Automating the integration and accessibility of diverse geospatial data cubes

Heather M. Savoy[1] and Debra P.C. Peters[2]

[1] New Mexico State University, Jornada Experimental Range, and Jornada Basin Long Term Ecological Research Program, Las Cruces, NM, USA
[2] USDA ARS, Jornada Experimental Range, and Jornada Basin Long Term Ecological Research Program , Las Cruces, NM, USA

Objective

Continental-scale geospatial data, such as daily precipitation or soil properties across North America, have the capability to serve the data needs of a wide variety of ecological projects, but their cost in terms of time and memory can hinder efficient integration. To reduce the time of downloading big data sets from online repositories, cropping the data to desired extents, harmonizing the resolutions and coordinate reference systems, and performing statistical analyses for multiple data sets, we aim to create a portal that automates these steps.

DASH: Data Access and Spatiotemporal Harmonization

Motivation: Ecological Studies

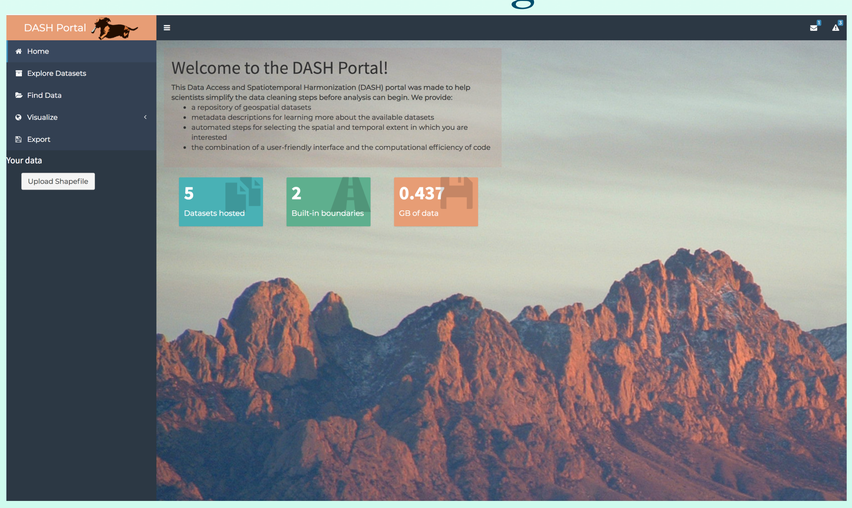
Ecology of Complex Disease Systems

- A USDA Grand Challenge project was created to use Vesicular Stomatitis Virus (VSV) as a model for predictive disease ecology.
- The disease ecology system for VSV is a complex network of interactions among the virus, multiple insect vectors, multiple livestock hosts, and the environment. A big data-model integration approach is being used for this transdisciplinary project (Peters et al., 2018).
- Predictive models are being built to serve as an early warning system (Peters et al., submitted), and these models incorporate many geospatial data layers, such as:
  - Climate (precipitation, temperature) [PRISM, NOAA]
  - Vegetation [LANDSAT]
  - Horse density [USDA]
  - Distance to streams with water [USGS]
- The DASH portal will serve as a data repository for this project to help scientists to access both the original and processed geospatial data layers used in this project.

Ecology of the Dust Bowl

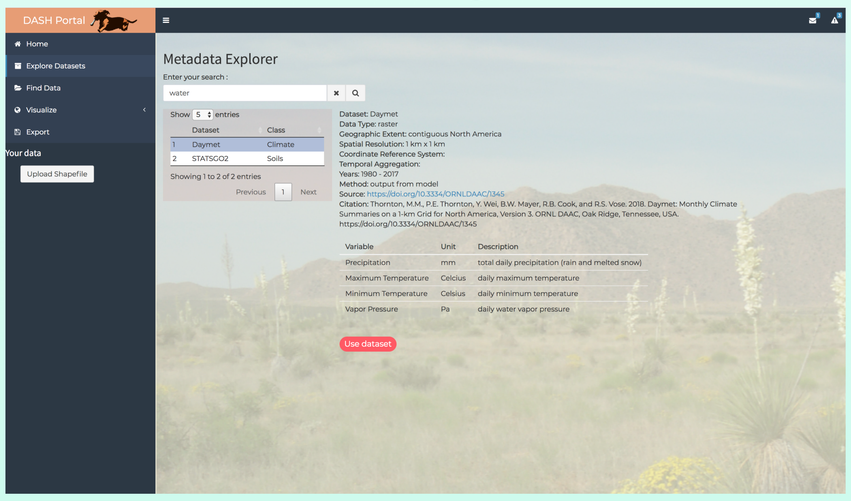
- This study is using historical agricultural data to predict ecological consequences of changes in multi-scale connectivity, climate, and land use.
- Data records from the 1933-1939 large-scale drought in the Central US, as well as before (1926-1932) and after (1940-1946), are being used to relate changes in agricultural production to climate, landscape, and broad-scale features of the land surface.
- Geospatial data involved include:
  - County-level corn yield and land use records
  - Climate (precipitation, temperature)
  - Soil texture properties
  - Simulated soil water content
  - Broad-scale dust and sand movement, and landform distribution
- The DASH portal will host the organized historical data and the derived variables from SOILWAT simulations for use by other scientists.

Home Page



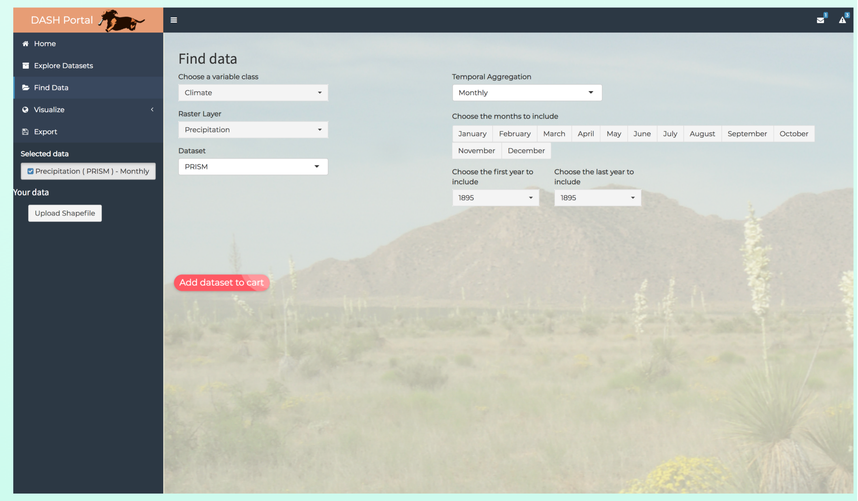
- Overview statistics about the data hosted are generated dynamically from the current metadata database and present files
  - Number of datasets
  - Number of built-in boundaries/shapefiles
  - Total amount of data hosted in GB
- The sidebar on the left provides navigation and is populated with data selections as the user performs tasks in the following screens

Dataset Explorer



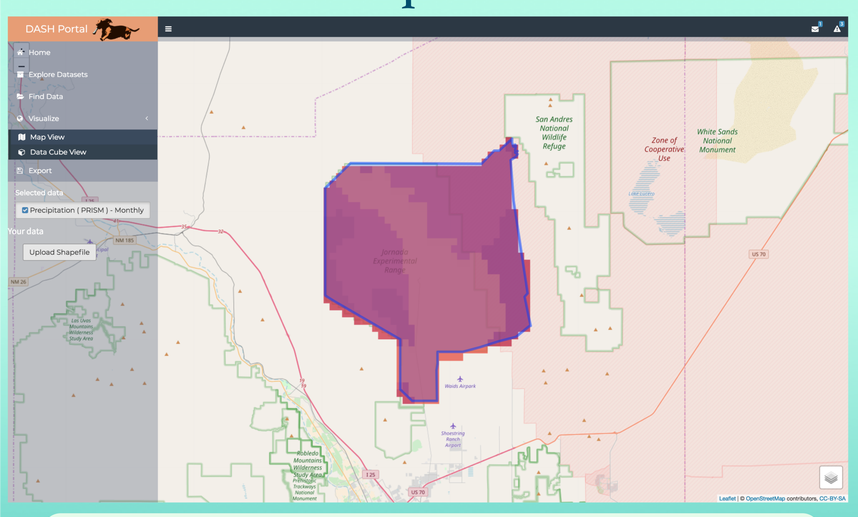
- Search for any dataset, variable, or property available in the portal
- The search is applied to the entire metadata database
- A table is generated listing all datasets that match the search
- The rows in the table can be selected to see formatted metadata about the dataset printed to the right
- A button to use data from the selected dataset is available to partially pre-populate the Data Selection screen

Data Selection



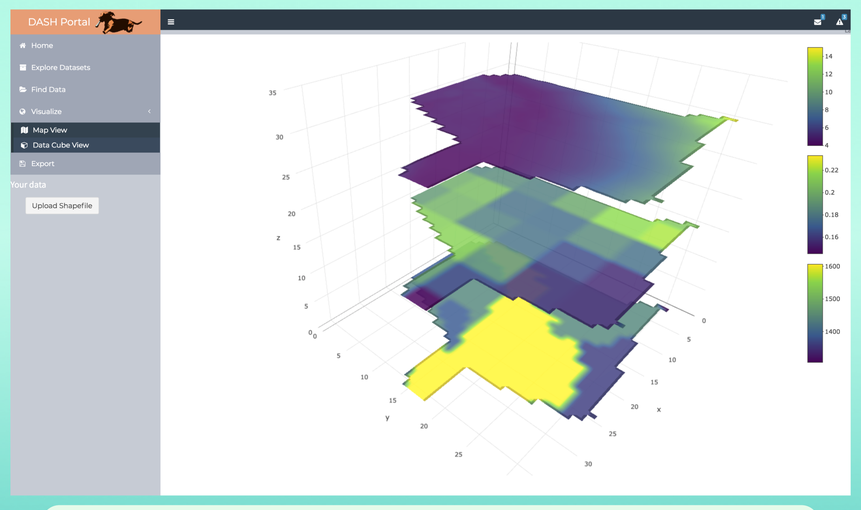
- Select data variables by class and/or type. Live search is supported, i.e. “temp” can be typed to reduce options to those related to temperature.
- If more than one dataset contains that variable, potential datasets are listed for selection
- If the variable has a temporal component, relevant temporal aspects will appear for selection
- Add selection to the list, which appears on the left sidebar
- Repeat as necessary until all desired data is selected

Map View



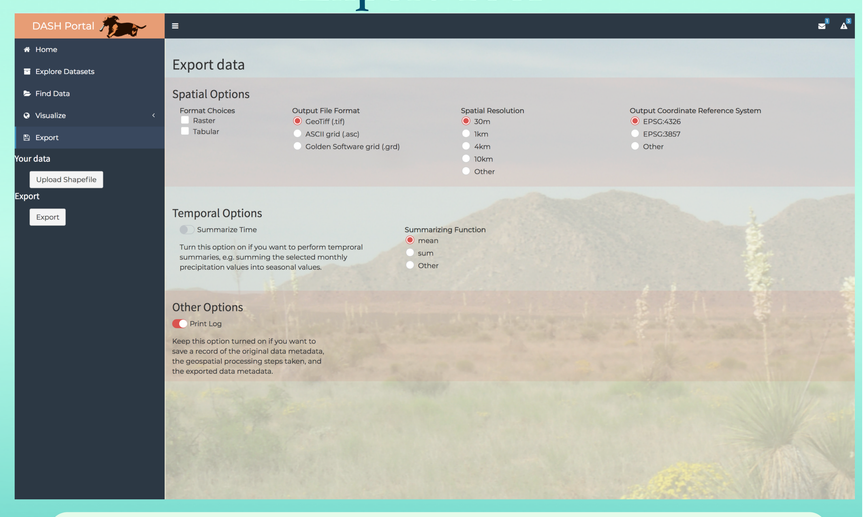
- Display selected data as either original extent, or clipped to a boundary.
- Boundaries can be chosen from the built-in library or uploaded
- The background map changes its level of detail depending on the zoom level to provide spatial context. There are different map options to choose from.
- The color scheme can be modified.
- This view is best for visualizing one data layer at a time.

Data Cube View



- Display selected data layers in a stacked view to visualize the data cube.
- The stack can be created with a single variable over time, multiple variables, or a combination.
- Hovering provides data from every layer at that location.
- This view supports interactive zooming and rotation to spin the data cube around for exploration.
- This view is best for visualizing multiple data layers together.

Export Tool



- Choose how the selected data should be exported.
- Spatial options exist to choose file formats, spatial resolution, and coordinate reference system.
- Temporal options exist to choose temporal aggregations, e.g. summing monthly precipitation into seasonal values.
- There is an option to export a log of operations performed throughout the session and metadata related to the original and exported data. This feature is designed to help in writing the methods section for publications.

Tools

Web Interface

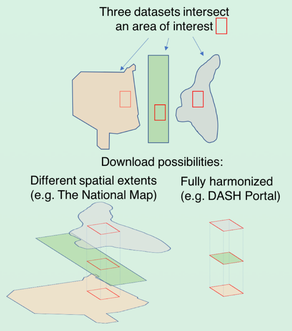
- The portal is built as a Shiny app with the `shiny` package in the R language (Chang et al., 2018; R Core Team, 2018). R provides data wrangling and statistical capabilities and Shiny allows for building web-based interactive tools to extend those capabilities.
- The map visualization feature relies on the `leaflet` package (Cheng et al., 2018). Leaflet is an opensource JavaScript library for making interactive maps.
- The data cube visualization feature relies on the `plotly` package (Sievert, 2018). Plotly is an opensource JavaScript library for making interactive plots in general.

Geospatial Processing

- For smaller datasets, R functions are used for transforming, cropping, etc.
- For larger datasets, external calls to the Geospatial Data Abstraction Library (GDAL, 2018) are used for geospatial manipulations for increasing computational efficiency.

Comparisons

- **Geographic Information Systems (GIS)** software programs exist to perform many of the operations in this tool, but the over-abundance of features can slow down the programs. Processing large datasets in GIS can cause unreasonable processing times or software crashes. This portal strives to support scientists needing large datasets or who potentially do not have GIS experience.
- **Online Data Portals** exist to host different datasets. For example, The National Map (https://nationalmap.gov/) provides multiple data layers for the US, but they are topography related and downloading multiple layers may result in different spatial extents. This portal will contain data from many fields to support complex, multi-faceted agro-ecosystem research and provide fully harmonized data cubes for downloading.



- **Google Earth Engine (GEE)** is a powerful tool for server-side geospatial processing and large dataset access. However, performing those processes requires knowledge of a GEE variant of JavaScript or python. Download options are limited to Google Drive or Google Cloud, which could incur fees for large datasets. This portal does not require any programming skill or fees.

References

Chang, W., Cheng, J., Allaire, J. J., Xie, Y., & McPherson, J. (2018). shiny: Web Application Framework for R. Retrieved from https://cran.r-project.org/package=shiny

Cheng, J., Karambelkar, B., & Xie, Y. (2018). leaflet: Create Interactive Web Maps with the JavaScript “Leaflet” Library. Retrieved from https://cran.r-project.org/package=leaflet

GDAL. 2018. GDAL - Geospatial Data Abstraction Library: Version 2.2.4, Open Source Geospatial Foundation, http://gdal.osgeo.org

R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Sievert, C. (2018). plotly for R. Retrieved from https://plotly-book.cpsievert.me

Peters, D.P.C., Burruss, N.D., Rodriguez, L.L., McVey, D.S., Elias, E.H., Pelzel-McCluskey, A.M., Derner, J.D., Schrader, T.S., Yao, J., Pauszek, S.J., Lombard, J., Archer, S.R., Bestelmeyer, B.T., Browning, D.M., Brungard, C.W., Hatfield, J.L., Hanan, N.P., Herrick, J.E., Okin, G.S., Sala, O.E., Savoy, H.M. & Vivoni, E.R. (2018). An integrated view of complex landscapes: a big data-model integration approach to trans-disciplinary science. *BioScience*, *68*(9), 653–669.

Peters, D.P.C, McVey, D.S., Elias, E.H., Pelzel-McCluskey, A.M., Derner, J.D., Burruss, N.D., Schrader, T.S., Yao, J., Pauszek, S.J., Lombard, J. & Rodriguez, L.L. (submitted). Developing an early warning strategy for a vector-borne disease using big data-model integration and machine learning. *Nature*.