Collaborative sharing of multidimensional space-time data in a next generation hydrologic information system

Tian Gan[a,c], David G. Tarboton[a], Jeffery S. Horsburgh[a], Pabitra Dash[a], Ray Idaszak[b], Hong Yi[b]

[a] Department of Civil and Environmental Engineering, Utah State University, 8200 Old Main Hill, Logan, UT 84322-8200, USA

[b] RENCI, University of North Carolina at Chapel Hill, Chapel Hill, NC 27517, USA

[c] Corresponding author: Institute of Arctic and Alpine Research, University of Colorado, Campus Box 450, Boulder, CO 80309-0450 USA. Phone (+1) 435-754-9720. Email: gantian127@gmail.com

Highlights

- Development of a framework for sharing multidimensional space-time data.
- Automatic harvesting of metadata to support data discovery and reuse.
- Easy to use metadata editing to add additional or missing information.
- Supports standard data services for programmatic access and subsetting.
- Users can share data through data services without requiring server setup.

This is the accepted version of the following published article:

Gan, T., D. G. Tarboton, J. S. Horsburgh, P. Dash, R. Idaszak and H. Yi, (2020), "Collaborative sharing of multidimensional space-time data in a next generation hydrologic information system," Environmental Modelling & Software, 129: 104706, https://doi.org/10.1016/j.envsoft.2020.104706.

Abstract

In hydrologic research, there is a need to manage, archive, and publish data in a discoverable way to increase data reuse, transparency, and reproducibility. Multidimensional space-time data are commonly used in hydrologic research, and systems are needed for sharing and exchanging such data. Simply exchanging files may result in loss of metadata information and can be challenging when files are large. We developed an approach to manage, share, and publish multidimensional space-time data in HydroShare, a next generation hydrologic information system and domain specific repository. This paper presents the design, development, and testing of this approach. We selected the Network Common Data Form (NetCDF) as the underlying data model. We defined specific metadata elements to store and manage multidimensional space-time data. We adopted and adapted existing software to automatically harvest, support entry of metadata, and establish standardized data services to serve and enhance access to the datasets shared in HydroShare.

Key words: multidimensional space-time data, NetCDF, collaborative data sharing, HydroShare, cyberinfrastructure

Software availability

The software created in this research is free and open source as part of the larger HydroShare software repository. The HydroShare software repository is managed through GitHub and is available at https://github.com/hydroshare/hydroshare

1 Introduction

With advances in hydrologic monitoring and model simulation technologies, hydrologic research has become data and computationally intensive, resulting in large volumes of scientific data generated or collected by individual researchers and organizations. Advances in hydrologic understanding now tends to require discovery, access, and integration of heterogeneous and dispersed data from multiple sources. Moreover, large-scale hydrologic problems often need to be solved by collaboration among researchers, thus working as a team to collaborate around data

has become indispensable. These emerging trends in hydrologic research are key drivers that demand new tools to support the entire research cycle of data creation, discovery, access, curation, publication, and analysis to help achieve new scientific breakthroughs (Horsburgh et al., 2015; Rajib et al., 2016; Morsy et al., 2017).

The Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) has devoted great effort to the development of cyberinfrastructure (CI) to satisfy this need, including HydroShare (http://www.hydroshare.org), a next generation Internet-based Hydrologic Information System (HIS) (Tarboton et al., 2014). HydroShare was developed to extend the capability of the earlier, server based CUAHSI HIS, which focused on the sharing of point observation time series data (Horsburgh et al., 2008, 2009; Tarboton et al., 2009). Given that the needs of hydrology researchers go well beyond time series data, HydroShare was established to add support for sharing a broader range of hydrologic datasets and models that are widely used in the hydrologic science community. These include data of varying dimensionality such as time series (i.e., observations varying over time at a single fixed location), geographic raster (i.e., a regular raster grid representing a spatial field at a single time), geographic feature (i.e., geospatial data represented by points, lines, or polygons representing a single point in time), and multidimensional space-time data (i.e., data that may vary in both space and time), as well as model instances, and model programs (Morsy et al., 2017). As a first step in HydroShare development, a resource data model was designed that enabled storing, transmitting, and cataloging of resources comprised of these diverse hydrologic data types and models to facilitate discovery (Horsburgh et al., 2015). Details for each of HydroShare's supported data types were specified, including required data format and content, metadata elements, and functions for data processing, analysis, or visualization. HydroShare's resource data model was designed to generalize the way datasets and models were managed and shared, while at the same time supporting specific metadata elements and functions required for enhancing the hydrologic analysis capability and interoperability for different data types and models.

Multidimensional space-time data (referred to as "MD data" in this paper) is widely used in hydrology for both observations and model results. Examples include weather and climate data such as spatially distributed precipitation, temperature, wind speed, and humidity used as model inputs, or snow water equivalent and soil moisture output from models. While commonly used,

there are several challenges associated with MD data that can make data sharing more difficult. One challenge is that there is no single, accepted data format for storing this type of data to support the interoperability needed for data sharing and analysis. Formats used include Tagged Image File Format (TIFF https://www.adobe.io/open/standards/TIFF.html), GRIdded Binary (GRIB https://www.wmo.int/pages/prog/www/WMOCodes/Guides/GRIB/), Common Data Format (CDF http://cdf.gsfc.nasa.gov/), Hierarchical Data Format (HDF https://www.hdfgroup.org/), and Network Common Data Form (NetCDF http://www.unidata.ucar.edu/software/netcdf/). File size can be another challenge. MD datasets can be large, making it inefficient to download and exchange entire large files when users may need only a subset or slice of them for visualization or analysis. It is also important to effectively capture, expose, and support the recording and editing of metadata associated with the MD data files. Recognizing these challenges, this paper describes our efforts to establish functionality to support the sharing of MD data in HydroShare.

Currently, several websites and software tools can be used for sharing MD data, and each has its own strengths and limitations. For example, Figshare (http://figshare.com/) is a website that enables users to manage their research output in the cloud to be stored, shared, published, and discovered. It supports permanent data publication and provides citation information for shared datasets to give the data provider credit and make their datasets citable. Figshare supports social functions such as commenting and access control to facilitate collaboration around the datasets. However, although Figshare has functions to capture simple metadata and preview file contents for commonly used formats such as Microsoft Word, PDF, and Microsoft Excel, no functions are provided to preview or edit the metadata or contents of the more advanced, scientific data formats used for MD data (e.g., NetCDF and CDF). These limitations hinder users' ability to describe, preview, access, and interpret file contents through the website, which can be a barrier to data sharing, inhibiting data reuse and the reproducibility of scientific analyses.

The Thematic Real-time Environmental Distributed Data Services (THREDDS) and Hyrax data servers can provide cataloging functionality and support access to metadata and data for scientific datasets through various data access protocols (OPeNDAP, 2017; Unidata, 2017a). However, this does require that a data provider have access to or be able to install and maintain server software and hardware. The THREDDS catalog and the Open-source Project for a

Network Data Access Protocol (OPeNDAP) service are common services supported by these two data servers. THREDDS catalogs are logical directories of available online datasets that help in data discovery. The OPeNDAP services enable users to subset or preview the contents of remote datasets and metadata. Moreover, OPeNDAP client software programs exist that can help retrieve remote datasets for analysis and visualization. These include the NetCDF4 Python package, the RNetCDF package for the R Statistical Computing Environment, NetCDF Operators (NCO) (Zender, 2008), Integrated Data Viewer (IDV) (Unidata, 2017b), and Panoply (NASA, 2018), etc. Thus, there is important value to data consumers made available through these services, motivating the need for this serving capability to be available for individual researchers or small research groups that do not have the capacity to set up THREDDS or Hyrax servers. Another issue with existing THREDDS and Hyrax server functionality is limited exposure to search capabilities, which may prevent or impede scientists from discovering datasets using search terms or the geolocation of the dataset, etc.

The Repository for Archiving, Managing and Accessing Diverse DAta (RAMADDA) (http://ramadda.org/) is another web-based application framework that provides a broad suite of services for content and data management, publishing, and collaboration. With RAMADDA, users can search, access, upload, or comment on datasets. The system incorporates the OPeNDAP service and data analysis tools to provide functions for file content preview, metadata capture and curation, and data subsetting and analysis for MD data. However, as with the THREDDS and Hyrax data servers, sharing MD data with RAMADDA requires setting up and maintaining the services, which may make sharing research datasets impractical for individual researchers or small research groups.

To address some of these limitations, in this paper we report the HydroShare MD data representation design and implementation. It provides functionality to help share MD data to promote data curation, publication, and reuse. We present a use case that demonstrates the new capabilities and contrasts them with capabilities of existing systems. When researchers store and share MD data in a format not widely used by the scientific community, current data sharing systems treat these files as generic file objects, which makes it difficult for others to know about the detailed metadata they contain. It also makes it challenging to directly subset, visualize, or process the datasets through the data sharing system. In contrast, our approach enables users to

share MD data in the NetCDF data format. Metadata from the file are exposed and presented in HydroShare. Users can collaboratively edit metadata in HydroShare and have these edits easily updated in the NetCDF file. When made public, the MD data are automatically registered in a HydroShare-connected THREDDS server that enables remote access using the OPeNDAP service without data providers being required to provision any server hardware or install or configure any software. With HydroShare's inherent data discovery, versioning, publication, and social functions, users can collaborate around datasets from initial data preparation to final data publication, and the sharing, discovery, and reuse of MD data is simplified.

In this paper, we describe the design, development, and testing of this approach. Section 2 provides a brief introduction of HydroShare system. Section 3 describes the functional design and implementation for the MD data type in HydroShare. Section 4 presents a use case that demonstrates functions in HydroShare that facilitate collaboration among users for data preparation, publication, and reuse. Discussion and conclusions are provided in the final sections.

2 Background

HydroShare is a web based hydrologic information system that provides functionality for metadata capture and curation, data manipulation, data publication, data discovery, and collaboration (Tarboton et al., 2014; Horsburgh et al., 2015). These functions represent a new paradigm in data sharing systems, supporting discovery through the integration of information from multiple sources, team work, collaboration, reuse of data, and transparency to enhance trust in research findings.

Fig. 1 shows a high level view of HydroShare's system architecture. The "Resource Sharing" functionality provides a web user interface to help users store and manage shared datasets and models in HydroShare. The "Actions on Resources" functionality includes web applications or web services from HydroShare or third-party organizations that enhance the capability for data analysis, visualization, or model simulation. The interactions between "Resource Sharing" and "Actions on Resources" are through HydroShare's Representational State Transfer (REST) application programming interface (API) and iRODS client interface (e.g., iRODS Python API). Heard et al. (2014) provide additional details of each open source component (e.g., Django and iRODS).

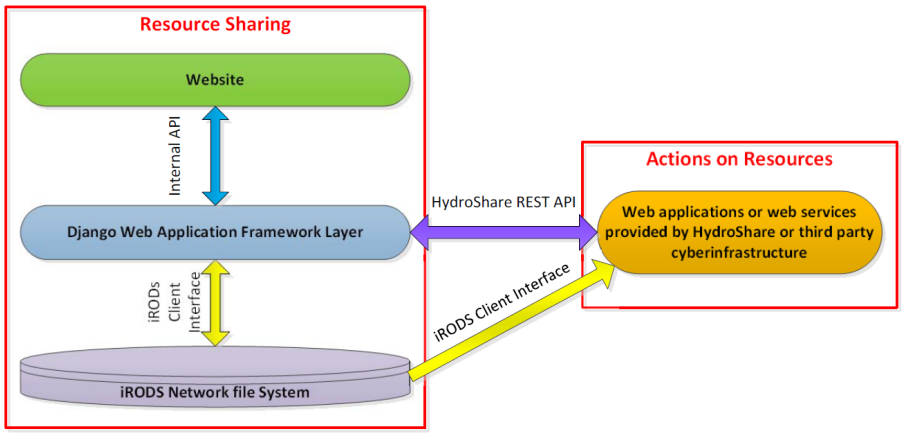


Fig. 1. High level system architecture of HydroShare.

The HydroShare resource data model was designed and implemented to manage various types of hydrologic datasets and models (Horsburgh et al., 2015). A HydroShare resource is the granular unit of shared content for access control, serialization for transport over the Internet, and cataloging for discovery within the system. Major concepts for HydroShare resource are listed below, and Fig. 2 shows an example to explain their relationships:

- **Resource content files:** single or multiple files uploaded into HydroShare by users that make up different hydrologic datasets and models or additional informational files.
- **Resource-level metadata file:** a file encoded using extensible markup language (XML) and generated by HydroShare for metadata that documents the whole resource.
- **Content type**: a widely used and well-known hydrologic data type (e.g., time series, geographic feature, and MD data) which is supported in HydroShare with advanced functions for metadata management and data analysis.
- **Content type files:** single or multiple resource content files within a resource that represent a dataset of a supported content type.
- **Content type metadata file:** a file encoded using XML and generated by HydroShare for metadata that documents a dataset of a supported content type.

In Fig. 2, an example HydroShare resource consists of a resource-level metadata file and multiple resource content files. These resource content files include an informational file (a Microsoft Word document) and multiple groups of files to represent different hydrologic datasets (e.g., time series and MD data). Each group of files, or "aggregation," includes a content type metadata file and one or more content type files to represent a hydrologic dataset for a supported content type in HydroShare.

HydroShare's resource data model allows for definition of a new content type through specification of a content type data model. The content type data model defines the data format and contents of the files along with content type metadata elements. Resource content files associated with a supported content type are grouped together into an aggregation by following the Open Archives Initiative Object Reuse (OAI-ORE) standard, which is used for the description and exchange of aggregations of web resources (Lagoze et al., 2008). Thus, an instance of a content type is referred to as an aggregation in HydroShare (e.g., MD aggregation in Fig. 2).

The advantage of the resource data model design is that HydroShare can manage (e.g., storage on disk, packaging for delivery over the Internet, access control, and cataloging for discovery) multiple types of datasets and models in the same way, regardless of the data formats and contents. Meanwhile, a content type data model enables users to standardize data formats and syntax and to add additional metadata to describe the hydrologic datasets. Developers can then use the standardized data formats and metadata to create advanced functions to facilitate metadata management, data analysis, or data visualization. This paper specifically reports the design of the content type data model for MD data and serves as an example demonstrating how to extend the HydroShare resource data model with a new content type.

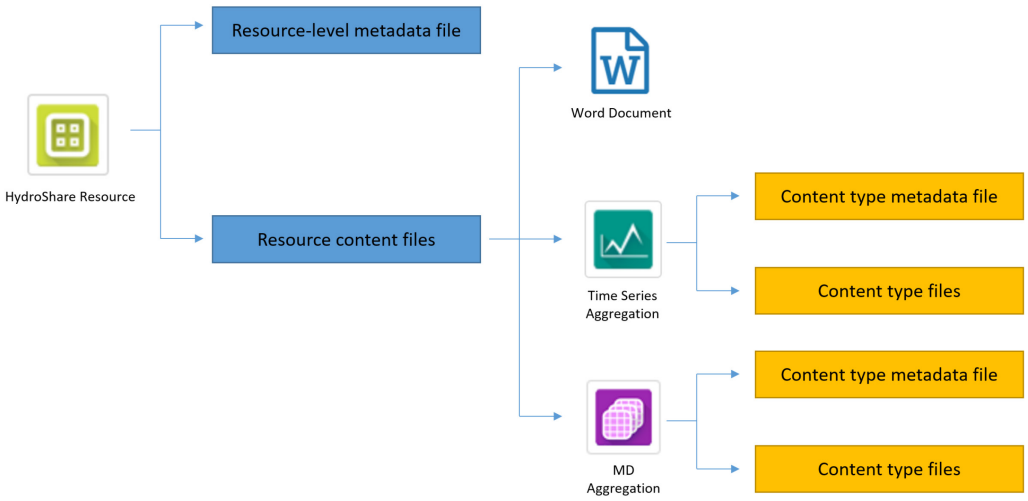


Fig. 2. Major components of an example HydroShare resource.

3 Methods

We designed the MD content type in HydroShare to support the sharing of MD data using the following steps: 1) we designed and implemented a content type data model for MD data; 2) we developed automated functions to support metadata extraction and editing for MD data; and 3) we set up the OPeNDAP web services to facilitate remote data access for data subsetting, analysis, and visualization. Detailed methods we used for each step are described in the following sections.

3.1 MD content type data model

3.1.1 Content type files

Since there are many scientific data formats capable of storing MD data, we evaluated the benefits and tradeoffs of these data formats and chose the one that we felt was most suitable for data storage and management in HydroShare. We established the following criteria to decide the data format used to represent MD data in HydroShare. First, the data stored in the file needed to be organized in a way that helps users understand the data structure and retrieve a subset for analysis. Second, the data format needed to be widely recognized and used in hydrologic research, with available open source software or libraries to help analyze or visualize the data. Third, widely accepted standards needed to be available to guide users in how to organize the data contents and metadata in the file to promote interoperability for data processing and sharing.

Based on these criteria, we compared several data formats that are widely used in the hydrologic science community, including TIFF, GRIB2, CDF, HDF5, and NetCDF, and adopted the NetCDF data format to store MD data in HydroShare. TIFF format is often used to store MD data with each file representing a physical phenomenon at a time slice over a spatial coverage. However, this format is inconvenient for data transfer and data subsetting when a MD dataset involves a large number of files. GRIB2, CDF, HDF5, and NetCDF data formats are able to store MD data and associated metadata within one file. Open source software programs for these data formats are available to support data analysis or visualization. The reasons for selecting the NetCDF data format were its wide use in modeling research in hydrology and aligned fields such as atmospheric science, its adoption as a standard (OGC, 2011), and support for standards for its metadata (Eaton et al., 2017; ESIP, 2017).

NetCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. The NetCDF software includes C, Fortran77, Fortran 90, and C++ interfaces for accessing NetCDF data. Programming interfaces are also available for Python, R, Java, and other languages. NetCDF utilities such as "ncdump" are available to facilitate simple data management tasks. The NetCDF data format can actually refer to multiple formats. In this paper, we specifically refer to the NetCDF classic formats and the NetCDF-4/HDF5 format, which are based on the NetCDF classic model and the enhanced data model. The NetCDF user's guide provided additional details (https://www.unidata.ucar.edu/software/netcdf/docs/user_guide.html).

The NetCDF data format is widely used to represent MD datasets as the input or output for hydrologic models (David et al., 2011; Sen Gupta et al., 2015; Thornton et al., 1997). It has also been used for data management and curation of data converted from other data formats (Guo et al., 2015). Moreover, many software programs and libraries available for NetCDF data processing, analysis, or visualization are widely applied among the research community (Unidata, 2017c). Thus, researchers with these tools can easily manipulate NetCDF files. This capability was an advantage that enabled us to develop the functions in HydroShare without starting from scratch. Furthermore, several conventions (https://www.unidata.ucar.edu/software/netcdf/conventions.html) are available to promote the processing and sharing of data in the NetCDF data format. For example, the NetCDF Climate

and Forecast (NetCDF-CF) convention (Eaton et al., 2017) specifies how to define the dimensions, variables, and attributes to represent MD data as regular grid data or point time series. The Attribute Conventions for Data Discovery (ACDD) (ESIP, 2017) were designed to define the metadata attributes needed to describe the whole NetCDF dataset to a discovery system.

With the selection of the NetCDF data format, we specified that a MD aggregation should include only one NetCDF file uploaded by the user and one metadata header information text file automatically generated by the system from the uploaded file to provide a brief summary of the contents in the NetCDF file. A HydroShare resource may contain one or many MD aggregations.

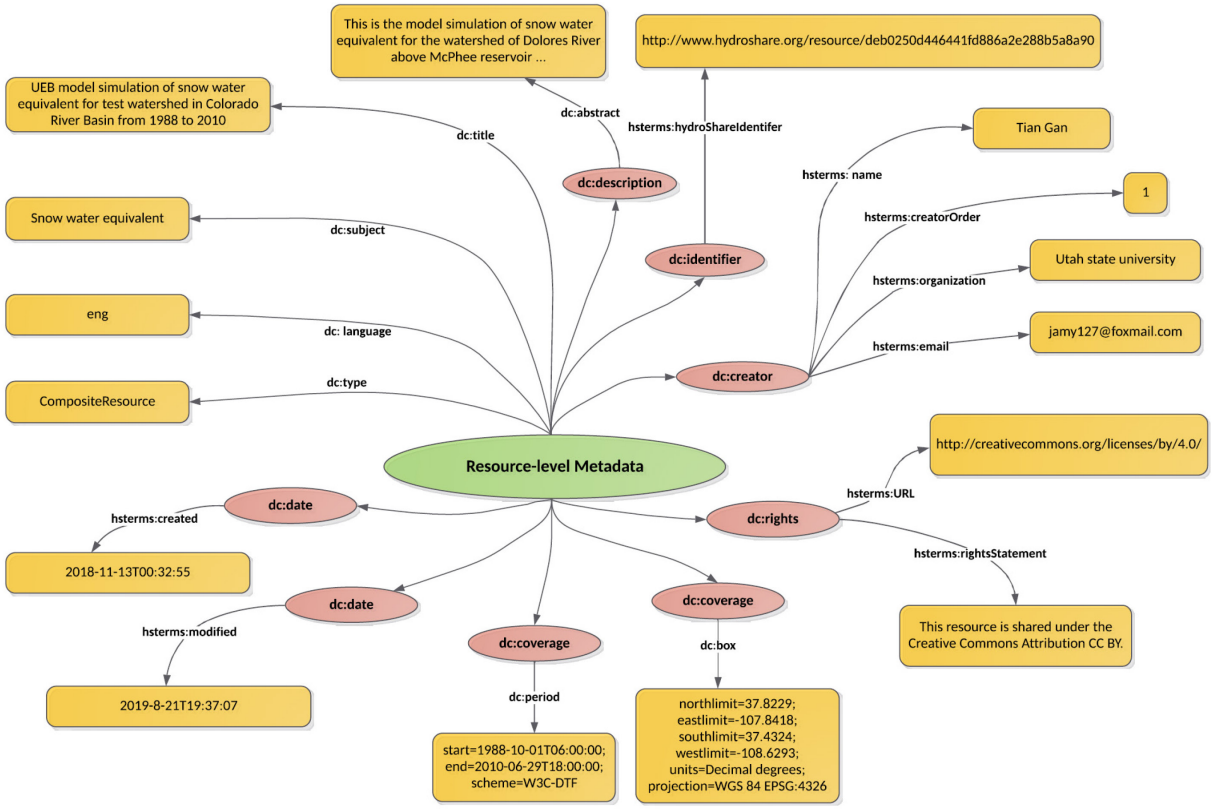
For the NetCDF file, it is recommended that users define the dimensions, variables, and attributes by following the NetCDF-CF and the ACDD conventions. HydroShare does not prevent users from sharing MD data in NetCDF files that do not follow these conventions. However, the functions developed to harvest metadata were based on these conventions. When they are not followed in the NetCDF file, metadata will not be automatically extracted and users will need to enter it manually.

The metadata header information text file is represented in a form called network Common Data form Language (CDL). It is a human-readable text representation of the metadata contained within the NetCDF file. This file includes information about the defined attributes and data structures extracted from the NetCDF file to provide users a brief summary of the file's contents.

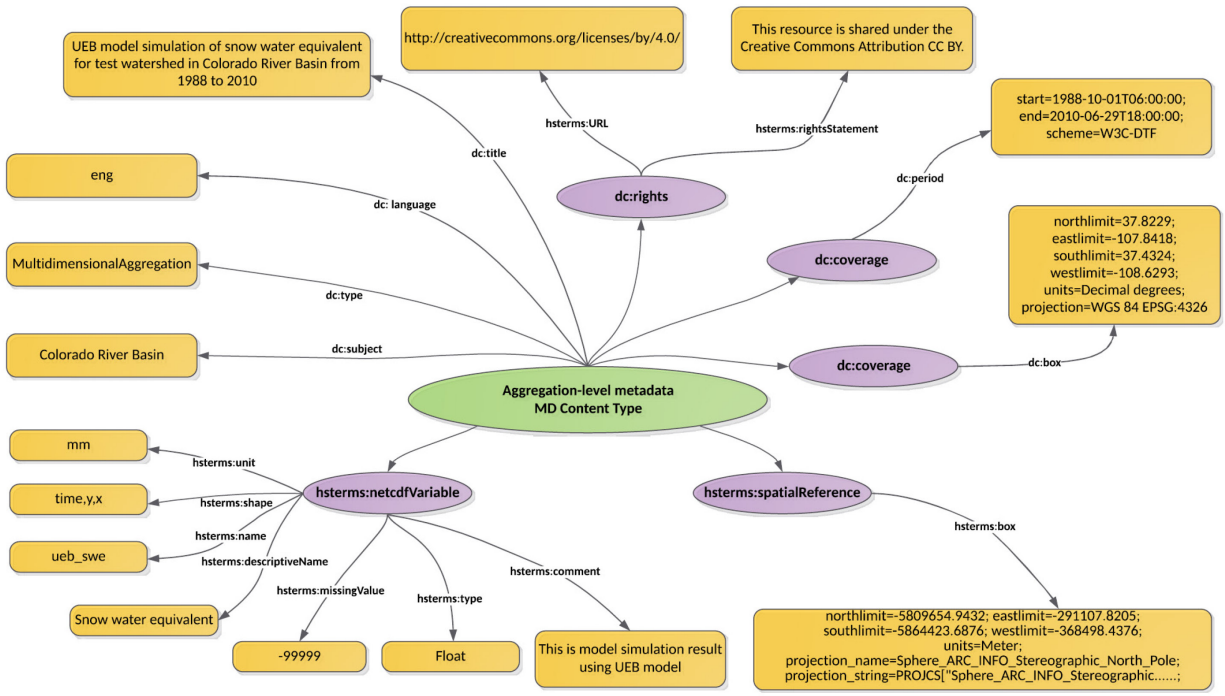
3.1.2 Content type metadata

A HydroShare resource holding a MD aggregation has two sets of metadata elements. Fig. 3(a) shows the resource-level metadata elements. They are based on the standard Dublin Core metadata elements which are common to all HydroShare resources describing the general attributes of a resource (e.g., title, creator, abstract). Fig. 3(b) shows content type metadata elements (or aggregation-level metadata) designed to describe the MD aggregation. The content type metadata includes general elements, which are based on Dublin Core metadata elements to capture the basic information of any content type (e.g., keywords and coverage), and extended elements, which are designed by content type developers to capture the data features in the

NetCDF file (e.g., spatial reference and variable information). Some metadata elements were designed to contain sub-elements. For example, the "creator" metadata element in Fig. 3(a) has sub-elements such as name, organization, and email that apply to the creator. Similarly, the "netcdfVariable" metadata element in Fig. 3(b) includes sub-elements to describe the name, data type, units, etc., for a given variable.



(a)



(b)

Fig. 3. Metadata elements for a HydroShare resource holding a MD aggregation. Panel (a) shows Dublin Core metadata elements held at the resource level. Panel (b) shows metadata elements specific to the MD content type. Each Dublin Core metadata element is prefixed with "dc"; each metadata element defined by HydroShare is prefixed with "hsterms." Individual metadata element names are labeled on the arrows, and examples of their values are shown in the rectangles.

In designing the MD content type, we chose to extract metadata elements held within the NetCDF file and explicitly list them as the resource-level or content type metadata for two reasons. First, this made it easier to present the full metadata description on a resource's landing page in HydroShare, which is the web page where users view and manage the resource, making it more accessible to potential users of the data (e.g., potential users are not required to download or open the NetCDF file to learn about its contents). Second, the resource-level metadata and content type metadata help HydroShare (and potentially other web services) catalogue the

information to enable data discovery, which allows users to search datasets in HydroShare based on desired data attributes.

3.1.3 Content type implementation

In HydroShare, a new content type can be created by inheriting from the abstract content type. Given this general extensibility pattern, we implemented a new content type to manage MD data in HydroShare. A UML diagram of the logical database design for the MD content type is shown in Fig. 4, which presents only major classes, attributes, and methods to demonstrate the organization of the MD content type in HydroShare.

In the UML diagram, there are two categories of classes: (1) the abstract classes that are inherited by any new content type, including the AbstractContentType class, the AbstractContentTypeMetadata class, and the AbstractMetadataElement class; and (2) the classes that define the MD content type, including the MDContentType class, MDMetadata class, SpatialReference class, and NetcdfVariable class. A brief description of each class is listed as follows:

- **AbstractContentType**: an abstract class that provides the interface to represent a content type. It includes the properties and methods for the system to manage a content type and provides a common interface to enable content type related functions. For example, the set_file_type() method is used to check the data format of uploaded datasets. If the data format is for a supported content type in HydroShare, an aggregation will be created.

- **AbstractContentTypeMetadata**: an abstract class used to manage the content type metadata. This class by default contains general metadata elements (e.g., keywords, coverages). The get_xml() method is used to generate the content type metadata XML file. The has_all_required_elements() method is used to check if the required metadata elements for a content type are provided by the user before the resource is shared to the public.

- **AbstractMetadataElement**: an abstract class used to represent a metadata element and define its sub-elements and methods. This class can be used to define extended metadata elements for any content type.

- **MDContentType**: a class that manages the MD content type and inherits from the AbstractContentType class. Functionality specific to the MD content type had to be developed by overriding some methods of AbstractContentType class. For example, the set_file_type() method was overridden to check if the uploaded MD data is in NetCDF data format to create a MD aggregation.

- **MDMetadata**: a class that manages the MD content type metadata and inherits from the AbstractContentTypeMetadata class. This class is composed of general metadata elements and extended metadata elements (e.g., variables and spatial_reference). The get_xml() method and has_all_required_elements() methods in the MDMetadata class override the corresponding methods from the abstract class for the MD content type.

- **SpatialReference**: a class that manages the "spatial reference" extended metadata element for the MD content type. This class inherits from the AbstractMetadataElement and is contained in the MDMetadata class. It includes attributes for storing the sub-elements for spatial reference metadata (e.g., projection_name, projection_string, and value).

- **NetcdfVariable**: a class that manages the "variable" extended metadata element for the MD content type. This class inherits from the AbstractMetadataElement class and is contained in the MDMetadata class. It includes sub-elements to describe a variable stored in the NetCDF file (e.g., name, unit, and type).

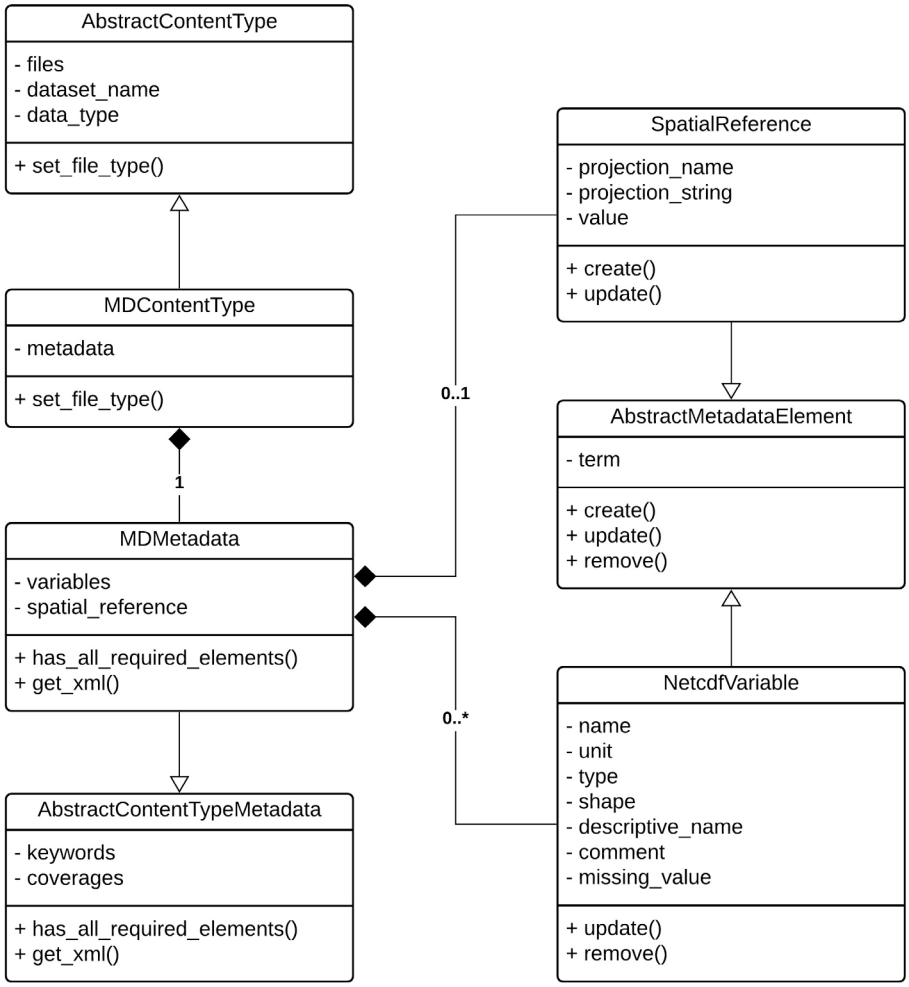


Fig. 4. UML class diagram for the MD content type data model in HydroShare.

3.2 Additional content type functions

As described above, HydroShare provides a base set of functionality for each resource that includes access control, publication, social functions, etc. However, one of the advantages of the design and implementation we describe here is that additional functionality can be developed for a specific content type to support specialized metadata management and sharing of the data via content type specific web services without affecting other content types. In the following subsections, we describe how this functionality was created for the MD content type.

3.2.1 Metadata management functionality

Two functions were designed to (1) extract information (where it exists) from the NetCDF file to populate the resource-level and content type metadata elements, and (2) generate the metadata header information text file. When a user uploads a file with the ".nc" extension, HydroShare will test whether the file holds valid NetCDF content, and if successful, execute these functions to create a MD aggregation from the file.

Aside from metadata extraction functions, we designed functionality for editing metadata in the NetCDF file through HydroShare. We established a mapping between HydroShare's metadata elements and the ACDD and NetCDF-CF conventions (Table 1). When a user edits the metadata in HydroShare, the system utilizes the metadata mapping to check for consistency between the NetCDF file and the HydroShare metadata. If there is a need to add or update the metadata in the NetCDF file, the system will notify the user, and the user can have the system update the file based on the new metadata edits. This functionality helps a user easily update the NetCDF file without having to download and manually edit it. When the initial file includes little metadata, this functionality makes it easy to create metadata in the file that follows NetCDF conventions.

We used the NetCDF4 Python library and NetCDF utility "ncdump" to implement the metadata extraction functions. For files that follow ACDD or NetCDF-CF conventions, the automated metadata extraction function retrieves and populates matched HydroShare metadata elements based on the metadata mapping (Table 1). For files without ACDD metadata elements, but with spatial or temporal coordinate variables given following the NetCDF-CF conventions, spatial and temporal coverage metadata elements determined by reading these data variables are populated in the content type and resource coverage metadata.

The metadata editing functionality was implemented using the NetCDF4 Python library and HydroShare iRODS client interface (Fig. 1). When metadata needs to be updated, the system first copies the original NetCDF file within the iRODS file storage system to a temporary folder. Second, the system writes HydroShare's metadata into the copied file using the NetCDF4 Python library. Then, the system generates a new metadata header information text file from the updated copied file. Finally, the system replaces the original NetCDF file and the metadata header information text file in iRODS with these newly created files.

Table 1. Mapping between HydroShare metadata terms and the NetCDF conventions metadata terms.

HydroShare metadata terms	**NetCDF conventions metadata terms**
creator: name	creator_name (ACDD)
creator: url	creator_url (ACDD)
creator: email	creator_email (ACDD)
contributor: name	contributor_name (ACDD)
coverage (temporal): start	time_coverage_start (ACDD)
coverage (temporal): end	time_coverage_end (ACDD)
coverage (spatial): northlimit	geospatial_lat_max (ACDD)
coverage (spatial): southlimit	geospatial_lat_min (ACDD)
coverage (spatial): eastlimit	geospatial_lon_max (ACDD)
coverage (spatial): westlimit	geospatial_lon_min (ACDD)
description	summary (ACDD)
relation: cites	references (ACDD)
rights	license (ACDD)
source	source (NetCDF-CF)
subject	keywords (ACDD)
title	title (ACDD)
identifier	id (ACDD)
netcdfVariable: unit	unit (NetCDF-CF)
netcdfVariable: descriptiveName	long_name (NetCDF-CF)
netcdfVariable: missingValue	missing_value (NetCDF-CF)
netcdfVariable: comment	comment (NetCDF-CF)
spatialReference: box	attributes for grid mapping variable (NetCDF-CF)

3.2.2 OPeNDAP service

OPeNDAP services help users learn about and work with the contents of the datasets without being required to download them first. Users are able to retrieve a subset of the data for use cases that require smaller spatial or temporal extent. To provide this capability for HydroShare users,

we automated the process of creating an OPeNDAP web service for all publicly shared MD data in HydroShare. Users can access and subset the dataset stored in HydroShare through an OPeNDAP data access form in a web browser or through existing OPeNDAP client software for data visualization or processing.

In HydroShare, support for OPeNDAP services was created by setting up a THREDDS data server to interact with HydroShare's iRODS file storage system. In the system architecture shown in Fig. 1, the data server plays the role of providing web services to enhance the capability for data analysis (orange frame). The data server requires direct file system access to the NetCDF files for its OPeNDAP services. Thus, we used existing iRODS client software to interface to the iRODS Network file system (yellow arrow connecting the orange and purple frames in Fig. 1). We developed a script that copies HydroShare public resources containing MD aggregations efficiently using the iRODS multi-thread parallel transfer "iget" command to a directory on the data server. This copying occurs: 1) when access control for a private resource is changed to public; and 2) when the time stamp of a public resource on the data server is older than that in HydroShare and a data update is needed. This takes advantage of iRODS' high performance parallel data transfers, but in the present implementation does require duplicate storage of NetCDF files. Moreover, since the data server does not support file level user access control as would be required for access to private files in HydroShare, the OPeNDAP service is limited to NetCDF files stored in public or formally published resources in HydroShare. This functionality saves users from the work that would be required to set up a server to host OPeNDAP services for their datasets and gives them the freedom to decide when to make their datasets accessible via OPeNDAP services by using HydroShare's access control settings.

4 Results

We validated the design with an experimental use case to demonstrate how sharing MD data in HydroShare can help users collaborate around datasets and to facilitate the activities involved in the data management life cycle shown in Fig. 5. We considered a use case where a researcher simulated the snowmelt process for the Dolores River watershed in the Colorado River Basin from 1988 to 2010 and shared model results in HydroShare. This was part of a study that the authors were involved in on snowmelt modeling and operational water supply forecasting within the Colorado River Basin (Gan, 2019a). We present the results to demonstrate how the

researcher shared the model output of snow water equivalent as MD data in HydroShare to support the activities from data creation, data publication, to data analysis. This use case involved multiple hypothetical users and was implemented by the first author acting as these users from separate HydroShare accounts.

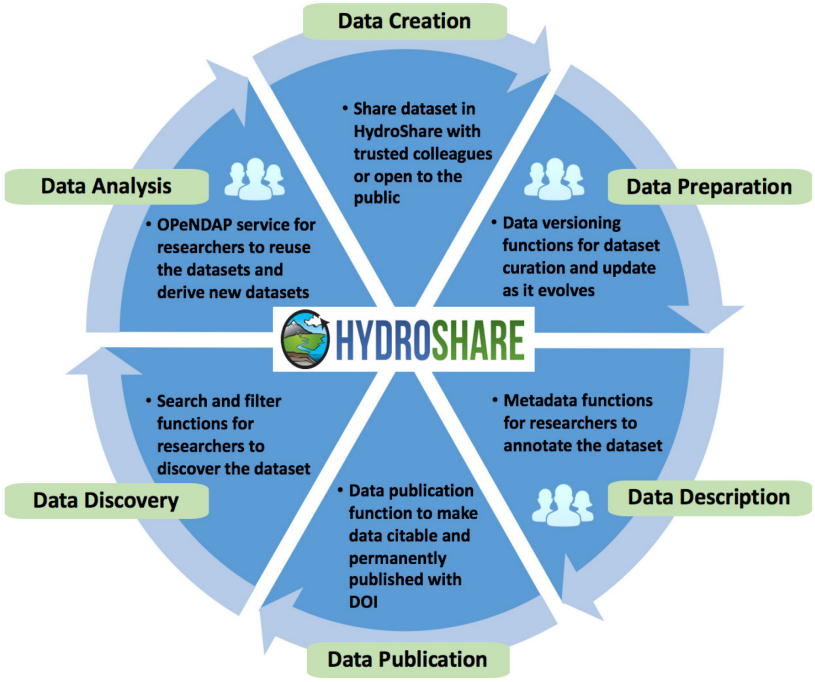


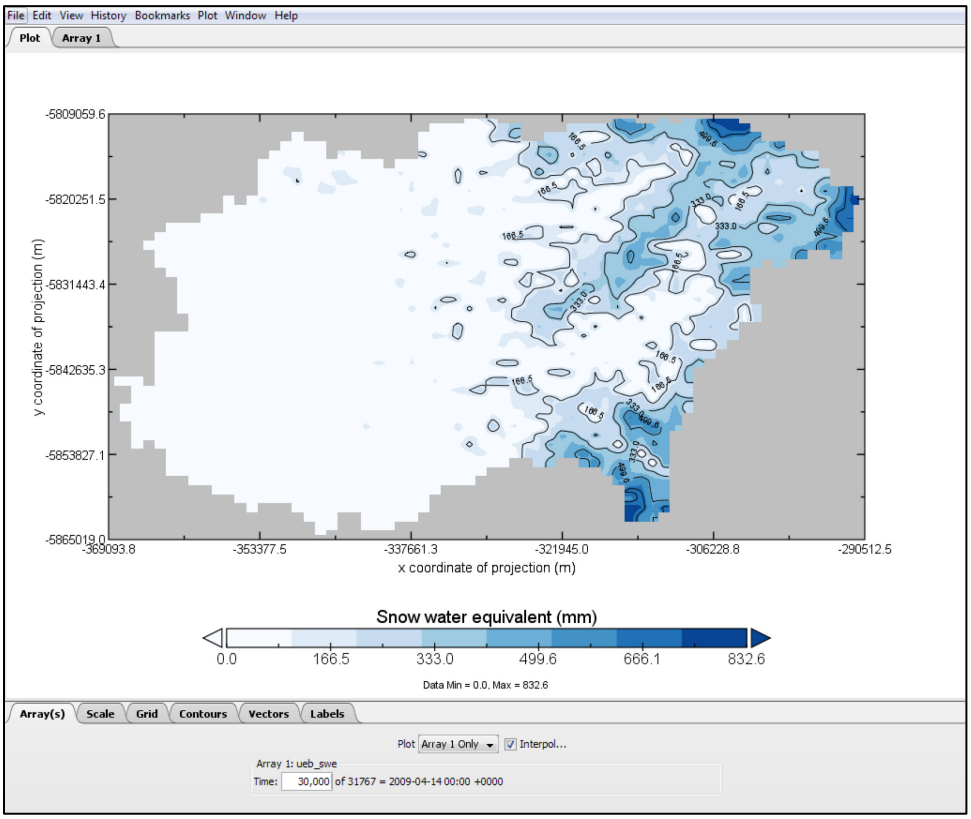
Fig. 5. HydroShare supports the collaborative sharing of MD data with multiple functions that facilitate the cycle of data sharing activities involved in collaborative research.

4.1 Data creation and preparation

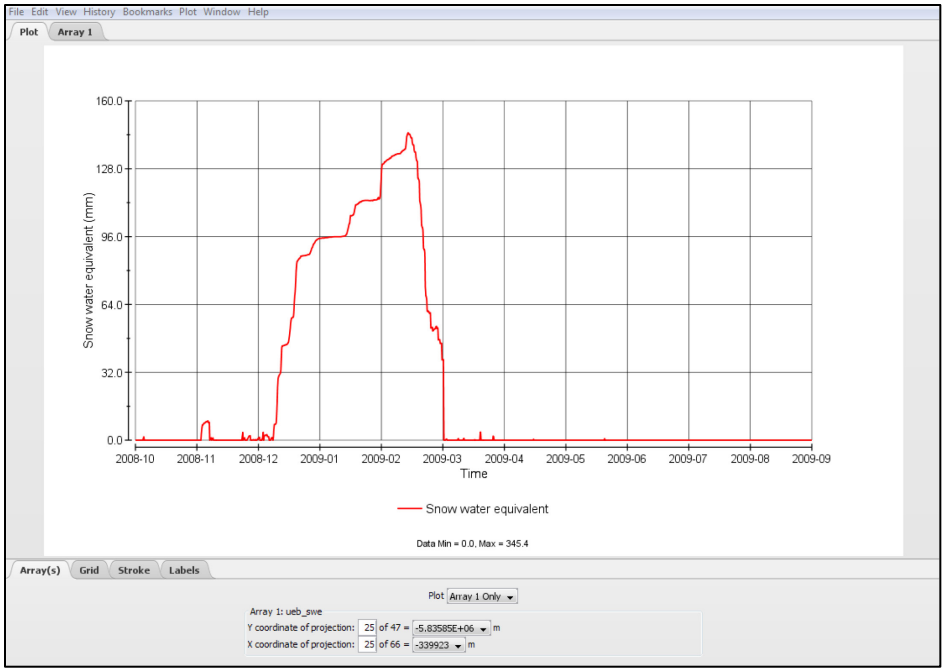
The simulated snow water equivalent datasets were initially stored as separate two-dimensional geospatial data files for each 6-hour time step by the model. This results in thousands of model output files for a 22-year simulation. Sharing of these original model output files has limitations that make data management and reuse difficult. First, information may be lost if any file is missed during the file transfer process. Second, when the original model output files are in a format not widely used by the research community, it is inconvenient to extract subsets that involve thousands of files and difficult to find available software for data analysis or visualization. Thus, the researcher developed a Python script to reorganize and convert the

multiple original model output files into one NetCDF file. Fig. 6 shows the visualization of the use case MD data.

Upon uploading the use case MD data into an empty HydroShare resource, the type of data file was automatically recognized, and a MD aggregation was created in the resource. HydroShare generated a resource landing page, which provides different functions for the user to manage the resource (Fig. 7 (a)) and shows the content type files and content type metadata for the MD aggregation (Fig. 7 (b)). For data preparation, the user can use the data access control and the data versioning functionality (Fig. 7(a)) to collaborate with trusted users to prepare the shared datasets with multiple versions if the original dataset evolves. Users can also edit, delete, copy, and formally publish the resource via its landing page.

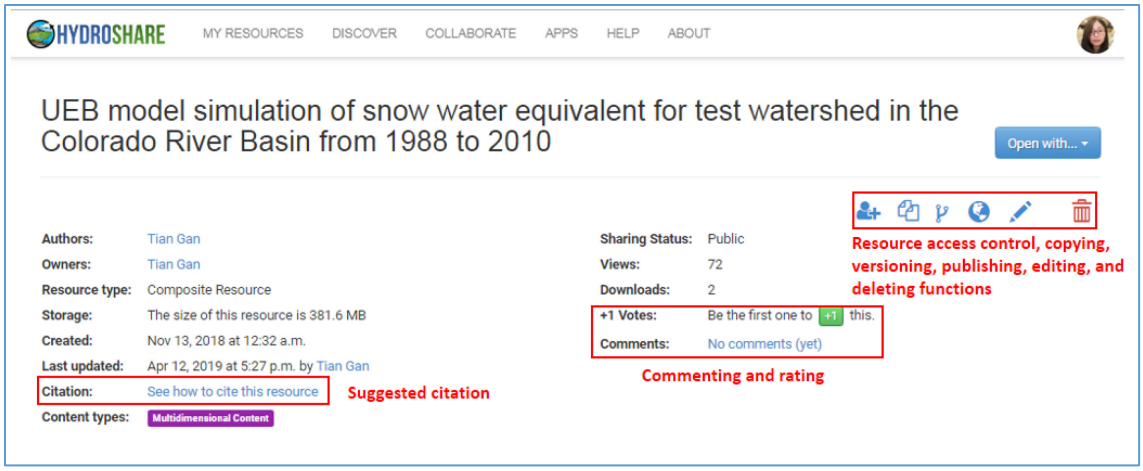


(a)

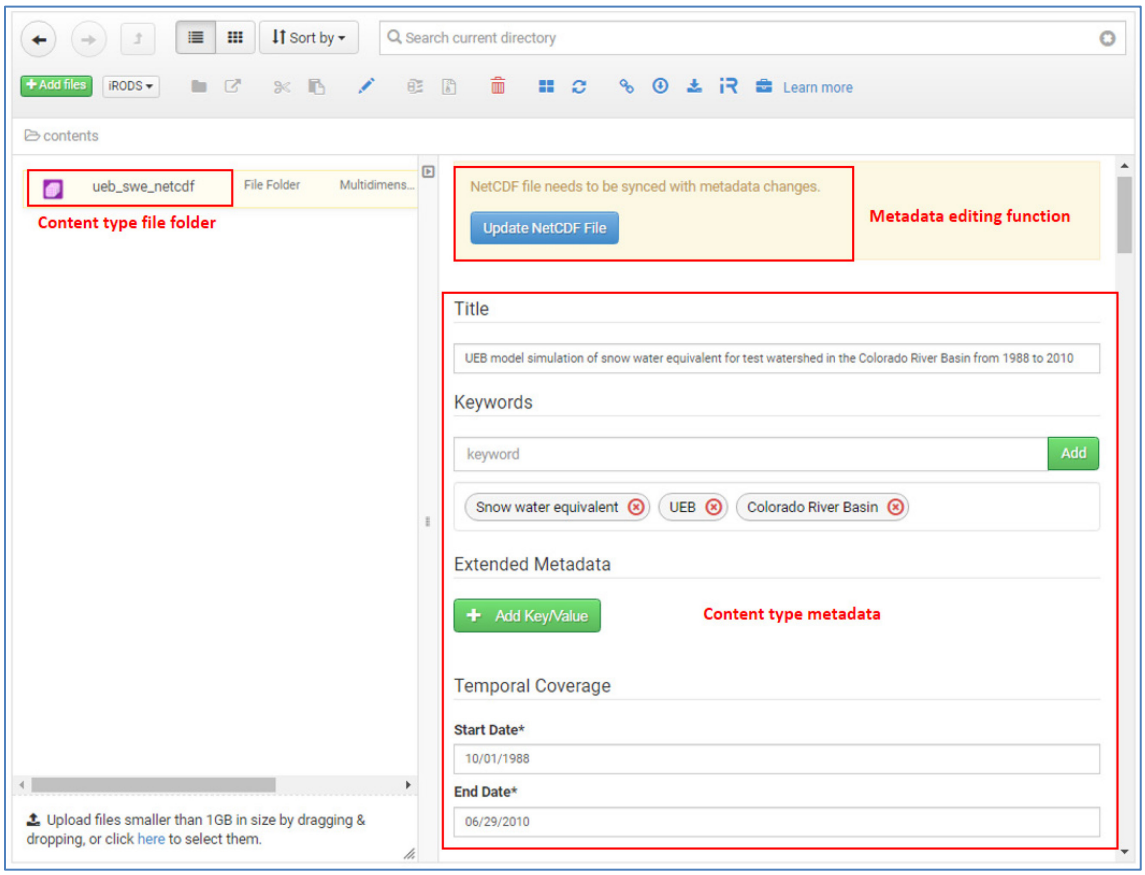


(b)

Fig. 6. Graphs generated in Panoply using the OPeNDAP service to access and subset the use case MD data shared in HydroShare. Panel (a) shows the 2D graph for snow distribution in the test watershed mapped by slicing the data at a specific time step. Panel (b) shows the time series graph for a single grid cell in the test watershed by slicing the data with fixed x and y coordinates.



(a)



(b)

Fig. 7. Resource landing page for the use case MD data. Panel (a) shows the basic data sharing functionality for the resource. Panel (b) shows the content type files and content type metadata for the MD aggregation.

4.2 Data description and publication

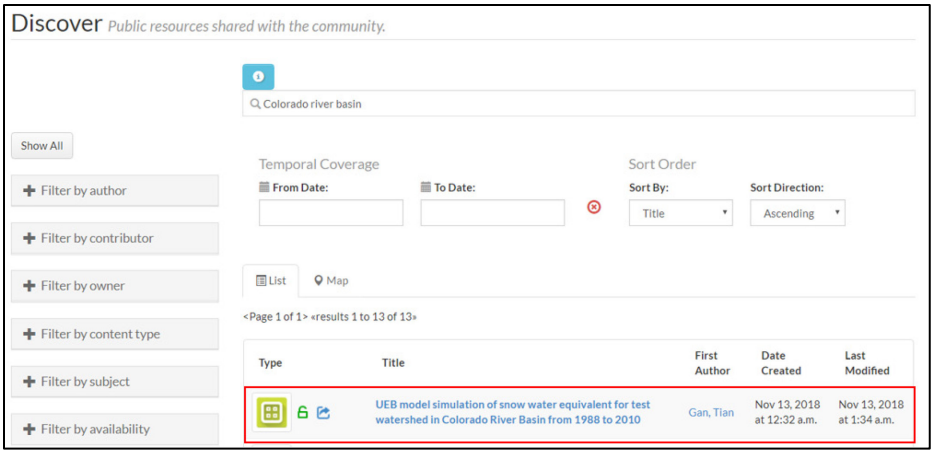
When the MD aggregation was created in HydroShare, two metadata extraction functions were executed. A metadata header information text file was automatically created and stored in the content type file folder (left panel in Fig. 7(b)). The content type metadata such as title, keywords, spatial/temporal coverage, spatial reference, and variable metadata were automatically created by extracting metadata from the NetCDF file (right panel in Fig.7 (b)).

The researcher collaborated with a trusted colleague (referred to as user 1) to update the content type metadata in HydroShare to better describe the data. HydroShare's metadata editing function was then used to update the metadata into the NetCDF file. For instance, when user 1 added a new keyword in the metadata panel, HydroShare's consistency check identified the presence of newly added metadata and showed an "Update NetCDF File" button (Fig. 7 (b)) to inform the user that the NetCDF file could be updated with the new information. Then, user 1 clicked the button to have HydroShare update the metadata in the NetCDF file. This is an example of how, using HydroShare, multiple users can collaborate to annotate the resource with metadata. This metadata editing function enhances NetCDF files to have more attributes that follow NetCDF conventions.

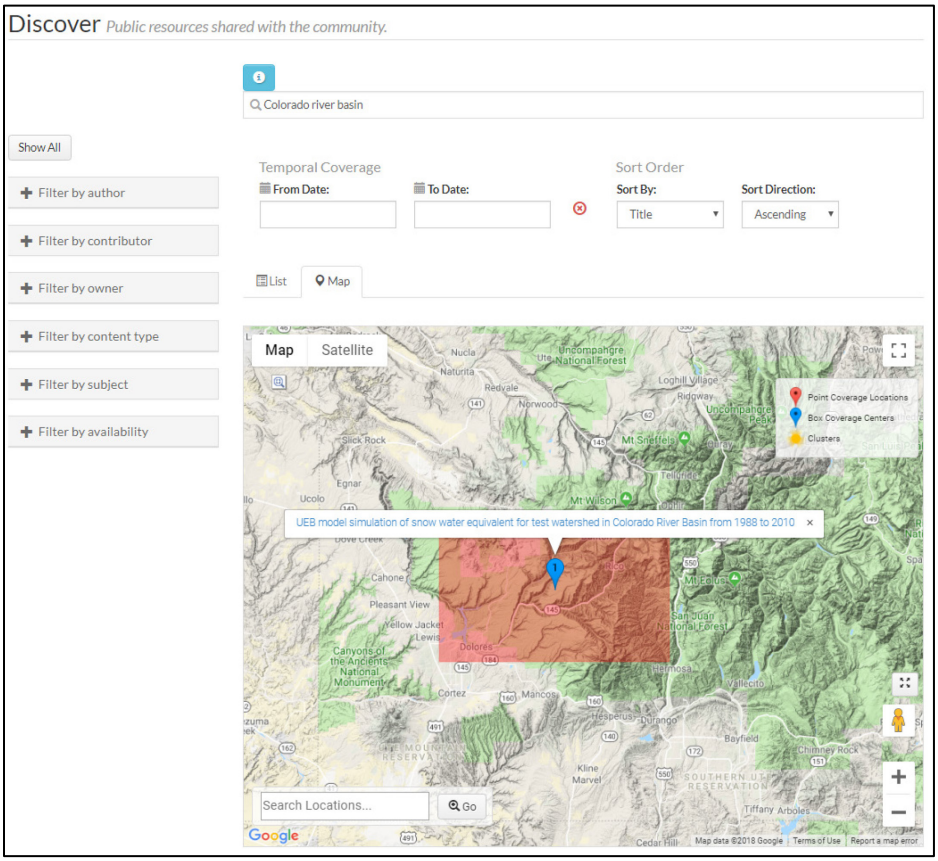
After the data description was completed, the researcher formally published the final data product with an assigned digital object identifier (DOI) in HydroShare (Gan, 2019b). The suggested citation information was generated in HydroShare to encourage proper citation of this dataset (Fig. 7(a)).

4.3 Data discovery and analysis

After the resource was formally published, anyone can discover this dataset using HydroShare's search and filter functions (Fig. 8). A HydroShare user (referred to as user 2) provided a search term ("Colorado river basin"), and HydroShare listed matching resources by querying the HydroShare metadata elements such as title, abstract, and keywords. The search results were filtered based on different metadata facets, such as content type, author, and subject. User 2 identified the use case resource and used HydroShare's map search function to determine the geographic location associated with this dataset.



(a)



(b)

Fig. 8. Data discovery of the use case dataset with the search and filter functions in HydroShare. Panel (a) shows the data discovery with a search term. Panel (b) shows the data discovery with geolocation of the dataset.

After discovering this resource, user 2 decided to reuse a subset of the use case MD data for data analysis. User 2 used the OPeNDAP service from HydroShare and the Panoply client software for data visualization without downloading the use case NetCDF file to a local computer (Fig. 6). Fig. 9 shows the NCO commands used to access, subset, and process the use case dataset using the OPeNDAP service. The code first subsets the data from January 1st to May 31st, 2009 to identify the maximum snow water equivalent for each grid cell, which provides the maximum snow accumulation (assumed to occur within this period) for that year (max.nc). The code then retrieves the data for April 1st and April 15th and evaluates the snow water equivalent difference between the two dates, which provides the analysis result for accumulation (increase) or ablation (decrease) during the period (diff.nc). Water managers often track such snow water equivalent changes in water supply forecasts.

User 2 then uploaded the data analysis code and the derived NetCDF files into HydroShare as a new resource (Gan, 2019c), which started a new cycle of activities involved in collaborative data publication and reuse to improve research reproducibility.

```
#!/bin/bash
ncwa -O -y max -a time -d time,29587,30190 http://hyrax.hydroshare.org:80/opendap/deb0250d446441fd886a2e288b5a8a90/data/contents/ueb_swe_netcdf/ueb_swe_netcdf.nc max.nc
ncks -d time,29947 http://hyrax.hydroshare.org:80/opendap/deb0250d446441fd886a2e288b5a8a90/data/contents/ueb_swe_netcdf/ueb_swe_netcdf.nc april_1.nc
ncks -d time,30006 http://hyrax.hydroshare.org:80/opendap/deb0250d446441fd886a2e288b5a8a90/data/contents/ueb_swe_netcdf/ueb_swe_netcdf.nc april_15.nc
ncbo -O april_1.nc april_15.nc diff.nc
```

Fig. 9. Data analysis for the use case MD data by using the OPeNDAP service and its client software program NCO.

5 Discussion

The use case illustrated how organizing MD data using the NetCDF data format and sharing it in HydroShare provided added value in terms of functionality for metadata management and data analysis. When compared with other MD data sharing methods, this approach has the following advantages:

- It provides functionality to capture, expose, and edit the metadata stored in the NetCDF file. The manage access function enables users to collaborate on metadata editing and thus improve its description of the data by following metadata standards. Other data sharing methods either do not automatically expose the metadata from a NetCDF file or make it difficult to collaboratively edit the metadata in the NetCDF file.

- It provides automated OPeNDAP services with access control for shared datasets to support data analysis that enhance opportunities for collaboration around the data. Other data sharing methods either do not provide an OPeNDAP service or require effort to set up and maintain a server and service.

- It provides better data discovery functionality for the shared datasets. It supports keyword and geolocation searches based on a catalog of metadata extracted from the NetCDF file or input by the data provider. Other data sharing methods provide limited discovery capability to search the MD data based on their attributes.

In our approach, several key factors make this advantageous functionality available: 1) we adopted a standard data format (NetCDF) to organize MD data, which has conventions that standardize how data and metadata are organized in the file to improve the interoperability of datasets; 2) we utilized existing tools and standard data services to develop automated functions for metadata management and data analysis to promote MD data sharing and reuse ; and 3) HydroShare's resource data model design helps improve consistent data discovery, access, and publishing across the broad range of data types used by scientists in the hydrology domain, while at the same time allowing value added functionality for specific data types.

However, there are limitations that need further improvements for sharing MD data in HydroShare. One limitation is that some users may not be familiar with the NetCDF data format. Users need to learn how to organize MD data in this data format for data sharing. Another limitation is web-based visualization. There is a need for additional functionality that provides researchers with greater capacity to process and visualize datasets directly without transferring the data or subsets of the data between the data sharing system and their local computers. To address these limitations, one possible solution is to create and share Jupyter Notebook examples in HydroShare to support MD data visualization and demonstrate how to convert MD data in various formats into NetCDF format.

In the future, we are considering support for other widely used standard variable ontologies such as the Community Surface Dynamics Modeling System (CSDMS) standard names (Peckham, 2004) in metadata management functions. Another enhancement we are considering is enabling web services beyond the OPeNDAP service for shared MD datasets to increase the value and

usability of NetCDF data, including Open Geospatial Consortium (OGC) Web Map Service (WMS), OGC Web Coverage Service (WCS), NetCDF Subsetting service (NCSS), etc.

6 Conclusions

HydroShare is a web based hydrologic information system that provides researchers with a platform to share their hydrologic data and models. As MD data is one of the widely used data types in hydrologic research, we developed an approach to support sharing of this data type within HydroShare. Our approach was aimed at overcoming challenges for sharing MD data, including: 1) lack of a single, accepted standard data format to support the interoperability needed for data sharing and analysis; 2) lack of advanced functions to preview or edit the metadata in the file; and 3) difficulty in subsetting data from large datasets for data visualization and processing. To address these challenges, we adopted a standard data format (NetCDF) and standard metadata elements to manage MD data in HydroShare, and we implemented value added functionality to manage metadata and support data reuse.

The use case presented demonstrates the new capabilities in HydroShare and shows that researchers can share MD data in a NetCDF file with the metadata automatically exposed in the system. Metadata can be edited collaboratively in HydroShare and automatically updated in the NetCDF file. Once publicly shared, users can subset the data with the automatically configured OPeNDAP service for visualization and analysis without effort to set up and maintain a server. In concert with existing HydroShare functionality (e.g., data discovery, data publishing, and access control), the work described here enables relatively straightforward sharing and formal publication of MD data. This increases transparency and reproducibility of the associated research and promotes reuse of data and the derivation of additional value from research data investments.

Beyond the context of the new functionality we have demonstrated, another contribution of this work is that the methods we developed for improving sharing of MD data can be used as examples for supporting other data types in HydroShare or for better supporting MD data in other systems. Cyberinfrastructure developers who are going to build or have built a data sharing system to support MD data sharing can use the recommendations of this work to organize data in a standard data format and document the datasets using the standards-based metadata. Using the patterns we established, they may be able to create standard data services or develop new

functionality to facilitate metadata management, data analysis, or visualization. Adopting standard formats and techniques across data repositories could lead to a level of interoperability that is worth considering in the future.

Funding:

This work was supported by the National Science Foundation [OCI-1148453, OCI-1148090].

Acknowledgements

This work was supported by the National Science Foundation under collaborative grants OCI-1148453 and OCI-1148090. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We are also thankful to others on the HydroShare development team for providing suggestions on function design and implementation for HydroShare system.

References

David, C.H., Maidment, D.R., Niu, G.Y., Yang, Z.L., Habets, F., Eijkhout, V., 2011. River network routing on the NHDPlus dataset. J. Hydrometeorol. 12, 913–934. https://doi.org/10.1175/2011JHM1345.1

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Blower, J., Caron, J., Signell, R., Bentley, P., Rappa, G., Höck, H., Pamment, A., Juckes, M., Raspaud, M., 2017. NetCDF climate and forecast (CF) metadata conventions. http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/cf-conventions.pdf (accessed 8.18.19).

ESIP, 2017. Attribute convention for data discovery 1-3. http://wiki.esipfed.org/index.php/Attribute_Convention_for_Data_Discovery (accessed 8.18.19).

Gan, T., 2019a. Advancing Cyberinfrastructure for Collaborative Data Sharing and Modeling in Hydrology (Doctoral dissertation). Utah State University, Logan, UT. https://digitalcommons.usu.edu/etd/7618/

Gan, T., 2019b. UEB model simulation of snow water equivalent for test watershed in the Colorado River Basin from 1988 to 2010. HydroShare. https://doi.org/10.4211/hs.deb0250d446441fd886a2e288b5a8a90

Gan, T., 2019c. Data analysis of snow water equivalent for test watershed in the Colorado River Basin in 2009. HydroShare. Gan, T. (2019). Data analysis of snow water equivalent for test watershed in the Colorado River Basin in 2009. HydroShare. https://doi.org/10.4211/hs.5f685bb517eb4343b1b293a921f14639

Guo, Q., Zhang, Y., He, Z., Min, Y., 2015. Web-based data integration and interoperability for a massive spatial-temporal dataset of the Heihe River Basin eScience framework. Adv. Meteorol. Volume 201. http://dx.doi.org/10.1155/2015/982062

Heard, J., Tarboton, D.G., Idaszak, R., Horsburgh, J.S., Bedig, A., Castronova, A.M., Couch, A., Frisby, C., Gan, T., Goodall, J., Jackson, S., Livingston, S., Maidment, D., Martin, N., Miles, B., Mills, S., Sadler, J., Valentine, D., Zhao, L., 2014. An architectural overview of HydroShare, a next-generation hydrologic information system. Proceedings of the 11th International Conference on Hydroinformatics, HIC 2014, New York City, USA.

Horsburgh, J.S., Morsy, M.M., Castronova, A.M., Goodall, J.L., Gan, T., Yi, H., Stealey, M.J., Tarboton, D.G., 2015. HydroShare: Sharing diverse hydrologic data types and models as social objects within a Hydrologic Information System,. J. Am. Water Resour. Assoc. 27517. https://doi.org/10.1111/1752-1688.12363

Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., Zaslavsky, I., 2008. A relational model for environmental and water resources data. Water Resour. Res. 44. https://doi.org/10.1029/2007WR006392

Horsburgh, J.S., Tarboton, D.G., Piasecki, M., Maidment, D.R., Zaslavsky, I., Valentine, D., Whitenack, T., 2009. An integrated system for publishing environmental observations data. Environ. Model. Softw. 24, 879–888. https://doi.org/10.1016/j.envsoft.2009.01.002

Lagoze, C., Sompel, H. Van de, Johnston, P., Nelson, M., Sanderson, R., Warner, S., 2008. Open archives initiative object reuse and exchange: ORE user guide – primer. http://www.openarchives.org/ore/1.0/primer (accessed 8.18.19).

Morsy, M.M., Goodall, J.L., Castronova, A.M., Dash, P., Merwade, V., Sadler, J.M., Rajib, M.A., Horsburgh, J.S., Tarboton, D.G., 2017. Design of a metadata framework for environmental models with an example hydrologic application in HydroShare. Environ. Model. Softw. 93, 13–28. https://doi.org/10.1016/j.envsoft.2017.02.028

NASA, 2018. Panoply netCDF, HDF and GRIB data viewer. http://www.giss.nasa.gov/tools/panoply/ (accessed 8.18.19).

OGC, 2011. OGC Network Common Data Form (NetCDF) core encoding standard version 1.0. http://www.opengis.net/doc/IS/netcdf/1.0 (accessed 8.18.19).

OPeNDAP, 2017. The Hyrax data server installation and configuration guide. https://opendap.github.io/hyrax_guide/Master_Hyrax_Guide.html (accessed 8.18.19).

Peckham, S.D., 2014. The CSDMS standard names: Cross-domain naming conventions for describing process models, data sets and their associated variables. Proc. - 7th Int. Congr. Environ. Model. Softw. Bold Visions Environ. Model. iEMSs 2014 1, 67–74

Rajib, M.A., Merwade, V., Kim, I.L., Zhao, L., Song, C., Zhe, S., 2016. SWATShare - A web platform for collaborative research and education through online sharing, simulation and visualization of SWAT models. Environ. Model. Softw. 75, 498–512. https://doi.org/10.1016/j.envsoft.2015.10.032

Sen Gupta, A., Tarboton, D.G., Hummel, P., Brown, M.E., Habib, S., 2015. Integration of an energy balance snowmelt model into an open source modeling framework. Environ. Model. Softw. 68, 205–218. https://doi.org/10.1016/j.envsoft.2015.02.017

Tarboton, D.G., Horsburgh, J.S., Maidment, D.R., Whiteaker, T., Zaslavsky, I., Piasecki, M., 2009. Development of a community hydrologic information system, in: 18th World IMACS / MODSIM Congr. 988–994.

Tarboton, D.G., Idaszak, R., Horsburgh, J.S., Heard, J., Ames, D., Goodall, J.L., Band, L., Merwade, V., Couch, A., Arrigo, J., Hooper, R., Valentine, D., Maidment, D., 2014. HydroShare: Advancing collaboration through hydrologic data and model sharing, in: 7th International Congress on Environmental Modelling and Software. International Environmental Modelling and Software Society (iEMSs).

Thornton, P.E., Running, S.W., White, M.A., 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. J. Hydrol. 190, 214–251. https://doi.org/10.1016/S0022-1694(96)03128-9

Unidata, 2017a. THREDDS data server (TDS). http://www.unidata.ucar.edu/software/thredds/current/tds/ (accessed 8.18.19).

Unidata, 2017b. Integrated data viewer (IDV). http://www.unidata.ucar.edu/software/idv/docs/userguide/toc.html (accessed 8.18.19).

Unidata, 2017c. Software for manipulating or displaying NetCDF Data. http://www.unidata.ucar.edu/software/netcdf/software.html (accessed 8.18.19).

Zender, C.S., 2008. Analysis of self-describing gridded geoscience data with netCDF Operators (NCO). Environ. Model. Softw. 23, 1338–1342. https://doi.org/10.1016/j.envsoft.2008.03.004