHO includes over 60 cohorts of children and their mothers, and is aimed to help better understand effects of environmental exposures on child health and development. Daily meteorological and long term climate conditions have been shown to have an adverse effect on health (Bell et al., 2018, Greenough et al., 2001, Rice et al., 2019, Sprangler et al., 2019, Zscheischler et al., 2014) and are thus one of the environmental exposures of interest to the investigators in the ECHO program. One of the ECHO cohorts is Project Viva, (Oken, et. al., 2014; Harvard Medical School. n.d.) a Boston, MA based longitudinal study including a cohort of some 2,000 mothers and children. For this project, the 800-meter resolution daily PRISM (PRISM n.d.) dataset was used to assign climate exposures to member address locations for the Project Viva cohort. The PRISM dataset spans the contiguous 48 United States, and contains daily observations from 1981 to present for seven climate variables: minimum, maximum, mean, and dewpoint temperature; precipitation; and minimum and maximum vapor pressure deficit. Each daily dataset is geo-referenced and published in raster “.bil” format. The full dataset (1981-2023) contains 15,695 rasters, each at a size of 85MB, presenting a total dataset size of around 15 Terabytes (TB).

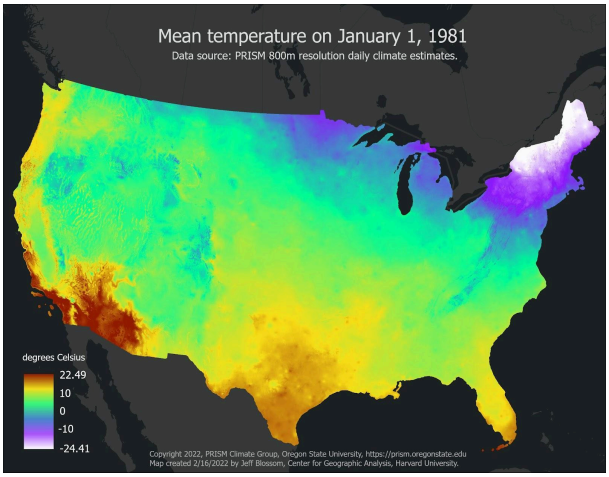

Figure 1. Mean temperature map for January 1, 1981, using 800m PRISM climate data

2.3 Methodology

For this particular use case we had 5,219 address locations spanning variable time periods from 1999 - 2023 producing a total of 18,131,095 extractions needed from the 15TB PRISM database. This magnitude of data necessitated substantial disk space and processing power, surpassing the capabilities of standard workstations or servers. To fulfill these high-performance computing requirements, we opted for an advanced cloud computing solution utilizing the technologies OpenShift, Crunchy Data, and PostGIS on the New England Research Cloud (NERC). NERC is a professionally operated regional resource that provides on-premises cloud service, including self-service Software-as-a-Service, (SaaS) automated Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS) which catalyzes hardware acceleration. NERC is commonly used by researchers in the Boston area to facilitate quality research using more-efficient computational environments (NERC. n.d., MOC. n.d.). NERC offers OpenShift technology - a comprehensive container platform that solves many challenges associated with deploying and managing containerized applications, offering a feature-rich and flexible solution for modern IT infrastructures. For RINX 2.0, each OpenShift pod was equipped with 300 GB of RAM and 8 virtual CPUs. For spatial data extractions, we utilized PostGIS software in a cloud environment with Crunchy Data support, which provides services such as automated failover, recovery, and alerting which are crucial for cloud environment maintenance (Crunchy Data. n.d., PostGIS. n.d.).

The default data security level on NERC is Level 3, which was determined sufficient to allow for storing the PRISM dataset. The other dataset needed for the extraction was Project Viva cohort address locations in latitude, longitude format. This data was determined to be security Level 3 (IRB protocol #951581, Data Use Agreement DAT22-0250 – ECHO, expires 8/20/2024). This determination required the data to be de-identified based on the Health Insurance Portability and Accountability Act (HIPPA) guidelines to allow for processing on the NERC. This de-identification for each cohort was performed on local computing systems, complying to the HIPPA “Safe Harbor” (HIPPA, n.d.) guidelines. The de-identified data containing latitude, longitude locations in CSV format was loaded onto the NERC for processing with the PRISM rasters, which were also loaded to NERC from a local workstation.

RINX 2.0 calculated climate exposure data for 5,000 patient addresses spanning different durations within a 26-year period, 1998 - 2023. The total patient days represented are equal to 17.5M, and with 7 observations per day this totals 123M total extractions. Additionally, relative humidity (rh) and absolute humidity (ah) were calculated, bringing the total observations to 158 million total variables calculated. The full dataset (1981-2023) contains 15,695 rasters for each variable, totaling 109,935 rasters, each at a size of 85MB, presenting a total dataset size of around 15 TB of data. This magnitude of data necessitated substantial disk space and processing power, surpassing the capabilities of standard workstations or servers. To fulfill these high-performance computing requirements, we opted for an advanced cloud computing solution utilizing OpenShift pods on NERC (NERC. n.d., Kubernetes. n.d., Red Hat. n.d.). Each pod was equipped with significant resources, consisting of 300 GB of RAM and 8 virtual CPUs (vCPUs). The cost of this processing system is estimated to be about $700/month on New England Research Cloud. This setup not only provided the requisite computational power but also offered the flexibility needed to efficiently process the large-scale data. In our approach, we seamlessly integrated PostGIS within a CrunchyData database environment to leverage its robust object-relational database system capabilities as shown in Figure 2 below. PostGIS, renowned for its extensive suite of over 500 spatial data processing tools and advanced raster processing features, played a pivotal role in managing and analyzing the vast PRISM dataset.

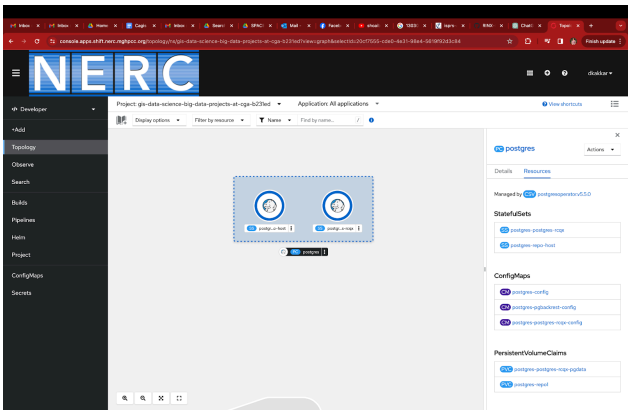

Figure 2. RINX2.0 system installed on NERC’s OpenShift Platform based on Crunchy PostGIS

The major processing steps of RINX 2.0 are the same as RINX V1.0: Database creation, data loading and data extraction (Kakkar, D. et al., 2022). The main difference between RINX 1.0 and 2.0 is that 2.0 is utilizing the cloud environment. This workflow for RINX 2.0 is presented in Figure 3 below. This method involved utilizing BASH scripts that 1) Loaded the PRISM rasters to the cloud and 2) Created the database on PostGIS in the cloud. Then, SQL scripts running on PostGIS in the cloud perform the climate data extractions, looping through all days in each date range specified for each address location for all climate variables. Scripts and specific commands we developed for this solution can be found on the Harvard CGA GitHub repository (Harvard CGA GitHub. n.d.).

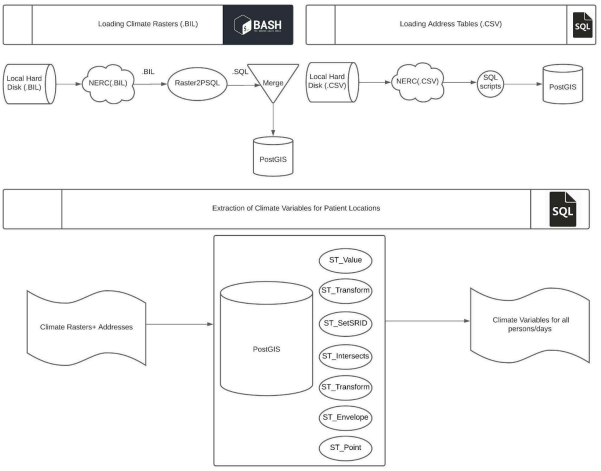

Figure 3. Workflow Diagram for RINX 2.0 on NERC OpenShift

Upon completion of the calculations, it became important to equip researchers with a suitable tool for analyzing and visualizing the extensive results of 163 million observations. Therefore, we developed a user interface that enables near real-time querying of results based on space, time, and climate variables using Heavy.AI (HeavyAI. n.d.). Heavy.AI is a state-of-the-art analytics platform specifically designed for handling large-scale datasets. It employs a combination of GPU and CPU processing to deliver exceptional performance. Key features of Heavy.AI include its open-source SQL engine, HeavyDB, which facilitates efficient data management. Additionally, it provides HeavyRender for server-side rendering and Heavy Immerse for web-based data visualization, making it a comprehensive tool for in-depth data analysis. Figure 4 below displays a snapshot of our RINX UI, offering researchers a platform to analyze and visualize processed climate-related data results in real-time. This real-time data analysis enables researchers to make informed decisions based on the latest information.

Figure 4. Heavy.ai based UI for real-time analysis and visualization of RINX results

2.4 Results

For the Project Viva cohort we used RINX 2.0 to extract 7 daily climate variables from the PRISM data for 5,219 address locations spanning variable time periods between from 1999 - 2023, for a total of 18,131,095 patient days. This produced a total of 126,917,665 climate observations, output into .csv format. This data is in the process of being combined with health outcome data, to analyze the effect climate may have on lung function and other systems. Further, the mean and dew point temperatures were used to calculate relative humidity and absolute humidity for each day, producing an additional 36,262,190 observations for a total of 163,179,855 observations. The entire process took 2 hours to load the rasters, and 1.5 days to calculate the 163M observations. This is a significant improvement in performance over RINX 1.0 which took 24 hours to load the data and 4 days to process the observations for 70M observations. It is estimated that traditional methods such as ArcGIS, QGIS, and R would have taken 2 months or more to extract the same amount of observations, thus demonstrating that RINX 2.0 saves considerable time and cost. Further, Heavy.ai based UI offers researchers a platform to analyze and visualize the processed climate-related data results in real-time. This real-time data analysis enables researchers to make informed decisions based on the latest information.

3. Conclusion

Utilizing RINX 2.0 in the cloud using the OpenShift environment allowed for 163M climate variable extractions at 5,219 individual locations, all calculated in less than two days. This architecture simplified the management and scaling of the RINX application and streamlined the development process by supporting continuous integration and continuous deployment. The platform easily supported scaling of our application, allowing for optimal resource allocation utilization. Additionally, this cloud-based system supported end-to-end application lifecycle management, from development and

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-391-2025 | © Author(s) 2025. CC BY 4.0 License.

testing to deployment and monitoring, ensuring a consistent and streamlined process. With RINX 2.0 it will now be possible to rapidly and efficiently enrich additional ECHO cohort address locations with climate exposure data. Additionally, by providing our open-source code and docker files for RINX v2.0 on GitHub, our solution can be deployed by others on a variety of computing environments to be used to extract point level data from any spatio-temporal raster datasets.

Acknowledgements

This work is sponsored by NSF grant #1841403, and by Dr. Diane R. Gold of the Harvard T.H. Chan School of Public Health (HSPH) within the NIH-ECHO program, grants UH3OD023286, R01HD034568, R24ES030894, and P30-ES000002. We would also like to acknowledge Dr. Chris Daly and Dr. Dylan Keon of the PRISM Climate Group at Oregon State University who provided helpful guidance on using the PRISM data. We would like to thank the New England Research Cloud (NERC) for providing the technical support and computing resources needed for the project.

References

Bell, Jesse E., Claudia L. Brown, Kathryn Conlon, Stephanie Herring, Kenneth E. Kunkel, Jay Lawrimore, George Luber, Carl
Schreck, Adam Smith, and Christopher Uejio. 2018. Changes in extreme events and the potential impacts on human health.
Journal of the Air & Waste Management Association 68 (4): 265-287. doi: 10.1080/10962247.2017.1401017.

Bloschl, G., and M. Sivapalan. 1995. Scale issues in hydrological modeling: A review. Hydrol. Processes, no. 9, 251–290. doi.org/10.1002/hyp.3360090305.

Brunsell, N. A., and R. R. Gillies. 2003. Scale issues in land atmosphere interactions: implications for remote sensing of the surface energy balance. Agric. For. Meteorol. 117:203–221.

Crunchy Data. n.d. PostGIS Solutions. Accessed January 11, 2024. https://www.crunchydata.com/solutions/postgis

ESRI. n.d. Esri: GIS Mapping Software, Location Intelligence & Spatial Analytics. Accessed June 14, 2022. https://www.esri.com.

Goodchild, M. F., and D. A. Quattrochi. 1997. Introduction: Scale, Multiscaling, Remote Sensing, and GIS. In Scale in Remote Sensing and GIS, edited by Dale A. Quattrochi, Michael F. Goodchild, and A. W. Goode, 1–12. N.p.: CRC-Press.

Greenough, G., M. McGeehin, S. M. Bernard, J. Trtanj, J. Riad, and D. Engelberg. 2001. The potential impacts of climate variability and change on health impacts of extreme weather events in the United States. Environ Health Perspectives 109, no. 2 (May): 191-198. doi.org/10.1289/ehp.109-1240666.

Harvard CGA Github. n.d. https://github.com/cga-harvard/RINX-Raster_INformation_eXtraction_System/tree/main/RINX2023

Harvard Medical School. n.d. Project Viva | A Study of Health for The Next Generation. Accessed June 14, 2022. https://www.hms.harvard.edu/viva/.

HeavyAI. n.d. HEAVY.AI PRODUCT OVERVIEW. Accessed January 22, 2024. https://www.heavy.ai/product/overview.

Henry Ford Health. n.d. The Childhood Allergy and the NeOnatal Environment (CANOE) study. Accessed January 25, 2024.
https://www.henryford.com/hcp/research/mom-and-baby-research-studies/our-studies/canoe

HIPPA. n.d. HIPPA“Safe Harbor” guidelines, accessed December, 2023. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#protected

Kakkar, D. et al., (2022) RINX: A SOLUTION FOR INFORMATION EXTRACTION FROM BIG RASTER DATASETS. *International archives of the photogrammetry, remote sensing and spatial information sciences.* [Online] XLVIII-4/W1-2022245–250.

Kubernetes. n.d. Accessed January 11, 2024. https://kubernetes.io/

Kumar, R. and Charu, S., 2015. Comparison between cloud computing, grid computing, cluster computing and virtualization. *International Journal of Modern Computer Science and Applications*, *3*(1), pp.42-47.

Laney, D. 2001. 3d Data Management: Controlling Data Volume, Velocity and Variety. Stamford, CT: META Group Inc.

Lee, Hwa-Jin, Oh-Sun Lee, Dong-Gul Woo, Han-Na Kim, Mark C. Wallace, and Yeong-Seok Jo. 2021. Current distribution and habitat models of the yellow-throated marten, Martes flavigula, in South Korea. Mammal Research 66:429-441. Liu, Peng. 2015. A survey of remote-sensing big data. Frontiers in Environmental Science. doi.org/10.3389/fenvs.2015.00045.

Liu, Peng. 2015. A survey of remote-sensing big data. Frontiers in Environmental Science. doi.org/10.3389/fenvs.2015.00045.

Md Mahbub Alam, Luis Torgo, and Albert Bifet. 2022. A Survey on Spatio-temporal Data Analytics Systems. ACM Comput. Surv. 54, 10s, Article 219 (January 2022), 38 pages. https://doi.org/10.1145/3507904

MOC. n.d. Massachusetts Open Cloud Alliance Accessed January 11, 2024. https://massopen.cloud/

National Institute of Health. n.d. Environmental influences on Child Health Outcomes (ECHO) Program. National Institutes of Health (NIH). Accessed June 15, 2022. https://www.nih.gov/research-training/environmental-influences-child-health-outcomes-echo-program

NERC, Massachusetts Green High-Performance Computing Center. n.d. Accessed January 11, 2024. https://nerc.mghpcc.org/

Oken, Emily, Baccarelli, Andrea A, Gold, Diane R, Kleinman, Ken P, Litonjua, Augusto A, De Meo, Dawn, Rich-Edwards, Janet, Rifas-Shiman, Sheryl L, Sagiv, Sharon, Taveras, Elsie M, Weiss, Scott T, Belfort, Mandy B, Burris, Heather, Camargo, Carlos A Jr, Huh, Susanna, Mantzoros, Christos, Parker, MargaretG, Gillman, Matthew W. 2014. Cohort profile: Project

Viva. International Journal of Epidemiology. Volume 44, Issue 1, February 2015. DOI: 10.1093/ije/dyu008

PostGIS. n.d. Accessed January 11, 2024. https://postgis.net/

PRISM. n.d. Accessed November, 2018. https://prism.oregonstate.edu/

QGIS. n.d. Welcome to the QGIS project! Accessed June 14, 2022. https://qgis.org/.

R. n.d. The R Project for Statistical Computing: R. Accessed June 14, 2022. https://www.r-project.org/.

Rice, Mary B., Wenyuan Li, Elissa H. Wilker, Diane R. Gold, Joel Schwartz, Antonella Zanobetti, Petros Koutrakis, et al., 2019. Association of outdoor temperature with lung function in a temperate climate. Mittleman European Respiratory Journal 53 (1800612). doi: 10.1183/13993003.00612-2018.

Red Hat. n.d. Accessed January 11, 2024. https://www.redhat.com/en

Reddy, G. P. O. 2018. Spatial Data Management, Analysis, and Modelling in GIS: Principles and Applications. In Geospatial Technologies in Land Resources Mapping, Monitoring and Management, edited by G. P. O. Reddy and S. K. Singh. Vol. 21. N.P.: Springer International Publishing. doi.org/10.1007/978-3-319-78711-4_7.

Sprangler, Keith R., Kate R. Weinberger, and Gregory A. Wellenius. 2019. Suitability of gridded climate datasets for use in environmental epidemiology. Journal of Exposure Science & Environmental Epidemiology volume 29:777–789. doi.org/10.1038/s41370-018-0105-2.

Talwalkar, A., S. Kumar, M. Mohri, and H. A. Rowley. 2013. Large-scale SVD and manifold learning. J. Mach. Learn. Res. 14 (3129–3152)

Tobias, Michele. 2014. Using R for Climate Raster Data Extraction. Monterey, CA: Presented at CaIGIS.

Wang, Zitao, Qimeng Liu, and Yu Liu. 1954. Mapping Landslide Susceptibility Using Machine Learning Algorithms and GIS: A Case Study in Shexian County, Anhui Province, China. Symmetry 12 (12). doi.org/10.3390/sym12121954.

Wu, Hua, and Zhao L. Li. 2009. Scale Issues in Remote Sensing: A Review on Analysis, Processing and Modelling. Sensors 9:1768–1793. doi.org/10.3390/s90301768.

Wylie, Bruce K., Neal J. Pastick, and Joshua J. Picotte. 2018. Geospatial data mining for digital raster mapping. GIScience & Remote Sensing 56 (3): 406-429. doi.org/10.1080/15481603.2018.1517445.

Zscheischler, J., A. M. Michalak, C. Schwalm, M. D. Mahecha, D. N. Huntzinger, and M. Reichstein. 2014. Impact of large-scale climate extremes on biospheric carbon fluxes: an intercomparison based on MsTMIP data. Glob. Biogeochem Cycles 28:585–600. doi.org/10.1002/2014GB004826.