RINX: A Solution for Information Extraction from Big Raster Datasets

Devika Kakkar[1*], Jeff Blossom[1], Wendy Guan[1]

Centre for Geographic Analysis, Harvard University, Cambridge MA

Commission IV, WG IV/4

KEY WORDS: Raster, Big Data, Climate, Open Source, PostGIS, PRISM

ABSTRACT:

Processing Earth observation data modelled in a time-series of raster format is critical to solving some of the most complex problems in geospatial science ranging from climate change to public health. Researchers are increasingly working with these large raster datasets that are often terabytes in size. At this scale, traditional GIS methods may fail to handle the processing, and new approaches are needed to analyse these datasets. The objective of this work is to develop methods to interactively analyse big raster datasets with the goal of most efficiently extracting vector data over specific time periods from any set of raster data. In this paper, we describe RINX (**R**aster **IN**formation e**X**traction) which is an end-to-end solution for automatic extraction of information from large raster datasets. RINX heavily utilises open source geospatial techniques for information extraction. It also complements traditional approaches with state-of-the-art high-performance computing techniques. This paper discusses details of achieving big temporal data extraction with RINX, implemented on the use case of air quality and climate data extraction for long term health studies, which includes methods used, code developed, processing time statistics, project conclusions, and next steps.

1. INTRODUCTION

RINX (**R**aster **IN**formation e**X**traction) is an end-to-end solution developed by the authors for automatic extraction of information from large rasters datasets. RINX heavily utilises open source geospatial techniques for information extraction. It also complements the traditional approaches with state-of-the-art high-performance computing techniques. The input for RINX is a set of rasters from which the information has to be extracted and a set of data point locations for which the information needs 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 pre-processing of the input datasets to RINX is accomplished using a combination of Bash and SQL scripting techniques for automation. This pre-processed input is then fed into the open source spatial database PostGIS to extract the required information by using multiple spatial techniques. Finally, the extracted output is post-processed for deduplication and standardisation of extracted information for research use. RINX is designed in a way that makes it easy to deploy and scale on any local, cloud, or cluster computing platform. RINX was created to aid the study of environmental conditions and how they affect the health of people over their lifespans for Project Viva which is described in detail in the following sections.

2. THE USE CASE OF PROJECT VIVA

The Environmental influences on Child Health Outcomes (ECHO) (National Institute of Health. n.d.) is a nation-wide program 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. Project Viva is 1 of 71 cohorts across the US, as shown in the Figure 1 below, that together form the ECHO Program.

Figure 1. Map showing 71 cohorts across US that together form the ECHO Program (Harvard Medical School. n.d.)

One of the ECHO cohorts in the Boston area is Project Viva (Harvard Medical School. n.d.), a Boston-based longitudinal study including a cohort of some 2,000 mothers and children. The goal of Project Viva is to find ways to improve the health of mothers and their children by looking at the effects of mother's diet as well as other factors during pregnancy and after birth. A key part of the analysis is calculating various social and environmental exposures at the Viva cohort member address locations over their life spans. 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

*Corresponding author
Email address: kakkar@fas.harvard.edu (Devika Kakkar)

are thus one of the environmental exposures of interest to the investigators of the Viva cohort and the ECHO program overall.

The Parameter-elevation Regressions on Independent Slopes Model (PRISM) climate group gathers a vast amount of daily weather and climate observations, and produces various models of short-term and long-term climate patterns across the contiguous 48 United States (Spangler et al., 2019). Figure 2 below shows a mean temperature map for January 1, 1981 using PRISM 800m climate data.

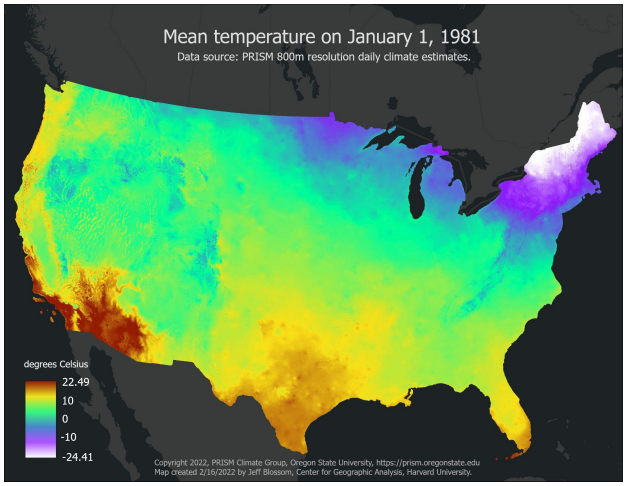

Figure 2. Mean temperature map for January 1, 1981 using 800m PRISM climate data

For this project, the 800-metre resolution daily PRISM dataset was chosen to assign climate exposures. This dataset contains daily observations from 1981 - 2021 for seven climate variables: minimum, maximum, mean, and dewpoint temperature; precipitation; and minimum and maximum vapour pressure deficit. By using the mean and dew point temperatures, relative and absolute humidity were also calculated. Birth years range from 1999 - 2001 for the Viva cohort, and daily climate exposure information was desired for all of childhood, up to age 18. This resulted in a total time period of 1999 - 2019 for the climate exposure calculations. Address histories were compiled for the cohort, producing a total of 4,796 unique cohort address locations during this time period. These addresses were geocoded, producing spatially resolved address histories for all cohort members.

Daily meteorological conditions for each cohort member allow for the finest possible analysis of climate effects on health. To describe this level of spatio-temporal data granularity, we use the term “patient/days”. This describes the total days of observations required to calculate climate exposures for all of the 4,796 patient locations during every day of childhood. For this project, the total number of patients/days ended up being 10,022,945. The PRISM dataset is published in BIL raster format, with one raster representing one climate variable per day. As there are 7 climate variables and 10,022,945 patient days, this produced a total number of 70,160,615 singular extractions needed from the PRISM rasters.

Calculating this amount of observations in a timely fashion required computing resources beyond the standard desktop computer. The problem calls for a system such as the RINX system to extract the data. For this use case, RINX was implemented on a high-performance computing cluster running a PostgreSQL database with the PostGIS extension. It accomplished the data extraction objective, saving 2-3 weeks of processing time as compared with existing traditional methods.

3. OVERVIEW OF EXISTING METHODS

Remote sensing big data is growing fast and presents with unique handling and processing challenges that are difficult to address using traditional techniques (Liu et al., 2015, Wu et al., 2009, Talwalkar et al., 2013, Bloschl et al., 1995, Brunsell et al., 2003). Extracting values from multiple rasters using vector point location input is a common operation that can be performed in a multitude of Geographic Information System software applications (Spangler et al., 2019, Jung, 2013, Lee et al. 2021, Wang et al., 1954, Reddy, 2018, Goodchild et al., 1997, Laney, 2001). These include ArcGIS Pro (ESRI. n.d.), QGIS (QGIS. n.d.), the R Project for Statistical Computing (R. n.d.), and the PostGIS spatial database extender for PostgreSQL (PostGIS. n.d.). The ability to perform extractions from thousands (or more) rasters is dependent on the amount of computing resources to load and process the rasters. Techniques described in the existing literature describe extracting information from all rasters at all point locations (Spangler et al., 2019, Jung, 2013, Tobias et al., 2014, Wylie et al., 2018, Wang et al., 2014). Applying this all to all extraction of our 4,796 locations across the entire 21-year time period (7,670 days) for all 7 climate variables would have resulted in 257,497,240 individual extractions (observations). As several cohort members only lived at certain addresses for a few years of the 21-year period, extractions were not needed at all locations for every day. As described above based on the varying time periods spent at each location, we only needed 70,160,615 individual observations. Performing only 70M instead of 257M extractions saves computing time and resources, but requires scripting to only feed in locations for extraction at the pertinent dates. For this specific use case we did not find any similar published research. To help determine the best solution for this big raster processing challenge, we gathered anecdotal descriptions of similar data extractions from colleagues at the T.H. Chan Harvard School of Public Health (HSPH) and Massachusetts Institute for Technology (MIT). One such case reported using the “extractextractr” and “exact_extract()” commands within R to process one climate variable for 14 years. The raster data extracted from was the PRISM 4km x 4km resolution rasters, with roughly 1,000 input vector polygons. This process was reported to take 2 - 3 weeks. Another project used the ArcGIS Pro “Extract multi-values to points" tool with Python scripting to process the PRISM monthly climate rasters. This project was estimated to take roughly one week to process 34 years of monthly data for the month of July for 2.4 million input address locations.

Our own tests of the QGIS (version 3.10) “SAGA Add Raster Values to Points” and PostGIS (version 3) ST_Value and ST_Point tools were conducted and found to produce correct results for extracting values from the PRISM rasters to our set of 4,796 input point locations. With this review of existing methods and our initial testing, we estimate our extraction would take 3 – 4 weeks using traditional methods.

4. SYSTEM DESIGN

With proof of concept that all 4 software packages described above were capable of performing the extraction, we thenevaluated the computing system needs for our project to determine the best methods to use. Our unique big data challenge involving extraction of 70 million observations over a 21-year period required processing all or part of 53,690 PRISM rasters that are of 800m x 800m resolution. Each raster is roughly 85MB in size, requiring a total of 4.5TB of available disk space. For optimal processing we realised that having computing processing and RAM resources available beyond what a typical workstation or server offers would be necessary. This led us to investigate using the Harvard Faculty of Arts and Sciences research computing cluster FASRC (FASRC. n.d.).

The FASRC computing environment contains 100,000 compute nodes running at 8 to 64 cores per node. The core software is CentOS 7, running the slurm job manager and Singularity. The system runs over 1,000 scientific tools and programs, including the PostgreSQL database with the PostGIS extension. As this was an HSPH sponsored project we had access to 8 TB of dedicated lab storage, with 2.4PB available as a global scratch space. With the available computing power, storage space, and the ability to use an open-source object-relational database system (PostGIS) with over 500 spatial data processing tools and advanced raster processing capabilities to process our data drive our decision was to develop RINX using the FASRC/PostGIS environment.

Major processing steps taken to execute our solution, as shown in Figure 3 of the RINX architecture diagram, included:

- Creation of the database,
- Data loading of climate rasters and patient address locations,
- Data extraction of 7 climate variables for all person/days,
- Calculation of additional humidity variables,
- Automation of the processes, and
- Scaling of solution on the cluster computing environment

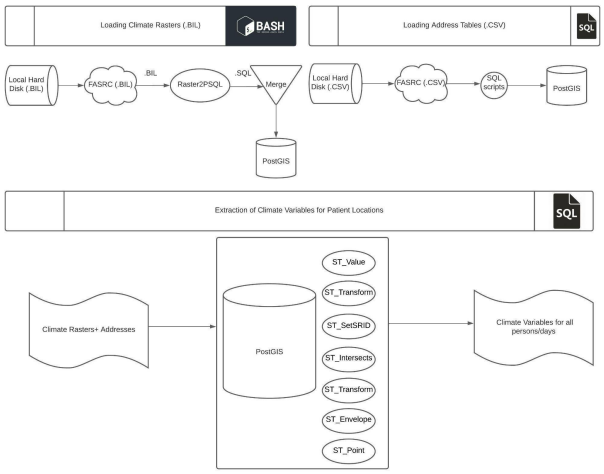

Figure 3. Architecture Diagram for RINX (Harvard CGA. 2022)

Database Creation: This involved launching the PostGIS database on FASRC. The database was enabled with the raster extension of PostGIS to support raster handling. The database schema was defined for both climate data and patient addresses.

Data Loading: After the creation of the database, the climate rasters are uploaded to the PostGIS database. The upload bash script reads the climate rasters from the disk and then using raster2pgsql functionality of PostGIS uploads it to the database. The raster2pgsql is a raster loader executable that loads GDAL supported raster formats into sql suitable for loading into a PostGIS raster table. It is capable of loading folders of raster files as well as creating overviews of rasters. The climate rasters are merged to one table using the UNION function of PostGIS. The input patient addresses are then uploaded to the database using another bash script. This automation script reads the input address list in .CSV format as shown in Table 1 below:

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

Table 1: Sample (made up) values for input address dataset

Longitude, Latitude locations in Table 1 are randomly determined, they are not actual patient locations. The address_id is a de-identified sequential number (001), with each address location appended with "_1", "_2", and so on for each patient address location through childhood. This data is uploaded to PostGIS, and is formatted as a table. This involves creation of an address table and copying of input data to that table. The table is then formatted in correct format to enable extraction of values from Raster data.

Data Extraction: After the upload of climate rasters and input addresses the next step was the extraction of values. For this a script was developed to achieve the extraction of values for every address point from climate raster using multiple PostGIS functions: ST_Value, ST_Transform, ST_SRID, ST_Intersect, ST_Envelope, ST_Point. The script performs these spatial operations in predefined order and automates the extraction process. The script was used for extraction of 7 climate variables for all patient locations.

Additional Calculations: Thereafter, two additional variables of absolute and relative humidity were calculated using the equations (1) and (2) respectively:

$$rh = 100 \times \left(\frac{e^{\frac{17.625 \times tdmean}{243.04 + tdmean}}}{e^{\frac{17.625 \times tmean}{243.04 + tmean}}}\right) \quad (1)$$

$$ah = \frac{6.112 \times e^{\frac{17.67 \times tmean}{243.5 + tmean}} \times rh \times 2.1674}{273.15 + tmean} \quad (2),$$

where rh = relative humidity (%)

ah = absolute humidity (grams/m^3)

tdmean = Daily mean dew point (°C)

tmean = Daily mean dew point (°C)

A SQL script was written to read the pre-calculated values of tmean and tdmean from the database and use it for calculation on humidity variables. Thus, a total of 9 climate variables are calculated using the system.

Process Automation: Due to the big data involved, the entire process was automated using a combination of various scripting techniques. BASH scripts were used to automate the data upload, pre-processing and formatting. SQL scripts were used to automate the extraction process from PostGIS.

Scaling: To enable fast and efficient calculation of the big data, the process was scaled on High Performance Compute Cluster as shown in Figure 4 below. The processing system for each climate variable consisted of PostGIS servers of 32 GB RAM and 2 vCPUs each. Two batches of patient addresses and climate Raster data was input in the processing system. The processing system was scaled to 7 PostGIS servers running in parallel to enable extraction of 1 climate variable for 19 years of raster data for two batches of 4,796 addresses each.

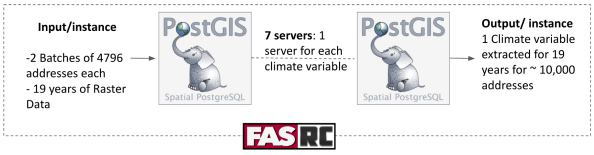

Figure 4. Scaling the process on High Performance Compute Cluster using PostGIS

5. RESULTS

Using RINX allowed us to extract 7 daily climate variables from the PRISM data for 10,022,945 patient days producing a total number of 70,160,615 climate observations. Additionally, relative humidity (rh) and absolute humidity (ah) were calculated by RINX, for a total of 9 variables. The 9 climate variables extracted are:

- ppt: Daily precipitation (mm)
- tmin: Daily minimum temperature in degrees Celsius (°C)
- tmax: Daily maximum temperature (°C)
- tmean: Daily mean temperature (°C)
- tdmean: Daily mean dew point (°C)
- vpdmin: Daily minimum vapour pressure deficit (hPa)
- vpdmax: Daily maximum vapour pressure deficit (hPa)
- rh: Relative humidity (%)
- ah: Absolute humidity (grams/m^3)

The entire process took 24 hours to load the rasters, and 4 days to process all observations. It is estimated that traditional methods would have taken 3 or more weeks to extract the same amount of observations, thus saving considerable time and cost, and enabling medical researchers to use the extracted data in a timely manner. The format of the output data included the address ID for the cohort member, the day of the observation in "yyyymmdd" format, and the 9 climate variables. Figure 5 below displays the data in its final format.

```
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
```

Figure 5. RINX Output Data Format

Exporting the results in .CSV text format allowed for medical health professionals to load this data into other databases and link it with patient health data for statistical analyses. This has enabled the analysis of climate effects on health outcomes. Regarding the use of this data, feedback from the medical researchers is summarised in this quote: "The PRISM climate data extracted by the CGA allowed us to study associations of precipitation, relative humidity and temperature with lung function in children. The climate data will also allow us to study similar associations in adults. There is a need to determine if short term exposure to these weather conditions affects the respiratory health of children and adults, especially in the context of a changing climate."

6. CONCLUSIONS AND BROADER APPLICATIONS

The RINX solution for extraction of spatio-temporal big raster data provides an essential first step to enable the study of climate effects on health. It is a scalable solution that uses exclusively open source software on a high-performance computing environment. RINX can be scaled to analyse much larger datasets, and can be implemented on any computing cluster, server, or workstation.

Our work used RINX to extract data from the PRISM 800m resolution daily climate data series, but we feel one of the greatest benefits of RINX is the ability to apply it to any spatio-temporal raster dataset. The high-resolution PRISM dataset is not free and had to be purchased, however PRISM also produces a free climate dataset at 4 kilometres resolution that our solution could be applied to. RINX has the potential to be used for any geospatial study involving extracting values from temporal raster data such as NDVI, land cover, night lights, etc.

The Viva cohort is just one of the 60 or so cohorts in the ECHO program, and for next steps our team will focus on enriching additional cohort datasets with climate exposure information using RINX. The Viva cohort analysed is Harvard based, allowing for processing data in a secure Harvard controlled environment by Institutional Review Board (IRB) approved personnel. Other cohorts are spread out among many Universities, with most restricting cohort data to residing on local environments, handled by IRB approved personnel for that University. Therefore, an upcoming challenge for our team will be to enable the use of RINX on non-Harvard computing environments. The open source code for RINX is available on our Github repository (Harvard CGA. 2022).

ACKNOWLEDGEMENTS

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

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.

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 modelling: 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.

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

FASRC. n.d. FAS Research Computing | Harvard University | Faculty of Arts and Sciences. Accessed June 15, 2022. https://www.rc.fas.harvard.edu/.

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. 2022. GitHub for RINX. https://github.com/cga-harvard/RINX-Raster_INformation_eXtraction_System.

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/.

Jung, Martin. 2013. *LecoS - A QGIS plugin for automated landscape ecology analysis*. N.p.: PeerJ Publishing. https://peerj.com/preprints/116v2.pdf.

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

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

Harvard Medical School. n.d. Viva Echo Fact Sheet. Accessed July 4, 2022. https://www.hms.harvard.edu/viva/Viva%20ECHO%20Fact%20Sheet.pdf.

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.

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.

PostGIS. n.d. About PostGIS | PostGIS. Accessed June 15, 2022. https://postgis.net/.

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.

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, L., K. Lu, P. Liu, R. Ranjan, and L. Chen. 2014. IK-SVD: Dictionary Learning for Spatial Big Data via Incremental Atom Update Publisher: IEEE Cite This PDF. *Comput. Sci. Eng.* 16 (4): 41-52. doi: 10.1109/MCSE.2014.52.

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.