2-2-2021

Collaborative Research: Quantifying Watershed Dynamics in Snow Dominated Mountainous Karst Watersheds Using Hybrid Physically Based and Deep Learning Models

Bethany T. Neilson
Utah State University, bethany.neilson@usu.edu

Data Management Plan

Types of data

We will generate databases containing: (1) data synthesized from external sources including results from previous hydrogeologic studies, LiDAR data, streamflow measurements, records at weather stations, gridded meteorological products, watershed physiographic data (soil maps, digital elevation models, and land cover), (2) data collected by our project team including river/stream/spring discharge, chemical and isotope information, (3) model outputs including snowmelt/rainfall rates predicted by the physically based snow model and streamflow generated by the deep learning karst model, and (4) the appropriate source code of models.

Besides the source code, the primary data types that will be collected include: non-spatial tabular data (e.g., previous tracer study results); time series of observations from fixed monitoring locations such as stream gages and weather stations as well as model simulated streamflow (0-D+time); raster and feature geospatial datasets such as digital elevation models, geological maps, and soil maps (2-D); time-varying spatial data such as gridded meteorological data and UEB model outputs (snowmelt).

Data and metadata formats, standards and organization

We will employ a tiered data management strategy to handle field, transformed, and model output data.

Field data include discharge and chemical data collected by the project team at gaging stations, measured water sample stable isotope ratios, tritium, and chlorofluorocarbon (CFC) concentrations. It will also include results from preceding hydrogeologic studies and externally sourced streamflow, watershed physiography, and meteorological data. Externally sourced data will be stored in the original format (digitized) along with metadata containing data collection methodologies or citations. USU PI Neilson has collected five years of streamflow at the sites outlined in Fig. 2. Additionally river and spring ion and isotope samples have been collected over a similar time frame. The raw and quality controlled data are posted on http://lro.usu.edu and via the Consortium of Universities for the Advancement of Hydrologic Sciences (CUAHSI)'s HydroShare data repository (http://www.hydroshare.org). New field data collected within the proposed project will similarly be integrated with the above existing archives. We will standardize our site naming conventions, variable names, and file naming conventions before integrating the current and future data sets. Raw data will go through already established QA/QC procedures developed as part of the Logan River Observatory.

Transformed data are quantities derived from field data and include temporally aggregated streamflow data, spatially interpolated and/or downscaled meteorological products, and watershed physiography data. Raw isotope ratios are calibrated to international standards following standard practices and reported as delta values ($\delta^{18}O$, $\delta^{2}H$) with one standard deviation uncertainties. Beyond maintaining the metadata for the raw data, we will prepare documentation so that anyone could produce a similar output given a variety of approaches or computational tools.

Model outputs include snowmelt/rain (2D+time) and discharge rates (0D+time) simulated by the UEB and deep learning karst models, respectively. The model output data will have similar documentation and metadata needs as transformed data.

Data will be formatted according to data types. Source code will be stored using the format of the corresponding programming language. Non-spatial tabular data and 0-D time series will be stored using easily readable data formats such as comma separated value (CSV) text files. For these data, we will use the Observations Data Model 2 (ODM2, *Horsburgh et al.*, [2016]), an existing standard supported in HydroShare. 2-D GIS data will be stored using the shapefile format, which is compatible with most GIS software. We anticipate that hierarchical databases (i.e., NetCDF) will be applied for 2-D+time data. The GIS data will conform to the data model implemented in HydroShare [*Horsburgh et al.*, 2015] that is an extension of the Dublin Core metadata standard [*DCMI*, 2012].

Data Storage and Sharing

Throughout the project timeframe, data produced will be stored at firewalled, cloud-based internal data storage system of each institution, with an extra copy at ASU. Because of the high temporal and spatial resolution, raw UEB output data will require a lot of storage space (~1TB per 10 year for the Logan River watershed). Therefore, we will coarsen the resolution for storage purposes. We plan to deposit and publish the final products from this project in CUAHSI's HydroShare data repository within six months from the conclusion of the project. Our datasets published on HydroShare will be assigned a permanent and citable digital object identifier (DOI) and become publicly accessible. Both the ASU internal data storage and HydroShare follow a multi-node structure. We anticipate that the two data repositories will meet the 10-year minimum data preservation standard without additional funds beyond the project period. Source code developed by this project will also be released and hosted in GitHub (https://github.com/), the leading software development and sharing platform.

All data collected and generated within this project will be made available for any intended use by other researchers in our fields and people beyond the original research community, such as water resources managers (see proposal Broader Impacts). A recommended formal citation to the HydroShare depository DOI and/or a relevant publication will be provided to facilitate reuse tracking and fair credit to data providers. In order to ensure data transparency, data will be distributed along with accurate and complete documentation that describes the data collection and generation methods.

Roles and Responsibilities

Our project uses each PI as the data manager for their part of the project to ensure that the collection-transformation-modeling-archiving processing stream for data is seamless and consistent from year to year throughout the project. The overall point of contact for all data sets archived on this project is PI Tianfang Xu, tianfang.xu@asu.edu. Undergraduate and graduate students will go through a training program for data reproducibility and proper storage.