Manuscript title:

- Automated retrieval, preprocessing, and visualization of gridded hydrometeorology data products for spatial-temporal exploratory analysis and intercomparison

Authors:

- Jimmy Phuong[1,2]; Christina Bandaragoda[2]*; Erkan Istanbulluoglu[2]; Claire Beveridge[2]; Ronda Strauch[3]; Landung Setiawan[4]; Sean D Mooney[1]

Author affiliation:

- [1]Biomedical and Health Informatics, School of Medicine, University of Washington, Seattle, Washington, USA; [2]Civil and Environmental Engineering, College of Engineering, University of Washington, Seattle, Washington, USA; [3]Seattle City Light, Seattle, Washington, [4]eScience Institute, University of Washington, Seattle, Washington, USA.
- *Correspondence should be addressed to Dr. Christina Bandaragoda. (cband@uw.edu), Civil & Environmental Engineering, University of Washington, 201 More Hall, Box 352700, Seattle, WA 98195-2700.

Highlights:

- OGH initializes data retrieval using the geospatial boundary defined by an ESRI shapefile.
- OGH facilitates intercomparison of multiple gridded data products within a user-defined extent.
- OGH provides a metadata template for annotation of 1D ASCII time-series gridded data products.
- OGH is designed to compute geospatial-temporal analysis in distributed docker or server environments.
- OGH use cases (4) illustrate the application usability for gridded hydrometeorological data processing.

Keywords: Python, cloud computing, Shapefile-based data retrieval, watershed hydrometeorology

Software and/or data availability:

- OGH can be installed from conda-forge. OGH v.0.1.11 is released on GitHub (https://github.com/Freshwater-Initiative/Observatory), and is freely available under an MIT license. This GitHub repository and the HydroShare resources are maintained by the corresponding author Christina Bandaragoda.
- This Python library was developed using Python 3.6 conventions within a JupyterHub-Unix docker environment hosted on the CUAHSI HydroShare server.
- Use-case notebooks can be found at the GitHub repository (https://github.com/Freshwater-Initiative/Observatory/tree/master/tutorials) and the HydroShare resource (https://www.hydroshare.org/resource/87dc5742cf164126a11ff45c3307fd9d/)

Abstract:

Spatially-distributed time-series data support a range of environmental modeling and data research efforts. A critical first step to any such effort is acquiring interpolated hydrometeorological data. Standardized tools to facilitate this process into analyses have not been readily available for watershed scale research. Here, we introduce the Observatory for Gridded Hydrometeorology (OGH), an open source python library that fills this critical software gap by providing a cyberinfrastructure component to fetch and manage distributed data processed from regional and continental-scale gridded hydrometeorology products. Our approach involves annotating metadata to make gridded data products discoverable and usable within the software, enabling interoperability and reproducibility of models that use the data. This paper presents the design, architecture, and application of OGH using four commonly practiced use-cases with gridded time-series data at watershed scales.

1. Introduction

Gridded data products are extensively used in Earth Science research [King *et al.,* 2013; Gampe *et al.,* 2016; Ledesma and Futter, 2017], social vulnerability analysis [Cutter *et al.,* 2014] and population risk and estimate studies [Lloyd *et al.,* 2017]. Gridded hydrometeorological data products are produced by interpolating local observations to predetermined spatial-temporal resolutions. The purpose of developing gridded products is to extend spatial information beyond point locations and provide space and time dimensions to observations such that spatial-temporal variability can be analyzed. Gridded data products also provide a means to compare and validate numerical weather prediction outputs (short term forecasts and long term climate change). Prior studies in hydrometeorology have highlighted the growing usefulness of gridded data products in Earth science modeling (most recently reviewed in Henn et al., 2018). In this paper, we introduce the Observatory for Gridded Hydrometeorology (OGH), an open-source python toolkit to streamline processes for interacting with gridded hydrometeorological data products at a user-defined spatial scale of interest (single location to regional watershed), designed to support watershed scale science applications. This tool fills a model pre-processing gap of processing large regional datasets (~1000 km2) for smaller scale geographic subsets (~1 km^2 - 100 km^2).

In the conterminous United States (CONUS), since the introduction of Parameter-elevation Regressions on Independent Slopes Model [PRISM, Daly *et al.,* 1994], gridded meteorological data products are routinely interpolated using daily measurements from over 20,000 NOAA COOP observation stations [Maurer *et al.,* 2002; Livneh *et al.,* 2013], with similar products in development for other regions around the world [Yanto *et al.,* 2017]. In these data products, each grid cell contains observation-interpolated, multivariate time-series [Maurer *et al.,* 2002]. It is common practice for the hydrologic community to incorporate gridded meteorological time-series variables as inputs to land surface hydrologic models such as the Variable Infiltration Capacity (VIC) model [Liang *et al.,* 1994]. A wealth of modeled land surface hydrologic states (e.g, soil moisture) and fluxes (e.g., latent and sensible heat) have been developed and used in Earth science research [Livneh *et al.* 2013; Livneh *et al.* 2015]. In mountainous regions, where ground-level observation collection is not feasible, the Weather Research Forecasting (WRF) atmospheric model has been downscaled and similarly used to produce hydrologic model outputs; however, this process inevitably requires bias-correction based on observational products [Salathé *et al.,* 2010; 2014]. Recently, gridded hydrometeorological data was

combined with geo-environmental and demographic surveys to yield gridded population data sets and new insight to enhance population-modeling resolution and accuracy [Cutter *et al.,* 2014; Lloyd *et al.,* 2017]. It can be expected that gridded data products will continue to increase in abundance and complexity in how they represent the impacts of geography and landscape morphology.

Before the potential usefulness of gridded data products can be realized for watershed-scale actionable research, several data and metadata access challenges need to be addressed. Continental-scale gridded data products, such as those from Livneh *et al.,* (2013; 2015) and Salathé et al., (2014), are increasingly being published as NetCDF files. For watershed researchers who are interested in studying physical processes, NetCDF files used for regional and continental-scale gridded products (1000 km^2 - 10,000 km^2) contain information that far exceeds the geographic extent needed for local, watershed-scale research (*e.g.,* 1-100 km^2 catchment area), adding computational resource burden in exploratory research. One alternative data product format is the 1D ASCII time-series files for a geographically-specific gridded cell. As observed in Livneh *et al.* (2013; 2015) and Salathé *et al.* (2014) data products, 1D ASCII time-series files are not self-described with column names, time-series dates, or value units, so annotations are needed in order to perform analyses with files in this format. Information to locate and use the data files may be confined to elusive publications and documentation files. Even so, the data files may be hosted in management structures for their study convenience (*e.g.,* Universal Transverse Mercator boundaries), making manual data retrieval non-trivial and not intuitive by human interpretation. Hence, annotating data provenance, metadata provenance, and file management structure are crucial steps towards making gridded data products findable, accessible, interoperable, and reusable [FAIR; Wilkinson *et al.,* 2016; Mons *et al.,* 2017] for secondary analyses.

Currently available tools for water data, such as WaterML, WOFpy, GSFLOW, Geoknife, offer access to time-series of observation data [Kadlec *et al.,* 2012, Gardner *et al.,* 2018; Read *et al.,* 2015]. However, in areas where observations are unavailable, such as heterogeneous landscapes at high elevation locations, data sparsity can be addressed with krigged and model-interpolated data products. Python libraries such as OpenClimateGIS offer access to NetCDF gridded data products, but these functionalities exclude legacy data sets provided in 1D ASCII

time-series format. More importantly, aside from data access, it is challenging to recognize differences between gridded data product - such as aggregation into different gridded cell schemas, the temporal resolution, time period, or long-term trends - for use in future modelling efforts and model validation operations.

To streamline the processes needed for interacting with gridded hydrometeorological data products in a FAIR manner, promote use in Earth modeling and interdisciplinary domains, and support decision-making for selecting gridded cells in a study site, we designed OGH as a open-source python toolkit to select gridded cells in a study site, data download, spatial-temporal analyses, and provide data visualization. In this paper, we describe the design of the OGH Python library, which contains functions to conduct basic data access and data processing operations to simplify gridded data product use in research. Users start with an ESRI shapefile that describes their study site (e.g., HUC12 units, county-boundaries, state-boundaries) to generate a comma-separated table that helps with file management of gridded cell data availability across gridded data products. Users can then conduct data retrieval in-parallel from a number of gridded hydrometeorological data products with automated file management for analyses. These functions are flexible to integrate new gridded products and publishing standards. To address the absence and/or variability in describing online gridded data products, we propose a set of minimum annotation criteria (metadata fields) for describing ASCII gridded hydrometeorological data products, and the decision steps needed to access these gridded datasets.

In the Methods section, we describe OGH software design for gridded cell selection and visualization, data download, spatial-temporal calculations, and applied statistics functionalities. OGH was designed to incorporate climatological and hydrometeorological gridded data products with comparable data structures to ASCII and NetCDF formats. Here, we emphasize watershed-scale applications using 1D ASCII time-series data products. In the Results section, we demonstrate these functionalities using three watersheds of end-member climatologies in CONUS: two high alpine glacierized watershed in the high-end of the precipitation gradient in the CONUS (> 3,000 mm), Sauk-Suiattle, WA and Elwha river basin, WA; and a desert watershed with a large precipitation gradient from valleys to peaks (200 mm - 3500 mm), Upper Rio Salado, NM. We compute precipitation and temperature spatial-temporal statistics and

exceedance probability calculations using gridded hydrometeorological data products from Livneh *et al.,* (2013) and Salathé *et al.,* (2014). OGH v.0.1.11 is publicly accessible at https://github.com/Freshwater-Initiative/Observatory and available by conda installation.

2. Methods

OGH is a python library designed to perform gridded cell selection, data download, data processing for desired space-time analytics, and visualization of spatial-temporal data. A proposed set of metadata annotations was developed to provide default information about data product capacities using two case study gridded products. We provide OGH examples reproducible on HydroShare, a cyber-infrastructure for sharing data and models [Heard *et al.,* 2014; Horsburgh *et al.,* 2016; Castronova, 2017]. OGH is not dependent on HydroShare, though in the example use-cases, HydroShare is a platform providing dockerized or local server environments for community software, including those used by OGH operations to manage, compute, and store directories of files.

User workflows, scenario use-cases, and key socio-technical needs were developed through key informant interviews, diagramming, and rapid prototype testing sessions (Baxter and Sommerville*,* 2011; Devi *et al.,* 2012). Four key principles of user-centered design and engagement were maintained throughout the rapid prototype testing process, described in Supplementary Table 1 (Devi *et al.,* 2012).

2.1. Software Design

We designed OGH using open-source Python 3.6 programming, which is interoperable with major operating systems and computing environments used by personal desktop computers, high performance computers, and supercomputers. OGH functions are written as modular components that leverage classes and methods from a number of Python libraries: time-series analysis from Pandas [McKinney *et al.,* 2011]; geospatial analytics from Geopandas, Fiona, Shapely, and Matplotlib-Basemap [McKinney 2011; Gillies 2007, 2011; Jordahl *et al.,* 2014]; and task management from Multiprocessing and Dask [Rocklin, 2015]. Operations from these libraries are assembled into scripted functions, which can be wrapped into sequential operations or applied in distributed computing practices.

OGH is intended to perform operations within computing environments, where input and output files are managed within data sharing platforms and community data repositories (e.g., HydroShare). While HydroShare has many features, we make use of the docker or server environment, where file storage, migration, and computation can be performed with multi-core resources. HydroShare (www.hydroshare.org) is a collaborative platform that supports data sharing and model reproducibility in hydrologic research, and it provides a cloud-computing environment through the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) JupyterHub server [Horsburgh *et al.,* 2016]. The HydroShare REST API Python library (*hs_restclient*) is used to migrate files contained in HydroShare resources in and out of the JupyterHub docker environments [HydroShare REST API Python client library, 2018]. File storage within the computing environment is intended to be temporary storage for the duration of the computations. The research workflows presented in this paper are designed to be guided by Jupyter notebooks, sharable code-execution interfaces that operate within the JupyterHub docker environments [Castronova, 2017].

2.2. Gridded data product annotations

We annotated seven daily, 1/16° (~6 km) gridded data products from three studies published and made available online (Livneh *et al.,* 2013, 2015; Salathé *et al.,* 2014). The datasets published by Livneh *et al.,* (2013) provided an interpolated climate-station meteorology for CONUS and a meteorology data product that was bias-corrected to the Columbia river basin regional climatology in the time-span from 1915 to 2011. Expanding on Livneh *et al.,* (2013), Livneh *et al.,* (2015) includes a PRISM-calibrated interpolated climate-station meteorology extends from Mexico to limited regions of Canada; however, the period has 33 total years less data with a time-span of 1950 to 2013. Both interpolated meteorology data products were used to predict macro-scale hydrologic fluxes at the 1/16°, daily resolution by the VIC model. The WRF gridded data product provides model downscaled daily precipitation, maximum and minimum air temperature, and wind speed for the Columbia river basin for the period of 1950 to 2010 [Salathé *et al.,* 2014]. The data products can be further differentiated according to their type of analysis and reported spatial coverage (Table 1).

Table 1: Summary of seven daily, 1/16° gridded data product. Descriptions in this summary compare data products by variable, date range, analysis type, intended spatial coverage, and their source publication.

Data set	Features and variables (in order)	Start date	End date	Analysis type	Spatial coverage	Publication
Climate station meteorology						
dailymet_livneh2013	PRECIP, TMIN, TMAX, WINDSPEED	1915-01-01	2011-12-31	raw	CONUS	[Livneh et al., 2013]
dailymet_bclivneh2013	PRECIP, TMIN, TMAX, WINDSPEED	1915-01-01	2011-12-31	bias-corrected	Columbia river	[Livneh et al., 2013]
dailymet_livneh2015	PRECIP, TMIN, TMAX, WINDSPEED	1950-01-01	2013-12-31	raw	CONUS	[Livneh et al., 2015]
WRF-NNRP model meteorology						
dailywrf_salathe2014	PRECIP, TMIN, TMAX, WINDSPEED	1950-01-01	2010-12-31	raw	Columbia river	[Salathé et al., 2014]
dailywrf_bcsalathe2014	PRECIP, TMIN, TMAX, WINDSPEED	1950-01-01	2010-12-31	bias-corrected	Columbia river	[Salathé et al., 2014]
Variable Infiltration Capacity						
dailyvic_livneh2013	YEAR, MONTH, DAY, EVAP, RUNOFF, BASEFLOW, SMTOP, SMMID, SMBOT, SWE, WDEW, SENSIBLE, LATENT, GRNDFLUX, RNET, RADTEMP, PREC	1915-01-01	2011-12-31	Physics-based model	CONUS	[Livneh et al., 2013]
dailyvic_livneh2015	YEAR, MONTH, DAY, EVAP, RUNOFF, BASEFLOW, SMTOP, SMMID, SMBOT, SWE, WDEW, SENSIBLE, LATENT, GRNDFLUX, RNET, PETTALL, PETSHORT, PETNATVEG	1950-01-01	2013-12-31	Physics-based model	CONUS	[Livneh et al., 2015]

The annotations describe the gridded data products published as ASCII files, where each file contains the gridded cell historic time-series data. Annotation features include the data set short name, information to locate the ASCII files, information about the file structure and sources of metadata, and metadata about the file variables (Table 2). File locations can be represented or reconstructed given by the web protocol (e.g., ftp, https), web domain and subdomain, decision steps within the subdomain to locate the data file subdirectory (e.g., centroid latitude given the the spatial resolution, bounding box bins), the filename structure, and the file format. The file structure is described by the variable list (left-to-right column order), time-series date range, temporal resolution, file delimiter, and the data types and unit increment for each variable. Full annotations are provided in the ogh_meta module.

Table 2: Minimum annotation criteria for gridded data products.

Metadata	Metadata descriptions
File location	
1. Dataset	name of the gridded data product
2. Spatial resolution	the distance between gridded cell centroids
3. Web protocol	the data transfer protocol
4. Domain	the web domain
5. Subdomain	the subdomain path
6. Decision steps	the file organization for locating data files
7. Filename structure	the standard components to the filename
8. File format	the file type at download
File structure	
9. Start date	the start date of the time-series
10. End date	the end date of the time-series
11. Temporal resolution	the unit increment for time-steps
12. Delimiter	the column separator within each line of data
13. Variable_list	the list of variables in order of appearance
14. Reference	the sources of metadata
Variable structure	
15. Variable_info	
● desc	the long name of the variable
● dtypes	the expected data type
● units	the unit increment of the data

2.2.1. Example use-cases

We present four example use-cases in the form of Jupyter Notebooks to demonstrate the OGH operations (Figure 1). In the first use-case, we identify the subset gridded cells of interest for four watershed study sites using the *treatgeoself* function (Figure 1A). The shapefiles for these watersheds are stored within a public HydroShare resource for ease of collaborative use [Sauk-Suiattle river basin available in Bandaragoda, C. (2017); Elwha river basin available in Beveridge, C. (2017); and Upper Rio Salado basin available in Bandaragoda, C. (2017)]. Watershed boundaries were defined in ArcGIS® using 12-digit Hydrologic Unit Code polygons from the National Watershed Boundary Database. In the second use-case, the time-series data files are retrieved, cataloged, then summarized for data availability (Figure 1B). In the third use-case, we focus on the Sauk-Suiattle watershed to determine the monthly meteorological spatial-temporal statistics computed using the Livneh *et al.,* (2013) Meteorology versus the Salathé *et al.,* 2014 WRF model output data products (Figure 1C). Finally, we compute potential runoff values using the VIC hydrologic data product from Livneh *et al.,* (2013) to approximate the 10% exceedance probability thresholds based on the daily time-series in each dataset (Figure 1D). Functions introduced in each use-case are summarized in Supplementary Table 2.

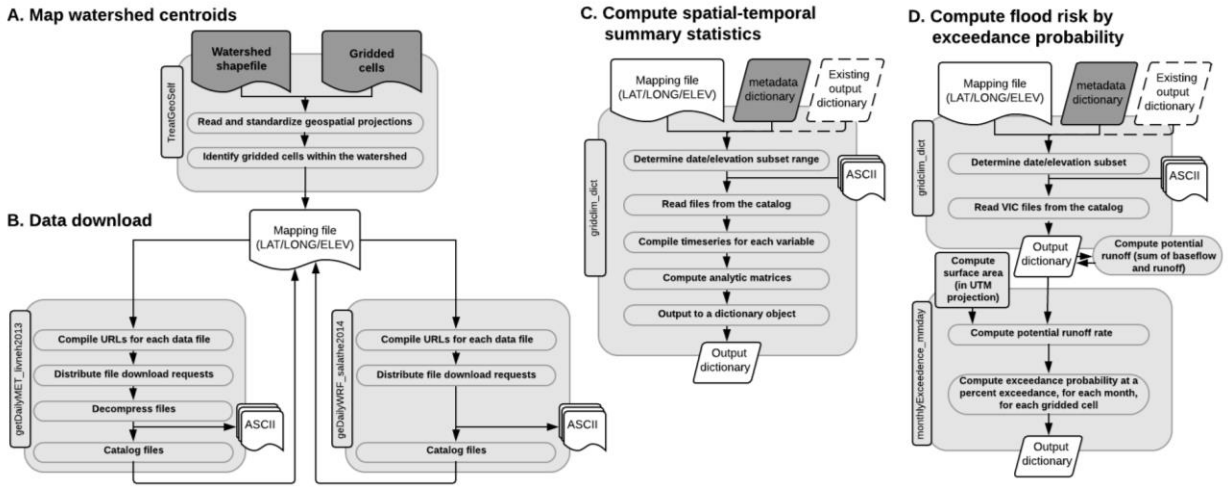

Figure 1: Scenario use-cases for OGH operations in cloud environments. A) With the user-defined watershed shapefile, spatial intersection with the 1/16º gridded cell centroid shapefile identifies the target gridded cell centroids, which are then documented within the mapping file output. B) The mapping file provides the Lat-Long coordinates for data download operations to produce a localized folder of ASCII files and a file catalog appended within the mapping file. C) For each gridded data product, spatial-temporal summary statistics are computed from the mapping file, file structure metadata, and ASCII files. The output dictionary can be reused to collect summary statistics for multiple data products. D) The hydrologic gridded product is used to compute exceedance probability at a statistical threshold for each gridded cell.

For illustration, five reference locations are used in the Sauk-Suiattle watershed examples. One gridded cell is identified at the highest average elevation value, 2216 meters above sea level. Two gridded cells were identified at the lowest average elevation value, 164 meters above sea level. The Darrington ranger station (COOP station 451992) is used as the reference source of meteorological observations, with data collected at 167 meters above sea level from Jan 1 1931 through Dec 31 2005. Sauk River Near Sauk, WA (USGS-12198500) used as the reference source of observed streamflow discharge using data collected between Jan 1 1950 through Dec 31 2011 [Konrad et al., 2012].

2.2.2. General workflow and required files

Workflows for the use-cases executed in the JupyterHub environment (Figure 2) are illustrated in detail in Figure 1. The general workflow begins with three HydroShare resources as sources of input files (Figure 2). Resource A - a HydroShare resource that contains a Jupyter Notebooks to execute code for each example use-case presented in this paper, Resource B - a HydroShare resource with a user-defined shapefile representing the region of interest (e.g., a watershed), and Resource C - a file of point-locations describing CONUS gridded cell centroids (only pre-requisite for ‘mapping watershed centroids’). From the web page for HydroShare Resource A, the Jupyter Notebooks are launched in the JupyterHub docker environment, wherein the HydroShare REST API functions migrate in requisite data files from Resource B

and C (Figure 2). Use-case notebooks 1 through 4 progress through OGH operations in Figure 2: 'identify watershed gridded cell centroids' (map watershed centroids),'download and display data availability' (data download), 'summarize monthly meteorology' (data processing), and 'compute exceedance probabilities' (another form of data processing) (Figure 2). Each use-case notebook produces output data files, plots data visualizations, and finally migrates these output to new shareable HydroShare resources to conclude the use-case demonstration.

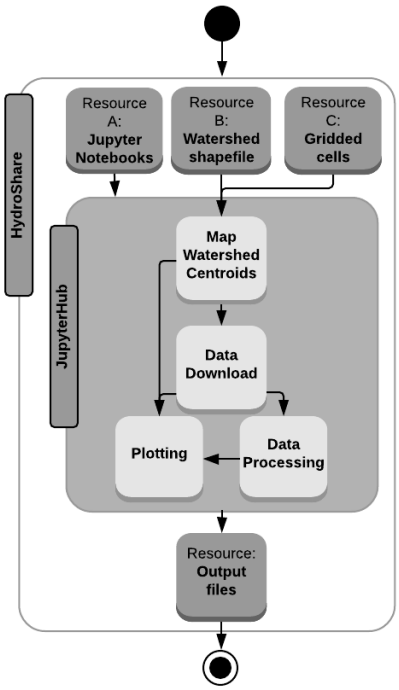

Figure 2: The general workflow for OGH in cloud-computing environments. HydroShare is the collaborative platform to facilitate file storage (Resources A, B and C), and HydroShare API rest-client migrates files in-and-out of the JupyterHub docker environment. Jupyter notebooks guide users through different use-cases like decision-making steps (e.g., map watershed centroids), data download, data processing, or generating visualization products. At the close of each notebook, research data products are migrated to HydroShare as new sharable resources.

2.2.3. Map watershed gridded cell centroids

For each of the three example watersheds, we generate a mapping file with the gridded cell centroids that spatially intersect these study sites (Figure 1a). Shapefiles were transformed into the 1984 World Geodetic System (WGS84) Lat-Long coordinates system as the standard projection. The study site was given a buffer region (default buffer distance of 0.06°) to include adjacent gridded cells. CONUS 1/16° (i.e., 0.06250°) gridded cell ESRI shapefile identifies each gridded cell by the 5-digit centroid latitude-longitude [Livneh *et al.,* 2013; Livneh, 2017]. Average elevation in the gridded cell (in meters above sea level) are based on the CONUS digital elevation model described in Livneh *et al.,* (2013).

The output from this use-case includes three mapping files, each denoting the latitude, longitude, and average elevation within the gridded cell. Other outputs include spatial visualizations of the study site (maps) and the elevation gradient among the gridded cells, and plots showing the data for select grid cell traces.

2.2.4. Summarize data download and availability

The mapping file guides data download from the seven gridded data products (Figure 1b). Target gridded cell files identified within the mapping file are web requested using data download wrapper functions (e.g., the *getDailyMET_livneh2013*). Request operations are distributed using multiprocessing pool operations. Downloaded files are cataloged into the mapping file (using *addCatalogToMap*). Data availability is determined for each gridded data product and watershed study site (using *mappingfileSummary*). Files that do not exist for retrieval are excluded from the catalog.

The output from this use-case is a summary table that describes data availability and seven folders containing the downloaded files for all the watersheds.

2.2.5. Summarize monthly meteorology

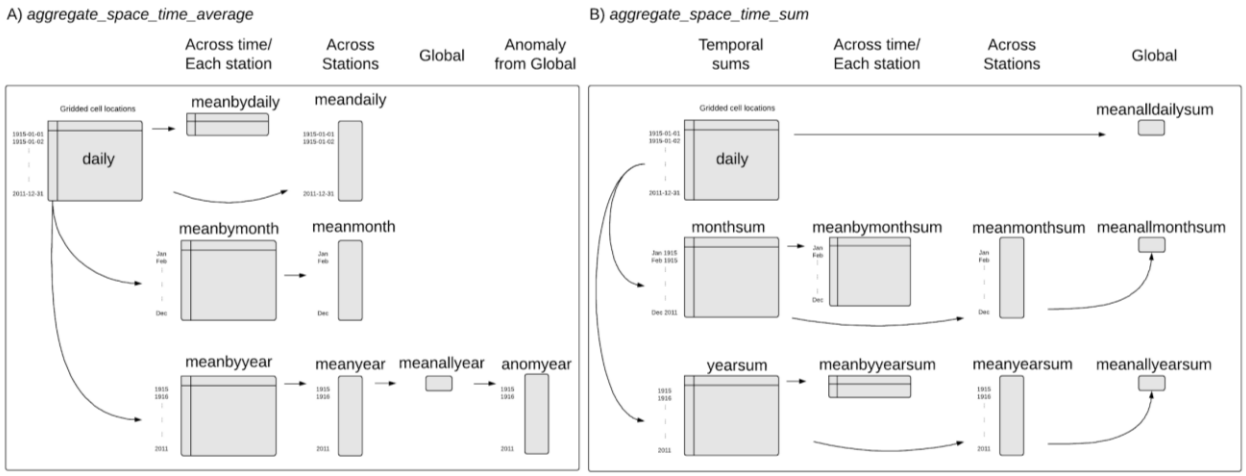

Figure 3: Spatial-temporal calculations (total sum and average). The *gridclim_dict* function reads and applies A) *aggregate_space_time_average* for each variable in the gridded data product. Variables to be considered by periodic sums (e.g. total annual precipitation) can be processed with B) *aggregate_space_time_sum*.

With the Livneh *et al.,* (2013) interpolated meteorology and Salathé *et al.,* (2014) WRF files, we compare the monthly meteorology variables for the Sauk-Suiattle river watershed using the 61-years of data from their mutual time-series period (i.e., Jan 1, 1950 through Dec 31,

2010).Using the Sauk-Suiattle mapping file, each variable from the ASCII gridded cell time-series are compiled into data frames, where rows are the daily time-series and columns are denoted with the gridded cell centroids. Temperature trends are interpreted using monthly mean, yearly mean, and global mean expected values (Figure 3A). Annual anomaly from the global mean value is used to identify years with extreme events (highest and lowest data values). Precipitation trends are interpreted using period sums (e.g., month-yearly sums and yearly sums) and mean of period sums (e.g., mean monthly sums, mean yearly sums, and the global mean of monthly sums) (Figure 3B).

The *gridclim_dict* function is a series of wrapped operations to return a dictionary object of spatial-temporal values across the ASCII gridded cell time-series (Figure 1C). *Gridclim_dict* provides parameters to specify elevation ranges or time-period selection, where defaults are all gridded cells and the full time-series. *Gridclim_dict* wraps *read_files_in_vardf*, which performs distributed file reading to generate a variable data frame for each variable in the data product, then applies *aggregate_space_time_average* to compute summary statistics using the prefix-suffix conventions for each variable (Figure 3A). The suffix represents the gridded data product, which can be user-defined or default to the annotated gridded data product dataset name (e.g., 'dailymet_livneh2013'). The first prefix appended to the suffix by underscore separation is the data product variable (e.g., 'PRECIP'). The second prefixes represent the statistical averages computed using the gridded cell dimensions (columns) and the temporal groupings (rows). The *aggregate_space_time_sum* produces outputs with the following second prefixes: "meanbydaily" (daily averages by each gridded cell), "meanbymonth" (daily averages by each month and gridded cell), "meanbyyear" (daily averages by each year and gridded cell), "meandaily" (average values across gridded cells by each date), "meanmonth" (daily averages across gridded cells for each month), "meanyear" (daily averages across gridded cells by each year), "meanallyear" (global mean of daily values across all years and gridded cells), and "anomyear" (the residual between each yearly mean and the global mean). To consider trends by period sums of daily events, the *aggregate_space_time_sum* function computes summary statistics of month-yearly and yearly sums (Figure 3B). The second prefixes here include "monthsum" (month-yearly sum of daily values by gridded station), "yearsum" (annual sum of daily values by gridded station), "meanbymonthsum" (mean of monthly sums for each calendar month and gridded cell), "meanbyyearsum" (mean of annual sums for each gridded cell), "meanmonthsum" (mean of month-yearly sums across gridded cells), "meanyearsum" (mean of annual sums across gridded cells),"meanalldailysum" (global mean of the daily sums across all

gridded cells), “meanallmonthsum” (global mean of month-yearly sums across gridded cells), ”meanallyearsum” (global mean of annual sums across gridded cells). Variations to these outputs are influenced by the gridded cells and time-period parameters.

The output for this use-case is a JSON dictionary object containing analytical data frames and data series, as shown in Figure 3, for each variable within a given time-frame. Other outputs include maps and monthly boxplots of the corresponding grid cell values.

2.2.6. Compute exceedance probabilities

For the Sauk-Suiattle study watershed, we approximate the 10% exceedance probability threshold using unrouted daily runoff for each calendar month and each gridded cell (Figure 1D). This is useful for visualizing the 10% highest daily streamflow generated for each grid cell in the dataset, which is a combined function of climate data, soils, land cover, and other model parameters in each grid cell. Potential runoff rates are computed as the sum of baseflow rate (mm/s) and surface flow runoff rate (mm/s) from Livneh *et al.,* (2013) VIC model outputs, converted the units to millimeters per day (mm/day) for comparison with daily precipitation rates. The same general operations were applied to Livneh *et al.,* (2015) VIC model outputs. For each calendar month (e.g., January) and each gridded cell (*e.g.,* centroid Lat-Long at 48.8723, -121.8974), daily potential runoff rates are compiled into a cumulative distribution function using data from 1 Jan 1950 through 31 Dec 2011 (62-years), the mutual/overlapping time-series period between Livneh *et al.,* (2013) and Livneh *et al.,* (2015) data products. Each distribution has approximately n=1800 VIC modeled observations. The 10% monthly exceedance probabilities (peak runoff threshold) for each gridded cell is estimated by linear interpolation as the 90th-percentile of the respective cumulative distributions [Vogel *et al.,* 2007]. An exceedance probability developed from a population of daily runoff in a given month should not be confused with annual flood statistics, which are developed by fitting a statistical distribution to a population of annual maximum daily streamflow. The 10% exceedance probability of observed streamflow discharge measured at Sauk River Near Sauk, WA (USGS-12189500) can be plotted with the modeled streamflow to provide an observed reference based on routed streamflow for relatively high flows.

The output for this use-case includes low, average, and high elevation analytical data frames at the 10% exceedance threshold for VIC results, compared to observations. Other outputs include maps and monthly boxplots of the exceedance probability for each gridded cell.

3. Results

3.1. Map Watershed Centroids

Functions used:

reprojShapefile, treatgeoself, multiSiteVisual, griddedCellGradient

Sauk-Suiattle, Elwha, and Upper Rio Salado watersheds were processed to generate mapping files and gridded cell gradient visualizations (Figure 4). Ninety-nine grid cells were identified for the Sauk-Suiattle river watershed, displaying the largest elevation difference (162 m - 2246 m) among the three watersheds (Table 3). Sauk-Suiattle river watershed is located in the northwestern region of the Cascade mountains in Washington state, USA, ranging from multiple high elevation areas in the southeast to a single outlet in the northwestern gridded cells. Fifty five gridded cells were identified for the Elwha river watershed, which has a comparable elevation difference to Sauk-Suiattle. The Elwha river watershed is located on the northern region of the Olympic Peninsula, where the elevation gradient (36 m - 1642 m) descends from the southern gridded cells by a single river draining to the northern gridded cells (Figure 4b). Thirty one grid cells were identified for the Upper Rio Salado watershed, with a higher elevation (1962 m - 2669 m) grid cells than the other two watersheds (Table 3). Upper Rio Salado's elevation gradient descends from the southwest-most to the northeast-most gridded cell.

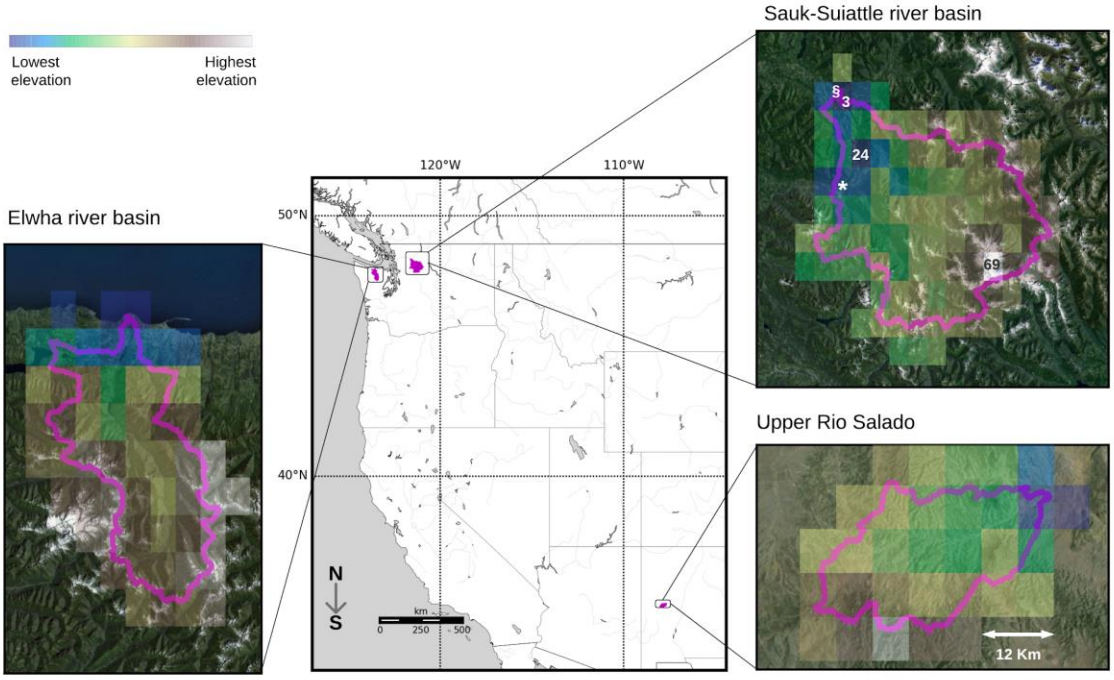

Figure 4. Aerial view of watersheds and gridded cells. The Sauk-Suiattle (164-2216 m), Elwha (36-1642 m), and Upper Rio Salado (1962-2669 m) watersheds located in western United States were visualized using the *multisiteVisual* function with an EPSG:3857 geospatial projection. Each watershed (outlined in magenta) and their gridded cells are visualized using the *griddedCellGradient* function at the 1/16º spatial-resolution (~6 km). In the Sauk-Suiattle watershed, five reference markers denote the highest elevation gridded cell (gridded cell 69, elevation: 2216 m), the lowest elevation gridded cells (gridded cells 3 and 24, elevation: 164 m), the Darrington Ranger Station site (*; COOP station 451992, elevation: 167 m) for observed meteorology data, and the Sauk River Near Sauk, WA streamflow gauge (§, USGS-12189500, elevation: 81 m) for observed streamflow discharge measured at the downstream-most tip of the watershed. While the numeric distributions can be conformed to a single scale, each watershed map uses a different numeric colorbar legend, so this figure is intended to provide a qualitative impression of the elevation gradient.

3.2. Summarize data download and availability

Functions used:

getDailyMET_livneh2013, getDailyMET_bcLivneh2013, getDailyMET_livneh2015, getDailyVIC_livneh2013, getDailyVIC_livneh2015, getDailyWRF_salathe2014, getDailyWRF_bcsalathe2014, mappingfileSummary

Among the seven gridded data products, 1D ASCII time-series files were fully-represented for Sauk-Suiattle, mostly represented for Elwha, and substantially limited in representation for Upper Rio Salado (Table 3). Download tasks for the full time-series ASCII files were distributed across 5-10 parallel worker CPUs. Computation efficiencies consisted of 693 Sauk-Suiattle files (10.0 Gb disk space) downloaded in 3 min 56 s wall time, 375 Elwha files (5.4 Gb) took 1 min 59

s, and 124 Upper Rio Salado files (2.7 Gb) took 48.8 s. All files were cataloged into their respective mapping files, organized by gridded data product short name and gridded cell centroid.

Elwha is located in the northwestern-most region of Washington state. Three gridded cells were available for all seven gridded data products, although they were available for the bias-corrected Livneh *et al.,* (2013) meteorology and Livneh *et al.,* (2015) meteorology and VIC model output products. Differences among the elevation gradient suggest that these three gridded cells were the northern-most low-elevation gridded cell, on the boundary of CONUS and Columbia River Basin extents (Figure 4). This poses certain limitations if multiple gridded data products for Elwha were used for intercomparison. These limitations are more obvious with Upper Rio Salado, which is located outside of the Columbia River Basin.

Livneh *et al.,* (2013) and (2015) gridded products were consistently spatially available in each of the watersheds, but the gridded products differed in the temporal extent (historic time-series included). The overlap period between Livneh *et al.,* (2013) and (2015) data products is Jan 1 1950 through Dec 31 2011 (62-years). Livneh *et al.,* (2013) and Salathé *et al.,* (2014) share the Jan 1 1950 through Dec 31 2010 (61-years), which is the same overlap period as Livneh *et al.,* (2013) and Salathé *et al.,* (2014). Despite the spatial availability of time-series data within a watershed, gridded data product intercomparisons should consider the historic time period represented as well as data variabilities such as correction methods and algorithms used to generate the gridded product.

Table 3: Counts of gridded cell ASCII files for each watershed by gridded data product. For the seven gridded data products, the downloaded files are summarized for each watershed as an inventory of the data availabilities and potential gaps due to spatial extent of the gridded data product.

	Watersheds		
	Sauk-Suiattle river	**Elwha river**	**Rio Salado**
Median Elevation in meters [range] (Number of gridded cells)	**1171[164-2216] (n=99)**	**1020[36-1642] (n=55)**	**2308[1962-2669] (n=31)**
dailymet_bclivneh2013	**1171[164-2216] (n=99)**	**1120[36-1642] (n=55)**	**0**
dailymet_livneh2013	**1171[164-2216] (n=99)**	**1146[174-1642] (n=52)**	**2308[1962-2669] (n=31)**
dailymet_livneh2015	**1171[164-2216] (n=99)**	**1120[36-1642] (n=55)**	**2308[1962-2669] (n=31)**
dailyvic_livneh2013	**1171[164-2216] (n=99)**	**1146[174-1642] (n=52)**	**2308[1962-2669] (n=31)**
dailyvic_livneh2015	**1171[164-2216] (n=99)**	**1120[36-1642](n=55)**	**2308[1962-2669] (n=31)**
dailywrf_salathe2014	**1171[164-2216] (n=99)**	**1142[97-1642] (n=53)**	**0**
dailywrf_bcsalathe2014	**1171[164-2216] (n=99)**	**1142[97-1642] (n=53)**	**0**

3.3. Summary monthly meteorology

Functions used:

findCentroidCode, overlappingDates, gridclim_dict, aggregate_space_time_sum, valueRange, saveDictOfDf, renderValueInBoxplot, renderValueInPoints

The function *gridclim_dict* generates a JSON dictionary for Sauk-Suiattle that contains 36 analytical data frames for Livneh *et al.,* (2013) meteorology and Salathé *et al.,* (2014) WRF outputs. Each data frame was named according to the analysis method (second prefix), variable (first prefix), and gridded data product short name (suffix). In Figure 5, average monthly total precipitation (*i.e., meanbymonthsum_PRECIP_dailymet_livneh2013* and *meanbymonthsum_PRECIP_dailywrf_salathe2014*) are depicted as boxplots to represent the distribution of values across the 99 gridded cells. The Livneh *et al.,* (2013) interpolated meteorology (Figure 5, top-left) indicates a greater variability of average monthly precipitation during the November through January months, while Salathé *et al.,* (2014) WRF model outputs (Figure 5, bottom-left) shows a higher median and greater variability from April through September. Average monthly precipitation for targeted high and low elevation grid cells (Figure 4) are plotted alongside the boxplots, as well as the point observations from Darrington Ranger Station (Figure 5). Comparison of observations and modeled precipitation shows that observed precipitation (at low elevations) is less than the monthly averages modeled by Salathé *et al.,* (2014) during spring and summer. Spatial variations observed with Livneh *et al.,* (2013) show large deviations between neighboring cells, especially in comparison to the smoother spatial trends organized with the elevation gradient can be observed from the Salathé *et al.,* (2014) (Figure 5A-F).

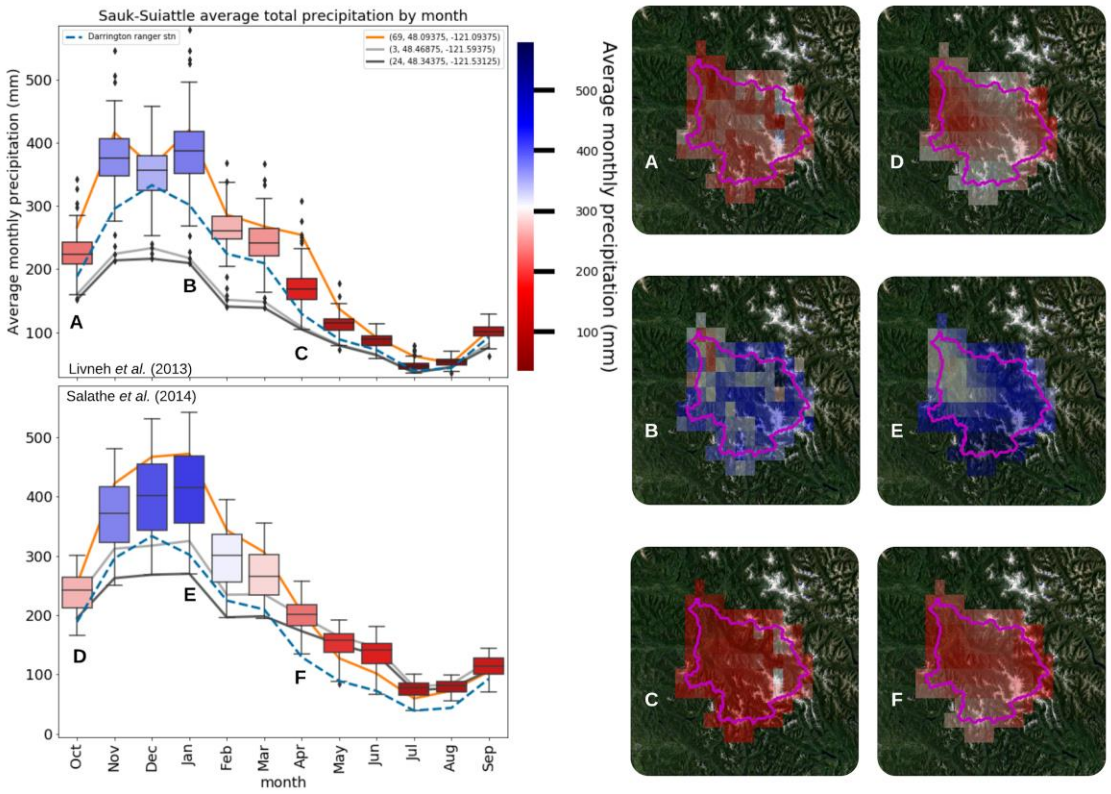

Figure 5: Comparison of the average monthly total precipitation for each gridded cell in the Sauk-Suiattle watershed. The boxplots compare the statistical distribution of the average monthly total precipitation (inches) to the spatial distribution of precipitation in each gridded cell (A-F), using data from Jan 1 1950 through Dec 31 2010. Created using the *renderValueInBoxplot* function, the boxplot colors represent the median value of the gridded cell distributions. Reference trend lines were included to illustrate Sauk-Suiattle's highest elevation gridded cell (#69; orange) and the lowest elevation gridded cells (#3 and #24; light and dark gray), found using the *findCentroidCode* function. The gridded cell distributions are rendered spatially with a basemap using the *renderValueInPoints* function. The spatial distribution of gridded cell values are rendered using the *renderValueInPoints* function for Livneh *et al.* (2013) interpolated meteorology for A) October, B) January, and C) April, compared with to Salathé *et al.* (2014) WRF model outputs for D) October, E) January, and F) April. All maps and boxplots use the same colorbar legend and numerical distribution shown in the top-left.

Monthly temperature statistics were computed for each grid cell using the daily minimum and maximum temperature between Jan 1, 1950 through Dec 31, 2010 (Figure 6). The distribution of mean maximum temperature shows that Livneh *et al.,* (2013) interpolated meteorology has greater variability than Salathé *et al.,* (2014) WRF model outputs. This effect is also observed when comparing the distribution of mean monthly minimum temperature, noting that Livneh *et al.,* (2013) has more extreme hot and cold trends, sometimes up to 5°C difference compared with the Salathé *et al.,* (2014) WRF model outputs. Reference meteorological observations from the Darrington Ranger Station closely resemble the Livneh *et al.,* (2013) interpolated

meteorology. Livneh *et al.,* (2013) values are dependent on source observations clustered around low elevation gridded cells (light and dark gray) with COOP stations; sparse observations limit the performance assessment for high elevation gridded cells. While average daily minimum temperature seems to be comparable, Salathé *et al.,* (2014) predicts colder maximum temperatures for low elevation areas for all months, and warmer temperatures for higher elevation areas (orange line) from November through April.

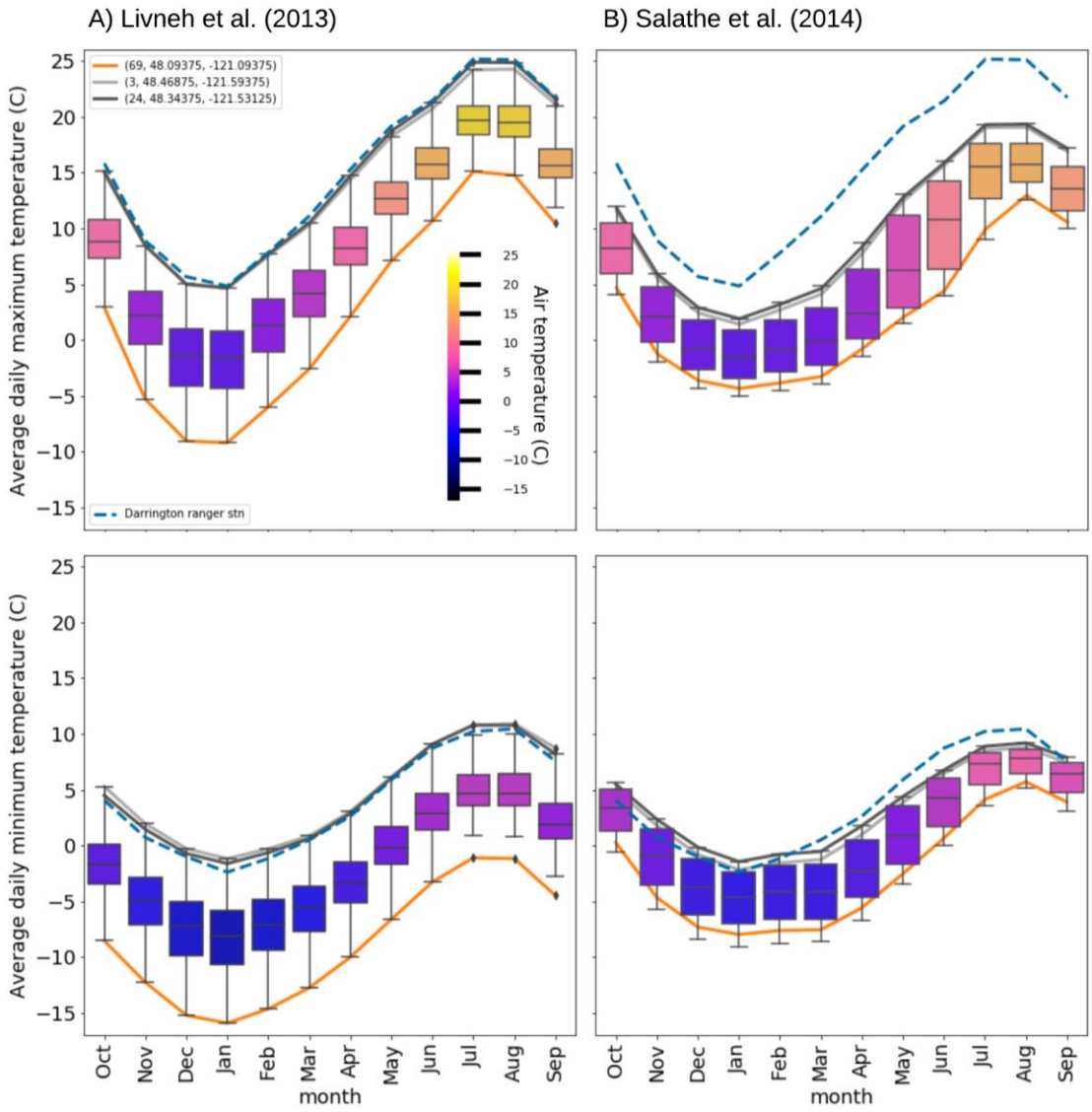

Figure 6: Comparison of monthly mean of daily minimum and maximum temperature. The monthly mean of daily maximum (top) and minimum (bottom) temperatures (in Celsius) were computed for each of the 99 Sauk-Suiattle gridded cells. The boxplots represent the observations from Livneh *et al.,* (2013) meteorology (left) and Salathé *et al.,* (2014) WRF model outputs (right). Reference trend lines were included to represent the highest elevation gridded cell (orange) and the lowest elevation gridded cells (light and dark gray) in Sauk-Suiattle. The field observations (blue dashed line) measured at Darrington Ranger Station (elevation: 167 m) indicates that maximum

daily temperature (top) are more closely represented by Livneh *et al.,* (2013) in the Sauk-Suiattle watershed, while there are no remarkable differences observable for minimum daily temperature (bottom).

3.4. Compute exceedance probabilities

Functions used:

monthlyExceedence_mmday, computeSurfaceArea, cfs_to_mmday

Figure 7 displays the monthly 10% exceedance probability and average (50%) thresholds for unrouted potential runoff using two VIC gridded data products for the Sauk-Suiattle watershed. Not to be confused with approximations like the 10,000-year flood which are based on empirical streamflow values, the probabilities generated by this function are based on empirical unrouted model outputs, which have limited numeric range and interpretation thereof. Each point in the distributions represent the potential runoff threshold at which there is only a 10% chance expectation of exceeding that value in that month. Both gridded data products display a slight increase in November where the muted impact of the highest extreme winter flood events on lower average monthly flows, followed by a high peak flow in June/July (from snowmelt runoff), using the same color scale and axis ranges (Figure 7, right side). The major contribution to potential runoff in any given year is from the Cascade mountain ranges during the snowmelt season (June-July). Although the 10% exceedance probability at the highest elevation are comparable between data products, the trends at the low elevation stations indicate that Livneh *et al.,* (2015) produces more potential runoff between November through April months than Livneh *et al.,* (2013). Livneh *et al.,* (2015) applies a bias correction (using PRISM data as a proxy for observations), which produces more precipitation from winter rainfall season compared to results without a bias correction. A comparison between the two gridded data products illustrates that Livneh *et al.,* (2015) boxplots have approximately 2-3 mm median increase in potential runoff between October to May compared to Livneh *et al.,* (2013) (Figure 7).

The monthly 10% exceedance probability with Sauk River Near Sauk, WA, discharge observations range from 4.2 to 15.3 mm/day across the calendar months. The exceedance probability threshold peaks in two months, November and June, corresponding with fall atmospheric river rainfall-dominated storm events, and early summer snowmelt. The first smaller peak is observed in November, which aligns with both Livneh *et al.,* (2013) and (2015) the boxplot distributions. The second larger peak occurs in June, where the mean decreases though the variance increases in July. The Sauk River Near Sauk gauge is approximate 81

meters above sea level, half of the average elevation in gridded cell 3 where it is located. The spatially averaged (mm/day) routed streamflow dynamics in Sauk River Near Sauk can be expected to fall within the elevation mean of unrouted VIC modeled runoff using the lower resolution 1/16° gridded cells. The observed data at the outlet is provided for context, and the modeled results is provided to demonstrate the spatial variability of the grid cells contributing to the modeled streamflow at the watershed outlet.

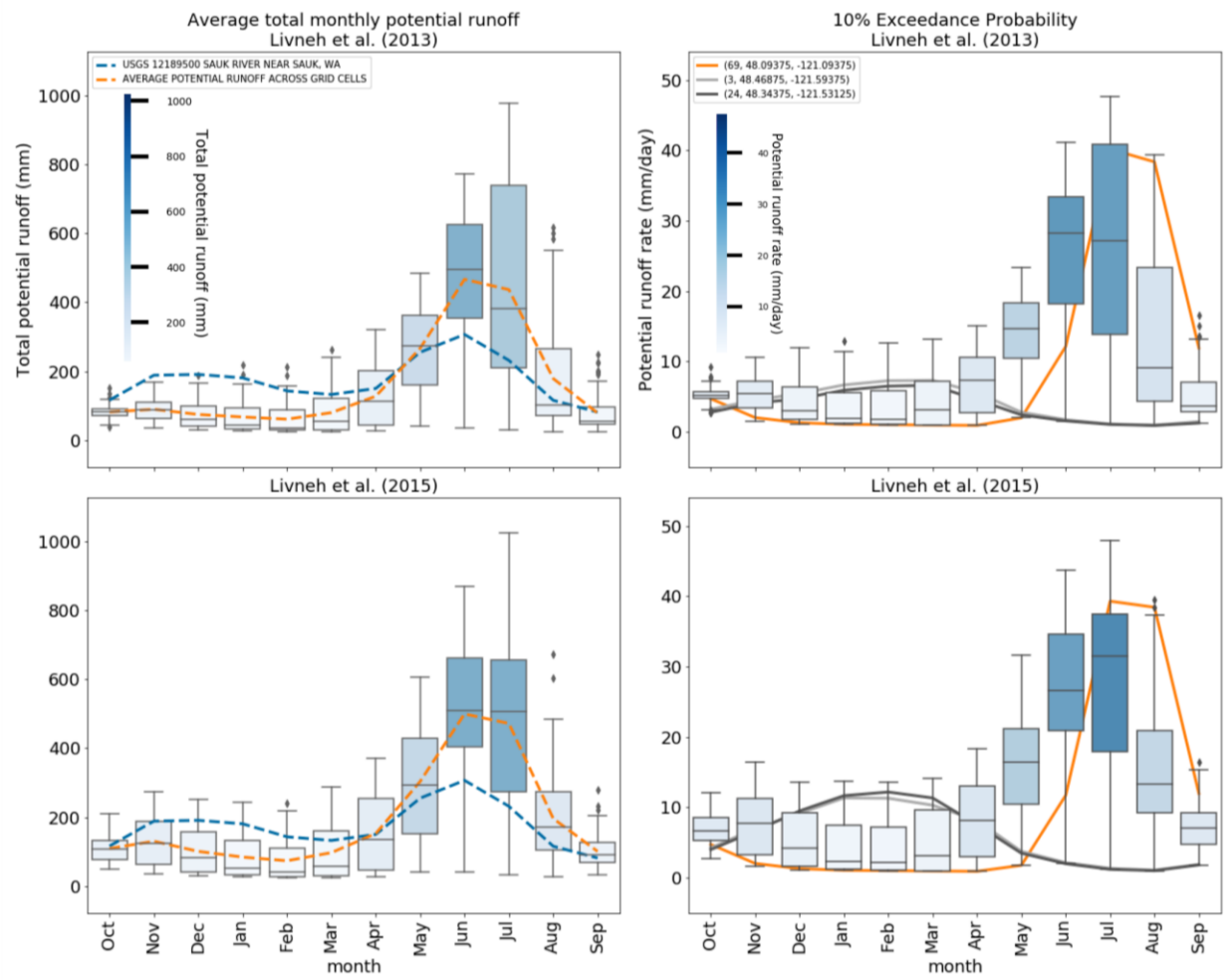

Figure 7: Average total monthly potential runoff (mm) and 10% exceedance probability for each monthly unrouted potential runoff (mm/day) within the Sauk-Suiattle watershed. The boxplots are comprised of 99 gridded cell values for each month. Peak of average total monthly potential runoff (left) occurs in November and June months shown by the observed USGS streamflow discharge (blue dash), and observable by the spatial average of the gridded cell (orange dash). The 10% exceedance probability for each gridded cell (right) is a function of the spatial average of peak flow occurs in November and July. The snowmelt season is the major period for expected runoff for highest elevation gridded cell (orange line), while the rainfall season is the major period contributing to runoff for lowest elevation gridded cells (light and dark gray lines).

4. Discussion

The primary step in the workflow presented here is *treatgeoself*, which enables users to control gridded cell inclusion and exclusion using shapefile-guided data selection. It generates the mapping file to guide distributed computing for data download and data processing; this catalogue allows for machine reading, selection and sorting of available data. The examples use watersheds defined by HUC12 boundaries, but the shape can be user-defined (*e.g.,* census block, legislative boundaries, fish species migration spatial clusters). Geopandas operations enable *treatgeoself* to transform shapefiles of varying spatial projections and to include buffer regions. In the early development stage, *treatgeoself* applied unfiltered spatial intersection with each shape polygon in the shapefile, resulting in slow mapping performance and interpretation difficulty when buffer regions were included. At present, study sites with multiple subpolygons are merged by spatial union into a single MultiPolygon object, simplifying *treatgeoself* into a first-order loop. Intercomparison of gridded data products of different gridded parcel schemas are expected to be enabled and more efficient with the use of the projection alignment and cross-mapping functionalities.

Data download operations are functionalized for distributed computing, but the concurrent queue and transfer rate are limited by the computing resources allocated by the user and the data content provider. Data download was found to be rate-limited to approximately 5 concurrent web requests to the Livneh *et al.,* (2013) web domain. All other gridded data product hosts enabled 10 or more concurrent web requests. A rate-limiter for the number of parallel data retrieval tasks was incorporated into the data download functions, but not for local data processing operations. The rate of data transfer would need to be assessed before OGH could integrate data servers with RESTful API such as ERDDAP, which could expedite mapping and retrieval of gridded data products and metadata [Simons and Mendelssohn, 2012]. Other limits include nuanced issues of data maintenance by the data publisher/provider. For example, during production and testing of workflow and functions, data products mentioned in Livneh *et al.,* (2013) were migrated to a new web domain, resulting in misdirected requests. Annotations and data retrieval functions may need updating over time.

We qualitatively described differences between two gridded data products of the same empirically estimated statistical exceedance probability approach. The VIC-modeled gridded

data products originate from unrouted flow modeling. In contrast to empirically estimated 10,000-year flood at-stream gauges, empirically estimated exceedance probability for grid cells without flow routing may be limited in interpretations to potential runoff. It is unclear how the different model simulations may affect these interpretations. These concerns regarding model comparison would merit further research and development of functionalities for in-watershed stream gauge selection and quantitative determinations for the goodness-of-fit between routed and unrouted modeled values relative to those estimated from at-stream observations.

While we designed OGH specifically for users who primarily import ASCII-formatted files into hydrologic and earth surface model software e.g. Landlab [Hobley et al., 2017], a noteworthy limitation of using ASCII file format is that as NetCDF adoption is increasing as a data standard, ASCII time-series may not be available for newer gridded data products. This limitation is addressed using the proposed minimum information criteria and having initial criteria for conducting gridded data product intercomparisons. NetCDF files are embedded with metadata, while ASCII files are unannotated. To inform the structure and use of the ASCII files, the proposed minimum information criteria serves as a road map for locating gridded data product files and considers the schema of the file organization and the features within each file. Gridded data products published by Livneh *et al.,* (2013) partitioned files by spatial bounding box subfolders denoted by the file prefix, West, East, South, and North cardinal limits. For Livneh *et al.,* (2013), *scrape_domain* and *mapToBlock* functions were designed to abstract the bounding boxes then decide the subfolder identity by spatial intersection. The spatial bounding box for gridded cells within British Columbia, Canada did not follow this folder naming structure; thus, a separate annotation was provided for the spatial boundary in British Columbia, Canada.

Among the annotated ASCII gridded data products, we've observed a variety of file organizations; different gridded cell schemas, spatial resolution, or NetCDF file organizations may be adaptable. Retrieval and data management of NetCDF files in cloud computing environments would benefit from further design assessments, as it is not yet clear how to conduct or evaluate NetCDF-to-ASCII intercomparison without *a priori* format preferences that may result in information loss. In addition, the development of a user-centered reference of controlled vocabulary would improve the usefulness and adoption of a minimum information criteria that can be used across data formats. These may help adapt climate and water resource information for researching interdisciplinary questions with other data products such as Air Quality or Population data sets [Wohlstadter *et al.,* 2016; Lloyd *et al.,* 2017]. For use by

researchers who are not hydrometeorology analyst, NetCDF files contain data outside the study area extent; 1D ASCII time-series files may be the preferred format for small study areas (1-100 km^2).

We tested and developed the examples using HydroShare for the computing and data sharing environment. An important benefit of HydroShare is that it hosts a REST API that enables data migration and the creation of new shareable data objects. Additionally, as a community repository for hydrologic science, FAIR publication of hydrologic data sets and software execution with reproducible workflows is demonstrated with the use-cases developed in this work. OGH operations are technically independent of HydroShare, and minor changes would allow the code to operate in other similar computing and data sharing environments such as local servers, cloud servers (Amazon AWS, Microsoft Azure), and dockerized virtual environments with a Jupyter instance (data.world, DataOne, PanGeo, ESIPhub).

5. Conclusions

OGH is a toolkit that makes download and processing of large climate datasets more efficient by leveraging distributed computing for watershed scale research and intercomparison of ASCII gridded data products, which extends climate modeling products to represent otherwise sparsely observed parts of the landscape. The mapping file output is the key data management tool, which catalogs the watershed gridded cells and downloaded files as a lens across gridded data products. Along with the proposed minimum information criteria to annotate ASCII gridded data products, these data management tools enable multiprocessing and dask-distributed operations comparable to the efficiency of Xarray for NetCDF gridded data products. This metadata component improves the standardization of gridded hydrometeorology products published for use by third-party researchers and scientists. The dictionary of analytical data frames is a key data management device that enables key-value pair retrieval and exporting of summary outputs. To address user needs for exploratory data analysis and visual control, various data frames were rendered into different geographic and temporal modes of human-readable visual inspection. Overall, OGH is equipped with metadata framework and workflow that makes it a useful introduction and training tool for watershed studies using gridded data products and ASCII time-series data sets. The data summary capabilities increase the efficiency of comparing multiple gridded hydrometeorology products without discontinuous use of different software. OGH and the four use-cases demonstrated are available for interactive use on

HydroShare (https://www.hydroshare.org/resource/87dc5742cf164126a11ff45c3307fd9d) and also available for open development from the University of Washington Freshwater Initiative Observatory repository: https://github.com/Freshwater-Initiative/Observatory).

6. Abbreviations

- Observatory for Gridded Hydrometeorology (OGH)
- conterminous United States (CONUS)
- Weather Research Forecasting (WRF)
- Findable, Accessible, Interoperable, and Reusable (FAIR)
- HydroShare REST API Python library (hs_restclient)
- Variable Infiltration Capacity (VIC)

7. Acknowledgements

This work benefited from the contributions from University of Washington (UW) Watershed Dynamics group who helped test and develop OGH, and members UW eScience Institute that helped with developing and using the Python toolkits. This project was supported in part by National Science Foundation HydroShare Cyberinfrastructure project (ACI 1148453), SI2:SSI Landlab project (ACI-1450412), National Center for Advancing Translational Sciences Institute for Translational Health Sciences grant (UL1-TR002319) and Clinical and Translational Sciences Award Program National Center for Data to Health (Grant U24TR002306), UW Civil & Environmental Engineering Department in collaboration with researchers and scientists of the Sauk-Suiattle Indian Tribe and the Skagit Climate Consortium. The project uses the HydroShare platform, which is supported by the Consortium of Universities for the Advancement of Hydrologic Sciences, Inc. (CUAHSI), a research organization supported by NSF cooperative agreement (EAR 1338606). Special thanks to early reviewers who contributed to the editing of this paper: Drs. Dan Ames, Emilio Mayorga, and Nicoleta Cristea.

8. References

Baxter G, Sommerville I. Socio-technical systems: From design methods to systems engineering. Interacting with computers. 2011 Jan;23(1):4-17.

Castronova AM, Brazil L, Seul M. Cloud-based Jupyter Notebooks for Water Data Analysis. In AGU Fall Meeting Abstracts 2017 Dec.

Cutter SL, Ash KD, Emrich CT. The geographies of community disaster resilience. Global environmental change. 2014 Nov 30;29:65-77. https://doi.org/10.1016/j.gloenvcha.2014.08.005

Daly C, Neilson RP, Phillips DL. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. Journal of applied meteorology. 1994 Feb;33(2):140-58.

Devi KR, Sen AM, Hemachandran K. A working framework for the user-centered design approach and a survey of the available methods. International Journal of Scientific and Research Publications. 2012 Apr;2(4).

Gampe D, Ludwig R. Evaluation of Gridded Precipitation Data Products for Hydrological Applications in Complex Topography. Hydrology. 2017 Nov 16;4(4):53. https://doi.org/10.3390/hydrology4040053

Gardner MA, Morton CG, Huntington JL, Niswonger RG, Henson WR. Input data processing tools for the integrated hydrologic model GSFLOW. Environmental Modelling & Software. 2018 Aug 9. https://doi.org/10.1016/j.envsoft.2018.07.020

Heard J, Tarboton D, Idaszak R, Horsburgh J, Ames D, Bedig A, Castronova A, Couch A. An Architectural Overview Of HydroShare, A Next-Generation Hydrologic Information System. CUNY Academic Works. 2014. https://academicworks.cuny.edu/cc_conf_hic/311

Henn B, Newman AJ, Livneh B, Daly C, Lundquist JD. An assessment of differences in gridded precipitation datasets in complex terrain. Journal of Hydrology. 2018 Jan 1;556:1205-19. https://doi.org/10.1016/j.jhydrol.2017.03.008.

Hobley, D.E.J., Adams, J.M., Nudurupati, S.S., Hutton, E.W.H, Gasparini, N.M., Istanbulluoglu, E., and Tucker, G.E., Creative computing with Landlab: an open-source toolkit for building, coupling, and exploring two-dimensional numerical models of Earth-surface dynamics. Earth Surface Dynamics, 2017, doi:10.5194/esurf-5-21-2017.

Horsburgh JS, Morsy MM, Castronova AM, Goodall JL, Gan T, Yi H, Stealey MJ, Tarboton DG. HydroShare: Sharing diverse environmental data types and models as social objects with application to the hydrology domain. JAWRA Journal of the American Water Resources Association. 2016 Aug 1;52(4):873-89. http://dx.doi.org/10.1111/1752-1688.12363.

hs_restclient: HydroShare REST API Python client library. Release v1.2.12. http://hs-restclient.readthedocs.io/en/latest/

Jordahl K. GeoPandas: Python tools for geographic data. 2014. https://github.com/geopandas/geopandas.

Kadlec J, StClair B, Ames DP, Gill RA. WaterML R package for managing ecological experiment data on a CUAHSI HydroServer. Ecological informatics. 2015 Jul 31;28:19-28. https://doi.org/10.1016/j.ecoinf.2015.05.002

Konrad, C.P., and Voss, F.D., 2012, Analysis of streamflow-gaging network for monitoring stormwater in small streams in the Puget Sound Basin, Washington: U.S. Geological Survey Scientific Investigations Report 2012–5020, 16 p.

King AD, Alexander LV, Donat MG. The efficacy of using gridded data to examine extreme rainfall characteristics: a case study for Australia. International Journal of Climatology. 2013 Aug 1;33(10):2376-87. https://doi.org/10.1002/joc.3588

Ledesma JL, Futter MN. Gridded climate data products are an alternative to instrumental measurements as inputs to rainfall–runoff models. Hydrological Processes. 2017 Aug 30;31(18):3283-93. https://doi.org/10.1002/hyp.11269

Liang X, Lettenmaier DP, Wood EF, Burges SJ. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. Journal of Geophysical Research: Atmospheres. 1994 Jul 20;99(D7):14415-28. https://doi.org/10.1029/94JD00483

Livneh B, Rosenberg EA, Lin C, Nijssen B, Mishra V, Andreadis KM, Maurer EP, Lettenmaier DP. A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions. Journal of Climate. 2013 Dec;26(23):9384-92. https://doi.org/10.1175/JCLI-D-12-00508.1

Livneh B, Bohn TJ, Pierce DW, Munoz-Arriola F, Nijssen B, Vose R, Cayan DR, Brekke L. A spatially comprehensive, hydrometeorological data set for Mexico, the US, and Southern Canada 1950–2013. Scientific data. 2015 Aug 18;2:150042. https://doi.org/10.1038/sdata.2015.42

Livneh B, Rajagopalan B. Development of a gridded meteorological dataset over Java island, Indonesia 1985–2014. Scientific data. 2017 May 23;4:170072. https://doi.org/10.1038/sdata.2017.72

Lloyd CT, Sorichetta A, Tatem AJ. High resolution global gridded data for use in population studies. Scientific data. 2017 Jan 31;4:170001. https://doi.org/10.1038/sdata.2017.1

Maurer EP, Wood AW, Adam JC, Lettenmaier DP, Nijssen B. A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. Journal of climate. 2002 Nov;15(22):3237-51. https://doi.org/10.1175/1520-0442(2002)015<3237:ALTHBD>2.0.CO;2

McKinney W. pandas: a foundational Python library for data analysis and statistics. Python for High Performance and Scientific Computing. 2011 Jun:1-9.

Mons B, Neylon C, Velterop J, Dumontier M, da Silva Santos LO, Wilkinson MD. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use. 2017 Jan 1;37(1):49-56. https://doi.org/10.3233/ISU-170824

Mote, P.W. & Salathé, E.P. Climatic Change (2010) 102: 29. https://doi.org/10.1007/s10584-010-9848-z.

Read JS, Walker JI, Appling AP, Blodgett DL, Read EK, Winslow LA. geoknife: reproducible web-processing of large gridded datasets. Ecography. 2016 Apr;39(4):354-60.

Rocklin M. Dask: Parallel computation with blocked algorithms and task scheduling. In Proceedings of the 14th Python in Science Conference 2015 (No. 130-136).

Salathé Jr EP, Hamlet AF, Mass CF, Lee SY, Stumbaugh M, Steed R. Estimates of twenty-first-century flood risk in the Pacific Northwest based on regional climate model simulations. Journal of Hydrometeorology. 2014 Oct;15(5):1881-99. https://doi.org/10.1175/JHM-D-13-0137.1

Sean Gillies. Fiona is OGR's neat, nimble, no-nonsense API. Toblerity.org. 2011. https://github.com/Toblerity/Fiona.

Sean Gillies. Shapely: manipulation and analysis of geometric objects. Toblerity.org. 2007. https://github.com/Toblerity/Shapely.

Simons RA, Mendelssohn R. ERDDAP-A Brokering Data Server for Gridded and Tabular Datasets. In AGU Fall Meeting Abstracts 2012 Dec.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3. http://doi.org/10.1038/sdata.2016.18

Wohlstadter M, Shoaib L, Posey J, Welsh J, Fishman J. A Python toolkit for visualizing greenhouse gas emissions at sub-county scales. Environmental Modelling & Software. 2016 Sep 1;83:237-44. https://doi.org/10.1016/j.envsoft.2016.05.016

Vogel RM, Matalas NC, England Jr JF, Castellarin A. An assessment of exceedance probabilities of envelope curves. Water resources research. 2007 Jul;43(7). https://doi.org/10.1029/2006WR005586

8.1. Data References

Bandaragoda, C. (2017). Upper Rio Salado Watershed Boundary - Cibola National Forest, HydroShare, http://www.hydroshare.org/resource/5c041d95ceb64dce8eb85d2a7db88ed7 [Last accessed 13 Aug 2018]

Bandaragoda, C. (2017). Sauk-Suiattle HUC12 17110006, HydroShare, http://www.hydroshare.org/resource/c532e0578e974201a0bc40a37ef2d284 [Last accessed 13 Aug 2018]Beveridge, C., C. Bandaragoda, J. Phuong (2017). Elwha Observatory- Public, HydroShare, http://www.hydroshare.org/resource/1de72928f573433290f6c8bb393523df [Last accessed 13 Aug 2018]

Livneh, B. (2017). Gridded climatology locations (1/16th degree): North American extent, HydroShare, http://www.hydroshare.org/resource/ef2d82bf960144b4bfb1bae6242bcc7f [Last accessed 13 Aug 2018]

National Water Information System. Daily streamflow discharge from the Sauk River Near Sauk, WA (USGS-12189500) from Jan 1 1950 through Dec 31 2011. https://waterdata.usgs.gov/nwis/dv?cb_00060=on&format=rdb&site_no=12189500&referred_module=sw&period=&begin_date=1950-01-01&end_date=2011-12-31. [Last accessed 13 Aug 2018].

Western Regional Climate Center - Darrington Ranger Station monthly climate summary. https://wrcc.dri.edu/cgi-bin/cliMAIN.pl?wa1992. [Last accessed 13 Aug 2018].

9. Figures and Tables:

1. Figure 1: The general workflow for OGH in cloud-computing environments
2. Table 1: Summary of seven daily, 1/16° gridded data product
3. Table 2: Minimum annotation criteria for gridded data products
4. Figure 2: Scenario use-cases for OGH operations in cloud environments
5. Figure 3: Spatial-temporal calculations (total sum and average)
6. Figure 4. Aerial view of watersheds and gridded cells
7. Figure 5: Comparison of spatial-temporal precipitation in Sauk-Suiattle
8. Figure 6: Comparison of annual mean of daily minimum and maximum temperature
9. Figure 7: Average total monthly potential runoff and 10% exceedance probability for each monthly unrouted potential runoff within Sauk-Suiattle.

10. Appendices:

Supplementary Table 1. Description of user centered design principles used (following framework in Devi et al., 2012).

Key Principles	Development Description	Users Involved
1- The active involvement of users and clear understanding of user and task requirements.	We worked with the Landlab project, UW Watershed Dynamics Lab, and other UW students to understand which datasets they needed.	Academic faculty (1) PhD students (7) Other students (10)
2- An appropriate allocation of function between user and system.	Limits were set based on using a JupyterHub server for compute resources, and available software.	Science domain users (3) Developers (4)
3- Iteration of design solutions.	Switching from manual processes with commercial software to open source software processes required more than 18 months of iterative development.	Team from 1 above.
4- Multi-disciplinary design teams.	Testing and presentation of the methods occurred in research groups in both College of Engineering and School of Medicine research lab groups.	Civil & Environmental Engineering (10), Biomedical informatics (1), Data Science (1)