1-9-2020

# HydroDS: Data Services in Support of Physically Based, Distributed Hydrological Models

Tseganeh Zekiewos Gichamo
*Utah State University*

Nazmus S. Sazib
*Science Application International Corporation*

David G. Tarboton
*Utah State University*

Pabitra Dash
*Utah State University*



## Recommended Citation

Gichamo, T. Z., N. S. Sazib, D. G. Tarboton and P. Dash, (2020), "HydroDS: Data Services in Support of Physically Based, Distributed Hydrological Models," Environmental Modelling & Software: 104623, https://doi.org/10.1016/j.envsoft.2020.104623

# HydroDS: Data Services in Support of Physically Based, Distributed Hydrological Models

Tseganeh Z. Gichamo [a] zacctsega@gmail.com, Nazmus S. Sazib [b] sazibap25@gmail.com, David G. Tarboton [a*] david.tarboton@usu.edu, Pabitra Dash [a] pabitra.dash@usu.edu

[a] Utah Water Research Laboratory, 8200 Old Main Hill, Utah State University, Logan, UT 84322-8200, USA.

[b] Science Application International Corporation, 8 Riverview Ct 302, Laurel, MD 20707, USA.

* Corresponding author.

## Abstract

Physically based distributed hydrologic models require geospatial and time-series data that take considerable time and effort to process into model inputs. Tools that automate and speed up input processing facilitate the application of these models. In this study, we developed a set of web-based data services called HydroDS to provide hydrologic data processing 'software as a service.' HydroDS provides functions for processing watershed, terrain, canopy, climate, and soil data. The services are accessed through a Python client library that facilitates developing simple but effective data processing workflows with Python. Evaluations of HydroDS by setting up the Utah Energy Balance and TOPNET models for multiple headwater watersheds in the Colorado River basin show that HydroDS reduces input preparation time compared to manual processing. It also reduces requirements for software installation and maintenance by the user, and the Python workflows enhance reproducibility of hydrologic data processing and tracking of provenance.

Keywords—HydroDS, web-based data services, distributed hydrologic modeling, Geographic Information Systems, Hydrologic Data Cyberinfrastructure.

---

This is the accepted version of the following article
Gichamo, T. Z., N. S. Sazib, D. G. Tarboton and P. Dash, (2020), "HydroDS: Data Services in Support of Physically Based, Distributed Hydrological Models," Environmental Modelling & Software: 104623, https://doi.org/10.1016/j.envsoft.2020.104623

## Highlights

- Web-based data services were developed for preparation of input data to selected distributed hydrologic models.
- Services are accessed through a Python client library and facilitate use in hydrologic data preprocessing workflows.
- Services reduce time for hydrologic model input preparation, enhance reproducibility, and enable tracking of data provenance.

## Software Availability

Program name: HydroDS

Description: A set of web-based, hydrologic data services for preparation of input data for selected physically based, distributed (grid or subwatershed model elements) hydrologic models. HydroDS comprises Python modules for watershed analysis, terrain and land cover data processing, climate data access and processing, and generating soil properties data. Individual service functions, accessed through a Python client library, may be chained together to form a Python workflow to perform a set of related tasks.

Platform: CentOS Linux for hosting the web services; Accessed from any platform.

License: 3 clause BSD license (open source)

Source code: https://github.com/CI-WATER/Hydro-DS/

Documentation:
https://github.com/CI-WATER/Hydro-DS/wiki/HydroDS-Web-API-Description

Developers: Tseganeh Z. Gichamo, Nazmus S. Sazib, David G. Tarboton, Pabitra Dash.

## 1. Introduction

Physically based, distributed hydrologic models are used for simulation of the hydrologic cycle to help answer questions related to water resource availability and quality, to assess the effect of changes in climate or land cover, and support water resources management, along with many other applications. An important challenge associated with the application of physically based, distributed hydrological models is

that they require more input data than their conceptual, often lumped, counterparts. While the rationale for high resolution physically based models is that better results can be achieved through detailed process representation, obtaining the extensive set of input data required by these models is a critical challenge. Leonard and Duffy (2013) call this set of input data "Essential Terrestrial Variables" (ETV). Obtaining ETV's in a format organized for use in distributed models is a significant bottleneck in distributed hydrologic modelling. The ability to configure and populate distributed models with data could enhance or hinder their use.

Prior work with regard to hydrologic data availability has focused on the task of enhancing access to data from different providers through web services (Horsburgh et al., 2009; Tarboton et al., 2009b; Ames et al., 2012) using standardized data formats (Taylor, 2012; Almoradie et al., 2013a). However, the data obtained remain the raw data provided by the sources and generally require further processing to generate suitable inputs to hydrological models, in terms of the data structure, file format, and variable type required by a model (e.g., relative humidity rather than specific humidity). As a result, in our experience, researchers and practitioners implementing hydrologic models spend a significant amount of time transforming the data available into the formats needed by models(Jones et al., 2013)(Jones et al., 2013)(Jones et al., 2013)(Jones et al., 2013)(Jones et al., 2013)(Jones et al., 2013)(Jones et al., 2013)(Jones et al., 2013)(Jones et al., 2013)(Jones et al., 2013)(Jones et al., 2013)(Jones et al., 2013)(Jones et al., 2013).

The data pre-processing tools currently available are generally desktop based and often limited by their customization to specific hydrologic models (e.g., Kumar et al., 2009). The increasing availability of cyberinfrastructure resources provides an

opportunity to extend such data pre-processing ability beyond the desktop environment (Wang et al., 2013) and adopt the paradigm of 'software as a service.' Developing data processing tools as web-based services will help to enhance access to these tools for users without necessarily requiring them to be a Cyber expert (Wright et al., 2013). Web services that can be accessed by multiple users facilitate better collaborative problem solving (Nyerges et al., 2013; Wang, 2010). In addition, they encourage the use of standardized data formats (e.g., WaterML and NetCDF) by multiple models.

In this paper, we introduce a set of web-based, hydrological data processing services called HydroDS. HydroDS provides a number of data processing functionalities including watershed delineation, terrain processing, estimation of canopy variables, retrieval and processing of weather forcing data, procuring soil data and generating soil properties. Data are stored and shared in three widely used data formats: GeoTiff raster, Shapefile, and multi-dimensional NetCDF. The data services are comprised of functions that can be used independently or form workflows that integrate a number of related tasks. The services are accessed through a Python client library that facilitates developing simple but effective data processing workflows with Python, providing access to data processing tools from an accessible and relatively easy to use programming environment. Data processed by HydroDS can be transferred to HydroShare, a platform for sharing of hydrologic data and models (Tarboton et al., 2014a). The objective here was to provide the means to setup Python workflows for preparation of input data for distributed hydrologic models. The services we developed support the Utah Energy Balance (UEB) snowmelt model (Tarboton et al., 1995) and TOPNET hydrologic model (Bandaragoda et al., 2004).

In the next section, we provide background information on prior work dealing with data access and processing for hydrologic modeling and the need for web-based data services that motivated this work. In Section 3, we report the required functionality, design, and implementation of HydroDS. In Section 4, we evaluate the data services using a case study of setting up instances of UEB and TOPNET models for multiple headwater watersheds in the Colorado River basin. Results and discussion are given in Section 5, followed by summary and conclusions in Section 6.

## 2. Background

### 2.1 *Input Data Processing for Hydrologic Models*

Providing access to hydrological data from different repositories through web services has been the focus of the Consortium of Universities for the Advancement of Hydrologic Science, Inc. - Hydrologic Information System (CUAHSI-HIS) (Horsburgh et al., 2009; Tarboton et al., 2009b). CUAHSI-HIS provides software tools for publishing and retrieving time series data through standardized web services in an XML format called WaterML (Tarboton et al., 2011; Valentine et al., 2012; 2007; Beran et al., 2009). WaterML2 was later developed as an Open Geospatial Consortium (OGC) standard for hydrologic time series data representation and exchange across multiple information systems (Taylor, 2012). Standardized web services and protocols facilitate interoperability between different data service providers and consumers (clients) for easy access to and retrieval of data.

Client applications can search for and download data made available through the CUAHSI-HIS data services. HydroDesktop, the CUAHSI-HIS data access client (Ames et al., 2012), provided an early ‘one-stop shopping’ platform to hydrologists by enabling

map based selection of a watershed (or the extent of the domain of interest) and data download, extraction, and analysis. This desktop functionality has now been replaced by the CUAHSI data client web tool (http://data.cuahsi.org/) for CUAHSI HIS data selection and extraction. Agencies such the U.S. Geological Survey (USGS: http://waterservices.usgs.gov/) and National Oceanic and Atmospheric Administration - National Centers for Environmental Information (NOAA- NCEI: http://www.ncdc.noaa.gov/cdo-web/webservices), and other data and model service providers have also made data from their repositories accessible using web services and data standards such as WaterML and other OGC web service standards (Almoradie et al., 2013b). These systems help reduce the time spent by researchers searching for and downloading data.

While availability of hydrological data through web services from sources such as CUAHSI-HIS, USGS, NOAA, or other organizations is growing, pre-processing is often needed to generate suitable inputs to hydrological models. In addition, CUAHSI-HIS compliant data services are currently limited to time series data at fixed geographic locations (e.g., points) using the Observations Data Model (ODM) (Horsburgh et al., 2008); no support is provided in CUAHSI-HIS services for multi-dimensional space-time data such as those stored in Network Common Data Form (NetCDF) data format (Rew et al., 2014). Hence, part of the data pre-processing tasks for distributed hydrological models involves organizing data in the input format suitable for the specific model (often arrays of space-time data).

Input data pre-processing often starts with geospatial analyses, including watershed delineation, stream network generation, and specification of modeling units

such as Hydrologic Response Units (HRU) or structured or unstructured grids of required spatial resolution. Then, input variables based on the watershed terrain, land cover characteristics, and climate forcing are mapped to the modeling units (Carlson et al., 2014). This mapping of continuous or discrete values to model units may require aggregation or interpolation in both space and time.

Extraction of hydrological variables from digital elevation models (e.g., terrain slope, aspect, topographic wetness index) is also part of the pre-processing required to develop model inputs. In addition, some model parameters need to be generated (or estimated) from observations. For example, land cover variables such as canopy indices have to be derived based on land cover type maps or from remote sensing images, and friction coefficients have to be estimated from the vegetation and geomorphological information of river reaches. These data pre-processing tasks can take a significant portion of the hydrological modeler's time and effort, and data pre-processor tools have been shown to considerably reduce the time required for model scenario setup and execution (Berry et al., 2014). An additional benefit is reproducibility, and the opportunity to support best of practice pre-processing methods, rather than expedient methods that may be selected by a user preparing model inputs manually using general purpose tools available to them.

## 2.2 *Web-based Data and Modelling Services*

At present, many geospatial data analyses are carried out using desktop-based GIS tools. Some of these GIS tools are stand-alone software products such as the ArcGIS software suite from ESRI (http://www.esri.com/) or the open source QGIS (http://www.qgis.org/en/site/) and GRASS (http://grass.osgeo.org/) software. Others are integrated with the hydrologic models they prepare inputs for. There is commercial and

open source modeling software that supports input data pre-processing as an integral part of hydrologic modeling. One example of commercial software is the MIKE SHE model's GIS-based graphical user interface and GIS database from DHI (http://www.dhigroup.com/). An example of open source model data processing tools is PIHMgis (Bhatt et al., 2008; Kumar et al., 2009; Bhatt et al., 2014), in which a GIS framework for model input pre-processing and input and output visualization is tightly coupled to the Penn State Integrated Hydrologic Model (PIHM - http://www.pihm.psu.edu/).

With the increasing availability of Cyberinfrastructure, there is an opportunity to extend model input data pre-processing tools to web-based services. Such web-based services could build on or provide additional services to the general-purpose geospatial and hydrologic data services such as CyberGIS, ArcGIS Online, and HydroTerre. CyberGIS (https://cybergis.illinois.edu) is a web-based approach to the delivery of GIS functionality as data and software services (Wang et al., 2013; Wang, 2010; Wright et al., 2013). CyberGIS supports large scale, data intensive modelling problems with spatial analysis tools that require more than just a few processing cores. ArcGIS Online (http://arcgis.com/) is an extension of ArcGIS to a web-based service that enables the rapid growth of available content, provides enhanced capability through being able to access cloud based resources, sharing and collaboration around geospatial developments, etc. (Wright et al., 2013). HydroTerre is a web-based hydrologic model data and visualization service (http://www.hydroterre.psu.edu/HydroTerre/Help/Ethos.aspx) that has made available about 200 TB of "Essential Terrestrial Variables (ETVs)," including elevation, soils, geology, land cover, precipitation and atmospheric conditions, sub-

watersheds and National Hydrography Dataset (NHD) stream reaches. Data are indexed by USGS NHD Hydrological Unit Code level-12 (HUC-12) sub-watersheds and can be downloaded to support detailed hydrologic modelling using PIHM or other models (Leonard and Duffy, 2013, 2014, 2016). HydroTerre data, model, and visualization workflows capture provenance and enable reproducibility (Leonard and Duffy, 2016).

There are also developments such as EcoHydrolib and RHESSysWorkflows (Miles and Band, 2015) and WaterHUB (http://water-hub.org/) that deal with specific models. EcoHydrolib was developed as a set of general data access and processing libraries that form building blocks for RHESSysWorkflows, the input data preparation workflows to generate instances of the Regional HydroEcological Simulation System (RHESSys) model (Miles and Band, 2015). RHESSysWorkflows preserve metadata that enable reproducibility (Miles, 2014). A web-based modeling service is provided by WaterHUB, which allows parameterized SWAT (Soil Water Assessment Tool) models and their input data to be uploaded, run on HPC resources, and shared among users (Merwade et al., 2012). This service also provides a web-based data preparation and modeling environment, access to existing models, their input/output datasets, and a mechanism to perform simultaneous simulations (Rajib et al., 2016).

The web-based technologies underlying these services enable taking advantage of high performance computation resources (provided they are available), distributed data storage facilities, analysis tools from multiple service/tool providers to deal with 'spatial big data' (Evans et al., 2013), and collaboration between researchers (possibly) remotely located from each other. From a user/client point of view, spatial analysis capabilities are readily accessible through the World Wide Web without requiring any local software

installation. This eliminates data size limitation of PCs, the need to install software, and operating system (platform) dependence. A web-based development environment facilitates better collaborative problem solving (Nyerges et al., 2013; Wang, 2010), eases access to analytic tools (e.g., geospatial analyses) for non-experts (Wright et al., 2013), and enables implementation of 'science gateway' functionalities that provide access to HPC centers (Wilkins-Diehr et al., 2008).

An example web-based development environment, integrating multiple of the above functionalities is HydroShare (http://www.hydroshare.org/). HydroShare is a collaborative environment for sharing hydrologic data and models taking advantage of modern information communication technology and Cyberinfrastructure. HydroShare supports the capability for users to store their work in the hydrologically oriented resource formats including time series, geographic features and rasters, and model programs and instances. HydroShare resources created by one user may be shared with others, and HydroShare's web service application programming interface (API) enables programmatic access to create and/or work directly with resources stored in the system (Tarboton et al., 2014a; 2014b; Horsburgh et al., 2016; Morsy et al., 2017).

## 3. Development of A set of Web-based Hydrologic Data Processing Services (HydroDS)

### *3.1 Required Functionality*

The first step in the development of HydroDS was identifying the functionality of data services required to support the input pre-processing for physically based gridded models commonly used in surface water hydrology (i.e., ETVs). This was influenced by input data pre-processing tasks, shown in Figures 1 and 2, for the UEB and TOPNET

models respectively. These figures represent workflows of tasks that are required to be executed to get the inputs for the UEB and TOPNET models for a given watershed and specific modeling period. Both these models require input data characterizing terrain slope and aspect, land cover / canopy type information, and weather forcing inputs of precipitation, temperature, relative humidity, wind speed, and solar radiation. In addition, the TOPNET model requires soil data, topographic wetness distribution data, and distance to stream distribution for runoff routing. In both cases, the input data preparation starts with the definition of the modeling domain that often requires watershed delineation based on digital elevation model (DEM) processing.

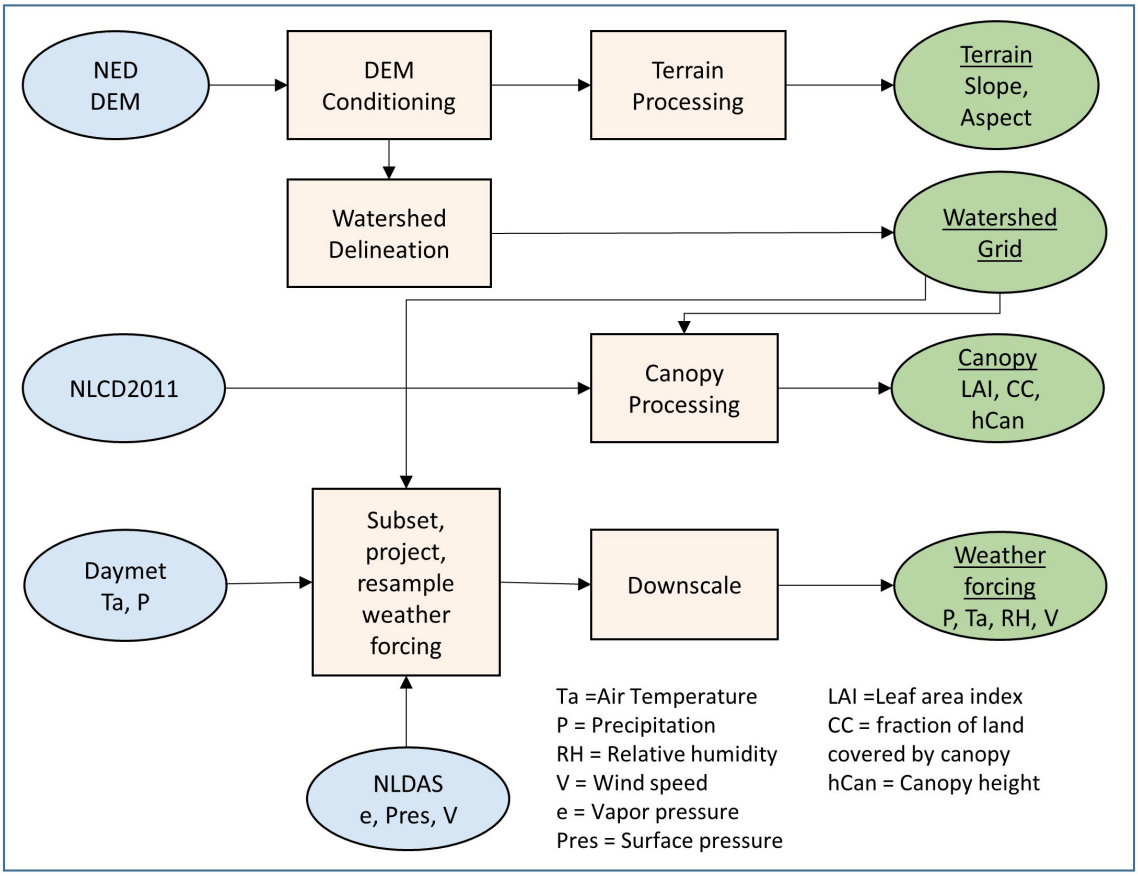


Figure1. Workflow for the Utah Energy Balance snowmelt model (UEB) input

preparation.

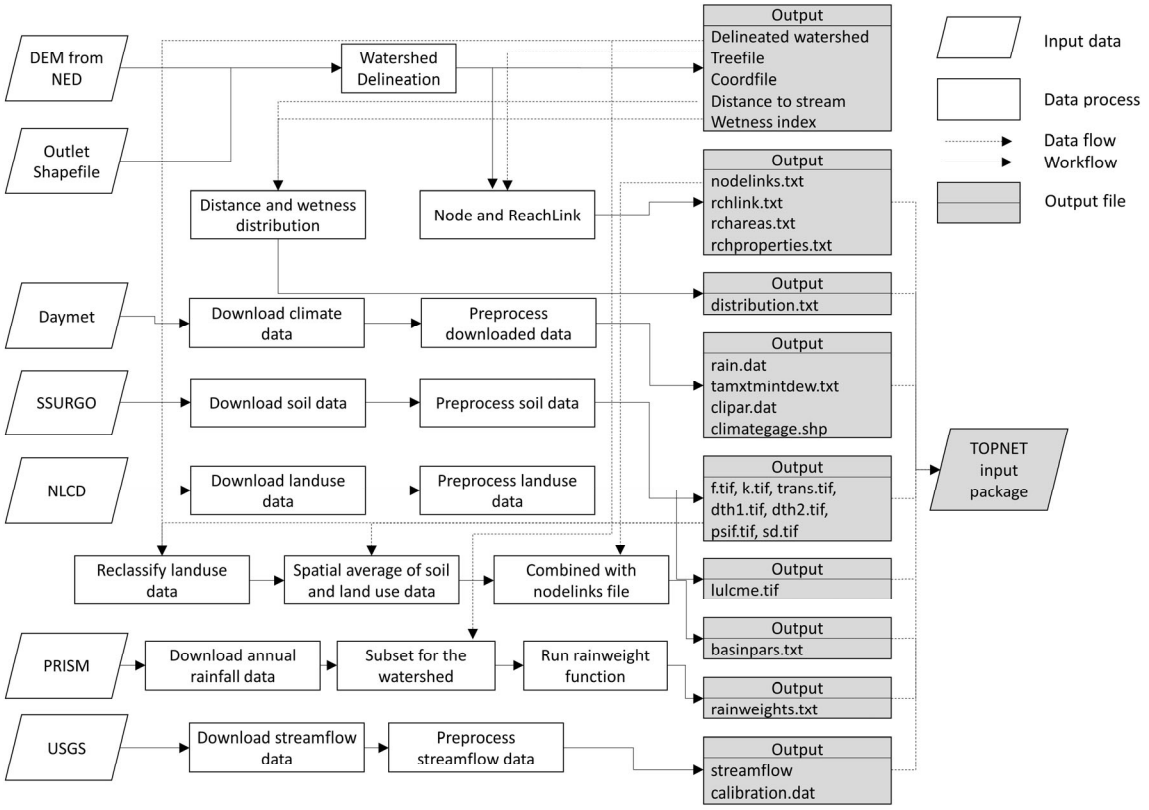


Figure 2. Input data pre-processing steps for TOPNET model

Prior to the development of HydroDS, the acquisition and preprocessing of input datasets for the UEB and TOPNET models had to be done manually. The steps involved include watershed delineation, generation of modeling elements (grids in UEB, subwatersheds in TOPNET), extraction of terrain variables from DEM, estimation of canopy variables based on datasets such as the National Land Cover Database (NLCD) (Homer et al., 2015) or satellite remote sensing products, e.g., MODIS. In addition, operations are required for spatial interpolation/aggregation and downscaling of weather forcing into model grid cells from gridded data sources or weighted interpolation of point

precipitation gages. Conversion of gridded spatial datasets into modeling parameters, computation of hydrologically relevant (model specific) variables from topography datasets, e.g., wetness index and distance to stream distributions, generation of soil properties from Soil Survey Geographic Database (SSURGO) (Soil Survey Staff, 2019), and file format conversions to the formats used by UEB and/or TOPNET models are also part of the input preprocessing tasks. Undertaking these tasks requires significant understanding of the data sources and GIS skills, and it takes considerable time and effort, especially for new users that must learn the data processing steps and complicated software configuration, as well as the requirement for documenting their work in a reproducible way.

The HydroDS data services are needed to support execution of workflows similar to Figures 1 and 2 as web services so that a user does not need to undertake these tasks manually on a desktop PC. The UEB and TOPNET models were selected as the starting models. This was because of the need to be able to efficiently set up multiple UEB models for use in water supply forecasting research (Gichamo, 2019) and the fact that TOPNET is already in use in streamflow forecasting applications (e.g., Clark et al., 2008) and was being used for hydrologic modeling examining the impact of climate change on streamflow regime (Sazib, 2016). These requirements provided impetus for developing general-purpose model setup capability. While developed for these specific models, the data required are also commonly used in other distributed hydrologic models (e.g., precipitation, temperature, relative humidity, wind speed, radiation) and the services developed here have potential to be more broadly applicable to other models. In addition, the three data formats used by HydroDS (shapefile, GeoTiff, and NetCDF) are among the

most widely used formats for representing these classes of data.

The required functionality identified included:

- Select a model domain (geographic location of watershed of interest) and, if necessary, delineate the watershed draining to an outlet point.
- Compute hydrological variables from a digital elevation model (DEM), including slope, aspect, topographic wetness index, etc.
- Estimate canopy variables and vegetation indices such as the leaf area index based on the National Land Cover Database (NLCD).
- Generate soil properties based on Soil Survey Geographic Database (SSURGO) data.
- Calculate wetness index and distance to stream distribution.
- Create node and reach link information for TOPNET model.
- Calculate weights used to interpolate precipitation from gage locations to model grid cells.
- Perform coordinate system conversions, resampling, and sub-setting to the desired model scale including grid spacing, support, and extent.
- Retrieve weather forcing data from national data sources (e.g., Daymet, NASA NLDAS) and process and map to model elements.
- Convert between data formats (e.g., GeoTiff raster to NetCDF and vice versa).
- Carry out arithmetic operations on array data stored in NetCDF or GeoTiff formats.
- Create HydroShare resources from data generated by HydroDS. The data

may be individual files such as a watershed delineated from a DEM or a set of model inputs and/or outputs. Also, support moving existing resources in HydroShare to HydroDS for processing.

- Create a model instance input package (e.g., all of the required input files to execute a model for a selected geospatial domain).
- Miscellaneous file manipulation services such as upload, download, delete, zip, show metadata of a resource, etc.
- Authentication and user access control for security.
- Saving work within a storage space allocated for a user and managing the contents of this storage.

### *3.2 Design and Architecture*

Figure 3 shows the high-level organization of HydroDS, including ***HydroDS Services*** and ***HydroDS Python Client Library***. The HydroDS Services are RESTful APIs (https://en.wikipedia.org/wiki/Representational_state_transfer). These services consist of data processing and user space and account management tools that were designed to meet the requirements listed above. These services can be categorized into two major types as (1) services providing general ETVs for the modeling domain, and (2) Model specific (UEB/TOPNET) services. The ETVs are variables that can be applied to other models that take as input gridded datasets in NetCDF or GeoTiff file formats. Examples of ETV services include watershed delineation, generation of stream networks, processing of weather and soil data. Model specific services include creation of node and reach link files (TOPNET), creation of wetness index (TOPNET), and creation of model parameters (UEB and TOPNET). Tables 1 and Table 2 show sample services for each category. In

addition, we added functions for common hydrological data processing tasks such as interpolation, resampling, and projection of geospatial data. Some of the data accessible through HydroDS are staged on the HydroDS servers for fast access. However, time variable data such as meteorological forcing need to be periodically updated by harvesting the data for recent years after it has become available.

The HydroDS services are comprised of tools implemented as a set of Python functions for accessing and processing of data in raster (GeoTiff), vector (shapefile), and multi-dimensional space-time (NetCDF) formats. Each tool contains one or more atomic data processing functions, each function with a single task. Thus, for the tools that comprise HydroDS, the design and implementation approach we followed was that each function is a stand-alone service that gets executed separately.

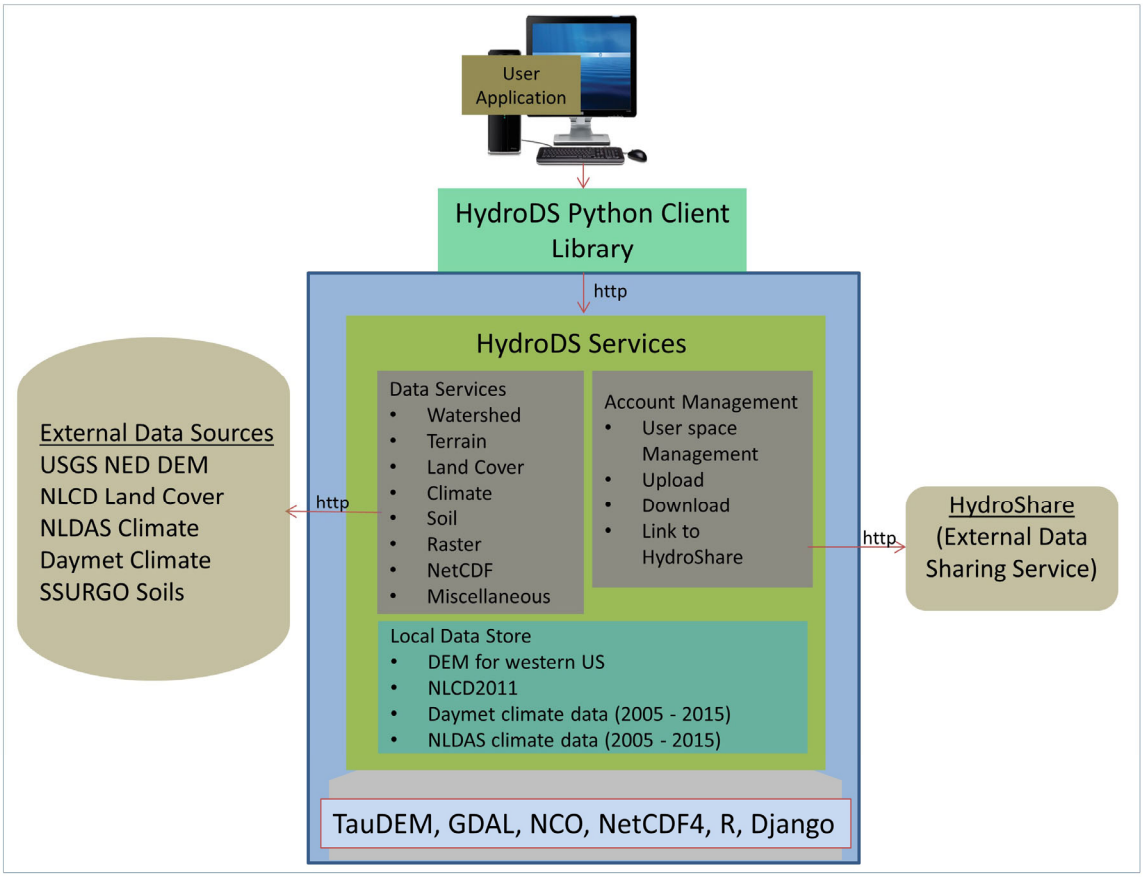

Figure 3. High-level architecture of HydroDS. HydroDS comprises HydroDS Services and HydroDS Python Client Library.

The Account Management functions provide user authentication services as well as ability for the users to manage the files in their user space, with functionality to upload or download data to or from their user space in HydroDS. With the linkage between HydroDS and HydroShare, a user is able to transfer data processed in HydroDS to HydroShare. This provides a mechanism by which data and model packages created by one user may be shared with others (Tarboton et al., 2014a; 2014b).

The HydroDS Python Client Library is a set of Python functions that can be invoked from user computer to make calls to HydroDS. For each data service function on the server side, a corresponding interface is implemented in the HydroDS client library. The HydroDS client library makes it easier to access these data services and thus facilitates scripting and execution of workflows that use the services from a programming environment on a desktop computer. The HydroDS client library can also be used by desktop applications to access the data services. Example client software that interacts with HydroDS through the client library is shown in Figure 4. This Google Map-based graphical user interface (GUI) program was developed using Python to enable calling HydroDS' watershed delineation function by graphically specifying the bounding box around the watershed of interest and watershed outlet location.

Upon a request from a user desktop through the Python client library, the data services are executed on the server side where needed service libraries and dependencies have been installed and configured, freeing the user from these dependency configuration challenges.

Table 1: Example HydroDS Essential Terrestrial Variable data services

| **Function** | **Description** |
|---|---|
| Delineate watershed and stream network | This function delineates watersheds and a stream network for a user selected spatial domain and outlet location. It also provides a stream network topology with network connectivity information, and stream network coordinates and attributes from each grid cell along the network. |
| Get soil | This function derives soil parameters (e.g., soil hydraulic conductivity, transmissivity, and porosity) based on data from SSURGO. |
| Get Daymet climate data | This function retrieves daily precipitation, maximum temperature, minimum temperature, vapor pressure, or shortwave radiation from Daymet data for the period 2005 - 2015 currently stored in the HydroDS server. |

Table 2: Example HydroDS TOPNET model specific services

| Function | Description |
|---|---|
| Create node and reach link | This function uses the stream network tree file and network coordinates file obtained from the 'Delineate Watershed' function for generating TOPNET node link, reach link and reach properties files. |
| Create wetness index and distance to stream distribution | This function is used to group grid values of topographic wetness index and distance to stream into bins for each subwatershed, tabulating the lower and upper bound of each bin and the proportion of area within each bin. |
| Create model parameters | This function uses soil, land use, and land cover data for estimating the time invariant model parameters for each subwatershed. |
| Create rainweight | This function calculates weights used to interpolate precipitation from gage locations to model grid cells. It also adjusts precipitation to account for topographic effects. |

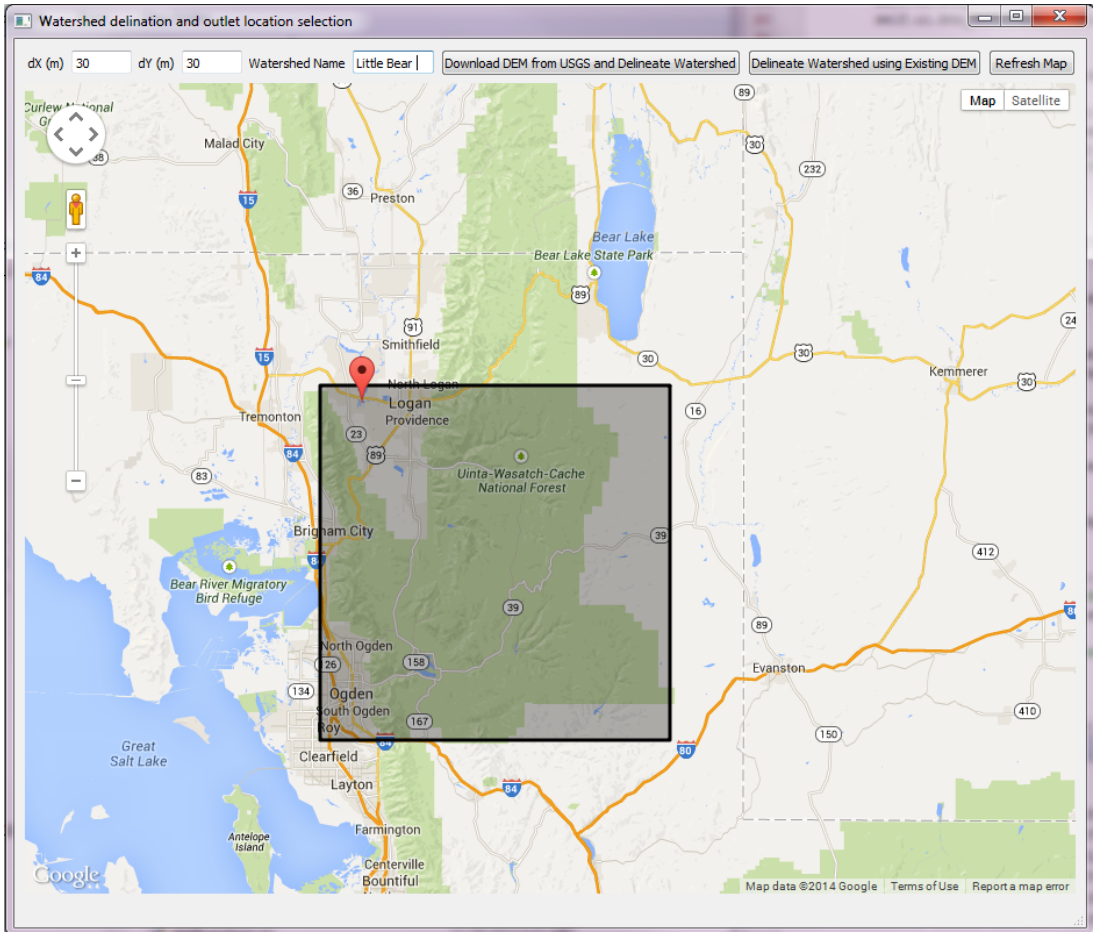


Figure 4. A desktop, Google Map-based GUI program for accessing USGS DEM and watershed delineation through HydroDS. Google Map drawing tools are used to specify the bounding box around the watershed of interest and Google Map marker is used to select an approximate watershed outlet location.

### *3.3 Implementations*

The watershed and terrain services are based on functions from the TauDEM (Tarboton, 2015; Tesfa et al., 2011; Tarboton et al., 2009a) and GDAL geospatial libraries (GDAL Development Team, 2014). The watershed and terrain functions deal with rasters and shapefiles, and, hence, functions for creating and editing these file formats also make up part of the services. The watershed tools delineate the watershed

upstream of the outlet location after extracting a subset of the DEM and resampling it to the required grid cell size. In addition to watershed delineation, a stream network is defined and delineated based on the TauDEM Peuker Douglas valley identification and stream drop approach (Peucker and Douglas, 1975; Tarboton and Ames, 2001). This approach chooses the appropriate threshold to delineate a stream network consistent with geomorphological properties. The outputs from this tool are stream network, subwatersheds draining to each stream network reach, wetness index, and distance to stream.

The terrain functions involve processing of a raw DEM and extraction of hydrological variables such as slope and aspect. Currently, a DEM data file containing the one arc-second (~ 30 m) spatial resolution National Elevation Dataset (NED DEM) covering the western U.S. (-128.0017 to -101.9983 longitude and 28.9983 to 50.0017 latitude) is available on the HydroDS server as the starting point for the watershed and terrain functions. The western U.S. was the focus of the research project supporting this work. To model a watershed outside of the western U.S. but in the Contiguous U.S. (CONUS), HydroDS has a wrapper function that is used to download, at run time, the one arc-second DEM from USGS web services (ftp://rockyftp.cr.usgs.gov/vdelivery/Datasets/Staged/NED/1/IMG/), based on user-specified boundary information in geographic coordinates. If a user wants to use different DEM data than those currently served by HydroDS, they can upload their own DEM, or move a raster resource from HydroShare to their user space in HydroDS.

The land cover services use the 2011 National Land Cover Database (Homer et al., 2015) together with a look-up table of canopy variables for each land cover category

to map the canopy variables into the watershed grid. These services are limited by the empirical canopy variable values available for each land cover class and currently apply only to the variables canopy height, fraction of grid cell area covered by vegetation, and leaf area index that are required by the UEB model. These can be updated when more and/or better information become available. For example, vegetation variables from remotely sensed Moderate Resolution Imaging Spectroradiometer (MODIS: https://modis.gsfc.nasa.gov/) products can be uploaded by the modeler into their working directory in HydroDS and used.

The Soil Data services provide rasters of soil properties such as soil hydraulic conductivity, transmissivity, and porosity for the delineated watershed based on data from the SSURGO Database (http://websoilsurvey.nrcs.usda.gov/). SSURGO segments the landscape into soil map-units, with each unit comprised of a number of components, each of which represents the soil as a number of layers (horizons). There is a NRCS Soil Data Access service that provides horizon level soil properties based the components in each map unit. We hosted the soil map unit key raster which is static information, but to obtain soil properties invoke the NRCS service on the fly to retrieve horizon soil properties for map units contained within the input watershed. A two-step weighting process for deriving soil unit average soil properties was implemented using R. First, the horizon level soil values are weighted by their thicknesses and then the component values are weighted by their percentage composition. The aggregate soil property values are converted into an R raster object with cell values containing soil properties. This function used functionality from existing R packages such as SoilDB, SSOAP, and raster. The results of this function are soil properties rasters for the watershed.

The climate services provide access to and processing capabilities for data in NetCDF format. Daily data for precipitation, maximum and minimum temperature, vapor pressure, shortwave radiation, snow water equivalent, and day length from Daymet (Thornton et al., 2014) with 1 km spatial resolution covering the CONUS for the period 2005 - 2015 are currently available in the HydroDS server to facilitate efficient access. There is also a wrapper function using the "DaymetR" codes and "raster" packages to download Daymet precipitation, temperature, and vapor pressure data for a specific time-period and a specific watershed. This function is based on the batch downloading utility of "DaymetR" to download weather variables at multiple points and convert them to daily-interpolated surface weather variables. Hourly data of precipitation, temperature, surface pressure, shortwave and longwave radiation, zonal and meridional wind speed, and specific humidity from the National Land Data Assimilation System (NLDAS) (Mitchell et al., 2004) with horizontal resolution of 0.125-degree geographic coordinates covering the CONUS are available for the period 2005 – 2015. The NLDAS data are organized in yearly NetCDF files for efficiency. The climate services include functions for downscaling and elevation adjustment of temperature, precipitation, and vapor pressure based on a downscaling methodology described by Sen Gupta and Tarboton (2016).

The model specific services rely on and build upon these general services. For TOPNET model the "Create Reach and Node Link" function, shown in Table 2 generates files and tables that define the association between model nodes, sub-catchments, river riches, and their properties. This function uses outputs obtained from the 'Watershed Delineation' service. Similarly, the "topographic wetness index distribution" inputs to

TOPNET are computed from outputs generated by the watershed delineation. The UEB and TOPNET model parameters are time invariant and describe the unchanging properties of the watersheds and subwatersheds. For TOPNET, these are expressed at the spatial scale of a subwatershed. These parameters are derived by averaging over the grid cells within the subwatershed (for TOPNET). The "Create model parameters" function uses extracted soil, land use, and land cover data as inputs, and aggregated parameter values for each subwatershed are written into the model parameters file. For UEB, parameters are assumed constant over the whole watershed, and generally taken to be transferrable across watersheds without requiring calibration.

TOPNET is configured to derive aggregated subwatershed precipitation inputs as a weighted sum of point precipitation measurements. The weights associated with each gauge for each subwatershed are calculated as part of the pre-processing by the "Create Rainweight" function using linear interpolation based on Delaunay triangles formed with a vertex at each rain gauge, adjusted using an annual rainfall surface to account for topographic effects. The method for determining precipitation weights is described in (Bandaragoda et al., 2004). This procedure provides a way to estimate precipitation as a smooth surface based on nearby surrounding gauges, while at the same time adjusting point gauge values for topographic effects. The adjustment for topographic effects is required because, often, precipitation is recorded at low elevation and hence may not accurately represent the precipitation in parts of the watershed with higher elevation.

The web services were written in Python and implemented in Django Python Web framework (https://www.djangoproject.com/). The service code uses existing functions as much as possible and provide Python or R wrappers to functions from TauDEM, GDAL,

NetCDF libraries (Rew and Davis, 1990; Rew et al., 2006), NetCDF Operators (NCO) (Zender, 2008), and national web services such as those from USGS, Daymet, EPA, SSURGO. Figures 5 and 6 illustrate the operations carried out for the "TauDEM Peuker-Douglas watershed delineation" and the "SSURGO soil data" services respectively.

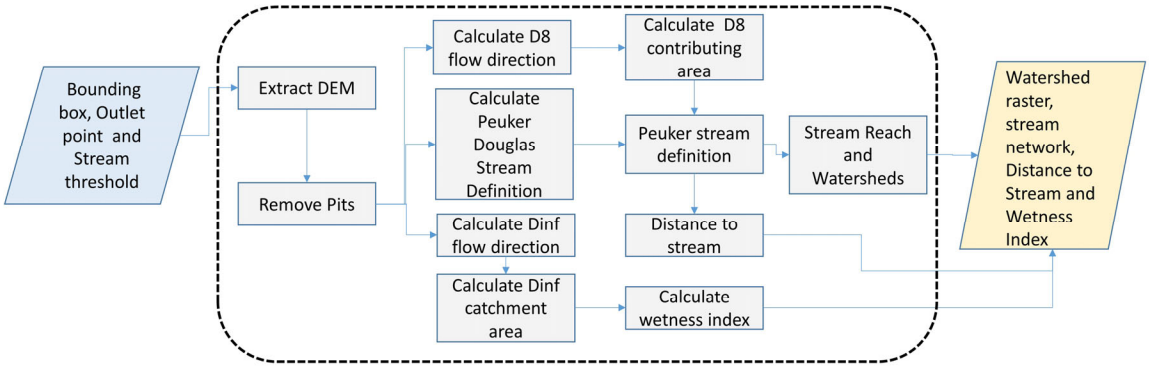


Figure 5: Watershed delineation processing using TauDEM. The boundary identified by the dotted line represents the steps in the "Delineate Watershed" function. Inputs are shown in the left box.

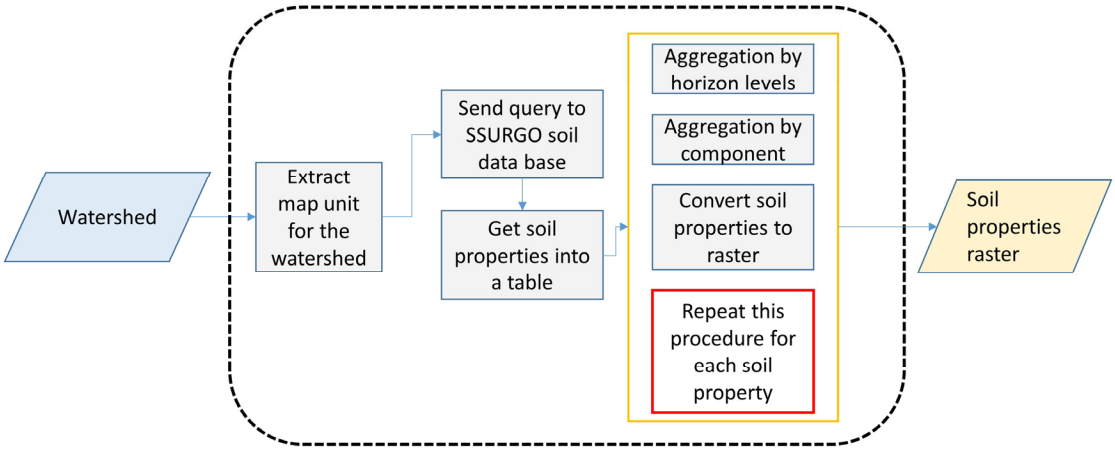


Figure 6: Steps required for retrieving soil data from the SSURGO soil data base. The boundary identified by the dotted line represents the steps in the "Get soil and land use" function.

As stated earlier, these services were implemented on a web server and they are

accessed from a user desktop through a single Python library (HydroDS Python Client Library). When using the Python client, the only software required by a user is a Python interpreting environment with the Python 'requests' module (http://docs.python-requests.org/en/latest/) installed. Transmission of function calls and data transfer between client and server uses HydroDS RESTful APIs over HTTP protocol.

## 4. Evaluation of HydroDS with Input Data Preparation for the Utah Energy Balance Snowmelt Model (UEB) and the TOPNET Hydrologic Model.

### *4.1 Motivation and Case Studies*

The Colorado Basin River Forecast Center (CBRFC) provides streamflow forecasts for watersheds in the Colorado River and Great Salt Lake basins (CBRFC basin) where a significant portion of the annual surface water input comes from snowmelt that primarily falls in the mountainous headwater watersheds. Currently, the CBRFC uses the National Weather Service River Forecasting System (NWSRFS) that consists of a temperature-index snowmelt model (Anderson, 2006; Anderson, 1973; Peck, 1976; Burnash and Singh, 1995). The motivation for this case study arose from the desire to evaluate the UEB snowmelt model (Tarboton et al., 1995) for inclusion in the NWSRFS. UEB is a physically based, point energy and mass balance model with a single ground snowpack layer and a vegetation component that accounts for major snow processes in forested watersheds (Mahat et al., 2013; Mahat and Tarboton, 2014, 2012; Luce and Tarboton, 2010; Mahat and Tarboton, 2013; You et al., 2014). As a single layer model, UEB is parsimonious, avoiding some of the complexities for more detailed multi-layered snowmelt models. In addition, the gridded version of the model has parallel processing capability using Message Passing Interface (MPI) and Graphics Processing Units (GPU)

methods to speed up simulation (Gichamo and Tarboton, 2020). These factors makes UEB a promising candidate for spatially distributed modeling in support of operational streamflow forecasting where computational time can be critical (Gichamo and Tarboton, 2019).

One of the issues that needed to be addressed in order to be able to use UEB in the streamflow forecasting system was whether the input data available for the energy balance model were of sufficient quality and could be efficiently prepared for forecast watersheds. In this study, we evaluated the HydroDS for preparation of the inputs to the UEB model for multiple forecast watersheds in the CBRFC basin. We quantified how much improvement was achieved by HydroDS when compared to desktop-based GIS tools in terms of the time taken to prepare input data using each approach. We also demonstrated the value of the data services to facilitate repeatability and reproducibility and the tracking of provenance through an automated workflow script. In addition, the use of web services reduces the need for individual users to have a local data copy and data organizing software.

In addition, we evaluated the services for preparation of input for the TOPNET model for one of the CBRFC forecast watersheds. TOPNET (Bandaragoda et al., 2004; Ibbitt and Woods, 2004) is a distributed hydrologic model in which topographically delineated subwatersheds (used as modeling units) discharge into the stream network. The stream network is then used to route streamflow to the watershed outlet. TOPNET was developed by combining TOPMODEL (Beven and Kirkby, 1979; Beven et al., 1995) with channel routing (Bandaragoda et al., 2004; Ibbitt and Woods, 2004). “A key contribution of TOPMODEL is the parameterization of the soil moisture deficit (depth to

water table) using a topographic index to model the dynamics of variable source areas contributing to saturation excess runoff" (Bandaragoda et al., 2004, p. 179). Additional enhancements in TOPNET beyond the original TOPMODEL include (1) calculation of reference evapotranspiration using the ASCE standardized Penman-Monteith method (ASCE-EWRI, 2005; Walter et al., 2000) and (2) calculation of snowmelt using the Utah Energy Balance Snowmelt model (Tarboton et al., 1995).

### *4.2 Study Watersheds and Input Data Preparation*

Currently, the CBRFC models are structured into watersheds that flow to NWS streamflow forecast points. As such, the modeling units are forecast watersheds, for which input data are structured independently. This makes the procedure manageable. To apply the UEB model for streamflow forecasting in the CBRFC basin, we needed to set up a model instance for each forecast watershed. Making a model setup for each watershed using desktop tools currently in use can be time-consuming, error prone, and hard to reproduce. Recognizing that the same set of data setup operations need to be carried out for each watershed, a workflow script to pre-process input data for one watershed can be reused for multiple watersheds.

A number of headwater watersheds in the Colorado River basin and the Great Salt Lake basin were selected to set up UEB inputs (Figure 7). These watersheds were selected because the CBRFC had an interest in evaluating potential improvements to their forecasts from using UEB. The HydroDS tasks required to get complete UEB model inputs for a given watershed are shown in the flowchart in Figure 8. This workflow is encapsulated in a single script file provided as a HydroShare resource (Gichamo et al., 2020). The inputs to this workflow script for a given watershed are the geographic coordinates of the bounding box of the domain holding the watershed, outlet location,

start and end time, model target cell size, and the spatial reference (projection in the form of EPSG Code http://spatialreference.org/ref/epsg/) to be used for the output. The commands in the workflow script can also be called interactively from any Python interpreter, or, as mentioned earlier, the service functions can be called from a user application such as shown in Figure 4.

TOPNET input preparation was tested for the Logan River Watershed (also a CBRFC forecast watershed). To setup a TOPNET model input package, a user needs to provide geographic coordinates for the bounding box around the watershed of interest, the approximate outlet location, a range of stream threshold values (from which an optimum threshold value is estimated for defining the stream network), and the modeling period. Note here that if the user provides incorrect inputs the services report an error and quit. Once complete, the user is provided with a link from which the processed data can be downloaded. An example script for the preparation and saving of TOPNET model input in HydroShare is provided in (Sazib and Tarboton, 2020). This script implements the steps shown in Figure 9.

The TOPNET input package was also generated manually for the Logan River following Figure 2 described above as part of the evaluation of HydroDS. The manually derived TOPNET input files were then compared with those from Hydro-DS services and found to match well, validating the data services. However, minor differences (1%-5%) were found in subwatershed soil properties values due to use of a gridded map unit key raster in HydroDS instead of the map unit key shapefile for extracting and processing soil properties from SSURGO. The gridded map unit key raster data have a 30 m cell size that approximates the vector polygon of the map unit key in an Albers Equal Area projection.

This approximation results in the small differences noted above.

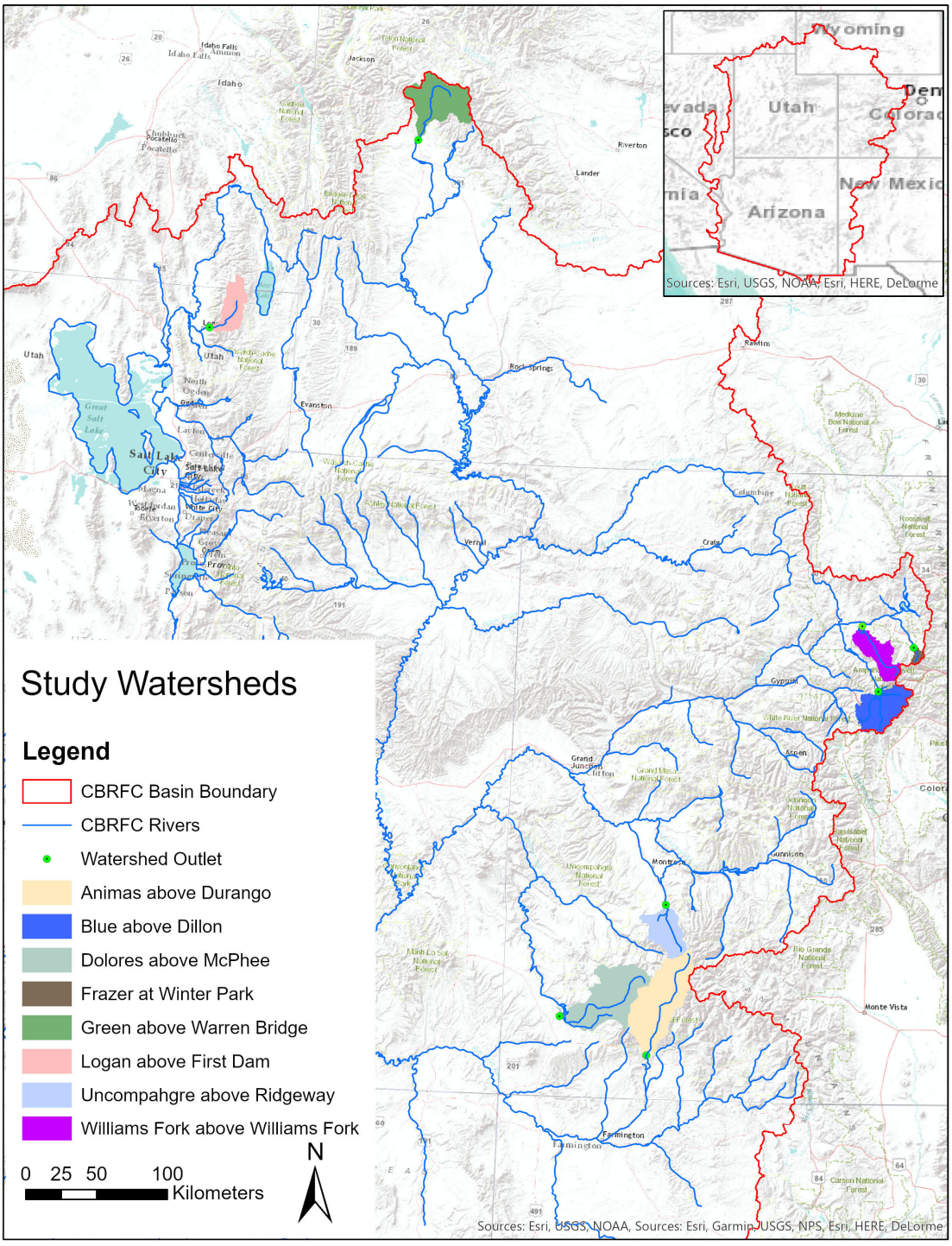


Figure 7. Map showing study watersheds draining to forecast points in the CBRFC where there was interest in evaluating UEB (prepared using ESRI ArcGIS www.esri.com).

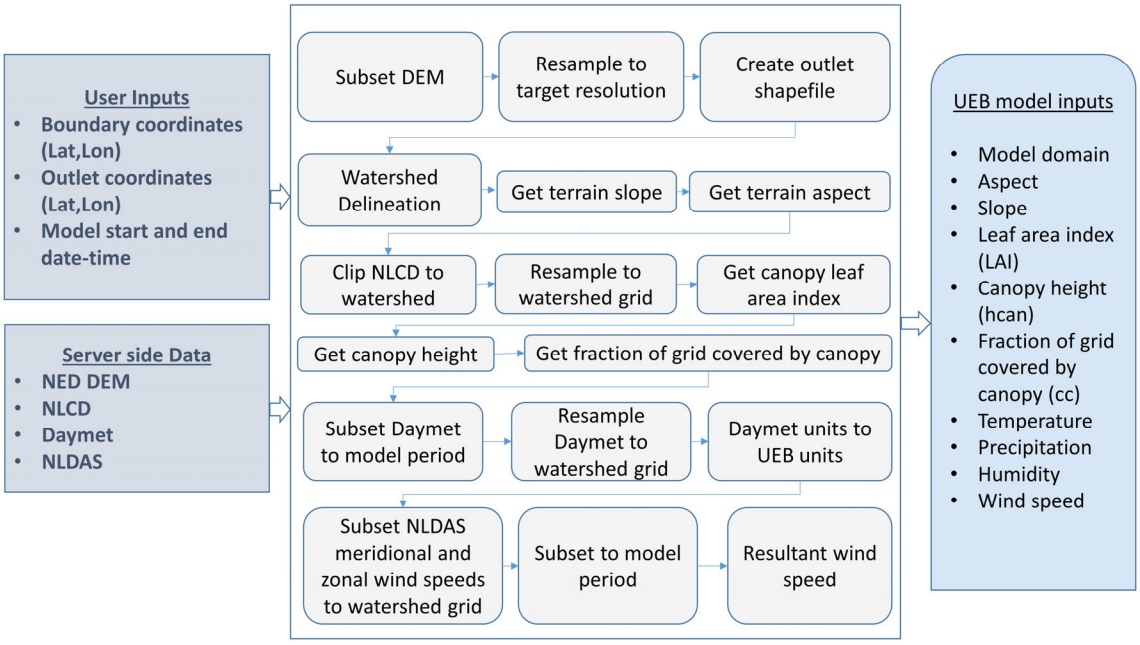


Figure 8. Flowchart of the Utah Energy Balance snowmelt model (UEB) input data preparation steps using HydroDS.

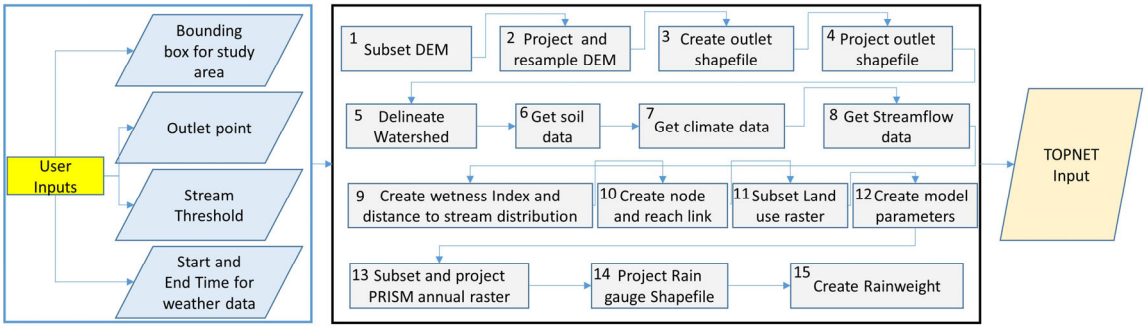


Figure 9: TOPNET data preparation steps using HydroDS data services.

## 5. Results and Discussion

### *5.1 Time Spent on Input Data Preparation*

Table 3 shows the time it takes for preparation of UEB input data for the Logan River watershed for the water year 2009 for three methods: manually on desktop PC; using automating scripts on desktop PC; and using HydroDS through a workflow script. Running the HydroDS script for a different watershed only requires modification of the

watershed boundary, location of the outlet, and projection information, as mentioned earlier. It took 10 minutes to run the pre-processing and package it and put it into HydroShare. It took comparable total time (between 9 and 15 minutes) to prepare inputs for the other study watersheds using HydroDS as shown in Table 4.

Preparing the inputs manually by the first author with multiple years of experience using desktop-based GIS software took more than 5 hours, which was cut to 2 hours and 45 minutes by simply automating the desktop tasks using scripts and that was further reduced to only 10 minutes when using HydroDS. The scripts that were used in the desktop environment are similar to those implemented in HydroDS. Thus, the difference between the time it takes HydroDS to prepare the inputs versus the time taken by scripts on a desktop PC can partly be attributed to the efficient organization of the data in HydroDS. On a desktop PC, even when using scripts that automate the processes, user intervention is necessary, for instance to locate the delineated watershed and point it to the scripts that run weather forcing pre-processing, because all the other inputs (terrain, canopy, weather forcing) have to be mapped onto the watershed grid file that defines the modeling domain.

Preparing the script to run the HydroDS took about 30 minutes (for someone who was already well familiar with the system), which is a one-time task, after which the same script can be re-used for different watersheds by only changing the user inputs shown in Figure 8. We also report in Table 3 the time it took to download data to a desktop PC separately, because, theoretically at least, this is a one-time operation—note most of the time here was taken by NLDAS weather data. We did not account for the time required to harvest the Daymet, NLDAS, NED DEM, and NLCD land cover data

into HydroDS data servers. This task of updating the HydroDS data stores with new data when they become available is a one-time task, which then makes data available to multiple users. However, we note that, while the HydroDS data disks, at the time of this writing, can store up to 10 TB of data, the desktop PC on which the test was carried out has a hard disk with a capacity of 500 GB. Thus, once the pre-processing of the inputs was finished, the intermediate files had to be deleted to free up storage space. Therefore, if we need to carry out similar operations, say in few months, downloading the data again might be necessary.

Another observation in Tables 3 and 4 is that the time for weather forcing data processing is dominated by the wind data from NLDAS. This is because the NLDAS data has hourly temporal resolution for the entire CONUS compared to the daily temporal resolution of the Daymet data. In addition, the hourly data for each NLDAS weather forcing variable comes in an individual NetCDF file. To increase efficiency of HydroDS, the NLDAS data in HydroDS were pre-organized so that one NetCDF file contains data for a year for each variable, which considerably reduces the amount of processing effort. Therefore, ignoring the time for downloading data into the desktop PC, much of the difference in the NLDAS data processing time between HydroDS and the desktop PC arises from the prior organization of NLDAS data in HydroDS. This is an optimal option because the NLDAS data, after harvesting from NASA servers, were processed and organized only once before being stored on the HydroDS server. Then multiple users can benefit from this organization, thus avoiding redundant and potentially error-prone data processing by different users or by the same user multiple times.

For TOPNET model, the work to prepare a model input package for the Logan

River Watershed using HydroDS took about 7 minutes (Table 5). When using HydroDS, the user does not need to remember the specific details of the sequence of steps to follow, as they were recorded in the workflow script—also helping reproducibility. Manually setting up the model for the Logan River Watershed took about 2 hours by the second author, representing a knowledgeable user familiar with the procedure. The difference between the time it takes using HydroDS and that on desktop PC can again be attributed to user intervention for downloading the geospatial and time series data manually from the data provider websites and the number of basic geospatial processing tasks that need to be carried out sequentially.

The TOPNET input data package created by HydroDS was then shared through HydroShare using the "create_hydroshare_resource" function. This sharing of the TOPNET input package enables collaboration among team members working on this model. It also facilitates publication of the data in support of research findings being published from the results, thereby enhancing research reproducibility and trust in the model results. Here the transfer to HydroShare occurred between servers, independent of the user's desktop system, a mode of working more amenable to large datasets, because data do not have to be copied into the user's desktop PC. The shared TOPNET input package was then downloaded from HydroShare to a local PC where parameters were calibrated and sensitivity analysis was performed. This demonstrated the suitability and usability of the HydroDS generated package in a typical hydrologic modeling exercise by a graduate student.

Table 3: UEB input pre-processing time for the Logan River Watershed for the 2009 water year

| Data preparation method | Time (min) | | | | | | |
|---|---|---|---|---|---|---|---|
| | Watershed Delineation | Preparing Terrain Variables | Preparing Canopy Variables | Preparing Weather Forcing Data | | Total | Downloading NLDAS Forcing data |
| | | | | Daymet | NLDAS (wind) | | |
| Manual on desktop PC[a] | 60 | 15 | 40 | 125[b] | 75[c] | 315 (5.25 hrs) | 120 |
| Using scripts on desktop PC[a] | 25 | 7 | 18 | 40 | 75 | 165 (2.75 hrs) | 120 |
| Using HydroDS[d] | 1 | 0.5 | 0.5 | 2 | 6 | 10 (0.17 hrs) | NA |

[a] Intel(R) Core™ i7-3770 PC with 4 cores, 3.40 GHz (maximum Turbo Boost frequency 3.90 GHz), hyper-threading, 32 GB RAM, 1 TB Disk configured as 500GB for applications and 500 GB for data storage on 1000 Gigabit internet connection.
[b] Not including time for data download and time spent troubleshooting the errors that occurred during the data processing. When including data downloading and error troubleshooting, this task (processing Daymet input data) alone took about 6 hours.
[c] Using scripts. It proved to be too laborious to process this manually—after about an hour of trying, decided to write the script, which took about two hours, then finish the work from the script.
[d] Preparing HydroDS Script takes 30 minutes for someone familiar with the system.

Table 4: UEB input pre-processing time using HydroDS for the 2009 water year

| Study Watershed | Time (min) | | | | | |
|---|---|---|---|---|---|---|
| | Watershed Delineation | Preparing Terrain Variables | Preparing Canopy Variables | Preparing Weather Forcing Data | | Total |
| | | | | Daymet | NLDAS (wind) | |
| Animas above Durango | 1 | 0.5 | 0.5 | 2 | 9 | 13 |
| Blue above Dillon | 0.5 | 0.5 | 0.5 | 1.0 | 9 | 11.5 |
| Dolores above McPhee | 1 | 0.5 | 0.5 | 1.5 | 11 | 14.5 |
| Frazer at Winter Park | 0.25 | 0.25 | 0.25 | 1.0 | 9 | 10.75 |
| Green above Warren Bridge | 0.5 | 0.5 | 0.5 | 1.5 | 6 | 9 |
| Logan above First Dam | 1 | 0.5 | 0.5 | 2 | 6 | 10 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Uncompahgre above Ridgeway | 0.5 | 0.25 | 0.25 | 1.0 | 12 | 14 |
| Williams Fork above Williams Fork | 0.25 | 0.25 | 0.25 | 1.0 | 13 | 14.75 |

Table 5: TOPNET input pre-processing time for the Logan Watershed for the 2009 water year

| Data preparation method | Time (min) | | | | | |
|---|---|---|---|---|---|---|
| | Watershed delineation and terrain derivatives | Soil data | Daymet weather forcing data | Streamflow | TOPNET specific data | Total |
| Manual on desktop PC | 40 | 25 | 10 | 3 | 45 | 123 (2.05 hrs) |
| Using HydroDS | 1.5 | 2.5 | 1 | 0.5 | 1.5 | 7 (0.12 hrs) |

### *5.2 Workflow Scripts, Reproducibility, and Provenance*

The services demonstrably reduced the time and effort required to prepare UEB and TOPNET inputs, which enables water scientists to spend less time extracting and formatting data. However, in the long run, a more useful benefit arises from the fact that the workflow script maintains the provenance of the data processing steps making it easier for modeling workflows to be shared and scientific results to be reproduced (Leonard and Duffy, 2014; Leonard et al., 2019; Miles and Band, 2015; Miles, 2014). For instance, few months after first using the script to prepare the Logan River watershed, we came back and used the script again with no additional work required, and obtained the exact same result. Thus, HydroDS facilitates reproducibility and repeatability of hydrologic data processing. In addition, by changing the user inputs shown in Figure 8, the same script can be used for a different watershed. This way, HydroDS facilitate speedy setup of models for the multiple forecast watersheds such as in the CBRFC basin. In addition, it provides the and ability to take advantage of a pre-configured system where the user need not be concerned about the organization of the server side functions,

data, software, and hardware where the dependencies are already sorted out. By providing the capability to automate the data processing steps, preserving provenance, and enhancing the reproducibility and repeatability of the hydrologic data processing, HydroDS thus provides a number of benefits of standard workflow systems (Goble and De Roure, 2009), while simplifying the responsibility of the user to handling a single Python workflow script.

More generally, the outcome of this work is a development of server-side data processing services, where lessons learned from the experience could be applied for other models. One lesson learned, based on our observations using the services, was that the provision of access to atomic functions through the HydroDS Python Client Library to call individual tasks appears to be not that useful, as workflow scripts combining multiple related tasks are often the ones that are applied. Therefore, provision of coarser grained convenience functions, e.g., providing "watershed delineation" but hiding the constituent functions such as "move outlets to streams", may be more productive.

The biggest limitation of HydroDS, as it stands currently, is the fact that the services are limited to gridded data such as those used in the UEB model and subwatershed based inputs customized for the TOPNET model. A number of hydrologic models use unstructured grids or other modeling units such as Hydrologic Response Unit (HRU). The data processing services need to accommodate for such modeling configurations if they are to be used by the wide range of models currently used by the hydrologic community. A related, but less critical, limitation of HydroDS is that it only supports GeoTiff, Shapefile and NetCDF file formats. The Hierarchical Data Format (HDF https://www.hdfgroup.org/) is as widely used as NetCDF and would add additional

flexibility to HydroDS if it were supported. An alternative is to add HydroDS functions for conversion of data from NetCDF to other standardized data formats such as HDF and vice versa.

Another limitation of this study is that all the watersheds evaluated were headwater watersheds whose final (pre-processed and ready to be used in the model) input data have relatively small file sizes (less than 2 GB). The work in this paper deals with large basins such as the Colorado River basin by breaking them down into CBRFC forecast watersheds and handling data processing for smaller, individual watersheds. Dealing with individual forecast watersheds with relatively small sizes was a design choice that keeps the size of the data and the computational resources for pre-processing of a single watershed easily manageable, while taking advantage of automation to address multiple watersheds. Characterizing how the services perform when increasing the sizes of the watersheds, for example by integrating multiple adjacent watersheds, may be an important next step. In such a scenario, the size of the weather forcing data increases more rapidly than the other data types, and weather data processing services, which currently use serial codes, may have to deal with large datasets in NetCDF format, which could necessitate implementation of parallel processing. Additional work is also required to deal with the potential increase in processing time due to increase in size of processed data. For example, a mechanism for queuing and batch processing of large operations with asynchronous notifications to a user that the batch of tasks from a workflow script is completed would be useful. This is because it would not be feasible for the user to wait for the web services to return when the execution time extends beyond the ~10 minutes reported in this paper.

As stated earlier, the HydroDS services were tested for selected watersheds in the CBRFC basin. And while the NLCD land cover data and the NLDAS and Daymet weather forcing data for the years 2005 – 2015 stored at the Hydro-DS server cover the whole CONUS, the NED DEM and soil map unit rasters hosted at HydroDS server are limited to the western US (-128.0017 to -101.9983 longitude and 28.9983 to 50.0017 latitude). This presents additional challenge to the applicability of the services for watersheds outside of western U.S. Still, a user can upload their own data and use the data processing tools—although this was not the primary mode of application envisioned for the services at the outset.

Currently, extending the available service functionalities requires obtaining appropriate credentials and familiarity with the development environment and the underlying technologies including Django, GDAL, TauDEM, NetCDF library, and NetCDF operators (NCO) in addition to Python programming skills. Future developments should consider a simplified way to extend the services to cover more geospatial processing tools and data. One way to enable a relatively easy extension of the services by addition of new functionalities is adding a Software Development Kit (SDK) as a component of the HydroDS services. The SDK could be as simple as providing sample source codes to modify for new functions or support more advanced features such as tools and libraries to serve as building blocks for new tools/functions.

Finally, while these results demonstrate that HydroDS helps reduce the time and effort required for accessing and pre-processing model input data, the task of deciding on what hydrological questions to ask depends on the researcher's prior experience. In this study, deciding the case study involved a number of iterations.

## 6. Summary and Conclusions

HydroDS, a set of web-based data services providing access to distributed (gridded and subwatershed based) hydrologic data and geospatial and temporal data analysis capabilities for hydrological models was introduced in this paper. The services comprise functions for important hydrologic data processing tasks such as watershed delineation, terrain processing, estimation of canopy variables based on the NLCD, generating soil properties based on data from SSURGO, and accessing and processing of climate data from Daymet and NLDAS. The services are composed of single task functions that can be used independently or can be chained together to form a Python workflow for complete generation of model inputs. A Python library, the HydroDS Client Library, provides access to the web services. Through the HydroDS Client Library, the services can be used in a Python script or desktop applications. Accessing the services requires only Python, which means that users can access them from any computing platform with Python support.

HydroDS was demonstrated by setting up instances of the Utah Energy Balance (UEB) and TOPNET models for watersheds in the Colorado River and Great Salt Lake basins. The cases demonstrate how HydroDS helps reduce the time and effort spent for accessing and pre-processing hydrologic model input data. A considerable part of the time saved by using HydroDS instead of desktop-based data processing comes from better organization of data in HydroDS. The Python scripting-based data processing workflows enhance reproducibility and repeatability because the same script can be re-used. The script needs to be modified only to specify few user inputs when used for a different watershed. As the workflow script also captures all the steps towards the final

model input, its provenance is preserved in the script. The 'software as a service' paradigm of the web services provides capability for multiple users and relieves users from concerns related to storage and organization of data, which is done in the server, and software and hardware dependencies which are sorted out when the software is configured on the server.

Based on our observations using the services, the provision of access, through the HydroDS Client Library, to the atomic functions to do individual tasks appears to be not that useful; rather the workflow scripts combining multiple coarser granular functions were more productive. The work in this paper deals with large basins such as the Colorado River Basin by breaking them down into CBRFC forecast watersheds and handling data processing for smaller, individual watersheds. This was a design choice that worked well for this study. Future studies should address the alternative approach of processing river basins such as the Colorado Basin as a whole. Future work should also extend the services to provide inputs for unstructured grid models and models using HRUs (or other equivalent tessellations of the landscape) for HydroDS to support a wider range of hydrologic models. Future development should consider provision of Software Development Kit (SDK) in HydroDS to enable (a relatively) easy extension of the services with new functionalities.

**Acknowledgments**

This work was supported by the National Science Foundation (NSF) under collaborative grants EPS 1135482 and 1135483, and by the Utah Water Research Laboratory (UWRL). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views

of the NSF or UWRL. Compute, storage, support, and other resources from the Division of Research Computing in the Office of Research and Graduate Studies at Utah State University and Advanced Research Computing Center at the University of Wyoming are gratefully acknowledged.

## References

Almoradie, A., Jonoski, A., Stoica, F., Solomatine, D., Popescu, I., 2013a. Web-based flood information system: case study of Somesul-Mare, Romania. J. Environ. Eng. Manage 12 1065-1070.

Almoradie, A., Popescu, I., Jonoski, A., Solomatine, D., 2013b. Web Based Access to Water Related Data Using OGC WaterML 2.0. International Journal of Advanced Computer Science and Applications (IJACSA) EnviroGRIDS Special Issue on "Building a Regional Observation System in the Black Sea Catchment", http://dx.doi.org/10.14569/SpecialIssue.2013.030310.

Ames, D.P., Horsburgh, J.S., Cao, Y., Kadlec, J., Whiteaker, T., Valentine, D., 2012. HydroDesktop: Web services-based software for hydrologic data discovery, download, visualization, and analysis. Environmental Modelling & Software 37 146-156.

Anderson, E., 2006. Snow accumulation and ablation model–SNOW-17, NWSRFS Users Manual Documentation, Office of Hydrologic Development, NOAA's National Weather Service, http://www.nws.noaa.gov/oh/hrl/nwsrfs/users_manual/htm/xrfsdocpdf.php.

Anderson, E.A., 1973. National Weather Service river forecast system--snow accumulation and ablation model. TECHNICAL MEMORANDUM NWS HYDRO-17, NOVEMBER 1973. 217 P.

ASCE-EWRI, 2005. The ASCE standardized reference evapotranspiration equation. Technical Committee Rep. to the Environmental and Water Resources Institute of ASCE from the Task Committee on Standardization of Reference Evapotranspiration.

Bandaragoda, C., Tarboton, D.G., Woods, R., 2004. Application of TOPNET in the distributed model intercomparison project. Journal of Hydrology 298(1-4) 178-201.

Beran, B., Goodall, J., Valentine, D., Zaslavsky, I., Piasecki, M., 2009. Standardizing access to hydrologic data repositories through web services, Advanced Geographic Information Systems & Web Services, 2009. GEOWS'09. International Conference on. IEEE, pp. 64-67.

Berry, P., Bonduá, S., Bortolotti, V., Cormio, C., Vasini, E.M., 2014. A GIS-based open source pre-processor for georesources numerical modeling. Environmental Modelling & Software 62(0) 52-64, http://www.sciencedirect.com/science/article/pii/S1364815214002357.

Beven, K., Lamb, R., Quinn, P., Romanowicz, R., Freer, J., 1995. TOPMODEL. Computer models of watershed hydrology. VP Singh 627-668.

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. Hydrological Sciences Journal 24(1) 43-69.

Bhatt, G., Kumar, M., Duffy, C.J., 2008. Bridging the Gap between Geohydrologic Data and Distributed Hydrologic Modeling, In: Sànchez-Marrè, M., Béjar, J., Comas, J., Rizzoli, A., Guariso, G. (Eds.), iEMSs 2008: International Congress on Environmental Modelling and Software Integrating Sciences and Information Technology for Environmental Assessment and Decision Making 4th Biennial Meeting of iEMSs. International Environmental Modelling and Software Society (iEMSs): Barcelona, Catalonia pp. 743-750.

Bhatt, G., Kumar, M., Duffy, C.J., 2014. A tightly coupled GIS and distributed hydrologic modeling framework. Environmental Modelling & Software 62(0) 70-84, http://www.sciencedirect.com/science/article/pii/S1364815214002266.

Burnash, R., Singh, V., 1995. The NWS river forecast system-Catchment modeling. Computer models of watershed hydrology. 311-366.

Carlson, J.R., David, O., Lloyd, W.J., Leavesley, G.H., Rojas, K.W., Green, T.R., Arabi, M., Yaege, L., Kipka, H., 2014. Data provisioning for the Object Modeling System (OMS), Proceedings - 7th International Congress on Environmental Modelling and Software: Bold Visions for Environmental Modeling, iEMSs 2014, pp. 230-237, http://www.scopus.com/inward/record.url?eid=2-s2.0-84911885427&partnerID=40&md5=59c7e70225b4ffd26414cd579d25022a.

Clark, M.P., Rupp, D.E., Woods, R.A., Zheng, X., Ibbitt, R.P., Slater, A.G., Schmidt, J., Uddstrom, M.J., 2008. Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. Advances in Water Resources 31(10) 1309-1324.

Evans, M.R., Oliver, D., Yang, K., Shekhar, S., 2013. Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities. CyberGIS: Fostering a New Wave of Geospatial Innovation and Discovery. Springer Book.

GDAL Development Team, 2014. GDAL - Geospatial Data Abstraction Library, Version 1.11.1. Open Source Geospatial Foundation, http://www.gdal.org.

Gichamo, T.Z., 2019. Advancing Streamflow Forecasts Through the Application of a Physically Based Energy Balance Snowmelt Model With Data Assimilation and Cyberinfrastructure Resources, Civil and Environmental Engineering. Utah State University, https://digitalcommons.usu.edu/etd/7463.

Gichamo, T.Z., Tarboton, D.G., 2019. Ensemble Streamflow Forecasting using an Energy Balance Snowmelt Model Coupled to a Distributed Hydrologic Model with Assimilation of Snow and Streamflow Observations. Water Resources Research 55.

Gichamo, T.Z., Tarboton, D.G., 2020. UEB parallel: Distributed snow accumulation and melt modeling using parallel computing. Environmental Modelling & Software 125 104614, http://www.sciencedirect.com/science/article/pii/S1364815219304050.

Gichamo, T.Z., Tarboton, D.G., Dash, P., 2020. HydroDS UEB model input setup Python scripts: HydroShare, https://doi.org/10.4211/hs.cad5dd0c4106489e87db2e8366dd66b1.

Goble, C., De Roure, D., 2009. The impact of workflow tools on data-centric research,

In: Tony Hey, Stewart Tansley, Tolle, K.i. (Eds.), The fourth paradigm: data-intensive scientific discovery. MICROSOFT RESEARCH: REDMOND, WASHINGTON, pp. 137 - 145.

Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N.D., Wickham, J., Megown, K., 2015. Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. Photogrammetric Engineering and Remote Sensing 81(5) 345-354.

Horsburgh, J.S., Morsy, M.M., Castronova, A.M., Goodall, J.L., Gan, T., Yi, H., Stealey, M.J., Tarboton, D.G., 2016. 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 52(4) 873-889, http://dx.doi.org/10.1111/1752-1688.12363.

Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., Zaslavsky, I., 2008. A relational model for environmental and water resources data. Water Resources Research 44(5).

Horsburgh, J.S., Tarboton, D.G., Piasecki, M., Maidment, D.R., Zaslavsky, I., Valentine, D., Whitenack, T., 2009. An integrated system for publishing environmental observations data. Environmental Modelling & Software 24(8) 879-888.

Ibbitt, R., Woods, R., 2004. Re-scaling the topographic index to improve the representation of physical processes in catchment models. Journal of Hydrology 293(1-4) 205-218.

Jones, N., Nelson, J., Williams, G., Ogden, F., Tarboton, D.G., Burian, S., 2013. CI-WATER: Cyberinfrastructure to Advance High Performance Water Resource Modeling, World Environmental and Water Resources Congress 2013. American Society of Civil Engineers, p. 2737.

Kumar, M., Bhatt, G., Duffy, C.J., 2009. An efficient domain decomposition framework for accurate representation of geodata in distributed hydrologic models. International Journal of Geographical Information Science 23(12) 1569-1596.

Leonard, L., Duffy, C.J., 2013. Essential Terrestrial Variable data workflows for distributed water resources modeling. Environmental Modelling & Software 50 85-96.

Leonard, L., Duffy, C.J., 2014. Automating data-model workflows at a level 12 HUC scale: Watershed modeling in a distributed computing environment. Environmental Modelling & Software 61 174-190.

Leonard, L., Duffy, C.J., 2016. Visualization workflows for level-12 HUC scales: Towards an expert system for watershed analysis in a distributed computing environment. Environmental Modelling & Software 78 163-178.

Leonard, L., Miles, B., Heidari, B., Lin, L., Castronova, A.M., Minsker, B., Lee, J., Scaife, C., Band, L.E., 2019. Development of a participatory Green Infrastructure design, visualization and evaluation system in a cloud supported jupyter notebook computing environment. Environmental Modelling & Software 111 121-133.

Luce, C.H., Tarboton, D.G., 2010. Evaluation of alternative formulae for calculation of surface temperature in snowmelt models using frequency analysis of temperature observations. Hydrology and Earth System Sciences 14(3) 535-543.

Mahat, V., Tarboton, D.G., 2012. Canopy radiation transmission for an energy balance snowmelt model. Water Resources Research 48(1) W01534,

http://dx.doi.org/10.1029/2011WR010438.
Mahat, V., Tarboton, D.G., 2013. Representation of canopy snow interception, unloading and melt in a parsimonious snowmelt model. Hydrological Processes n/a-n/a, http://dx.doi.org/10.1002/hyp.10116.
Mahat, V., Tarboton, D.G., 2014. Representation of canopy snow interception, unloading and melt in a parsimonious snowmelt model. Hydrological Processes 28(26) 6320-6336, http://dx.doi.org/10.1002/hyp.10116.
Mahat, V., Tarboton, D.G., Molotch, N.P., 2013. Testing above- and below-canopy representations of turbulent fluxes in an energy balance snowmelt model. Water Resources Research 49(2) 1107-1122, http://dx.doi.org/10.1002/wrcr.20073.
Merwade, V., Feng, W., Zhao, L., Song, C.X., 2012. WaterHUB: a resource for students and educators for learning hydrology, Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond. ACM, p. 59.
Miles, B., Band, L.E., 2015. Ecohydrology Models without Borders?, International Symposium on Environmental Software Systems. Springer, pp. 311-320.
Miles, B.C., 2014. Small-scale residential stormwater management in urbanized watersheds: A geoinformatics-driven ecohydrology modeling approach. The University of North Carolina at Chapel Hill, p. 217, https://cdr.lib.unc.edu/indexablecontent/uuid:84f67003-6421-4b27-9a3a-39f367a1bc8c.
Mitchell, K.E., Lohmann, D., Houser, P.R., Wood, E.F., Schaake, J.C., Robock, A., Cosgrove, B.A., Sheffield, J., Duan, Q., Luo, L., Higgins, R.W., Pinker, R.T., Tarpley, J.D., Lettenmaier, D.P., Marshall, C.H., Entin, J.K., Pan, M., Shi, W., Koren, V., Meng, J., Ramsay, B.H., Bailey, A.A., 2004. The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. Journal of Geophysical Research: Atmospheres 109(D7) D07S90, http://dx.doi.org/10.1029/2003JD003823.
Morsy, M.M., Goodall, J.L., Castronova, A.M., Dash, P., Merwade, V., Sadler, J.M., Rajib, M.A., Horsburgh, J.S., Tarboton, D.G., 2017. Design of a metadata framework for environmental models with an example hydrologic application in HydroShare. Environmental Modelling & Software 93 13-28.
Nyerges, T.L., Roderick, M.J., Avraam, M., 2013. CyberGIS design considerations for structured participation in collaborative problem solving. International Journal of Geographical Information Science 27(11) 2146-2159, http://dx.doi.org/10.1080/13658816.2013.770516.
Peck, E.L., 1976. Catchment modeling and initial parameter estimation for the National Weather Service river forecast system. Office of Hydrology, National Weather Service.
Peucker, T.K., Douglas, D.H., 1975. Detection of surface-specific points by local parallel processing of discrete terrain elevation data. Computer graphics and Image processing 4(4) 375-387.
Rajib, M.A., Merwade, V., Kim, I.L., Zhao, L., Song, C., Zhe, S., 2016. SWATShare–A web platform for collaborative research and education through online sharing, simulation and visualization of SWAT models. Environmental Modelling &

Software 75 498-512.
Rew, R., Davis, G., 1990. NetCDF: an interface for scientific data access. IEEE computer graphics and applications 10(4) 76-82.
Rew, R., Davis, G., Emmerson, S., Davies, H., Hartnett, E., Heimbigner, D., Fisher, W., 2014. NetCDF Documentation (http://www.unidata.ucar.edu/software/netcdf/docs/). Unidata, University Corporation for Atmospheric Research (UCAR) Community Programs (UCP). , http://www.unidata.ucar.edu/software/netcdf/docs/.
Rew, R., Hartnett, E., Caron, J., 2006. NetCDF-4: Software implementing an enhanced data model for the geosciences, 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanograph, and Hydrology.
Sazib, N., 2016. Physically Based Modeling of the Impacts of Climate Change on Streamflow Regime Civil and Environmental Engineering. Utah State University: Logan, p. 161, http://digitalcommons.usu.edu/etd/5067/.
Sazib, N., Tarboton, D., 2020. HydroDS TOPNET model input setup Python scripts: HydroShare, https://doi.org/10.4211/hs.bcae759c38b844c7aae3bf62fb35211f.
Sen Gupta, A., Tarboton, D.G., 2016. A tool for downscaling weather data from large-grid reanalysis products to finer spatial scales for distributed hydrological applications. Environmental Modelling & Software 84 50-69.
Soil Survey Staff, N.R.C.S., United States Department of Agriculture. , 2019. Soil Survey Geographic (SSURGO) Database. Available online at https://sdmdataaccess.sc.egov.usda.gov, https://sdmdataaccess.sc.egov.usda.gov.
Tarboton, D., Schreuders, K., Watson, D., Baker, M., 2009a. Generalized terrain-based flow analysis of digital elevation models, Proceedings of the 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, Cairns, Australia, pp. 2000-2006.
Tarboton, D.G., 2015. Terrain Analysis Using Digital Elevation Models (TauDEM). Utah Water Research Laboratory, Utah State University, Logan, Utah, http://hydrology.usu.edu/taudem/taudem5/index.html.
Tarboton, D.G., Ames, D.P., 2001. Advances in the mapping of flow networks from digital elevation data, World water and environmental resources congress. Am. Soc Civil Engrs USA, pp. 20-24.
Tarboton, D.G., Chowdhury, T.G., Jackson, T.H., 1995. A Spatially Distributed Energy Balance Snowmelt Model, In: Tonnessen, K.A., Williams, M.W., Tranter, M. (Eds.), Biogeochemistry of Seasonally Snow-Covered Catchments (Proceedings of a Boulder Symposium). IAHS, pp. 141-155.
Tarboton, D.G., Horsburgh, J., Maidment, D., Whiteaker, T., Zaslavsky, I., Piasecki, M., Goodall, J., Valentine, D., Whitenack, T., 2009b. Development of a community hydrologic information system, 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, ed. RS Anderssen, RD Braddock and LTH Newham, Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, pp. 988-994.
Tarboton, D.G., Idaszak, R., Horsburgh, J., Heard, J., Ames, D., Goodall, J., Band, L., Merwade, V., Couch, A., Arrigo, J., 2014a. HydroShare: Advancing Collaboration through Hydrologic Data and Model Sharing, Proceedings of the

7th International Congress on Environmental Modelling and Software, San Diego, California, USA, International Environmental Modelling and Software Society (iEMSs), ISBN, pp. 978-988.
Tarboton, D.G., Idaszak, R., Horsburgh, J.S., Heard, J., Ames, D., Goodall, J.L., Band, L.E., Merwade, V., Couch, A., Arrigo, J., Hooper, R., Valentine, D., Maidment, D., 2014b. A Resource Centric Approach for Advancing Collaboration Through Hydrologic Data and Model Sharing, 11th International Conference on Hydroinformatics, HIC 2014: New York City, USA.
Tarboton, D.G., Maidment, D., Zaslavsky, I., Ames, D., Goodall, J., Hooper, R.P., Horsburgh, J., Valentine, D., Whiteaker, T., Schreuders, K., 2011. Data Interoperability in the Hydrologic Sciences, Proceedings of the Environmental Information Management Conference, pp. 132-137.
Taylor, P., 2012. OGC WaterML 2.0: Part 1-Timeseries. Open Geospatial Consortium Implementation Standard, OGC 10-126r3, 149pp.
Tesfa, T.K., Tarboton, D.G., Watson, D.W., Schreuders, K.A., Baker, M.E., Wallace, R.M., 2011. Extraction of hydrological proximity measures from DEMs using parallel processing. Environmental Modelling & Software 26(12) 1696-1709.
Thornton, P.E., Thornton, M.M., Mayer, B.W., Wilhelmi, N., Wei, Y., Devarakonda, R., Cook, R.B., 2014. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2. Data set. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, USA., p. Medium: X, http://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1219.
Valentine, D., Taylor, P., Zaslavsky, I., 2012. WaterML, an information standard for the exchange of in-situ hydrological observations, EGU General Assembly Conference Abstracts, p. 13275.
Valentine, D., Zaslavsky, I., Whitenack, T., Maidment, D.R., 2007. Design and implementation of CUAHSI WATERML and WaterOneFlow Web services, Proceedings of the Geoinformatics 2007 Conference, San Diego, California, pp. 5-3.
Walter, I.A., Allen, R.G., Elliott, R., Jensen, M., Itenfisu, D., Mecham, B., Howell, T., Snyder, R., Brown, P., Echings, S., 2000. ASCE's standardized reference evapotranspiration equation, Watershed management and operations management 2000, pp. 1-11.
Wang, S., 2010. A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. Annals of the Association of American Geographers 100(3) 535-557.
Wang, S., Anselin, L., Bhaduri, B., Crosby, C., Goodchild, M.F., Liu, Y., Nyerges, T.L., 2013. CyberGIS software: a synthetic review and integration roadmap. International Journal of Geographical Information Science 27(11) 2122-2145, http://dx.doi.org/10.1080/13658816.2013.776049.
Wilkins-Diehr, N., Gannon, D., Klimeck, G., Oster, S., Pamidighantam, S., 2008. TeraGrid science gateways and their impact on science. Computer 41(11) 32-41.
Wright, D.J., Kopp, S., Brown, C., 2013. Esri Position Paper – 2013 CyberGIS '13 All Hands Meeting.
You, J., Tarboton, D., Luce, C., 2014. Modeling the snow surface temperature with a

one-layer energy balance snowmelt model. Hydrology and Earth System Sciences 18(12) 5061-5076.
Zender, C.S., 2008. Analysis of self-describing gridded geoscience data with netCDF Operators (NCO). Environmental Modelling & Software 23(10-11) 1338-1342.