Toward Reproducible and Interoperable Environmental Modeling: Integration of HydroShare with Server-side Methods for Exposing Large-Extent Spatial Datasets to Models

Article *in* Environmental Modelling & Software · October 2024

DOI: 10.1016/j.envsoft.2024.106239

14 authors, including:

Young Don Choi
K-water
11 PUBLICATIONS **139** CITATIONS
SEE PROFILE

Iman Maghami
University of Virginia
11 PUBLICATIONS **75** CITATIONS
SEE PROFILE

Jonathan L. Goodall
University of Virginia
180 PUBLICATIONS **4,240** CITATIONS
SEE PROFILE

Larry Band
University of North Carolina at Chapel Hill
48 PUBLICATIONS **934** CITATIONS
SEE PROFILE

Toward Reproducible and Interoperable Environmental Modeling: Integration of HydroShare with Server-side Methods for Exposing Large-Extent Spatial Datasets to Models

Young-Don Choi[a,b], Iman Maghami[b,c,*], Jonathan L. Goodall[b], Lawrence Band[b,d], Ayman Nassar[e], Laurence Lin[b], Linnea Saby[b], Zhiyu Li[f], Shaowen Wang[g], Chris Calloway[h], Hong Yi[h], Martin Seul[i], Daniel P. Ames[c], David G. Tarboton[e]

[a] AI Research Laboratory, R&D Management Department, K-water Research Institute, South Korea
[b] Department of Civil and Environmental Engineering, University of Virginia, Charlottesville, Virginia, USA
[c] Department of Civil and Construction Engineering, Brigham Young University, Provo, Utah, USA
[d] Department of Environmental Science, University of Virginia, Charlottesville, Virginia, USA
[e] Department of Civil and Environmental Engineering, Utah Water Research Laboratory, Utah State University, Logan, Utah, USA
[f] Center for High Performance Computing, University of Utah, Salt Lake City, Utah, USA
[g] Department of Geography & Geographic Information Science, University of Illinois at Urbana-Champaign, IL, USA
[h] Renaissance Computing Institute, University of North Carolina at Chapel Hill, North Carolina, USA
[i] Consortium of Universities for the Advancement of Hydrologic Science, Inc, Arlington, Massachusetts, USA

* To whom correspondence should be addressed (E-mail: im3vp@virginia.edu; Address: University of Virginia, Department of Civil and Environmental Engineering, University of Virginia, 151 Engineers Way, P.O. Box 400747, Charlottesville, VA, 22904, USA)

This is the accepted version of the article published in final form at:

Choi, Y.-D., I. Maghami, J. L. Goodall, L. Band, A. Nassar, L. Lin, L. Saby, Z. Li, S. Wang, C. Calloway, H. Yi, M. Seul, D. P. Ames and D. G. Tarboton, (2025), "Toward reproducible and interoperable environmental modeling: Integration of HydroShare with server-side methods for exposing large-extent spatial datasets to models," Environmental Modelling & Software, 183: 106239, https://doi.org/10.1016/j.envsoft.2024.106239.

Highlights

- Environmental modelling studies often require use of massive spatial datasets.
- Sharing these data through data services can support replicable modelling.
- HydroShare facilitates large data sharing via GeoServer and THREDDS data services.
- We test model replicability using HydroShare with a data-intensive hydrologic model.
- The methods and modeling approach consistently produced replicable model outputs.

Abstract

Reproducible environmental modelling often relies on spatial datasets as inputs, typically manually subset for specific areas. Yet, models can benefit from a data distribution approach facilitated by online repositories, and automating processes to foster reproducibility. This study introduces a method leveraging diverse state-scale spatial datasets to create cohesive packages for GIS-based environmental modelling. These datasets were generated and shared via GeoServer and THREDDS Data Server Connected to HydroShare, contrasting with conventional distribution methods. Using the Regional Hydro-Ecologic Simulation System (RHESSys) across three U.S. catchment-scale watersheds, we demonstrate minimal errors in spatial inputs and model streamflow outputs compared to traditional approaches. This spatial data-sharing method facilitates consistent model creation, fostering reproducibility. Its broader impact allows scientists to tailor the method to various use cases, such as exploring different scales beyond state-scale or applying it to other online repositories using existing data distribution systems, eliminating the need to develop their own.

Keywords

Large spatial datasets, Server-side data access, Reproducible environmental modeling, HydroShare, GeoServer, THREDDS Data Server

1. Introduction

Reproducibility is fundamental to accumulating knowledge and advancing science (Baker, 2016; National Academies of Sciences, 2019; Stagge et al., 2019; Wilkinson et al., 2016). In one survey, 1500 researchers were polled to assess the issue of reproducibility in scientific research. The results showed that about 70% of researchers across different science fields had failed to reproduce another researcher's results, and 50% failed to reproduce their own research results (Baker, 2016). Although data management and stewardship are essential elements to improve reproducibility (Wilkinson et al., 2016), another survey related to the success and challenges of data scientists reported that data scientists spent 19% of their time collecting data (finding and accessing) and 60% of their time cleaning and organizing data to fit data into their models (CrowdFlower, 2016). That leaves only 21% of their time for core analysis. To overcome these problems, the Findable, Accessible, Interoperable, and Reusable (FAIR) principles have been presented as high-level guidelines to improve scientific data management and access (Wilkinson et al., 2016). Ongoing efforts on FAIR guiding principles have advanced online data repositories with identifier mechanisms, data management plans, policies, and standards (Hodson et al., 2018).

In recent years, the number of online repositories adopting FAIR principles has grown (Crosas, 2020; Wilkinson et al., 2017). For example, Dataverse (https://dataverse.org/) provides the ability to create DOIs, share metadata and data files, and access data with public licenses on their data landing pages. Similar to Dataverse, FigShare (https://figshare.com), Mendeley

(https://mendeley.com), and Zenodo (https://zenodo.org/) offer repositories that support these capabilities, aligning with the FAIR principles. These online repositories have been developed for researchers focused on data publication through small files, typically under 1 GB. Often these files are subsets of larger datasets, such as in the case of hydrologic and environmental modelling geospatial datasets of digital elevation, land use, soils, or other geographic information. By focusing only on sharing data through small files rather than providing the ability to Find, Access, Interoperate with, and directly Reuse subsets of larger datasets, online data repositories are missing an opportunity to foster more reproducible science. Accommodating seamless data access, where model inputs can be reproducibly subset for an area of interest from larger-extent datasets via server-side methods, enhances reusability and interoperability of common datasets across multiple applications and case studies.

In the hydrologic and environmental science community, the types of data and models, and thus the data and file-sharing needs of the modelling community, are diverse (Horsburgh et al., 2016). HydroShare (https://www.hydroshare.org) helps to meet these diverse needs by providing an online repository to support sharing these multiple types of data and models across various environmental disciplines (Maghami et al., 2024; Morsy et al., 2017; Tarboton et al., 2024). The platform accommodates a wide array of data types, including time series, geographic features (Shapefile), geographic raster data (GeoTIFF), and multidimensional space-time datasets such as Network Common Data Form (NetCDF). In addition, HydroShare defines a computational model with two components to support model and model data sharing: 1) model program, i.e., source code or compiled software with related metadata such as version, programming language, and release date, and 2) model instance, i.e., model input and output data with related metadata such as application methods and a relationship to a model program. After creating a HydroShare

resource, users can share it using a unique uniform resource locator (URL) or DOI (after the resource is published).

Moreover, the HydroShare Python Client, hsclient ("HydroShare/hsclient: HydroShare Python Client," 2021), facilitates programmable interaction with HydroShare resources and JupyterHub computational environments (CUAHSI JupyterHub and CyberGIS-Jupyter for Water) for various analyses such as modelling and big data analysis using Jupyter notebooks (Choi et al., 2021). As such, HydroShare can be regarded as one of the most versatile online data repositories supporting FAIR principles in the field because it provides users with different features/tools: 1) unique identifiers (Findable), 2) metadata of multiple data and model types (Accessible), 3) hsclient (Interoperable), and 4) JupyterHub (Reusable). The platform is equipped to handle a broad range of data and models, making it suitable for various types of environmental modeling, with particular strength in hydrological studies. HydroShare is mainly used for data publication of a collection of smaller files at the file level and is not commonly used to distribute larger datasets (over a few gigabytes in size), sharing the "small files" limitation as other online repositories such as FigShare. One example of such support in HydroShare is for a national-scale network of time series data types through an external web service application to support access to the National Water Information System (NWIS) as a seamless dataset. Time series data rarely reach gigabyte sizes, and efforts to provide similar seamless access to gridded datasets, which can easily reach gigabyte sizes, through HydroShare are limited.

For national-scale spatial data sharing in the United States, several government-sponsored organizations and research centers have open-web data distribution systems. For example, the United States Department of Agriculture (USDA) National Resources Conservation Service (NRCS) provides over 100 high-resolution raster and vector data such as Census, Digital Elevation

Model (DEM), Hydrography, Land Cover, Soil, and Transportation in the Geospatial Data Gateway ("Geospatial Data Gateway," 2021). The United States Geological Survey (USGS) 3D Elevation Program (3DEP) ("USGS 3DEP," 2021) provides various elevation maps such as DEM and Lidar point clouds. The MRLC (Multi-Resolution Land Characteristics Consortium) ("MRLC," 2021) currently provides land cover, tree canopy, urban imperviousness, and other related data from 2001 to 2021. Also, while many web-based data distributed systems support Application Programming Interfaces (APIs) to access data and metadata for data interoperability programmatically, it is not uncommon for researchers to opt for graphical user interfaces (GUIs) to download needed data from these systems and then assemble and process it locally to support their modelling and analysis needs. This choice may be influenced by factors such as the ease of use of GUIs due to visualization advantages, the need for programming skills to effectively use APIs, and the potential for APIs to change or require updates over time, which may introduce additional maintenance requirements for the code.

The vision of using a Service-Oriented Architecture, which supports standardized and reusable data interchange components and open web-based data-sharing technologies, has been put forth as more convenient data access approaches and approaches for integrating certain environmental models (Chen et al., 2020). Achieving this vision is difficult given the complex and heterogeneous data requirements in hydrologic and the larger environmental modelling communities. Miles and Band (2015) provided one solution to the problem: the Ecohydrolib Python library for managing spatial data distribution and preparation workflows for ecohydrology modelling. Their approach was to access data from standardized, national data providers and to use data processing software to map these data into specific environmental model needs for a modeled watershed, eliminating the need to curate the spatial data in a unified intermediate

database. HydroTerre (Leonard, 2015) offered a different solution to the problem. It involved creating a centralized database of essential terrestrial variables and inputs that simulation models can use more directly. The database was built by gathering and processing data from multiple national-scale spatial datasets. While generic and model agnostic, the idea was demonstrated for creating inputs to the Penn State Integrated Hydrologic Modelling (PIHM) system (Kumar et al., 2010). Neither approach took advantage of the growing adoption of online data repositories like HydroShare to act as a standard gateway for data and model sharing, which has only been advanced in recent years.

Beyond these government-sponsored datasets, a growing number of datasets collected and maintained by national and international scientists could benefit from ways to share data online in a machine-readable way. Large sample hydrology studies (Addor et al., 2020) have become popular to cover large areas with consistent and robust high-quality datasets to "balance depth with breadth" (Gupta et al., 2014). For example, the Model Parameter Estimation Experiment (MOPEX) project provided hydrometeorological observation and attribute data for 438 catchments across the USA (Duan et al., 2006). Another example is the European catchments of Hydrological Predictions for the Environment model (E-HYPE), providing flow signatures and catchment attributes in 35,215 catchments and 1,366 river streamflow gauges in Europe (Kuentz et al., 2017). In a recent effort, Catchment Attributes and Meteorology for Large sample Studies (CAMELS, 671 catchments in the USA) (Addor et al., 2017; Newman et al., 2015) and CAMELS-Chile (516 catchments in Chile) (Alvarez-Garreton et al., 2018) were created to provide climate data and catchment attribute data. Computational platforms are beginning to be developed to support direct access to these geosciences' datasets. For example, PANGEO (https://pangeo.io) supports a community platform with big data in the climate, hydrologic, and ocean fields (Hamman et al.,

2018). Outside such systems, data must be downloaded manually and processed for use in particular environmental models.

A critical technological advancement that addresses this need for FAIR access to large-scale datasets in online data repositories is machine-to-machine protocols for subsetting and transmitting big data. Two examples are GeoServer (http://GeoServer.org) (Crawley et al., 2017; Youngblood, 2013) and Thematic Real-time Environmental Distributed Data Services (THREDDS) Data Server (TDS) (Gan et al., 2020; Unidata, 2024). GeoServer supports data access, display, and processing of geographic raster and feature data using the Open Geospatial Consortium (OGC) Web Map Service (WMS) (Michaelis and Ames, 2017; Wenjue et al., 2004). TDS is advanced client/server software that provides remote access to data and metadata stored in various geo-temporal datasets. These services are integrated into HydroShare (Lippold, 2019). This means users can more easily share, access, retrieve, and subset large geographical data via GeoServer and various types of scientific data such as NetCDF via TDS. Automatic metadata harvesting and data sharing functionalities in HydroShare enable user-uploaded NetCDF, geographic raster data, and feature data to be available through its connected GeoServer and TDS instances. This allows functionalities provided by GeoServer and TDS to be leveraged to visualize and analyze spatial data stored in HydroShare, such as clipping geographic features and interpolating elevation data. We have observed that these capabilities of GeoServer and TDS are underused in HydroShare. GeoServer is primarily used to visualize geographic raster and feature data online, and TDS is primarily used to share and visualize grid-based multidimensional climate data (Gan et al., 2020).

Ensuring data accuracy and reproducibility in environmental modeling is crucial, particularly when working with large-scale datasets from multiple sources. One challenge in data

integration involves managing spatial datasets that span various coordinate systems, which can lead to misalignments. Although georeferencing is a standard method for correcting these issues, its significance is often overlooked in large-scale models. Merging datasets across different coordinate systems can result in location shifts that impact data accuracy, potentially leading to errors in model outputs. Addressing georeferencing as a critical step in data integration is essential for maintaining consistency and reliability, particularly when working with spatial datasets that cover extensive areas. This consideration is especially relevant in large-scale environmental studies where precise data alignment is vital to accurate modeling.

This research aims to explore and demonstrate how HydroShare, with its integrations of GeoServer and TDS, can support more complex use cases required in environmental modelling. Specifically, we focus on hydrological challenges like integrating and analyzing spatial data at varying scales, where georeferencing is critical. In the remainder of this paper, when we refer to large datasets or similar terms, we specifically mean datasets at the state scale unless otherwise noted. By "state scale," we refer to the administrative boundaries of individual states within the United States. The states we use in this study range from 32,000 km^2 to 140,000 km^2. While these state-scale spatial datasets, obtained from national-scale datasets provided by government agencies, are smaller than datasets covering continental or regional extents, they are still valuable for our analysis. We use server-side methods, such as GeoServer and TDS, to expose large-extent spatial datasets to models, simulating watershed dynamics through streamflow simulation using the RHESSys model. The data collection, integration, and availability challenges we address in this study are part of the broader technical challenges faced in environmental modeling. Thus, our modeling objective tackles a small but essential subset of broader environmental modeling efforts.

This approach is a proof of concept and can be adapted to geographic boundaries, such as watershed boundaries or ecological regions, or applied to larger datasets in future studies.

With this goal in mind, we aim to answer the following research questions: 1) Can the GeoServer and TDS implementations with HydroShare enable more seamless datasets access to support more reproducible environmental modelling? 2) Can HydroShare, with GeoServer and TDS, provide a more sustainable and scalable solution for sharing machine-readable, large-extent spatial (LES) datasets? 3) Can the approach approve the consistency between a conventional approach and the use of GeoServer and TDS data server? The remainder of the paper is organized as follows. The background section provides information about GeoServer and TDS within the context of this study. The methods section presents the procedures for creating LES datasets and sharing them on GeoServer and TDS in HydroShare in the context of the Regional Hydro-Ecologic Simulation System (RHESSys) model (Tague and Band, 2004) as it combines diverse, interdisciplinary sources of spatial data that are available on national servers for application to smaller watershed extents. The results section explores the results of this effort with applications in North Carolina, Maryland, and Virginia. The discussion section reviews the advantages and limitations of using the HydroShare integration of GeoServer and TDS for LES data access. We conclude with a summary of the contributions of this research in the larger context of reproducible environmental modelling and suggest pathways for future research to further advance spatial data analysis in support of reproducible environmental modelling.

2. Background

2.1 GeoServer

GeoServer is a Java-based open-source software developed to publish and visualize spatial data online. Open Geospatial Consortium (OGC), a non-profit organization, has released standards for sharing spatial data online, including the Web Map Service (WMS), Web Feature Service (WFS), and Web Coverage Service (WCS). Using a simple HTTP interface, WMS provides geo-registered spatial images such as JPEG or PNG. WFS provides direct interoperability to discover, retrieve, and subset feature geographic data rather than sharing geographic data at the file level, such as downloading data. WCS is similar to WFS except that WCS provides direct interoperability to the raster geographic data while WFS provides direct interoperability to feature data. We used OWSLib ("OWSLib," 2021) to visualize, retrieve, and subset spatial data using various formats such as Shapefiles, ArcGRID, and GeoTIFF through OGC web services (i.e., WMS, WFS, and WCS). GeoServer has been used within HydroShare to support spatial data visualization. In earlier versions of HydroShare, GeoServer served as the spatial backend for the HydroShare GIS web app based on the Tethys framework (Crawley et al., 2017). This web app has since been deprecated in favor of providing preview maps of spatial data directly on the resource landing page using GeoServer WMS. This feature allows the sharing and visualizing public resources in HydroShare that contain spatial data. Every HydroShare resource that becomes public and contains geographic raster or feature content is automatically registered with GeoServer using customized middleware.

2.2 THREDDS Data Server (TDS)

The Thematic Real-time Environmental Distributed Data Services (THREDDS) Data Server (TDS) is open-source software distributed by the Unidata community program of the University Corporation for Atmospheric Research (UCAR) (Unidata, 2024). TDS provides web services that provide remote access to data and metadata stored in a variety of well-known, geo-temporal dataset formats used for environmental research such as GRIB (Gridded Binary), HDF5 (Hierarchical Data Format version 5), and, most commonly, NetCDF (Network Common Data Form) as viewed through a Common Data Model (CDM) (Nativi et al., 2008). TDS presents gridded, point, and time series datasets organized into thematic catalogs and provides data access through web services.

The TDS specifies two types of web service requests and responses, Data Access Protocol 2 (DAP2) and Data Access Protocol 4 (DAP4), for remotely accessing CDM datasets. DAP2 is widely used for its simplicity and compatibility with many existing tools, making it suitable for standard data access tasks like subsetting and requesting dimensional constraints. DAP4, on the other hand, is a newer version that supports more complex data structures and enhances performance, but it is not yet as widely supported as DAP2. This study focuses on DAP2 due to its compatibility and efficiency in handling our specific data access needs. Remote access to TDS-hosted datasets through DAP2 client software enables placing dimensional constraints on the CDM variable arrays transported in a response. The DAP2 request contains the constraints and effectively creates a subset of the TDS-hosted dataset for transport. The requesting client, therefore, need not concern itself with the size of the TDS-hosted dataset but only with the size of the data response.

As an additional advantage, DAP2 clients, such as Unidata's NetCDF libraries or higher-level software utilizing those libraries such as xarray (Hoyer and Hamman, 2017), initially make requests for only the metadata contained in the CDM header via a DAP2 Data Descriptor Structure (DDS) request and do not further request data until the variable array data is instantiated in the client via a DAP2 Distributed Oceanographic Data Systems (DODS) request. This "lazy-loading" behavior allows remotely opening the entire dataset but only transports portions of the dataset as needed programmatically, thereby reducing network load, transmission time, and memory consumption.

3. Methods

The overall modelling workflow used in the study is presented in Figure 1. The figure shows how datasets can be shared through HydroShare with different data management and distribution systems. All files in HydroShare are stored in the Integrated Rule-Oriented Data System (iRODS), a distributed data storage and management system that allows for the fast parallel transfer of large datasets (Yi et al., 2018). Public spatial datasets are also automatically replicated into GeoServer and TDS in HydroShare. One goal of this study is to explore how to retrieve spatial datasets from GeoServer and TDS to support more reproducible environmental modeling. To demonstrate this, we present an example application using LES datasets within RHESSys on CyberGIS-Jupyter for Water computing gateway, which is a well-tailored CyberGISX (Yin et al., 2017) instance to support data-intensive and reproducible research in the environmental modeling community. Further details on these steps are provided in the following subsections.

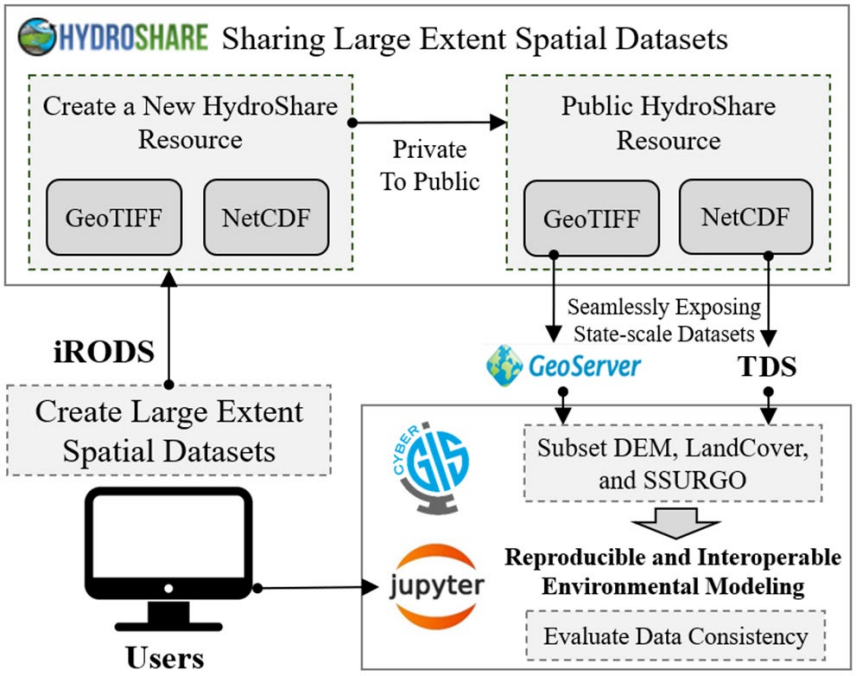

Figure 1. The workflows to create, share, subset, apply, and evaluate LES datasets for more reproducible environmental modelling.

3.1 Create and Share LES Datasets

3.1.1 Collect Spatial Data

Environmental models often require spatial datasets, including Digital Elevation Models (DEMs), land cover, and soil maps, to generate model inputs (DeVantier and Feldman, 1993). DEM data are used to delineate a watershed and extract spatial attributes such as flow direction, slope, aspect, and flow accumulation. Land cover is used to calculate surface roughness, evaporation, and transpiration according to the different land cover types, including impervious surfaces and vegetation type and extent. (e.g., agricultural crops, pasture, forest communities). Soil map data are used to generate parameters required to calculate water movement through soil, including infiltration, groundwater recharge, lateral transport, and soil hydraulic information

required to estimate evaporation and transpiration. Many web-based spatial data distribution systems currently provide low to high-resolution spatial data. They provide low-resolution data services such as the 90 m resolution Shuttle Radar Topography Mission (SRTM) ("SRTM," 2021) and the Google Earth Engine Datasets ("Google Earth Engine," 2021) for various earth science data and analysis. National-scale data is supported by federal government organizations such as USGS and USDA, as mentioned in the introduction section. High-resolution data services, most often hosted by specific research centers or state government organizations, provide 1-m or even higher resolution DEM and land cover data.

Depending on the specific application and the size of the watersheds, hydrologists often use 10-m or 30-m resolution DEM data in GeoTIFF format, especially when modeling state or larger scale areas where finer resolutions may not be practical or add significant value. For collecting DEM data to support modelling, we tested different data distribution interfaces and selected the Geospatial Data Gateway (GDG), which is operated by inter-government cooperation between the three service center agencies: Natural Resources Conservation (NRCS), Farm Service Agency (FSA), and Rural Development (RD). We selected GDG because it distributes data by various spatial units, including states, counties, bounding box, and custom areas of interest. For collecting land cover data, we emphasized data continuity for various applications such as land cover change. Therefore, we selected MRLC, a group of federal agencies, because this product provides consistent and reliable land cover information from 2001 to 2016 at the national scale. Finally, to collect soil data, we selected GDG. There are three soil datasets in GDG: 1) the National Soil Geographic Database (NATSGO), which is a very general soil map of the entire U.S.; 2) the State Soil Geographic Database (STATSGO), which is a more detailed state-wide map; and 3) Soil Survey Geographic Database (SSURGO), which is the most detailed county-level data. We

obtained the most detailed soil data from SSURGO to support watershed modeling use cases. In addition, we collected attribute data of SSURGO relevant to environmental models. SSURGO GeoTIFF datasets have a Mukey (Map unit key, the index to link different soil metadata tables) for each cell. We selected five SSURGO attribute tables that RHESSys required to link with the MUKEY in the GeoTIFF: 1) mapunit (mukey table), 2) chorizon (horizon table), 3) chtexgrp (horizon texture group table), 4) chtextur (horizon texture table), and 5) comp (component table). These tables are distributed at the county level through the Web Soil Survey web distribution system (Soil Survey Staff, 2021), which is associated with the GDG. Therefore, we downloaded county-level SSURGO metadata for each of the three states and merged them into a single SSURGO attribute table that can be joined to the GeoTIFF through the Mukey attribute.

While DEM, Land Cover, and SSURGO soil datasets are systematically available and regularly updated, there are important considerations regarding their availability and quality. For example, depending on the source, DEMs can vary in resolution and accuracy, which may affect the precision of watershed delineation and other spatial analyses. Land Cover datasets, although typically updated every few years, may only partially capture rapid land use changes, potentially leading to outdated or less accurate modeling inputs. Similarly, while detailed, SSURGO soil data may have varying completeness levels across regions, which could influence the accuracy of soil-related parameters in the model. These factors highlight the importance of selecting and evaluating datasets to meet the specific hydrological modeling objectives. A fuller exploration of the availability and quality of these datasets is outside the scope of this paper; as such, a discussion could form the basis for a separate study on data limitations and quality.

After selecting a source for obtaining the spatial datasets, we next selected the best scale (national, state, or local) for storing the datasets in HydroShare. In making this decision, we

considered 1) the file size of the spatial datasets, 2) the capabilities of GeoServer and TDS at handling different-sized datasets, and 3) the reusability of applications across different watersheds. Ultimately, we decided that aggregating the spatial data at the state scale would be best for the following reasons. Initially, we considered the feasibility of using national-scale spatial data within GeoServer and TDS because this would allow for truly reproducible environmental modelling. The size of the national-scale 30 m resolution DEM, land cover dataset, and SSURGO dataset at the national scale are 44.6, 20, and 3.7 GB, respectively. We could not find specific guidelines that specify the maximum size of datasets that can be distributed using GeoServer and TDS. Based on our experience, due to memory limitations, GeoTIFF files exceeding 1 GB can be challenging to manipulate and convert to NetCDF on most personal computers. Therefore, while a national-scale dataset is technically feasible, it may be impractical to work with and maintain. Our decision to avoid using national-scale datasets was driven by these technical constraints, which limit practical processing rather than the typical file size limitation for sharing data in many online repositories, as discussed in the introduction. Given the challenges with the national scale, we then considered state-scale data aggregations. In this case, using Virginia as an example, the DEM is 951 MB, land cover is 342 MB, and SSURGO is 157 MB. The process of uploading and subsetting the data on GeoServer and TDS was simplified at this file size. For datasets smaller than state-scale, it will be difficult to support more reproducible modelling because many watersheds cross county boundaries. For these reasons, we chose state-scale as a useful spatial aggregation for distributing LES datasets, while other similarly sized spatial aggregations, like large watersheds, could also have been used. We obtained 10-m or 30-m DEM and 30-m SSURGO spatial data from GDG. Also, we collected 30-m land cover spatial data in 2001, 2003, 2006, 2008, 2011, 2013, and 2016 from MRLC to create the LES datasets (Figure 2). The following subsections describe how these

data were processed to have a consistent spatial reference system and uploaded and shared through HydroShare.

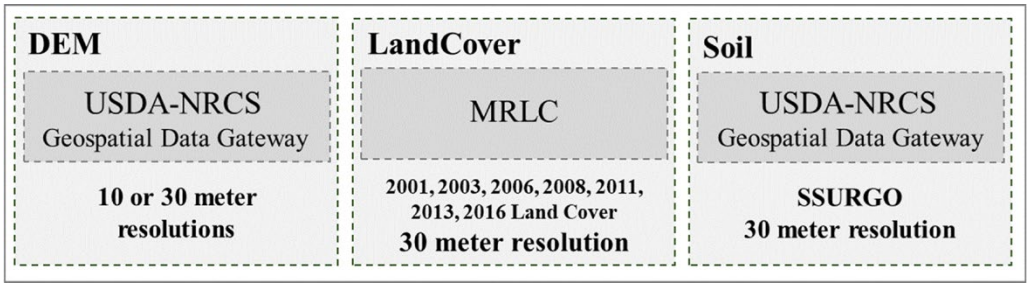

Figure 2. The selection of data distribution systems and spatial data to create LES datasets

3.1.2 Create Consistent LES Datasets

In this step, we considered the following factors in creating consistent state-scale spatial datasets for use in GeoServer and TDS: 1) using data types optimal for GeoServer and TDS, 2) adopting a consistent coordinate system across all spatial data, 3) applying georeferencing transformations to adjust shifted locations in the merged state-scale DEM, and 4) having complete and meaningful metadata compatible with HydroShare for each dataset.

We adopted GeoTIFF as the spatial data type for the LES datasets on GeoServer and NetCDF for TDS. GeoTIFF is an OGC implementation standard for raster data commonly used to store grid-based spatial data with geographic metadata that describes the spatial location, including spatial extent, coordinate reference system, and resolution. Most data providers serve DEM and land cover data as GeoTIFF, which is also one of the primary data types GeoServer supports. For SSURGO data, the information is typically served at the county scale using a shapefile format with soil attribute metadata. We explored merging the county-scale SSURGO shapefiles into a state-

scale shapefile and serving the data through GeoServer. Doing so resulted in a large dataset that is difficult to manage and feed into environmental models. Fortunately, since Feb 2021, the USDA NRCS National Soil Survey Center has started to provide national-scale SSURGO data as a 30-m resolution GeoTIFF. The GeoTIFF format can store a collection of 2D arrays, which makes it easily transferrable into a NetCDF multidimensional array. Given that TDS supports the NetCDF format in the background, we decided to use GeoTIFF in GeoServer and NetCDF in TDS as the data types for distributing the LES copies of the DEM, land cover, and soil spatial datasets at a state scale.

Each dataset obtained from the federal agencies was originally provided in a different spatial coordinate system. Adopting a consistent coordinate system is essential to create a unified LES dataset to support environmental modelling. We adopted the Universal Transverse Mercator (UTM) geographic coordinate system for consistency at the state-scale. Resampling is required when transforming coordinate systems. For resampling of the LES datasets, we used a bilinear interpolation resampling method for the DEM because it is continuous data and used a nearest neighbor resampling method for the land cover and soil data because they are categorical data, meaning they are organized into discrete categories or classes (e.g., land cover types or soil types).

When multiple DEMs are merged into one DEM, there are often overlapping areas at the edges of each original DEM. These areas made the merged DEM (GeoServer and TDS LES DEM) shift about 0.3-m to 1.0-m horizontally compared to the original raw DEM. If users delineate a watershed using the merged LES datasets without recognizing these changes, the watershed boundary may not match the boundary delineated using the original data products. This may sometimes allow environmental models to execute without producing any fatal errors, depending on the specific application. To address these potential issues, we applied a georeferencing tool to

transform spatial data in ArcGIS to linearly shift the merged DEM back to its original location using control points. It is essential to note that the importance of georeferencing becomes even more apparent when non-fatal errors are less noticeable. Unlike cases where changes in the watershed's shape are easily recognized, modelers may overlook discrepancies without prior studies for comparison and validation. This oversight can lead them to assume that the LES datasets are identical to the original data, potentially resulting in erroneous model setup and parameter tuning. This underscores the critical need for a rigorous georeferencing step in the process.

Figure 3 shows the complete workflow using the steps described above to create the LES datasets in the GeoTIFF and NetCDF formats. First, each state's data was projected using the appropriate UTM zone coordinate system. We used a projected coordinate system because environmental models use length units, such as meters, instead of degrees used in a geographic coordinate system. After creating the state-scale merged DEM, we applied a georeferencing tool to adjust its spatial alignment to match its original location. Land cover data are distributed at the national scale; therefore, we extracted state-scale land cover datasets and projected the data to the same coordinate system as the state-scale DEM. Finally, SSURGO data, distributed at the national scale using a USA Contiguous Albers Equal-Area Conic USGS version, was clipped to the state scale and projected to the same coordinate system as the DEM and land cover data.

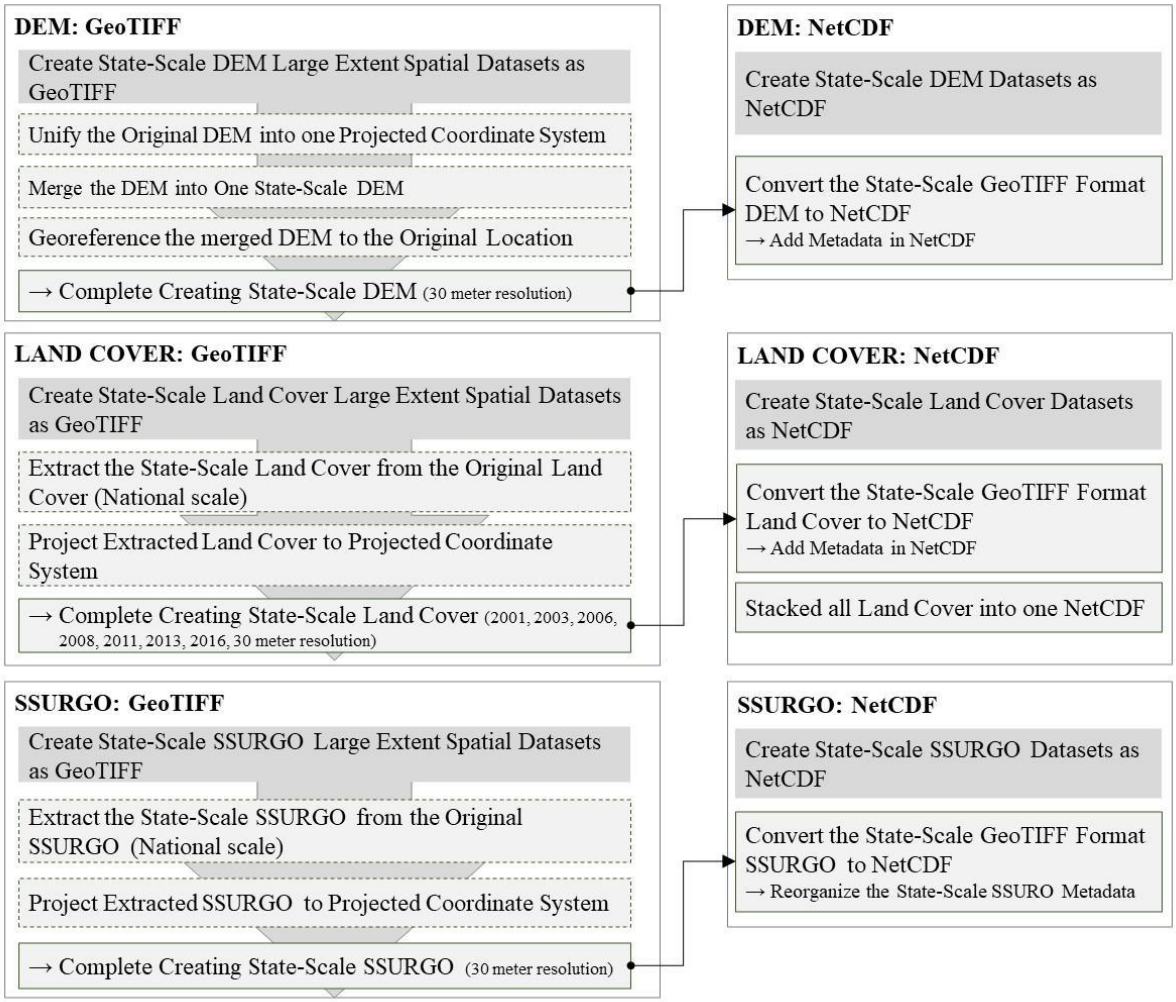

Figure 3. The workflows to create the LES in GeoTIFF and NetCDF format.

After creating the LES datasets, we added metadata directly within the NetCDF datasets to capture essential information about the changes made in the original metadata file distributed by GDG. We added 16 metadata fields, including data title, bounding coordinates, grid coordinate system name, UTM Zone number, scale factor at central meridian, and horizontal datum name.

3.1.3 Share Datasets in HydroShare

For this step, we reviewed and used three tools to share the LES datasets through HydroShare: 1) iRODS to fast transfer large datasets (over 1 GB) into HydroShare, 2) OWSLib

Python libraries to make the LES datasets interoperable via GeoServer, and 3) xarray (Hoyer and Hamman, 2017) Python libraries to make the LES datasets interoperable via TDS.

The first step was to upload the datasets into a new HydroShare resource. This step is trivial if the size of datasets is under 1 GB as datasets can then directly be uploaded through the HydroShare web browser user interface. For datasets exceeding 1 GB, users must employ iRODS for fast transfer of large data into HydroShare. Various iRODS clients, such as icommands and Cyberduck, serve as effective tools for parallel transferring substantial datasets via multiple threads into the HydroShare iRODS user space. We used Cyberduck for its user-friendly interface and convenience. After uploading the LES datasets into the HydroShare resource, datasets were automatically recognized with the proper aggregation type of geographic raster (GeoTIFF) or multidimensional contents (NetCDF). The content type metadata, such as title, keywords, spatial/temporal coverage, and spatial reference, and variable metadata, were automatically extracted by HydroShare as part of the data upload process. When a HydroShare resource is made public, the spatial datasets are automatically provided through GeoServer and TDS.

After the LES datasets are available on HydroShare and through the linked GeoServer and TDS access points, users can quickly discover and programmatically interact with the LES datasets from GeoServer and TDS using the newly created HydroShare resource ID. Users can use OWSLib to request subsets of data from GeoServer. In TDS, users can use xarray to subset specific data of interest. Subsequently, users can convert the NetCDF output from TDS into GeoTIFF format using the rioxarray package(“rioxarray,” 2021). This conversion allows the data to be used as input for environmental models that expect GeoTIFF raster inputs.

3.2 Example Application for an Environmental Model Use Case

We used an example application to demonstrate the data service and how it supports reproducible environmental modelling workflows. Figure 4 depicts the workflow steps for seamlessly applying LES datasets as model inputs, which include the DEM, extracted land cover, and soil texture maps. The development of parameters from the spatial data, execution and visualization modeling workflow was modified from Miles and Band's (2017) RHESSys Workflow. As part of our evaluation of data consistency between LES datasets (TDS and GeoServer) and the conventionally accessed spatial data provided by federal agencies, we incorporated daily streamflow outputs from the RHESSys model (Tague and Band, 2004) as an illustrative example. RHESSys is a GIS-based hydro-ecological modelling framework for simulating carbon, water, and nutrient fluxes. In this evaluation, we also used the newly prototyped pyRHESSys(Choi and Lin, 2021), an API for RHESSys that provides programmatic control over model input creation, manipulation, execution, and output visualization and analysis.

We compared three approaches for accessing the spatial data required to parameterize the RHESSys model: 1) the data manually subset from data provided by federal agencies and conventionally shared at file level, 2) the data processed and distributed through GeoServer in the GeoTIFF format, and 3) the same data processed and distributed through TDS in a NetCDF format. We compared the watershed DEM, extracted land cover, and SSURGO data to evaluate data consistency across the three approaches. To ensure a fair comparison, all model parameters and parameterization processes remained constant across the three data input approaches. The hydrological processes in RHESSys were driven by daily forcing datasets, including precipitation, maximum and minimum temperature, vapor pressure deficit, relative humidity, and direct downward shortwave radiation. These forcing datasets were sourced from Daymet (Thornton et

al., 2022) for Scotts Level Branch, MD, and Spout Run, VA, while data for Coweeta Subbasin 18, NC, was obtained directly from the Coweeta Hydrologic Laboratory (personal communication, November 12, 2021). After executing RHESSys using data from these three approaches as input, we also assessed the impact of the three data access approaches on model output, particularly on daily streamflow outputs. The goal was to confirm that the proposed seamless data distribution approach and reproducible workflows are consistent with a careful, manual modelling process.

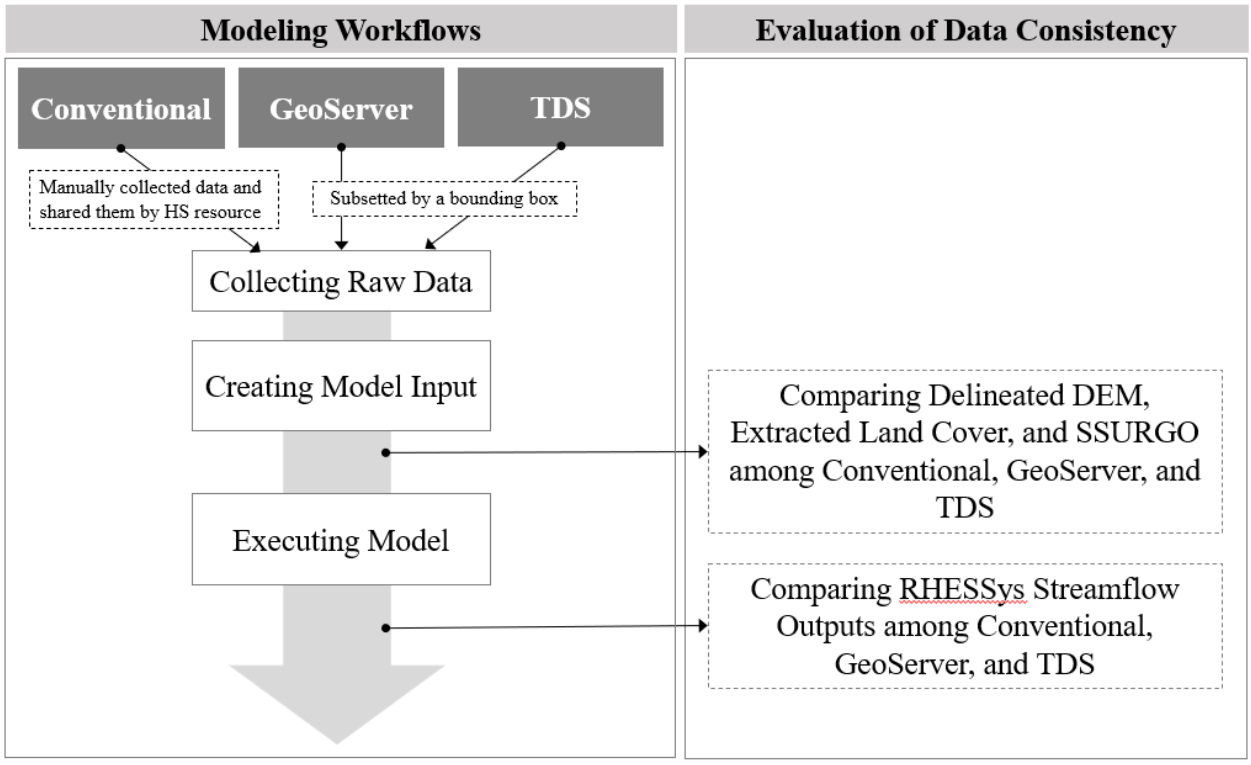

Figure 4. Workflows for reproducible RHESSys modelling and evaluation of data consistency using LES datasets.

4. Results

This section presents results from the example application where the proposed reproducible workflows are applied to three different watersheds using the RHESSys model on CyberGIS-Jupyter for Water computing gateway. First, we describe the three example study watersheds. Next, we present the HydroShare resources for sharing the datasets and workflows. Then, we present the developed workflows as Jupyter notebooks that show the steps to create and subset LES datasets based on Figure 1 in the methods section. Finally, we present results from the method evaluation described in Figure 4 that aims to measure data consistency across three approaches for distributing spatial data and how each method impacts the results of a reproducible RHESSys model run.

4.1 Example Watersheds

The three study watersheds are 1) Coweeta Subbasin18, NC (A = 0.126 km^2, DEM resolution: 10 m), 2) Scotts Level Branch, MD (A = 8.36 km^2, DEM resolution: 30 m), and 3) Spout Run, VA (A = 55.42 km^2, DEM resolution: 60 m), all of which are shown in Figure 5. The United States Forest Service Coweeta Hydrologic Laboratory has been measuring hydrologic and ecologic variables for gaged catchments since 1934 ("USFS Coweeta Hydrologic Laboratory," 2021). Subbasin18 is a forest-dominated basin. Coweeta catchment with steep topography that is well-studied in hydrologic studies, leading to long-term records of several hydrologic and environmental parameters and variables, including observed streamflow in the subbasin. For this reason, we chose it as one of our study watersheds. Scotts Level Branch, located near Baltimore, Maryland, has a USGS streamflow observation station (USGS 01589290) and represents a suburban watershed. Spout Run is a mixed land-use watershed with substantial agricultural land cover in Northern Virginia and a USGS streamflow observation station (USGS 01636316). The watershed contains the Blandy Experimental Forest, which has one of the National Science

Foundation's National Ecological Observatory Network (NEON) sites. Using these three different-sized watersheds, we evaluated the applicability of the LES dataset distribution method as part of a reproducible modeling workflow. The simulation periods for each watershed are as follows: 1/1/2004 - 10/31/2014 for Coweeta Subbasin18, NC, 1/1/2007 - 9/30/2020 for Scotts Level Branch, MD, and 1/1/2004 - 12/31/2020 for Spout Run, VA (the dates reported herein follow the U.S. formatting convention of MM/DD/YYYY).

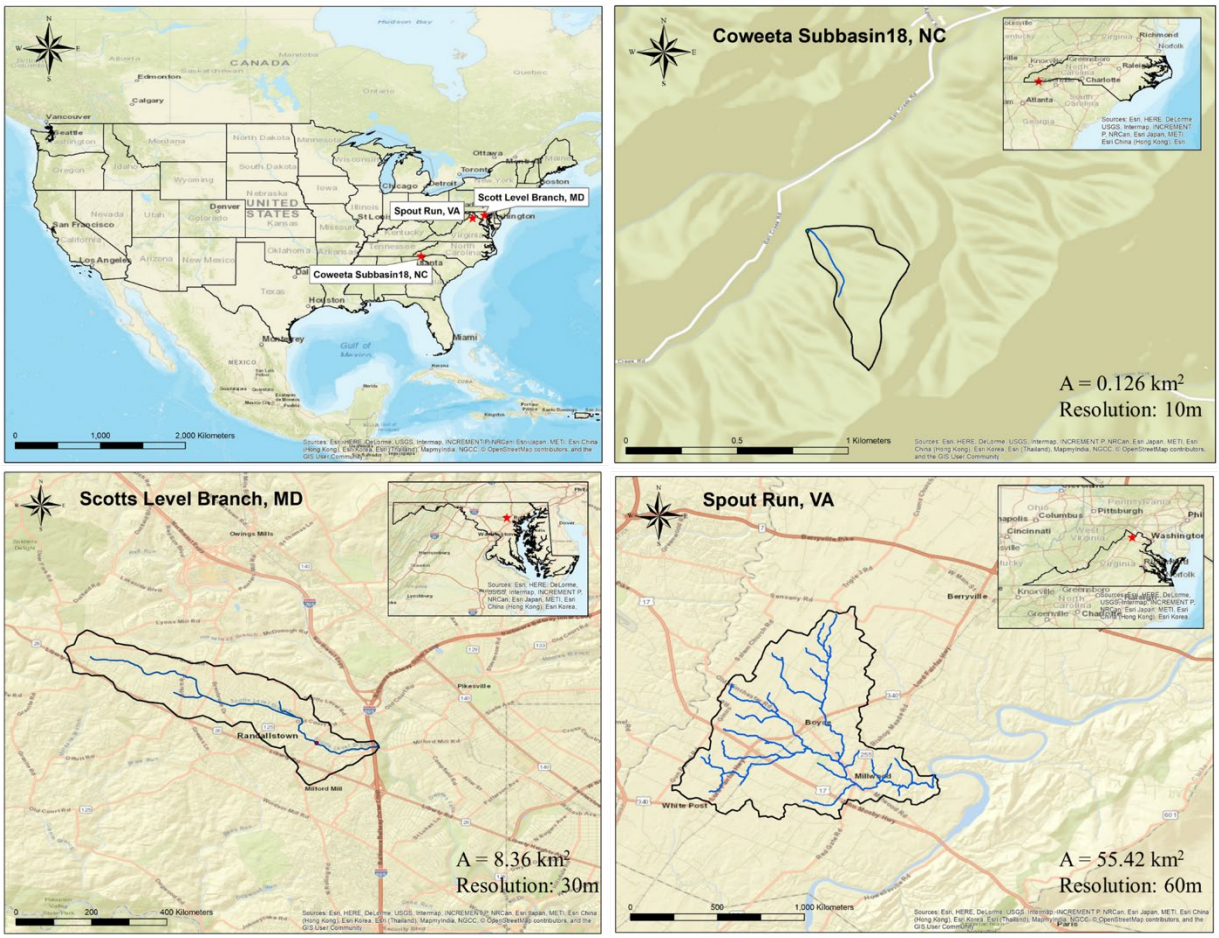

Figure 5. Three different sized watersheds to evaluate data consistency in different resolutions using LES datasets: 1) Coweeta Subbasin18, NC (A = 0.126 km^2, DEM resolution: 10 m), 2) Scotts Level Branch, MD (A = 8.36 km^2, DEM resolution: 30 m), 3) Spout Run, VA (A = 55.42 km^2, DEM resolution: 60 m)

4.2 Sharing the Datasets and Workflows in HydroShare

To share the datasets and workflows we used and developed, we created a HydroShare collections resource ((Choi et al., 2024)) referred to as HS 1. This collection resource comprises seven HydroShare resources (HS 2-8), each containing different datasets or workflows. To simplify referencing each HydroShare resource, we assigned an identifier known as the "HS

number." The interactions and dynamics of these eight HydroShare resources are elucidated in the Data Availability section using a visual diagram (Figure D1) that illustrates their interrelationships.

4.3 Creating the LES Datasets

We created Jupyter notebooks to automate the data processing workflow required to create the LES datasets for the three states (Choi et al., 2024) (HS 2). GIS processing was first done in these workflows to merge, extract, and project GeoTIFF data. We used ArcPy, a GIS Python package, to perform geographic data analysis, conversion, and data management in ArcGIS (Toms, 2015). After creating LES datasets in GeoTIFF format to be exposed by GeoServer, we converted GeoTIFF to NetCDF using xarray and rioxarray Python packages so that the same data used by GeoServer can be exposed by TDS which requires data to be in the NetCDF format. We used xarray to manipulate data type and add metadata in the NetCDF file and rioxarray to save GeoTIFF to NetCDF format. These procedures created three composite HydroShare resources (HS 3-5) to share LES datasets(Choi et al., 2024).

The automated workflows consist of three parts (DEM, land cover, and SSURGO), as we mentioned in Figure 3 in the methods section. In this subsection, we demonstrated how these general methods can be applied to create Virginia LES DEM in GeoTIFF format (Figure 6). Before starting this procedure, we created ArcPy Conda virtual environments from ArcGIS Pro 2.1. Then, we created Jupyter notebooks to capture these automated workflows (Figure 6). We imported required libraries such as ArcPy, xarray, and NumPy. After collecting 30-m resolution DEM from GDG, we unified the multiple projected coordinate systems of the original DEM into one projected coordinate system using the ProjectRaster_management module in ArcPy. In the case of Virginia, the DEM has UTM Zone 17N and 18N projected coordinate systems, so we unified them to UTM Zone 17N. We then merged each DEM into one state-scale DEM in GeoTIFF format using the

MosaicToNewRaster_management module in ArcPy. After that, we read the Virginia GeoTIFF file and created an xarray data format using the rasterio Python library. Then, we added metadata of the original DEM, such as spatial domain, UTM detail information, and geodetic model information, into NetCDF. Finally, we saved xarray format data to NetCDF. Following similar procedures, we created state-scale land cover and SSURGO using this Jupyter notebook. Due to the proprietary nature of ArcGIS Pro software, we developed these Jupyter notebooks (HS 2) on the Windows operating system, which means they cannot be directly used in the CyberGIS-Jupyter for Water computing gateway. These Jupyter notebooks can be employed by researchers interested in creating LES datasets, including for other states ore regions.

```
# Import required libraries
import arcpy
import glob
import xarray as xr
import os, shutil
import numpy as np

# Unify the original DEM into one projected coordinate system
proj_path = "C:/Users/Choi/Documents/large_sample_datasets/DEM_STATES/Virginia \
             /elevation_NED30M_va_3902660_02/Project_UTM17"
for i, value in enumerate(raw_dem):
    dataset = xr.open_rasterio(value)
    CRS = dataset.crs.split(":")[1]
    if CRS == '26918':
        arcpy.ProjectRaster_management(value, proj_path+"/"+ raw_dem[i].split("\\")[-1],
            "PROJCS['NAD_1983_UTM_Zone_17N',GEOGCS['GCS_North_American_1983',\
            DATUM['D_North_American_1983',SPHEROID['GRS_1980',6378137.0,298.257222101]],\
            PRIMEM['Greenwich',0.0],UNIT['Degree',0.0174532925199433]],\
            PROJECTION['Transverse_Mercator'],PARAMETER['False_Easting',500000.0],\
            PARAMETER['False_Northing',0.0],PARAMETER['Central_Meridian',-81.0],\
            PARAMETER['Scale_Factor',0.9996],PARAMETER['Latitude_Of_Origin',0.0],\
            UNIT['Meter',1.0]]", "NEAREST", "30 30", "", "", "PROJCS['NAD_1983_UTM_Zone_18N',\
            GEOGCS['GCS_North_American_1983',DATUM['D_North_American_1983',\
            SPHEROID['GRS_1980',6378137.0,298.257222101]],PRIMEM['Greenwich',0.0],\
            UNIT['Degree',0.0174532925199433]],PROJECTION['Transverse_Mercator'],\
            PARAMETER['False_Easting',500000.0],PARAMETER['False_Northing',0.0],\
            PARAMETER['Central_Meridian',-75.0],PARAMETER['Scale_Factor',0.9996],\
            PARAMETER['Latitude_Of_Origin',0.0],UNIT['Meter',1.0]]", "NO_VERTICAL")
    elif CRS == '26917':
        shutil.copy(value, proj_path)
    else:
        print("Need to add more coordinate system to project DEM")

# Merge the DEM into one state-scale DEM
merged_dem_name = "VA_DEM30m_UTM17.tif"
collect_path = "C:/Users/Choi/Documents/large_sample_datasets/Large_Datasets/Virginia"
arcpy.MosaicToNewRaster_management(proj_dem, collect_path, merged_dem_name, "", \
                                   "32_BIT_FLOAT", "", "1", "LAST", "FIRST")

# Open GeoTIFF using rasterio library
dem_dataset = xr.open_rasterio(os.path.join(collect_path, merged_dem_name))
dem_dataset

# Read metadata from a text file and add them into NetCDF
dataset1 = []
filename = os.path.join(raw_dem_path, 'gway_3902660_02_NED30M.txt')
with open(filename, 'r') as f_in:
    for items in f_in:
        dataset1.append(items.split("\n"))
new_dem_dataset.attrs.update({dataset1[4][0].split(":")[0].split()[0]: dataset1[4][0].split(":")[1]})

# Save xarray to NetCDF
new_dem_dataset.to_netcdf(os.path.join(collect_path,'VA_DEM30m_UTM17.nc'))
```

Figure 6. A workflow example: creating Virginia LES DEM in GeoTIFF and NetCDF formats (HS 2). Note that the figure displays a modified single cell for clarity and illustrative purposes, while the original code in the corresponding notebook in HS 2 is presented in multiple cells.

Using these automated workflows, we created GeoTIFF and NetCDF LES datasets for the DEM, land cover, and SSURGO in North Carolina, Maryland, and Virginia. For the DEM, we selected different resolutions for each state. Given the small drainage area of Coweeta Subbasin18 and to evaluate higher resolution data consistency, we generated 10 m resolution DEM GeoTIFF for North Carolina. We could not convert 10 m resolution DEM GeoTIFF (a DEM with higher resolution compared to other states) to NetCDF LES datasets due to computing memory limitation in a local computer and CyberGIS-Jupyter for Water. As a result, we could not use the TDS approach in North Carolina; instead, we had to employ LES DEM exclusively in GeoTIFF format via GeoServer. In Maryland and Virginia, we created 30 m resolution LES DEM in GeoTIFF format and converted it to NetCDF. We applied 30 m resolution DEM to Scotts Level Branch as an example for general resolution application. Considering the larger drainage area of Spout Run and for evaluating lower resolution data consistency, we resampled DEM from 30 m to 60 m resolution. For the land cover from MLRC and SSURGO from GDG, we created 30 m LES datasets for all three states because they only come with 30 m resolution GeoTIFF. After creating GeoTIFF LES datasets for DEM, land cover, and SSURGO, we converted them into NetCDF LES datasets. Given the NetCDF capacity to create time or variable stacked datasets within the same domain using dimension and coordinate structures, we created a consolidated land cover NetCDF LES dataset by stacking seven GeoTIFF LES datasets from 2001 to 2016 (2001, 2003, 2006, 2008, 2011, 2013, and 2016). This approach offers enhanced convenience and efficiency, enabling researchers to manage seven years of land cover data within a single NetCDF file.

Table 1 shows the file sizes and resolutions of GeoTIFF and NetCDF LES datasets in the three states. The original LES datasets retrieved from the federal agencies are not compressed files and are very difficult to handle. To minimize the file sizes, we used GeoTIFF and NetCDF

compressed formats. We used a LZW compression algorithm (Akoguz et al., 2016) to calculate GeoTIFF in ArcGIS. The nccopy tool, a command-line utility to compress NetCDF files, supports multiple levels of compression (level 0-9, where a high value represents high compression and requires more time) for variable data in NetCDF. We used the level 1 compression command "nccop -d1 input.nc output.nc," which effectively reduced the size of the original NetCDF. Specifically, the original size of the Virginia GeoTIFF DEM (30 m) was 1.53 GB, and the compressed size was 0.95 GB. The original size of Virginia NetCDF DEM (30 m) was 1.60 GB, and the compressed size was 0.78 GB. Therefore, using the compressed format makes it more convenient to share the LES dataset online.

Sharing LES datasets is convenient for seamless access to spatial data, though it is challenging to suggest appropriate dataset sizes considering the varying memory capacities of different computers. In the case of North Carolina GeoTIFF DEM (with a 10 m resolution), although the file size was 5.66 GB, there was no problem in creating a GeoTIFF DEM. Yet, as mentioned earlier in this section, we could not convert DEM from GeoTIFF to NetCDF for North Carolina due to a memory limitation preventing us from using TDS to access DEM. The memory capacity of the computer used to prepare datasets and the server hosting the data is crucial for creating, uploading, and interoperating large datasets. The servers hosting GeoServer and TDS require significant testing, considering the large volume and size of file transfers required concurrently across different users' requests. In addition to the guidelines mentioned in the methods section regarding challenges in manipulating and converting GeoTIFF to NetCDF due to memory limitations, our experimental findings are based on the specific experiments conducted during this research, further emphasizing the importance of considering data size for stable data interoperability.

Based on these experiments, it is apparent that maintaining stability for data interoperability via GeoServer is achievable for datasets below 5 GB. Similarly, for TDS, datasets up to 1 GB demonstrate stable performance. Despite the servers being able to handle larger sizes, adhering to these specified limits ensures more robust performance, particularly when faced with concurrent user requests. Suppose a user chooses to work with finer-scale data compared to our study. In that case, they are expected to encounter similar limitations more prominently and potentially face additional challenges we did not encounter. Depending on the specific computing resources available, these challenges could further complicate the practical use of finer resolutions.

Table 1. Compressed file sizes and resolutions of GeoTIFF and NetCDF in the three states

States		**DEM**		**Land Cover (7 Years)***		**SSURGO**	
		GeoTIFF	**NetCDF**	**GeoTIFF**	**NetCDF**	**GeoTIFF**	**NetCDF**
North Carolina	**File Size (MB)**	5,659	-	257	304	112	80
Maryland		358	294	134	165	52	44
Virginia		951	783	342	422	157	121
North Carolina	**Resolution (m)**	10	-	30	30	30	30
Maryland		30	30	30	30	30	30
Virginia		60	60	30	30	30	30

*The land cover GeoTIFF file size (MB) mentioned for each state in the table are the summation of sizes of the seven GeoTIFF files each of which represents one specific year but the NetCDF file size for each state are for one file which is stacked land cover of the seven years

4.4 Subsetting the LES Datasets

After creating three states' LES datasets and sharing them on the HydroShare GeoServer and TDS, we subset these datasets to collect spatial model input for specific watersheds for use in RHESSys preprocessing workflows. In Figures 7 and 8, we used Scotts Level Branch, MD, as a demonstrative example to illustrate the subsetting LES datasets on GeoServer and TDS within HydroShare. This subsetting procedure is shared through the RHESSys workflow notebooks (HS 7) in HydroShare (Choi et al., 2024).

4.4.1 Subset LES Datasets from GeoServer

We used OWSLib within the CyberGIS-Jupyter for Water environment to perform the GeoTIFF DEM subsetting in GeoServer. Figure 7 shows the process using the Maryland LES Datasets (GeoTIFF) in HydroShare as a specific example. First, we imported the required Python libraries to use the WCS service in GeoServer. Second, we requested GeoTIFF as an object using a WebCoverageService module in OWSLib. Then, we subsetted certain areas using a getCoverage method with a bounding box. Finally, we saved the subsetted object to GeoTIFF format.

```
# Import required libraries
from owslib.wcs import WebCoverageService
import shutil
import os

# Set name and HS resource ID
PROJECT_NAME = "SLB_30m"
name = "geoserver_dem30m.tif"
gis_folder = os.path.join(os.getcwd(), PROJECT_NAME, "gis_data")
resource_id = "4f5a33d96a004bd496747956c45cae7a"
dem_name = "MD_DEM30m_UTM18"

# Subset GeoTIFF using a WCS request from HydroShare-provisioned GeoServer
url = "https://geoserver.hydroshare.org/geoserver/
       wcs?service=WCS&version=1.1.0&request=GetCapabilities&namespace=HS-"+resource_id+":"+dem_name
dem = WebCoverageService("https://geoserver.hydroshare.org/geoserver/wcs", version='2.0.1')

dem_subset=dem.getCoverage(identifier=['HS-'+resource_id+":"+dem_name],
                           subsets=[('E',x_min, x_max), ('N',y_min,y_max)], format='image/tiff')

dem_tif=dem_subset.read()

f = open('./'+name, 'wb')
f.write(dem_tif)
f.close()
```

Figure 7. Subsetting Example: Scotts Level Branch from MD LES GeoTIFF Dataset from GeoServer using OWSLib (HS7: Jupyter notebookStep_2_Retrieve_Spatial_Inputs_and_TimeSeries_Data.ipynb. Note that the codes presented in this figure are simplified. In the original notebook, there are more details, including conditional statements based on the selected data access approach and study area, and the codes are spread across multiple cells)

4.4.2 Subseting LES Datasets from TDS

Subsetting land cover LES Datasets (NetCDF format) in TDS is more straightforward than in GeoServer (GeoTIFF format) because users can directly access TDS via xarray. Therefore, users can easily create xarray array format (*xarray.DataArray*) objects using an xarray *open_dataset* module. Then, they can subset the land cover data in their watershed and simulation year of interest, e.g., the Scotts Level Branch watershed for 2006, using slicing ranges for x and y to specify the UTM corner coordinates and years (Figure 8). Finally, users can convert xarray data array objects to GeoTIFF format using the rioxarry library. This conversion is a crucial step because the

subsetted data obtained from TDS serves as RHESSys model inputs, and it specifically requires GeoTIFF format for its input data.

```
# Import required libraries
import xarray as xr
import rioxarray
import shutil

# Read NetCDF using xarray
nlcd = xr.open_dataset('http://thredds.hydroshare.org/thredds/dodsC/hydroshare/resources/ \
                        4f5a33d96a004bd496747956c45cae7a/data/contents/MN_NLCD30m_UTM18.nc')

# Subset NetCDF using x, y range and land cover year
nlcd_subset = nlcd.sel(y=slice(4362188, 4357920), x=slice(340347, 349049), years=2006)

# Save xarray to GeoTIFF format
nlcd_subset.rio.to_raster("opendap_nlcd_2006_30m.tif")
```

Figure 8. Subsetting Example: Scotts Level Branch from MD LES NetCDF Dataset from TDS using xarray (HS 7: Jupyter notebook: Step_2_Retrieve_Spatial_Inputs_and_TimeSeries_Data.ipynb. Note that the codes presented in this figure are simplified. In the original notebook, there are more details, including conditional statements based on the selected data access approach and study area, and the codes are spread across multiple cells)

4.5 Evaluating Data Consistency

To evaluate the data consistency in three different watersheds with different spatial data resolutions: 10, 30, and 60 m, we created RHESSys input from LES datasets and executed RHESSys using a reproducible RHESSys Jupyter notebook (HS 7) (Choi et al., 2024). In these procedures, we created nine case studies using three different datasets resulted from the three data distribution approaches (conventional, GeoServer, and TDS) and three watersheds (Coweeta Subbasin18, Scotts Level Branch, and Spout Run). For the conventional approach, where spatial datasets were manually collected to represent a scenario where a user does not use an API to

retrieve data, we created three resources that contain model instances for each watershed in HydroShare (Choi et al., 2024) (HS 7). We then presented the evaluation results of data consistency in three different watersheds using Jupyter notebooks (Choi et al., 2024) (HS 8). For this evaluation, we employed difference maps for model inputs (watershed DEMs, extracted land covers, and SSURGO maps), comparing the data obtained from the conventional approach to LES datasets (GeoServer and TDS) (Figure 9) and regression plots for model outputs (RHESSys daily streamflow outputs) (Figures 10 and 11). These results are presented in the following subsections.

4.5.1 Evaluate Spatial Model Input

To emphasize the significance of georeferencing, we contrasted RHESSys daily streamflow outputs before and after applying them to the LES datasets with those derived from georeferenced conventionally distributed data. Figures 9(1a) to 9(3c) show the spatial model input difference between the conventional approach and LES (GeoServer or TDS) approaches without applying georeferencing. In contrast, Figures 9(4a) to 9(4c) present the same comparisons with georeferencing applied. As mentioned earlier, GeoServer and TDS datasets are the same data but only in different file formats, meaning they both produce the same results. Therefore, only one of them is used to create the difference maps for spatial model inputs in ArcMap (Figures 9(1) to 9(4)).

Figure 9(1) shows the differences in the DEM elevation between the conventional approach shared as HS 7 resource and the subsetted LES (GeoServer or TDS) approaches at (a) Coweeta Subbasin18 (10 m), NC, (b) Scotts Level Branch (30 m), MD, and (c) Spout Run (60 m), VA before applying the georeferencing corrections described in the methodology section. The difference is defined as LES datasets DEM subtracted from conventionally accessed DEM. In each

map, red and green cells show the highest positive and negative DEM elevation differences, respectively, between the data conventionally distributed and the LES datasets, while the yellow ones indicate the lowest values. The differences in elevations in Figure 9(1) primarily stem from horizontal shifts due to the lack of georeferencing in the state-scale DEM compared to the georeferenced conventionally distributed data. These shifts lead to varying changes in elevation values across cells, with more significant discrepancies observed in some cells (highlighted in red and green for the highest positive and negative differences, respectively). Notably, these shifts are most pronounced between discrete hillslopes, particularly in Coweeta Subbasin18 (Figure 9(1a)) and Scotts Level Branch (Figure 9(1b)), potentially reflecting horizontal shifts of more than one pixel resolution. Coweeta Subbasin18, characterized by its forest-dominant watershed, exhibits the most significant differences in elevation compared to the other two watersheds, which are urban and suburban. Smaller elevation changes (depicted by yellow cells) occur in areas with gentler topography, predominantly seen in Spout Run (Figure 9(1c)), a suburban watershed with milder hillslopes. Furthermore, when merging multiple DEMs to create a state-scale DEM, inevitable resampling of cell values (elevation) may result in further differences between the DEM values of georeferenced conventionally distributed DEM and non-georeferenced subsetted LES DEM.

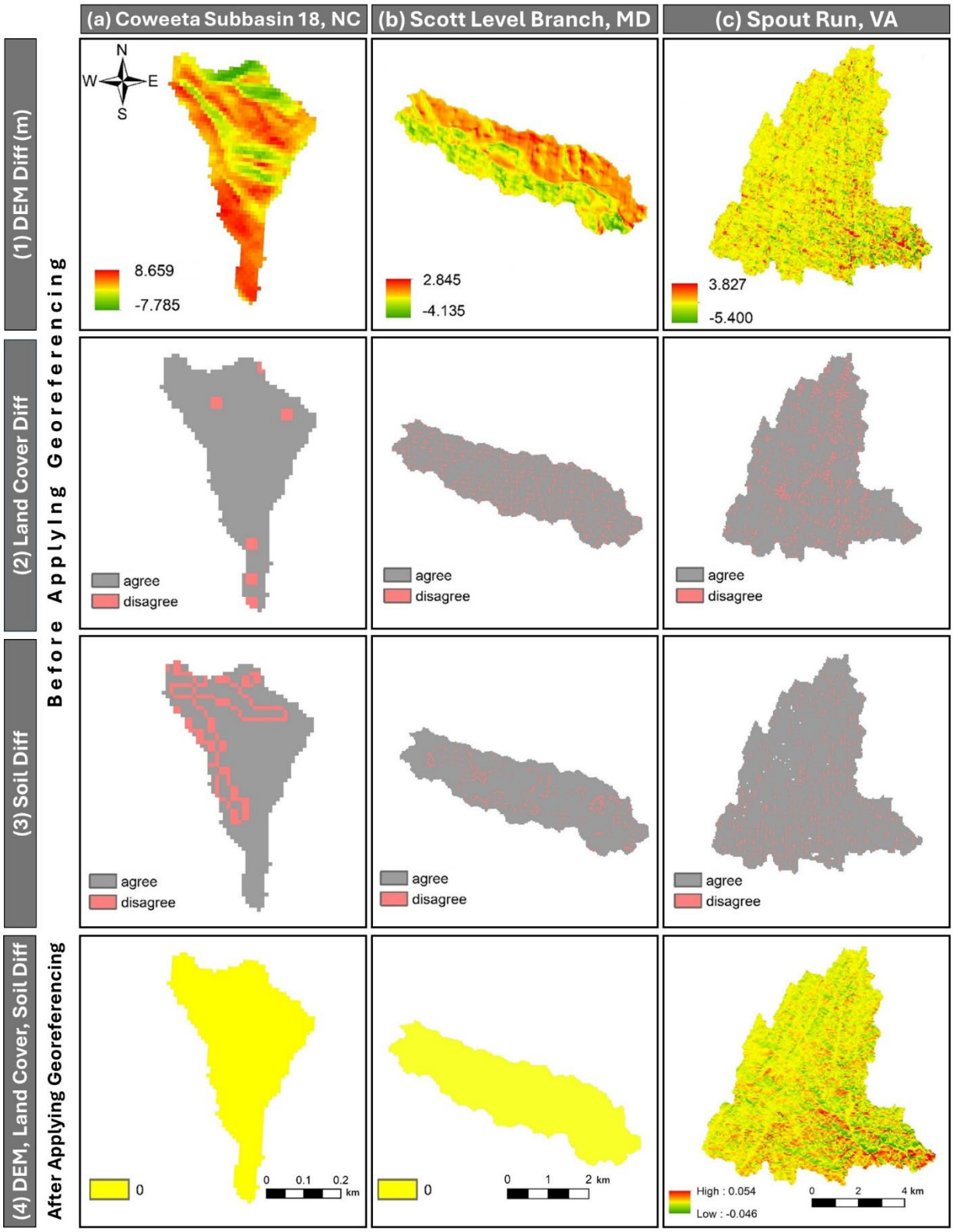
(a) Coweeta Subbasin 18, NC
(b) Scott Level Branch, MD
(c) Spout Run, VA
(1) DEM Diff (m)
(2) Land Cover Diff
(3) Soil Diff
(4) DEM, Land Cover, Soil Diff
Before Applying Georeferencing
After Applying Georeferencing
N
W
E
S
8.659
-7.785
2.845
-4.135
3.827
-5.400
agree
disagree
0
0 0.1 0.2 km
0 1 2 km
High : 0.054
Low : -0.046
0 2 4 km

Figure 9. Comparison of differences in DEM elevation (in meters), extracted land cover classification code, and extracted SSURGO soil texture between the conventionally accessed data and subsetted LES datasets (GeoServer or TDS) before and after georeferencing for (a) Coweeta Subbasin 18, NC, (b) Scotts Level Branch, MD, and (c) Spout Run, VA. Panels 9(1a) to 9(1c) show DEM differences, 9(2a) to 9(2c) show extracted land cover classification code differences, and 9(3a) to 9(3c) show extracted SSURGO soil texture differences, all before applying georeferencing. Panels 9(4a) to 9(4c) show results after georeferencing for DEM, land cover, and soil. Soil and land cover difference maps for Spout Run, VA, are zero and not shown in 9(4c).

Figure 9(2) compares to Figure 9(1), focusing on the differences in the extracted land cover classification code. The differences are shown in terms of whether the LES datasets' land cover values for each cell agree with the conventionally accessed values (gray cells) or not (red cells) because these maps used land cover classification codes as discrete integer values; for instance, code 41 corresponds to deciduous forest. The conventionally accessed land covers were extracted from national-scale land cover maps. Therefore, theoretically, the value of conventionally accessed data and LES data in the same location should be the same. In the model preprocessing, the watershed boundaries delineated using the DEMs are used to extract land cover and soil maps. Because the watershed boundaries are slightly shifted due to a lack of georeferencing, as stated earlier, the land cover classification codes are slightly changed in the LES datasets. Figure 9(2a), (2b), and Figure 9(2c) illustrate that 4.2%, 9.1%, and 9.4% of land cover changed, respectively, due to the creation of LES datasets and RHESSys preprocessing. This change percentage is calculated as the count of altered cells divided by the total number of cells multiplied by 100. The higher differences in land cover changes in Spout Run and Scotts Level compared to Coweeta Subbasin18 can be attributed to the higher diversity in land cover classifications among adjacent

neighboring cells within these two watersheds. This makes them particularly sensitive to slight shifts in the watershed boundary, resulting in comparatively higher differences in land cover changes.

Figure 9(3) provides a similar comparison to Figures 9(1) and 9(2) but focuses explicitly on differences in the extracted SSURGO soil texture. Similar to land cover maps, the differences are shown in terms of whether the soil values for each cell are changed (LES datasets agree/disagree with conventionally accessed data values) because these maps used soil texture code as discrete integer values. The reason for the difference in the soil maps is the same as in the land cover discussed earlier. Figures 9(3a), (3b), and (3c) show that 18.0%, 4.9%, and 5.3% of the soil maps were altered, respectively, during the creation of LES datasets and RHESSys preprocessing. The higher differences in soil texture observed in Coweeta Subbasin18, compared to the other two watersheds, can be attributed to the greater diversity in SSURGO classifications among adjacent neighboring cells. This heightened diversity makes Coweeta Subbasin18 particularly sensitive to slight shifts in the watershed boundary, resulting in higher observed differences in soil maps.

To demonstrate the importance of georeferencing, we applied it to all three datasets across the three states, and the results are presented in Figure 9(4). This figure displays the differences in the watershed DEM elevation, extracted land cover classification code, and extracted SSURGO soil texture between the conventionally accessed data and the subset LES datasets (GeoServer or TDS) at the three watersheds: (a) Coweeta Subbasin18, NC, (b) Scotts Level Branch, MD, and (c) Spout Run, VA, after applying georeferencing (only watershed DEM is shown for Figure 9(4c). Typically, the georeferenced method uses at least three points to transform a raster or shapefile; in our case, resampling and merging multiple GeoTiff files only altered cell values and shifted cell

locations without introducing distortion. Therefore, we found that using a single control point was sufficient to align the merged DEM with the original raw DEM accurately. As a result of this georeferencing, we were able to eliminate the DEM differences except for the DEM of Spout Run (60 m resolution), where slight differences, within a range of ± 6 cm, between the DEM of the conventionally accessed data and subsetted LES datasets (GeoServer or TDS) exist, as shown in Figure 9(4c). These slight differences exist because we resampled the DEM from 30 m to 60 m resolution to evaluate the applicability of lower resolution for larger watersheds. The difference between the conventionally accessed and LES datasets DEMs, land covers, and soil maps at Coweeta Subbasin18 (Figure 9(4a)) and Scotts Level Branch (Figure 9(4b)), and the difference between the conventionally accessed and LES datasets land cover, and soil at Spout Run (not shown in the figure) are zero.

4.5.2 Evaluate Model Output

To further emphasize the importance of georeferencing, the RHESSys model outputs before and after georeferencing are compared. Figures 10 and 11 show three regression analyses, each comparing two RHESSys daily streamflow outputs from the three different data input approaches: conventional, GeoServer, and TDS. The results are provided for the three watersheds: Coweeta Subbasin18 in North Carolina, Scotts Level Branch in Maryland, and Spout Run in Virginia. Figure 10 explains the performance results of RHESSys outputs without applying the georeferencing tool (conventional vs. GeoServer: Nash-Sutcliffe Efficiency (NSE) 0.684, conventional vs. TDS: NSE 0.647) for the Coweeta 18 watershed. As explained earlier, applying georeferencing significantly improved the model inputs. As a result, the RHESSys outputs are also expected to be improved. Figure 11 (a) and (b) showed perfect agreement between the model

outputs generated from the conventional and LES inputs (conventional vs GeoServer: NSE 1.0, conventional vs TDS: NSE 1.0) for Coweeta Subbasin18 in North Carolina. Also, other subplots in Figure 11 showed perfect agreement after applying the georeferencing tool for the other two watersheds. These results demonstrate the importance of georeferencing for achieving data consistency; without georeferencing, spatial input data will not result in the expected modeled daily streamflow output.

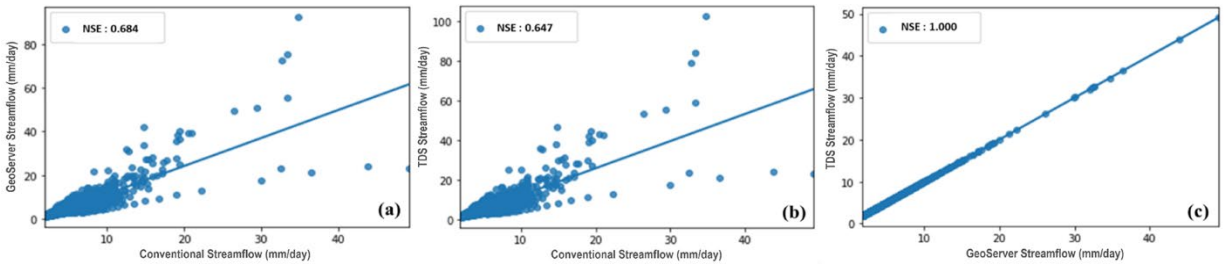

Figure 10. Comparing RHESSys daily streamflow outputs through regression analyses using three data distribution approaches in Coweeta Subbasin18, North Carolina, before georeferencing: (a) Conventional vs GeoServer, (b) Conventional vs TDS, and (c) GeoServer vs TDS

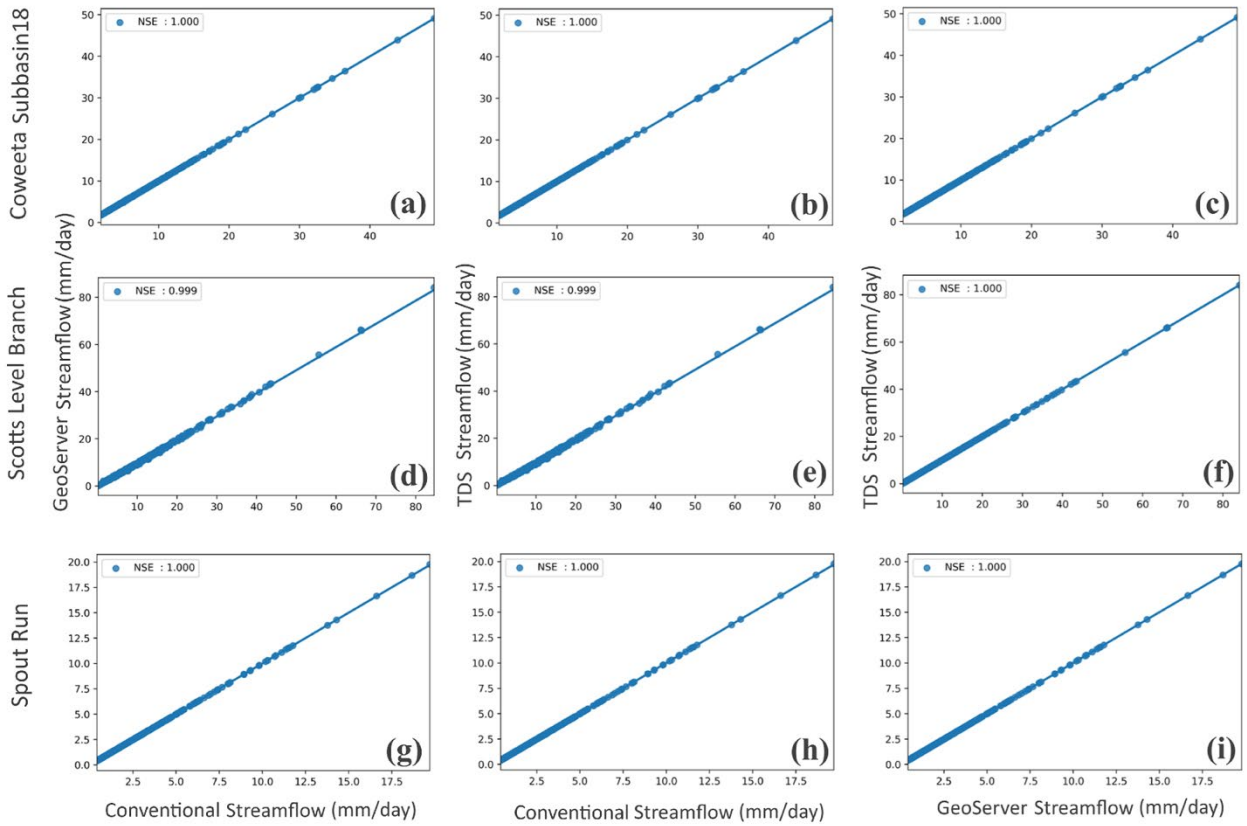

Figure 11. Comparing RHESSys daily streamflow outputs through regression analyses using three data distribution approaches after applying georeferencing in three watersheds: Coweeta Subbasin18 in North Carolina - (a) Conventional vs. GeoServer, (b) Conventional vs. TDS, and (c) GeoServer vs. TDS; Scotts Level Branch in Maryland - (d) Conventional vs. GeoServer, (e) Conventional vs. TDS, and (f) GeoServer vs. TDS; and Spout Run in Virginia - (g) Conventional vs. GeoServer, (h) Conventional vs. TDS, and (i) GeoServer vs. TDS (HS 8)

5. Discussion

This research focuses on leveraging GeoServer and Thematic Real-time Environmental Distributed Data Services (THREDDS) integrated into HydroShare to expose spatially large datasets to environmental models for open and reproducible environmental modelling. This approach addresses the limitations of previous spatial data sharing for hydrologic research. Beyond

addressing these limitations, a key benefit of using GeoServer and THREDDS data services (TDS) in our proposed approach is the significant reduction in manual effort required for data collection, cleaning, and integration. The approach streamlines workflows by automating these traditionally time-consuming tasks, reducing the time needed for preparing datasets and minimizing the potential for human error. As data scientists typically spend a large portion of their time on such preparatory tasks, this method offers a considerable improvement in efficiency, enabling researchers to focus more on analysis and model building. This efficiency gains particular importance when dealing with large datasets, where the manual approach would be more cumbersome and error-prone.

We acknowledge that our approach may still face some other challenges. The first challenge concerns API-based data access: when APIs are updated or changed, it can render hard-coded scripts obsolete over time. As mentioned in the introduction, this is a common issue that might contribute to why some researchers prefer manual data collection over APIs. It also underscores the need for ongoing maintenance and adaptability in data-driven workflows. The second challenge involves limitations in using various types of geographic datasets. While GeoServer supports Raster (GeoTIFF) and Vector (Shapefile) formats for spatial data, our primary choice was the GeoTIFF format with GeoServer. This preference is rooted in the practical consideration that subsetting SSURGO shapefiles can be resource-intensive due to the heterogeneous details of soil attributes, making it a computationally demanding process compared to grid-based geographic data. As we look to use LES datasets more effectively in the future, it becomes apparent that exploring alternative approaches or capabilities to work with shapefiles and their attributes (dbf table) in a manner well-suited for environmental modelling yet resource-efficient would be a valuable pursuit.

We selected the state scale as the spatial unit for distributing LES datasets, though other spatial aggregations can also be considered. For some hydrologic applications, the Hydrologic Unit Code (HUC) (Seaber et al., 1987) might be a more suitable scale than the state scale as often used by administrative maps. We opted for the state scale because federal web-based distribution systems readily provide data at this aggregation level, requiring minimal data processing for GeoServer and TDS distribution. The presented approach is general and agnostic to the specific spatial aggregation unit. Therefore, the presented steps for creating and sharing the datasets can be used or adopted by other researchers seeking to establish datasets for inter-state boundary watersheds.

Regarding the tools used, GeoServer and TDS each have strengths and limitations. TDS is better suited for handling multidimensional datasets, such as NetCDF, commonly used in environmental modeling due to their ability to represent complex temporal and spatial relationships. However, for very large files, such as the 10m resolution GeoTIFF dataset for North Carolina, TDS may encounter performance issues, as demonstrated in this study. On the other hand, GeoServer is highly efficient in managing and serving geospatial raster and vector data, such as GeoTIFF and Shapefiles. It excels in subsetting and visualizing geographic data but lacks TDS's multidimensional handling capabilities. The choice between these two tools depends on the specific data formats and use cases the study requires.

In the introduction, we compared our approach to two other approaches for supporting more reproducible environmental modelling in the literature: EcoHydroLib and HydroTerre. In more recent years, a third approach has emerged that is important to consider: Google Earth Engine (GEE) (Gorelick et al., 2017). GEE is a cloud-based platform for planetary-scale geospatial analysis that supports applications such as climate change, disease surveillance, environmental

protection, and water management. Over 450 articles published in 150 journals have used GEE datasets (Kumar and Mutanga, 2018), and the datasets available through GEE are continuously updated at a rate of nearly 6000 scenes per day (Gorelick et al., 2017). One example of a system built on GEE is Climate Engine (http://ClimateEngine.org), which leverages cloud computing to process, visualize, and download climate and remote sensing data for resource monitoring and analysis (Huntington et al., 2017). However, the use of GEE has two disadvantages. First, datasets like SSURGO are not natively included in GEE, requiring researchers to manually upload their data. Also, GEE is not open source, and it would be valuable to have an open and fully transparent alternative to GEE to support scientific modelling where users have control over the spatial data used as model input, and the data can be easily shared with appropriate metadata from HydroShare. The proposed approach by our study can address both GEE disadvantages.

Containerizing proprietary software such as ArcPy, part of ArcGIS, poses a barrier to reproducibility because not everyone may have access to this software. We used ArcPy to create the LES datasets. ArcPy is currently compatible only with the Windows operating system. While Docker Containers on Windows is experimental and available only on some versions, such as Windows Server 2019 and Windows 10 Professional and later editions, it is feasible to containerize proprietary software like ArcPy that operates exclusively on Windows. Thus, if we can install the proprietary software with containerization tools in the same operating system and if the software license is successfully recognized, it is possible to containerize proprietary software. The issue remains regarding access to the software license so that anyone, and not only those with access to the software, can reproduce the work. Also, the application of open-source GIS software such as GRASS GIS or QGIS will be a future work to achieve end-to-end environmental modeling.

Data availability to support environmental modelling is increasing rapidly, but data replication across distributed systems can present problems. For one, some data may have copyright issues because these data represent intellectual property (Abubahia and Cocea, 2017). Even if data can be freely used and copied it will become increasingly difficult to understand if verified data are being used to support a study. We had to manipulate the raw data provided by the federal agencies to give the data a consistent and accurate spatial coordinate system. For reproducibility, it is important to document such changes and associate the procedure for making the changes with the new data product. As a result, geographic data ownership and provenance are crucial concepts to consider (National Research Council, 2004). In the broader technology landscape, similar challenges are being addressed through blockchain technology, where a distributed digital ledger can track changes to a digital object. Related to blockchain technology, the concept of Non-Fungible Tokens (NFTs) (Farnaghi and Mansourian, 2020; Franke et al., 2020), where digital objects are uniquely identified within the blockchain, could prove valuable for identifying digital objects (National Research Council, 2004) used in environmental modelling (e.g., both data and processing scripts) and tracking the provenance of these objects in a consistent, globally accessible, and secure way. An extension of this work could involve treating spatial datasets as NFTs with ownership and provenance, thereby incorporating blockchain technology into the existing HydroShare data management capabilities. This approach would provide a means to precisely clarify the attributes and provenance of the increasing volume of raw and processed datasets necessary for reproducible and open environmental modelling.

6. Conclusions

Spatial data is an essential component for open and reproducible environmental modelling. Recently, there have been many efforts to improve the use of spatial data as model input. HydroShare provides a means for easily sharing datasets, including spatial datasets. It also supports exposing spatial data stored in HydroShare through APIs for programmatic data access within environmental modelling. Currently, HydroShare provides the capability for spatial datasets to be distributed using GeoServer and TDS, which can be accessed using APIs. These capabilities have been used mainly for data visualization use cases. This research demonstrates how these capabilities can support seamless, reproducible environmental modelling workflows.

The primary contribution of this research is the methods developed for integrating HydroShare with GeoServer and TDS for exposing Large Extent Spatial datasets to models for open and reproducible environmental modelling. We demonstrate how to create, share, and subset large datasets as inputs to an environmental model, thereby advancing the concept of seamless modelling, where model inputs can be constructed using reproducible workflows and common base datasets. We show the applicability of developed methodologies through three different watershed applications in three different states at three different spatial scales. Using the RHESSys model for each watershed, we show that no significant error is introduced when using the new data distribution system compared to traditional approaches where a user manually retrieves the original data from the federal agencies.

Georeferencing is critical in ensuring spatial data accuracy, especially when dealing with large-scale datasets spanning multiple coordinate systems. In this study, we applied georeferencing to correct spatial misalignments, demonstrating its importance in maintaining data consistency. Our results highlight that, without proper georeferencing, even small shifts in data can lead to

considerable inaccuracies in model outputs. By emphasizing the need for georeferencing, particularly in large-scale environmental modeling, we ensure that datasets from different sources are correctly aligned, ultimately improving the reliability of the results.

We discussed ways the proposed approach can be further advanced, for example, using other spatial aggregations of large data beyond the state-scale aggregation used in this paper. Finally, we discussed the challenge of data tracking and provenance, especially across systems and in the context of environmental modelling, where data from multiple sources is needed and each dataset requires extensive preprocessing.

Beyond sharing large, national-scale datasets maintained by federal agencies, individual scientists can also deploy the methods used in this work. Current data sharing through online repositories allows for data publication, a crucial step toward reproducibility. The proposed approach builds on this step to show how technologies like GeoServer and TDS when integrated with an online repository, provide a means for scientists to create and share file-based scientific data in a way that provides programmatic access without the need to deploy their own web-based data distribution systems. Scientists can easily share, update, and extend their data through such systems, including HydroShare, as demonstrated in this research, to support reproducibility and replicability through robust API-based access to their data. Creating and sharing datasets online using this approach offers a powerful means for scientists to achieve FAIR guiding principles, including reusability of data for multiple applications in different case studies and interoperability for programmatic access to multiple data collections using a consistent access protocol.

Data and Software Availability

All data and codes used are available through eight HydroShare resources (see Figure D1 for a visual representation of these resources). We published all data with persistent digital object identifiers (DOI's) on HydroShare and shared all data in a collection resource in HydroShare (Choi et al., 2024). This collection resource provides the links for all HydroShare resources as "Collection Contents." Eight HydroShare resources consist of the following: one collection resource (HS 1), one resource with Jupyter notebooks for automated workflows to create LES datasets (HS 2), three resources for three LES datasets (HS 3-5), one resource for RHESSys model instances (i.e., input) of the conventional data distribution approach and the observation time series in three different watersheds (HS 6), one resource with Jupyter notebooks for three different approaches and three different watersheds (HS 7), and one resource with Jupyter notebooks for evaluation of data consistency (HS 8). The Jupyter notebooks in HS 7 utilize the "RHYESsys-2024-02" kernel on CyberGIS-Jupyter for Water. The libraries and dependencies in the kernel will remain the same. Still, if a user needs to use it on an environment other than CyberGIS-Jupyter for Water, they may extract an environment.yml file and, with some effort, can reproduce and build on our study for a different environment.

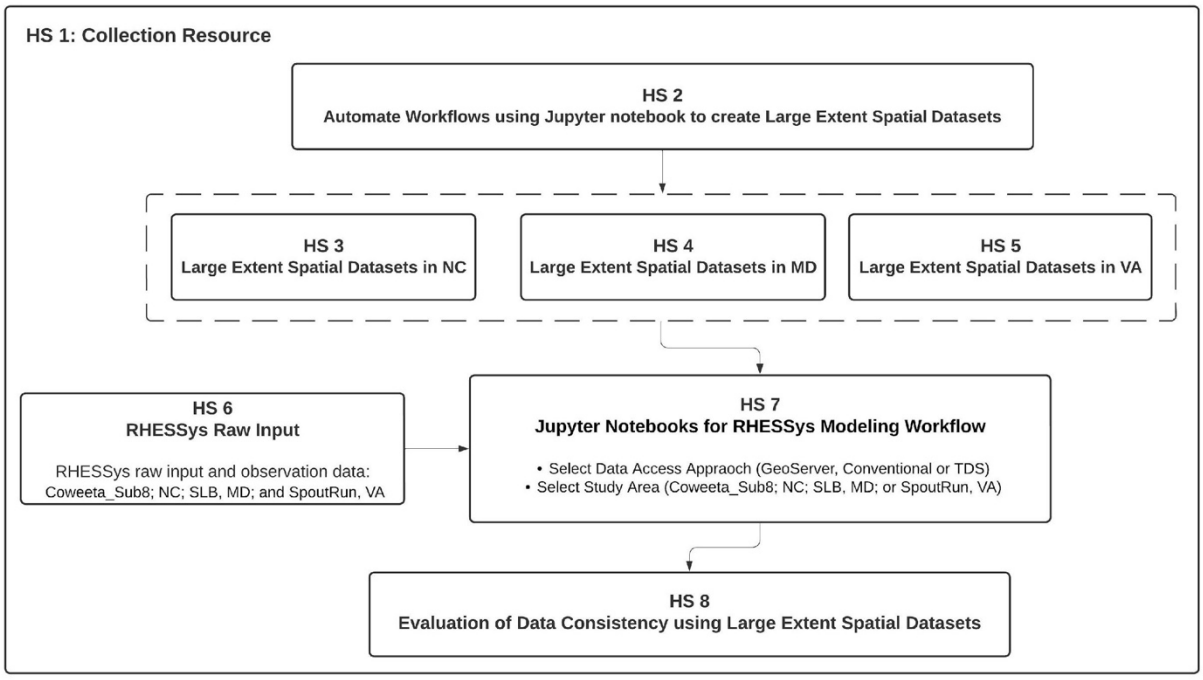

Figure D1. Visual representation of the eight HydroShare resources created in the study

List of Relevant URLs

Chesapeake Conservancy conservation innovation center:

https://www.chesapeakeconservancy.org/conservation-innovation-center

Creating Python Conda virtual environment (ArcPy) in ArcGIS Pro:

https://pro.arcgis.com/en/pro-app/latest/ArcPy/get-started/work-with-python-environments.htm

CUAHSI JupyterHub: https://jupyterhub.cuahsi.org

CyberGIS-Jupyter for Water: http://go.illinois.edu/cybergis-jupyter-water

CyberDuck: https://cyberduck.io

Cyberduck application: https://help.hydroshare.org/creating-and-managing-resources/accessing-hydroshare-irods-from-a-windows-pc-or-mac

icommands: https://help.hydroshare.org/creating-and-managing-resources/accessing-hydroshare-irods-from-linux

MRLC (Multi-Resolution Land Characteristics Consortium): https://www.mrlc.gov

National-scale SSURGO 30 m resolution GeoTIFF data:

https://nrcs.app.box.com/v/soils/folder/132131296196

nccopy: https://www.unidata.ucar.edu/software/netcdf/workshops/2011/utilities/Nccopy.html

OGC implementation standard: http://docs.opengeospatial.org/is/19-008r4/19-008r4.html

OWSLib: https://github.com/geopython/OWSLib

pyRHESSys: https://github.com/uva-hydroinformatics/pyRHESSys

rioxarray: https://github.com/corteva/rioxarray

SSUGRO Mukey Grids (GeoTIFF): https://nrcs.app.box.com/v/soils/folder/132131296196

USDA NRCS Geospatial Data Gateway: https://datagateway.nrcs.usda.gov

USGS 3D Elevation Program (3DEP): https://www.usgs.gov/core-science-systems/ngp/3dep

Web Soil Survey web distributed system:

https://websoilsurvey.sc.egov.usda.gov/App/WebSoilSurvey.aspx

xarray: http://xarray.pydata.org

Acknowledgement

We thank Natalie Thompson and the consultants of the University of Virginia Graduate Writing Lab for their helpful feedback in preparing the manuscript. This material is based upon

work supported by the U.S. National Science Foundation (NSF) under awards 1664018, 1664061, 1664119, 1849458, and 2118329. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Abubahia, A., Cocea, M., 2017. Advancements in GIS map copyright protection schemes - a critical review. Multimed Tools Appl 76. https://doi.org/10.1007/s11042-016-3441-z

Addor, N., Do, H.X., Alvarez-Garreton, C., Coxon, G., Fowler, K., Mendoza, P.A., 2020. Large-sample hydrology: recent progress, guidelines for new datasets and grand challenges. Hydrological Sciences Journal 65. https://doi.org/10.1080/02626667.2019.1683182

Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrology and Earth System Sciences Discussions. https://doi.org/10.5194/hess-2017-169

Alvarez-Garreton, C., Mendoza, P.A., Pablo Boisier, J., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., Ayala, A., 2018. The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies-Chile dataset. Hydrol Earth Syst Sci 22. https://doi.org/10.5194/hess-22-5817-2018

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. Nature 533, 452–454. https://doi.org/10.1038/533452a

Chen, M., Voinov, A., Ames, D.P., Kettner, A.J., Goodall, J., Jakeman, A.J., Barton, M.C., Harpham, Q., Cuddy, S.M., DeLuca, C., Yue, S., Wang, J., Zhang, F., Wen, Y., Lü, G.,

2020. Position paper: Open web-distributed integrated geographic modelling and simulation to enable broader participation and applications. Earth Sci Rev. https://doi.org/10.1016/j.earscirev.2020.103223

Choi, Y., Lin, L., 2021. Python Model API for RHESSys Model [WWW Document]. URL https://github.com/uva-hydroinformatics/pyRHESSys

Choi, Y.-D., Goodall, J., Band, L., Maghami, I., Lin, L., Saby, L., Li, Z., Wang, S., Calloway, C., Seul, M., Ames, D., Tarboton, D., Hong, Y., 2024. (HS 1) Toward Seamless Environmental Modeling: Integration of HydroShare with Server-side Methods for Exposing Large Datasets to Models [WWW Document]. https://doi.org/10.4211/hs.afcc703d884e4f73b598c9e4b8f8a15e

Choi, Y.-D., Goodall, J., Sadler, J.M., Castronova, A.M., Bennett, A., Li, Z., Nijssen, B., Wang, S., Clark, M.P., Ames, D.P., Horsburgh, J.S., Yi, H., Bandaragoda, C., Seul, M., Hooper, R., Tarboton, D.G., 2021. Toward open and reproducible environmental modeling by integrating online data repositories, computational environments, and model Application Programming Interfaces. Environmental Modelling and Software 135. https://doi.org/10.1016/j.envsoft.2020.104888

Crawley, S., Ames, D., Li, Z., Tarboton, D., 2017. HydroShare GIS: Visualizing Spatial Data in the Cloud. Open Water Journal 4.

Crosas, M., 2020. Fair Principles and Beyond: Implementation in Dataverse. Septentrio Conference Series. https://doi.org/10.7557/5.5334

CrowdFlower, 2016. Data Science Report - 2016 [WWW Document]. Crowd Flower.

DeVantier, B.A., Feldman, A.D., 1993. Review of GIS Applications in Hydrologic Modeling. J Water Resour Plan Manag 119. https://doi.org/10.1061/(asce)0733-9496(1993)119:2(246)

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., Gusev, Y.M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O.N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., Wood, E.F., 2006. Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, in: Journal of Hydrology. https://doi.org/10.1016/j.jhydrol.2005.07.031

Farnaghi, M., Mansourian, A., 2020. Blockchain, an enabling technology for transparent and accountable decentralized public participatory GIS. Cities 105. https://doi.org/10.1016/j.cities.2020.102850

Franke, L., Schletz, M., Salomo, S., 2020. Designing a blockchain model for the paris agreement's carbon market mechanism. Sustainability (Switzerland) 12. https://doi.org/10.3390/su12031068

Gan, T., Tarboton, D.G., Horsburgh, J.S., Dash, P., Idaszak, R., Yi, H., 2020. Collaborative sharing of multidimensional space-time data in a next generation hydrologic information system. Environmental Modelling and Software 129. https://doi.org/10.1016/j.envsoft.2020.104706

Geospatial Data Gateway [WWW Document], 2021. URL https://datagateway.nrcs.usda.gov

Google Earth Engine [WWW Document], 2021. URL https://earthengine.google.com

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens Environ 202. https://doi.org/10.1016/j.rse.2017.06.031

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., Andréassian, V., 2014. Large-sample hydrology: A need to balance depth with breadth. Hydrol Earth Syst Sci 18. https://doi.org/10.5194/hess-18-463-2014

Hamman, J., Rocklin, M., Abernathy, R., 2018. Pangeo: A Big-data Ecosystem for Scalable Earth System Science, Geophysical Research Abstracts.

Hodson, S., Jones, S., Collins, S., Genova, F., Harrower, N., Laaksonen, L., Mietchen, D., Petrauskaité, R., Wittenburg, P., 2018. Turning FAIR data into reality: interim report from the European Commission Expert Group on FAIR data (Version Interim draft). Interim report from the European Commission Expert Group on FAIR data.

Horsburgh, J.S., Morsy, M.M., Castronova, A.M., Goodall, J., 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. J Am Water Resour Assoc 52. https://doi.org/10.1111/1752-1688.12363

Hoyer, S., Hamman, J., 2017. xarray: N-D labeled Arrays and Datasets in Python. J Open Res Softw 5, 10. https://doi.org/10.5334/jors.148

Huntington, J.L., Hegewisch, K.C., Daudert, B., Morton, C.G., Abatzogloum John T., McEvoy, D.J., Erickson, T., 2017. Climate Engine: Cloud Computing and Visualization of Climate and Remote Sensing Data for Advanced Natural Resource Monitoring and Process Understanding. Bull Am Meteorol Soc 98. https://doi.org/https://doi.org/10.1175/BAMS-D-15-00324.1

HydroShare/hsclient: HydroShare Python Client [WWW Document], 2021. URL https://github.com/hydroshare/hsclient

Kuentz, A., Arheimer, B., Hundecha, Y., Wagener, T., 2017. Understanding hydrologic variability across Europe through catchment classification. Hydrol Earth Syst Sci 21. https://doi.org/10.5194/hess-21-2863-2017

Kumar, L., Mutanga, O., 2018. Google Earth Engine applications since inception: Usage, trends, and potential. Remote Sens (Basel) 10. https://doi.org/10.3390/rs10101509

Kumar, M., Bhatt, G., Duffy, C.J., 2010. An object-oriented shared data model for GIS and distributed hydrologic models. International Journal of Geographical Information Science 24. https://doi.org/10.1080/13658810903289460

Leonard, L.N., 2015. HydroTerre: Towards an expert system for scaling hydrological data and models from hill-slopes to major-river basins. ProQuest Dissertations and Theses Global.

Lippold, K.J., 2019. Improving HydroShare and Web Application Interoperability Through Integrated GIS and HIS Data Services (Master's Thesis). Brigham Young University.

Maghami, I., Morsy, M.M., Sadler, J.M., Horsburgh, J.S., Dash, P.K., Choi, Y., Chen, K., Seul, M., Black, S., Tarboton, D.G., Goodall, J.L., 2024. An extensible schema for capturing environmental model metadata: Implementation in the HydroShare online data repository. Environmental Modelling & Software 172, 105895. https://doi.org/10.1016/j.envsoft.2023.105895

Michaelis, C.D., Ames, D.P., 2017. Web Feature Service (WFS) and Web Map Service (WMS), in: Encyclopedia of GIS. Springer International Publishing, Cham, pp. 2485–2488. https://doi.org/10.1007/978-3-319-17885-1_1480

Miles, B., Band, L.E., 2017. RHESSysWorkflows [WWW Document]. URL https://github.com/selimnairb/RHESSysWorkflows?tab=readme-ov-file (accessed 4.12.22).

Miles, B., Band, L.E., 2015. Ecohydrology Models without Borders? https://doi.org/10.1007/978-3-319-15994-2_31

Morsy, M.M., Goodall, J., 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 and Software 93, 13–28. https://doi.org/10.1016/j.envsoft.2017.02.028

MRLC (Multi-Resolution Land Characteristics Consortium) [WWW Document], 2021. URL https://www.mrlc.gov

National Academies of Sciences, 2019. Reproducibility and Replicability in Science. https://doi.org/https://doi.org/10.17226/25303

National Research Council, 2004. Licensing Geographic Data and Services, Licensing Geographic Data and Services. National Academies Press, Washington, D.C. https://doi.org/10.17226/11079

Nativi, S., Caron, J., Domenico, B., Bigagli, L., 2008. Unidata's Common Data Model mapping to the ISO 19123 Data Model. Earth Sci Inform 1. https://doi.org/10.1007/s12145-008-0011-6

Newman, A.J., Clark, M.P., Sampson, K., Wood, A., Hay, L.E., Bock, A., Viger, R.J., Blodgett, D., Brekke, L., Arnold, J.R., Hopson, T., Duan, Q., 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. Hydrol Earth Syst Sci 19, 209–223. https://doi.org/https://doi.org/10.5194/hess-19-209-2015, 2015

OWSLib [WWW Document], 2021. URL https://github.com/geopython/OWSLib

rioxarray [WWW Document], 2021. URL https://github.com/corteva/rioxarray

Seaber, P.R., Kapinos, F.P., Knapp, G.L., 1987. Hydrologic Unit Maps (USA). US Geological Survey Water-Supply Paper 2294. https://doi.org/10.3133/wsp2294

Soil Survey Staff, 2021. Web Soil Survey: Natural Resources Conservation Service, United States Department of Agriculture [WWW Document]. URL https://websoilsurvey.nrcs.usda.gov

SRTM (Shuttle Radar Topography Mission) [WWW Document], 2021. URL https://srtm.csi.cgiar.org

Stagge, J.H., Rosenberg, D.E., Abdallah, A.M., Akbar, H., Attallah, N.A., James, R., 2019. Assessing data availability and research reproducibility in hydrology and water resources. Sci Data 6, 1–12. https://doi.org/10.1038/sdata.2019.30

Tague, C.L., Band, L.E., 2004. RHESSys: Regional Hydro-Ecologic Simulation System—An Object-Oriented Approach to Spatially Distributed Modeling of Carbon, Water, and Nutrient Cycling. Earth Interact. https://doi.org/10.1175/1087-3562(2004)8<1:rrhsso>2.0.co;2

Tarboton, D.G., Ames, D.P., Horsburgh, J.S., Goodall, J.L., Couch, A., Hooper, R., Bales, J., Wang, S., Castronova, A., Seul, M., Idaszak, R., Li, Z., Dash, P., Black, S., Ramirez, M., Yi, H., Calloway, C., Cogswell, C., 2024. HydroShare retrospective: Science and technology advances of a comprehensive data and model publication environment for the water science domain. Environmental Modelling & Software 172, 105902. https://doi.org/10.1016/j.envsoft.2023.105902

Thornton, M.M., Shrestha, R., Wei, Y., Thornton, P.E., Kao, S.,, Cook, R.B., 2022. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4. ORNL DAAC.

[WWW Document]. URL https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1840 (accessed 12.11.21).

Toms, S., 2015. ArcPy and ArcGIS–Geospatial Analysis with Python. Packt Publishing Ltd.

Unidata, 2024. THREDDS Data Server (Version 5.4). https://doi.org/https://doi.org/10.5065/D6N014KG

USFS Coweeta Hydrologic Laboratory [WWW Document], 2021. URL https://www.fs.usda.gov/research/srs/forestsandranges/locations/coweeta

USGS 3DEP (The United States Geological Survey 3D Elevation Program) [WWW Document], 2021. URL https://www.usgs.gov/core-science-systems/ngp/3dep

Wenjue, J., Yumin, C., Jianya, G., 2004. Implementation of OGC web map service based on web service. Geo-Spatial Information Science 7. https://doi.org/10.1007/BF02826653

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. https://doi.org/10.1038/sdata.2016.18

Wilkinson, M.D., Verborgh, R., da Silva Santos, L.O.B., Clark, T., Swertz, M.A., Kelpin, F.D.L., Gray, A.J.G., Schultes, E.A., van Mulligen, E.M., Ciccarese, P., Kuzniar, A., Gavai,

A., Thompson, M., Kaliyaperumal, R., Bolleman, J.T., Dumontier, M., 2017. Interoperability and FAIRness through a novel combination of Web technologies. PeerJ Comput Sci 2017. https://doi.org/10.7717/peerj-cs.110

Yi, H., Idaszak, R., Stealey, M., Calloway, C., Couch, A.L., Tarboton, D.G., 2018. Advancing distributed data management for the HydroShare hydrologic information system. Environmental Modelling and Software. https://doi.org/10.1016/j.envsoft.2017.12.008

Yin, D., Liu, Y., Padmanabhan, A., Terstriep, J., Rush, J., Wang, S., 2017. A CyberGIS-Jupyter Framework for Geospatial Analytics at Scale, in: Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact. pp. 1–8. https://doi.org/https://doi.org/10.1145/3093338.3093378

Youngblood, B., 2013. GeoServer Beginner's Guide. Packt Publishing Ltd.