Computers & Geosciences 159 (2022) 105024

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

BASIN-3D: A brokering framework to integrate diverse environmental data

Charuleka Varadharajan [a,*], Valerie C. Hendrix [b], Danielle S. Christianson [b], Madison Burrus [a], Catherine Wong [b], Susan S. Hubbard [a], Deborah A. Agarwal [b]

[a] Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA, 94010, USA
[b] Computing Sciences Area, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA, 94010, USA

ARTICLE INFO

Keywords:
Data integration
Multiscale diverse data
Synthesis
Environmental data

ABSTRACT

Diverse observational and simulation datasets are needed to understand and predict complex ecosystem behavior over seasonal to decadal and century time-scales. Integration of these datasets poses a major barrier towards advancing environmental science, particularly due to differences in the structure and formats of data provided by various sources. Here, we describe BASIN-3D (Broker for Assimilation, Synthesis and Integration of eNvironmental Diverse, Distributed Datasets), a data integration framework designed to dynamically retrieve and transform heterogeneous data from different sources into a common format to provide an integrated view. BASIN-3D enables users to adopt a standardized approach for data retrieval and avoid customizations for the data type or source. We demonstrate the value of BASIN-3D with two use cases that require integration of data from regional to watershed spatial scales. The first application uses the BASIN-3D Python library to integrate time-series hydrological and meteorological data to provide standardized inputs to analytical and machine learning codes in order to predict the impacts of hydrological disturbances on large river corridors of the United States. The second application uses the BASIN-3D Django framework to integrate diverse time-series data in a mountainous watershed in East River, Colorado, United States to enable scientific researchers to explore and download data through an interactive web portal. Thus, BASIN-3D can be used to support data integration for both web-based tools, as well as data analytics using Python scripting and extensions like Jupyter notebooks. The framework is expected to be transferable to and useful for many other field and modeling studies.

1. Introduction

It is important to predict how watersheds and other ecosystems will respond to changing environmental conditions for optimal management of natural resources. The study and modeling of natural environments requires integration of diverse observations that sample different properties of these complex systems (Fig. 1). Often, data are required at multiple spatial and temporal scales to enable both quantification of fine-scale processes and their aggregation to larger scales. For example, heterogeneous data on water cycle processes, such as precipitation, river discharge, soil moisture, infiltration, evapotranspiration, need to be integrated with data on topography, soil properties, and water use to predict water availability and quality at seasonal to decadal scales (e.g. Krysnova and Arnold, 2008; Maxwell et al., 2015; USGS Water Resources, 2020).

The development of long-term monitoring networks and research data management requirements have led to an unprecedented volume and diversity of available environmental data (Rode et al., 2016). Advances in data analytics, scale-aware mechanistic modeling, Machine Learning (ML) and other computational methods create opportunities to integrate diverse environmental data into a predictive framework (Hubbard et al., 2020). Yet, these data are underutilized in studies that seek to improve the understanding and predictions of environmental systems, in part due to major challenges associated with the discovery and integration of relevant datasets for scientific analysis and modeling. Currently data integration across different providers is an arduous, time-intensive task, needing considerable harmonization efforts (refer to Table 1 for a list of common data integration terms). Most data sources are organized around the provider's requirements and not the user's needs to discover, integrate and utilize the data. For example, in the United States (US), several monitoring networks have been established across federal, state and local agencies that have not adopted common data or metadata standards. Different suites of physical, chemical and biological parameters are measured across regions and time periods

* Corresponding author.
E-mail address: cvaradharajan@lbl.gov (C. Varadharajan).

https://doi.org/10.1016/j.cageo.2021.105024
Received 15 October 2020; Received in revised form 12 November 2021; Accepted 13 December 2021
Available online 15 December 2021

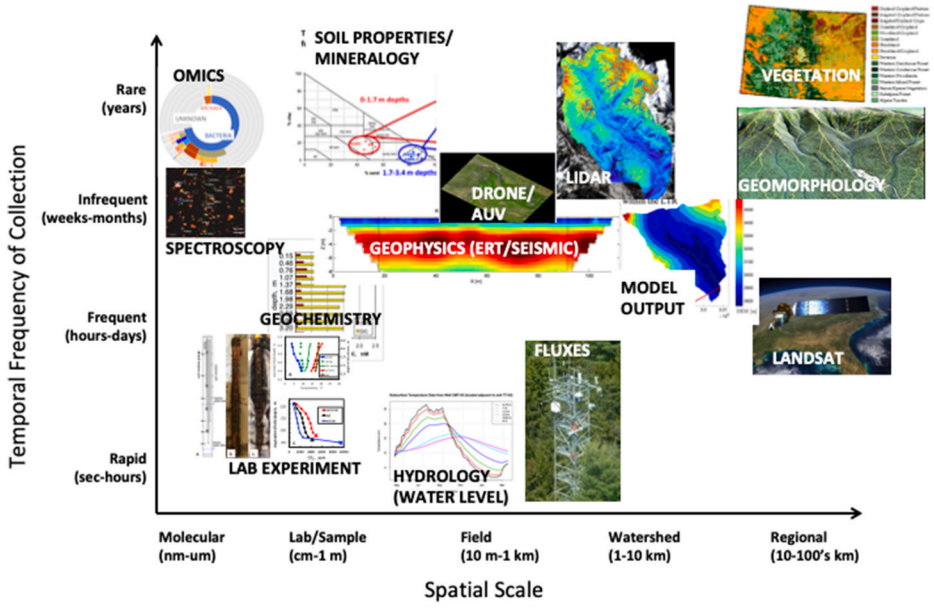

Fig. 1. Gaining an understanding of multi-scale ecosystem processes often requires acquisition and integration of a variety of data that sample different properties across multiple temporal and spatial scales. Examples shown here include data associated with fields of hydrology, meteorology, geology, geochemistry, ecology and biology, which can all be useful for understanding watershed behavior.

using a variety of methods. Data are served using an assortment of formats, variable names, units and schemas (e.g. Sprague et al., 2017). As an example, stream discharge and water level data collected by the US Geological Survey (USGS), US Bureau of Reclamation (USBR), several state agencies, local efforts (e.g. water management districts, watershed associations) and numerous research networks such as the Long Term Ecological Research (LTER), Critical Zone Observatories (CZO), National Ecological Observatory Network (NEON) are published with heterogeneous data formats and access mechanisms. Thus researchers typically create manual, one-off integrated products for their scientific needs, which become outdated when the data change.

Several approaches have been used to integrate disparate datasets over the past few decades. One method, known as data warehousing, integrates data from various sources into a centralized database, which can be queried to retrieve synthesized data streams. Some systems using this approach, like the National Water Quality portal (https://www.waterqualitydata.us/) and Ameriflux (https://ameriflux.lbl.gov/), combine structurally similar data by requiring sources to provide data in particular formats (Blodgett et al., 2015; Pastorello et al., 2017). However, centralized databases can become outdated and are difficult to maintain as the number of desired sources, data types and volumes grow.

Data federation, also known as the hub and spoke model (Haas et al., 2002), is an alternate approach that has gained traction. Here, data are left at the original sources and an intermediate brokering software maintains a catalog and retrieves data on demand (Genesereth, 2010; Nativi et al., 2013). This allows users to access the latest version of the data from different sources as though it were available in a central location. The brokering approach has been adopted by systems such as the Group on Earth Observations (GEOSS) Discovery and Access Broker (Nativi et al., 2014) and the related BCube brokering framework (Khalsa, 2017). The US National Groundwater Monitoring Network (https://cida.usgs.gov/ngwmn/index.jsp) uses an advanced brokering

Table 1
A list of commonly-used terms that have specific definitions in this paper and BASIN-3D. The first four terms are hierarchical in that the prior term is a component of the following term. For example harmonization is the transformation of different data formats into the same formats. The definitions and examples are provided to illustrate the intent of the term's usage and are not necessarily capabilities available in the current version of BASIN-3D.

Term	Definition	Example
Transform/Transformation	Convert or translate data into different units, variable names, and/or structural formats.	Air temperature data with variable name T_{air} and units deg F are translated into variable name TA and units deg C.
Harmonize/Harmonization	*Transform* (see definition above) data of the same type collected with comparable methodology into the same variable names, unit terms, and structural formats. In some cases, units may be converted into a preferred standard.	River discharge gage data from different sources are harmonized by transformations into a common BASIN-3D controlled vocabulary term 'RDC' and OGC representation 'time-value pair', with conversion to common SI units of 'm^3/sec'.
Synthesize/Synthesis	Combine data of different types from different sources into the same overarching data structure. Data of similar types from different sources are *harmonized* (see definition above). Data are matched in space (by monitoring feature) and by time (using the timestamp provided by the data source).	Geochemical data from different sources (e..g DOE databases, National Water Quality Portal) are harmonized into the same units, variable names, and structural formats. These geochemical data are then combined with other types of data, like meteorological and hydrological data, using the same units, variable names, and structures across different data types when possible. Data with the same time stamps and locations are aligned (e.g. in the same row or column).
Integrate/Integration	Provide a single point of access in a unified view for *synthesized* (see definition above) data from different sources.	Water quality data synthesized across different sources and data types are available through the BASIN-3D synthesis web or python APIs.
Aggregate/Aggregation	Group data by a defined spatial entity and/or temporal period. The grouped data may be represented by "aggregate" value(s) that are often statistical calculations of the grouped data.	Data collected at 15-min intervals are aggregated to hourly or daily values.

approach to synthesize datasets from various sources to support a portal with interactive visualizations; however, this system only handles specific types of groundwater monitoring data (water level, quality, lithology), and requires cooperative agreements with providers to standardize their data to facilitate data exchange (https://acwi.gov/sogw/ngwmn_framework_report_july2013.pdf). One of the more successful implementations of a brokering approach is the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) Hydrologic Information System (HIS; https://hiscentral.cuahsi.org), which transforms diverse time-series data using the WaterOneFlow web services into a standardized WaterML format with common variable and unit names from the CUAHSI controlled vocabularies (Horsburgh et al., 2009, 2016). The HIS enables unified access to data synthesized from over 95 providers via the Hydroclient interactive portal. However, the WaterOneFlow web services need to be hosted and maintained by the provider or CUAHSI, which limits its application to data sources that belong to the HIS ecosystem. The HIS system also does not support large data downloads as the Hydroclient limits search and access to 25,000 results. This tends to be problematic for intensive data-driven applications such as ML, where programmatic access to large amounts of data from sources outside of HIS may be needed.

More recently, federated computational tools have emerged to provide streamlined access to big datasets across different data sources on the cloud. Examples include the Pangeo platform (http://pangeo.io), which creates interactive and reproducible open source workflows to discover, analyze and visualize large geoscience datasets on cloud resources, and is integrated with data discovery tools such as intake (https://github.com/intake) and the SpatioTemporal Asset Catalog (https://stacspec.org/). However, these federated approaches and tools only produce integrated data catalogs or co-located datasets. They do not parse or translate data from different sources into an integrated view.

Thus, despite the advances made by these systems, the most essential and difficult data integration tasks—the conceptual reconciliation of the various ways data is served by the providers and harmonization of data formats, units, and semantics—are still left to the end users. There remains a critical need for generalizable frameworks that can be easily used by environmental scientists or practitioners to integrate vastly diverse data types, structures and formats. Such frameworks would automate integration of disparate, multiscale data on-demand from heterogeneous databases, as they are being dynamically updated on the original sources, and enable users to easily search, subset and retrieve synthesized data (BERAC, 2013).

To address this gap, we present BASIN-3D (Broker for Assimilation, Synthesis and Integration of Diverse, Distributed Datasets), a generic, extensible data-brokering software designed to integrate heterogeneous, multiscale data into a coherent framework and create synthesized datasets that scientists can easily utilize. We aimed to develop a broker that would allow its users to create a single access point for diverse environmental databases and data types by retrieving data from the sources on-demand and harmonizing the data streams to provide an integrated view. This would help bridge the gap between the user-centric objective, to easily find, subset and synthesize relevant data, and the provider-centric objective, to organize and publish their entire data collection. BASIN-3D enables users to access the most recent version of the data from each of the sources, as if the data were available from a single provider, without losing the data provenance.

We demonstrate the utility of BASIN-3D with two prototype implementations that require integration of diverse time-series measurements at watershed to regional scales, from multiple sources that use different data formats, organization and terminologies. BASIN-3D is the first brokering software to our knowledge that enables custom integration of heterogeneous time-series data from users' preferred data sources, without requiring coordination with providers. Thus, it is intended to make the typical scientific workflow of acquiring and harmonizing data from diverse sources more repeatable and reproducible. Users of BASIN-3D can remain agnostic to the location, format, structure and authorization requirements of the sources. Multiple clients, such as web portals, visualization tools, and analytical and modeling codes, can connect to the BASIN-3D Application Programming Interfaces (APIs) to access synthesized datasets. The flexibility provided by this approach is essential for modern environmental data science applications, where there is a need to synthesize disparate data on demand, to address evolving science questions or stakeholder needs.

This paper is organized as follows. In Section 2, we describe the features and architecture of BASIN-3D. Section 3 provides a demonstration of BASIN-3D to integrate diverse observations at watershed to regional scales for two U.S. Department of Energy (DOE) projects. Section 4 discusses the BASIN-3D approach along with its advantages and limitations, as well as opportunities for building generalizable frameworks that integrate large, complex environmental data.

2. Methods: Data synthesis constructs and features of BASIN-3D

We followed a *scientist-centered design* approach to develop BASIN-3D, by first understanding the end product desired through use cases. Then, we worked backward to identify the corresponding data sources, synthesis and transformations, and interfaces needed (Ramakrishnan et al., 2014). The BASIN-3D data integration approach uses a central abstract data-model schema with mappings to different sources (Genesereth, 2010). BASIN-3D utilizes the Open Geospatial Consortium (OGC) and International Standards Organization (ISO) standard "Observations and Measurements; OGC 10-004r3/ISO 19156: 2013" and "OGC Timeseries Profile of Observations and Measurement; OGC 15-043r3" schemas (Cox, 2011; Tomkins and Lowe, 2016).

The following sections describe the key concepts, architecture and usage models of BASIN-3D. Although the current implementations of BASIN-3D use concepts tailored to time-series data used in watershed science, the framework is applicable for a variety of environmental data types and applications.

2.1. Multiscale spatial and temporal representations

Environmental observations made at different spatial scales are often grouped using hierarchical relationships to facilitate data management, synthesis and aggregation. The definition of spatial extents and groupings can be domain or study specific, and can include multiple hierarchical relationships (Christianson et al., 2017). For example, an observation made at a point location can be represented as a member of a USGS Hydrologic Unit Code (HUC) hierarchy of hydrological features that includes the river basin and watershed, as well as part of other spatial hierarchies like the study plot and site (Fig. 2).

BASIN-3D uses constructs from the OGC standard to represent multiscale spatial elements with their location features, associated groupings and hierarchies. In particular, BASIN-3D uses '*Monitoring Feature*' entities which inherit components of the OGC entities '*Feature*', '*Sampling Feature*', '*Spatial Sampling Feature*' (Appendix 1) [Cox, 2011; Tomkins and Lowe, 2016]. *Monitoring Features* are classified by a controlled list of *Feature Types* that represent spatial features at different scales relevant to watershed sciences: '*Region*', '*Subregion*', '*Basin*', '*Subbasin*', '*Watershed*', '*Site*', '*Plot*', '*Horizontal Path*', '*Vertical Path*' and '*Point*' (e.g. Fig. 2); this list is specific to each BASIN-3D implementation and can be expanded to include additional *Feature Types* relevant to other disciplines. *Monitoring Features*, as an extension of *Spatial Sampling Features*, are geographic entities that have a shape property to describe their spatial geometry as one of four types specified by OGC: '*point*' (e.g., point, specimen), '*curve*' (e.g., river, well, tower), '*surface*' (e.g., river basin, watershed, site, plot), and '*solid*' (e.g. lidar cloud). The physical location coordinates of a *Monitoring Feature* are represented using the Federal Geographic Data Committee (FGDC) data standard (FGDC 1998), which provides support for multiple spatial reference systems including geographic (latitude/longitude), grid (Universal Transverse Mercator), and planar (distance/bearing representation) coordinates.

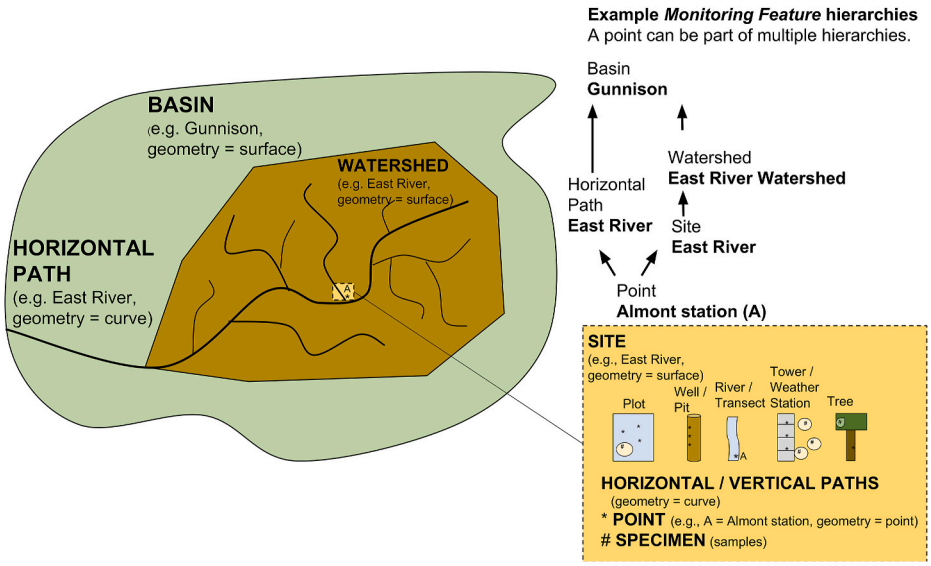

Fig. 2. Example *Monitoring Feature* groupings and hierarchies for a watershed study. The site attributes that are indicated in the features are described in Section 3.1.

Monitoring Features can be infinitely nested using parent-child spatial relationships. For example, a plot containing multi-level wells will have three types of *Monitoring Features* defined as a surface (plot) > curve (well) > point (sensors at different depths in the well; e.g. Fig. 2).

BASIN-3D also supports observations collected at different temporal scales and resolutions. Synthesis over different temporal resolutions is enabled via three parameters defined in the OGC observation-based framework: (1) phenomenon time (the time of the observation); (2) aggregation duration (a qualitative description of the duration over which the observation was acquired, e.g., annual, day, hourly); and (3) time reference position (the position of the time value within the observation time, e.g., start, middle, end, instant). These parameters can accommodate different temporal resolutions provided by the data source. For time-series observations, an aggregation duration parameter is typically specified by the data source, and the retrieved data is included in the *Observation Result* (Section 2.2) as a Time Value Pair (TVP).

2.2. Abstracted representations for diverse observation types

BASIN-3D also supports diverse observation types with the OGC concepts. It uses a generically-defined '*Observation*' entity with three components: the '*Observed Property*' describes the measurement (e.g. river discharge, stream chemical concentrations), the '*Feature of interest*' defines the subject of observation (e.g. river), and the '*Observation Results*' defines the results of the observation using abstracted data structures (e.g., a time-series of coupled timestamps and values). The components of the *Observation* are linked as follows: *Observation Results* of an *Observed Property* are reported for a *Feature of interest*, typically specified as a representative *Monitoring Feature* (Section 2.1).

BASIN-3D maintains a controlled list of *Observed Properties* and a preferred set of units to which the data provider's variables and units are mapped (Appendix 2). This enables BASIN-3D to harmonize variable names across data used for the same measurement (e.g., Al, Aluminium) into a single *Observed Property* for a diverse set of observations (e.g., physical, chemical and biological measurements) with standardized units.

BASIN-3D also uses abstracted *Observation Result* types, (e.g. time-series, image, grid), instead of discipline-specific representations (e.g., discharge, water quality parameter). For our applications so far, we have implemented an *Observation Result* for time-series data. Thus time-series data from different science disciplines or research applications (e.g. meteorology, geochemistry) are transformed into *Observation Results* in the uniform format '*Measurement Timeseries Time Value Pair (TVP)*', an array of paired timestamp and data values. All BASIN-3D timestamps follow the ISO 8601 standard (https://www.iso.org/iso-8601-date-and-time-format.html). The *Observation Result* can be extended to represent additional data types beyond time-series measurements, such as categorical data, imagery, and remote sensing (Section 4.1). Other metadata such as the measurement units are provided within the *Observation Result* object.

2.3. Framework architecture and implementations

BASIN-3D presents two approaches for data integration – first as a Python library, and secondly as a Django web-based application (https://www.djangoproject.com/) leveraging its vast extensions to provide mechanisms for site administration, authentication and authorization, and documentation. It has a modular architecture consisting of (1) a 'Data Acquisition Layer' that connects to data sources and retrieves data dynamically, (2) a 'Data Synthesis Layer' where data are transformed to the BASIN-3D schema (hereafter referred to as the Synthesis models), and (3) either a 'Data Output Layer' that transforms synthesized data into specified data structures in the Python version (Fig. 3a) or a 'Web Service Layer' that receives requests from and returns results to different clients in the Django version (Fig. 3b). It also contains an extensible internal catalog to maintain a list of data sources with their authentication information, and a controlled vocabulary of *Observed Properties*.

The Data Acquisition Layer provides functionality to customize data source connections as required for the application using a plugin architecture containing extensible Python classes. The base Plugin classes enable connection to any network-accessible source such as a database, web service or a remote or local filesystem. They also include an extensible HTTP connection module with support for some common authentication methods such as the Hypertext Transfer Protocol (HTTP) authentication API that supports OAuth2 (https://oauth.net/2/) and token-based authentication. Custom data access plugins consist of 2 components: 1) a python module, and 2) a csv file with a mapping of data source variables to BASIN-3D variables. In the plugin python module, the developer extends the base plugin classes and implements the authentication required by the source (if any), constructs queries to retrieve data and metadata required by BASIN-3D (Table 2), and maps the structure, format and semantics of the returned data to the Synthesis models. In particular, information about measurement locations (via a mapping to *Monitoring Feature* objects) and time-series data (via a

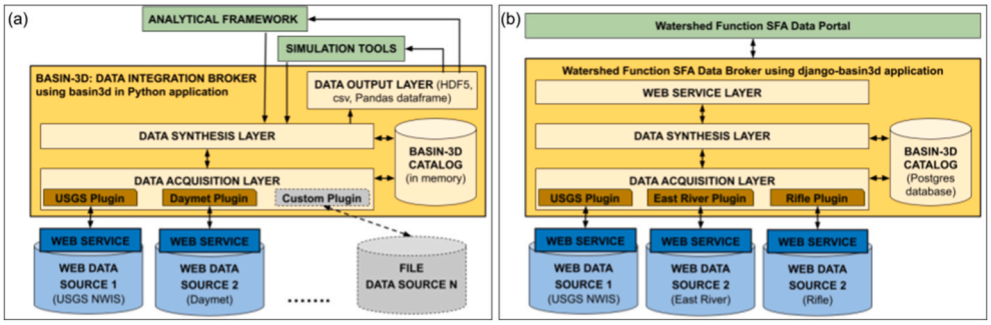

Fig. 3. BASIN-3D architecture showing (a) the Python implementation where users connect to plugins as needed and results of a query to the Synthesis Layer are returned as HDF5 files or a Python Pandas dataframe via the Data Output Layer; and (b) the Django-based web implementation (django-basin3d) where the Watershed Function SFA (WFSFA) Broker is a custom implementation containing plugins to access two Watershed SFA field databases and USGS NWIS. The Web Service Layer enables the WFSFA Broker to provide a single web service that clients can use to specify queries and access results via the Synthesis Layer. Here BASIN-3D refers to the entire ecosystem, basin3d refers to the core Python software and django-basin3d refers to a Django wrapper around basin3d that enables the web framework (see https://github.com/BASIN-3D). Note that the Python library (basin3d) is an install requirement in django-basin3d.

Table 2
Mapping of Data Source Objects across plugins to the BASIN-3D Synthesis Model for the two BASIN-3D implementation approaches. Web API calls shown here require configuring a custom broker implementation (e.g. WFSFA). Python API calls require installing the BASIN-3D package, configuring the appropriate data source plugin (s), and instantiating a "synthesizer" object. Note that the API calls for the Measurement TVP Timeseries object are examples of USGS NWIS Daily Values data synthesis. Documentation for the Python API calls are available at https://basin3d.readthedocs.io/(in the Key Functions section). Documentation for the Web API calls are available at https://django-basin3d.readthedocs.io/(in the REST API section).

USGS NWIS Data Source	East River Data Source	Rifle Data Source	BASIN-3D Synthesis Model	BASIN-3D Synthesis Web API Calls	BASIN-3D Synthesis Python API Calls
NWIS Sites	Locations	Locations	MonitoringFeature Type: Point	/monitoring_features/points	synthesizer.monitoring_features (feature_type 'POINTS')
HUC: Watershed	Directly coded	Directly coded	MonitoringFeature Type: Watershed	/monitoring_features/watersheds	synthesizer.monitoring_features (feature_type='WATERSHED')
Hydrological Unit (HUC) ● Region ● Subregion ● Basin ● SubBasin	No Mapping	No Mapping	MonitoringFeature Type: ● Region ● Subregion ● Basin ● Subbasin	/monitoring_features/regions/monitoring_features/subregions/monitoring_features/basins/monitoring_features/subbasins	synthesizer.monitoring_features (feature_type='REGION') synthesizer.monitoring_features (feature_type='SUBREGION') synthesizer.monitoring_features (feature_type='BASIN') synthesizer.monitoring_features (feature_type='SUBBASIN')
NWIS Daily Values	Location Data	Location Data	Measurement_TVP_Timeseries	/measurment_tvp_timeseries/?monitoring_features=USGS-0911000&observed_property_variables=RDC,WT&start_date=2019-10-25&end_date = 2019-10-30&aggregation_duration=DAY	synthesis.get_timeseries_data (synthesizer, monitoring_features=['USGS-09110000'], observed_property_variables=['RDC','WT'], start_date='2019-10-25', end_date='2019-10-30')

mapping to *Measurement TVP Timeseries* objects) is configured in the python plugin module. Mapping to the BASIN-3D synthesis models is open-ended and accommodates a range of scenarios depending on the availability of data or metadata from the data source. Plugin developers can choose to return mapped data from the data source, return data from other supplementary local or remote sources, or return nothing if no relevant information is provided by the data source. The WFSFA implementation (Section 3.2) describes additional examples of plugin configuration. Plugins can be shared between the Python library and Django implementations. Currently both versions are bundled with a plugin to the public USGS National Water Information System (NWIS; https://nwis.waterdata.usgs.gov) that can be used out-of-the-box to access the NWIS data and also used as a template to create new plugins to connect to new sources. After the plugins are created, it is trivial to query the BASIN-3D APIs for integrated search and access of data across all configured sources. Any custom data access plugin that extends the BASIN-3D Plugin classes can be registered for use with the Data Synthesis Layer using a simple function call.

The Data Synthesis Layer transforms data received from the sources into two BASIN-3D synthesis objects that represent time-series data, namely the *Monitoring Feature* (e.g., Watershed, Site, Point) and *Measurement Timeseries TVP* (Sections 2.1 and 2.2). The variables used by the data sources are harmonized to a controlled vocabulary of *Observed Properties*. The synthesis layer supports common query parameters to filter synthesized data by time, location, observed properties and data quality (Table 2).

The Data Output Layer translates the BASIN-3D synthesis objects into commonly used analysis formats. BASIN-3D currently supports Python Pandas dataframe objects and csv files for *Measurement Timeseries TVP* data as well as Python dictionary objects and csv files for the *Monitoring Feature* and *Observed Property* metadata. It also supports HDF5 file formats that contain both the data and metadata. The output formats are specified by the data user and can be extended to additional types.

The Web Services Layer provides a scalable, fault tolerant and extensible Representational State Transfer (REST) API powered by the Django Application Framework (https://django-basin3d.readthedocs.io/). Clients can use the Synthesis Web API to view the catalog of locations and observed properties, make data requests and retrieve results in a Javascript Object Notation (JSON) format.

3. Results: Applications of BASIN-3D to integrate water data

3.1. Python application to synthesize regional-scale water data

A first application of BASIN-3D is for a Python-based, data-driven framework iNAIADS (iNtegration, Artificial Intelligence, Analytical Data Services) developed for a DOE project that aims to predict how hydrologic disturbances will change water quality in large river corridors of the United States using ML models that account for differences in watershed characteristics, including geomorphology, geology, climate and land use. Thus, the study needs to utilize hydrological, geochemical, climate time-series and other spatial datasets at point to regional scales from a vast array of sources such as NWIS, Daymet (https://daymet.ornl.gov/) and the National Hydrography dataset (https://www.epa.gov/waterdata/nhdplus-national-hydrography-dataset-plus). The project requires flexible programmatic access to data sources for exploratory analysis, subsetting and retrieval of data on-demand, and repeatable integration of large datasets into an easily useable format. Here, BASIN-3D is used to integrate data as standardized inputs for the analytical and ML modules, and insulate them from any changes in how data is served by the sources.

The BASIN-3D Python library (basin3d) is a standalone software that can be deployed with the simple "pip install" command. It is easy to integrate with typical data wrangling, analysis and plotting capabilities within Jupyter notebooks (http://jupyter.org) or other Python scripts (Fig. 4). To request data, users specify the set of monitoring features, observed properties and time period of interest as parameters within a Python function call (see basin3d documentation: https://basin3d.readthedocs.io/). The output is returned as a Python *Pandas dataframe* with harmonized units and aligned timestamps where each time-series is a column named by the combination of Monitoring Feature ID and the BASIN-3D Observed Property variable name. The Pandas missing value is used if a particular time series does not have a measurement for a given timestamp. The integrated data can be saved in HDF5 file formats for offline retrieval. The outputs also separately return associated metadata as a Python *dict* object with elements for each time series (i.e., column), which includes metadata such as the *Observed Property*, and results metadata such as units, statistics, result quality and temporal aggregation. These metadata attributes can be modified as needed, and future extensions will include the original variable name and monitoring feature geolocation details, if provided by the data source, to enable transparency in the plugin mappings.

As an example application, basin3d was used to integrate 70 years (1950–2020) of daily stream discharge and temperature data from NWIS and meteorological data from Daymet (version 4) for stations in all hydrologic regions within the Continental United States (CONUS) using its Synthesis Python API calls with just a few lines of code (https://basin3d.readthedocs.io/en/latest/quick_guide.html). The integrated datasets were used to create input and validation datasets for ML models to predict stream temperature. By using basin3d, it would be trivial to extend the integrated datasets to add new years of data. Any changes to the underlying data sources such as a new version of the Daymet dataset being made available or changes in variable semantics or data access mechanisms can be handled within the BASIN-3D plugin, without requiring changes in the analytical or ML modules.

3.2. Web-based application for an integrated data portal at the watershed-scale

The second application of BASIN-3D is for the DOE's Watershed Function Science Focus Area (SFA; http://watershed.lbl.gov), which seeks to gain a predictive understanding of how perturbations like early snowmelt influence hydrobiogeochemical processes in mountainous watersheds and impact downstream water availability and quality at subseasonal to decadal timescales (Hubbard et al., 2018). The project has two field sites in the Upper Colorado River Basin of the US: the headwaters East River catchment near Crested Butte, CO (ongoing since 2014), and a floodplain site on the Colorado River near Rifle, CO (active from 2009 to 2017). The project, in partnership with several collaborating organizations, generates a vast amount of diverse data, including hydrogeological, geochemical, climate, metagenomic and remote sensing observations (Kakalia et al., 2021). The SFA maintains two

CAMELS_WT.ipynb | EDA_plots.ipynb

Code | Python 3

Retrieving data for a specified station using BASIN-3D

```
[ ]: from basin3d.plugins import usgs
     from basin3d import synthesis
```

```
[228]: synthesizer = synthesis.register()
       usgs_df_CAMELS, usgs_metadata_CAMELS= synthesis.get_timeseries_data(synthesizer, monitoring_features=['USGS-02077200','USGS-13340600','
                                                observed_property_variables=['WT'],
                                                start_date='1980-01-01', end_date='2014-12-31')
```

```
[229]: usgs_df_CAMELS
```

[229]:

	TIMESTAMP	USGS-01466500_WT	USGS-02011400_WT	USGS-02077200_WT	USGS-13340600_WT	USGS-14138870_WT
1980-01-01	1980-01-01	5.9	3.6	NaN	NaN	4.3
1980-01-02	1980-01-02	5.8	3.4	NaN	NaN	4.2
1980-01-03	1980-01-03	5.8	3.9	NaN	NaN	4.0
1980-01-04	1980-01-04	5.5	3.3	NaN	NaN	4.4
1980-01-05	1980-01-05	5.4	2.3	NaN	NaN	4.0
...	...	...	...	...	...	...
2014-12-27	2014-12-27	6.2	4.7	6.4	2.5	5.6
2014-12-28	2014-12-28	6.4	6.1	8.3	2.0	5.1
2014-12-29	2014-12-29	6.6	6.9	9.2	1.5	4.6
2014-12-30	2014-12-30	6.1	6.2	7.7	1.5	3.2
2014-12-31	2014-12-31	5.3	3.3	5.4	1.5	3.1

12784 rows × 6 columns

Fig. 4. BASIN-3D can be easily used in a Python-based Jupyter notebook to retrieve synthesized data with a few lines of code.

private databases for each field site, the East River (ERDB) and Rifle (RiDB) databases that hold the project's sensor- and sample-based observations and provide APIs to access the data (Varadharajan et al., 2019). Data from the remote sensing campaigns are maintained separately on SFA or other data systems such as the National Snow and Ice Data Center (NSIDC). Metagenomic data are held in a specialized database, ggKBase (http://ggkbase.berkeley.edu). The SFA also utilizes data from pre-existing infrastructure, such as USGS gaging stations, the Environmental Protection Agency's (EPA) Clean Air Status and Trends Network (CASTNET) and National Resources Conservation Service (NRCS) Snow Telemetry (SNOTEL) sites.

The project required integration of these diverse data held in different sources to minimize redundant and inconsistent efforts by scientists to retrieve and synthesize data. A critical need was for a software to integrate data from the two SFA private databases that required authentication with data from public sources such as the USGS NWIS and EPA (Fig. 5). Hence, the web version of BASIN-3D was used to support integration across the SFA's East River and Rifle field sites and USGS sites across the East-Taylor Watershed, and to support serving the data through a user-friendly interactive web portal.

In addition to the core BASIN-3D framework, the implementation needed two additional software components to enable a web-based view of the integrated data: (1) the Watershed SFA Broker Service (subsequently referred to as WFSFA); and (2) the SFA data portal to access the integrated datasets. The WFSFA is a custom Django-based broker implementation of BASIN-3D with plugins to connect to three remote data sources - the ERDB, RiDB and NWIS (Fig. 3b). In each plugin, a mapping between BASIN-3D synthesis objects and the data source objects are defined (see Section 2.3). For example, *Monitoring Feature* with type *point* in BASIN-3D is mapped to Location objects in the ERDB and RiDB, and to the site object in NWIS (Table 2). The *Monitoring Feature* with type *watershed* is mapped to the HUC watershed object in NWIS, and is explicitly defined for the ERDB and RiDB since they do not provide the watershed information as part of the location metadata. The plugins also transform source variable names to BASIN-3D controlled vocabularies (e.g. Al in the East River database to Aluminium (Al)). Data synthesized by the WFSFA are accessible through Synthesis Web API calls (Table 2).

Authenticated SFA users can access and download integrated datasets using a web portal with search and interactive visualization capabilities (Fig. 6). To request data, users select the locations, observed properties (parameters or measurement variables), time period and data sources (sites) of interest on the portal's selection widgets. The portal features filtering capabilities to help users identify observed properties measured at a set of locations and vice versa, which dramatically reduces the number of futile searches for which no data is available. Data search requests trigger Synthesis API requests via the WFSFA to dynamically retrieve data from the relevant sources, and return a *Measurement Timeseries Time Value Pair (TVP) Observation Result* (Fig. 7). The portal displays results as a downloadable table of values, as well as interactive Javascript visualizations (https://dygraphs.com/). The use of BASIN-3D enables the portal code to be maintained separately from the plugins accessing the data. Updates to ERDB and RiDB are handled by modifying the plugins in the WFSFA, which minimizes maintenance required for the portal code.

4. Discussion

4.1. The BASIN-3D brokering approach

BASIN-3D is a unique tool that serves different purposes than any prior broker-based environmental data integration system. In the era of data-intensive science and ML, there is a need for tools that enable users to integrate data across providers of their choosing, in a manner that makes most sense for their applications. In some cases users may want to integrate data differently, even when connecting to the same providers. It is impossible for any centralized system to anticipate the myriad data sources and ways in which users want to pull and synthesize data.

Hence, we developed BASIN-3D to provide a flexible means for users to customize their data integration and to give them greater control on how data across providers are mapped into a common format. In particular, it was designed to address the following needs from our use cases, for which there was no pre-existing solution.

1. **Data integration for modeling and analysis applications** - There was a need to pull together large datasets for exploratory analysis and ML (Section 3.1), which requires the ability to query and retrieve data using a programmatic API without any prior coordination with

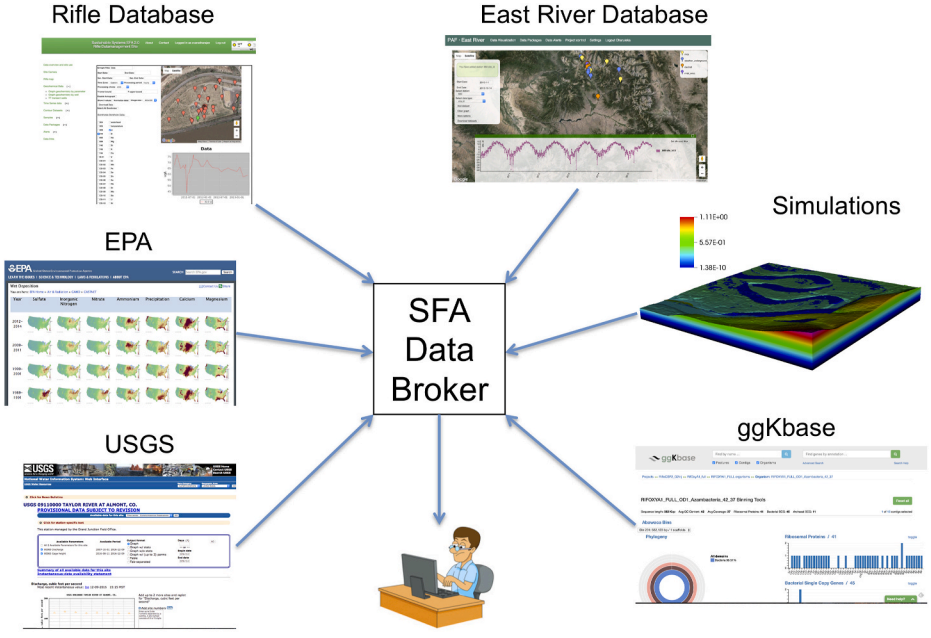

Fig. 5. Conceptual diagram of the functionality desired from the BASIN-3D supported data brokering system for the Watershed Function SFA. The broker would connect to various data sources across organizations and present an integrated view of the data to the user.

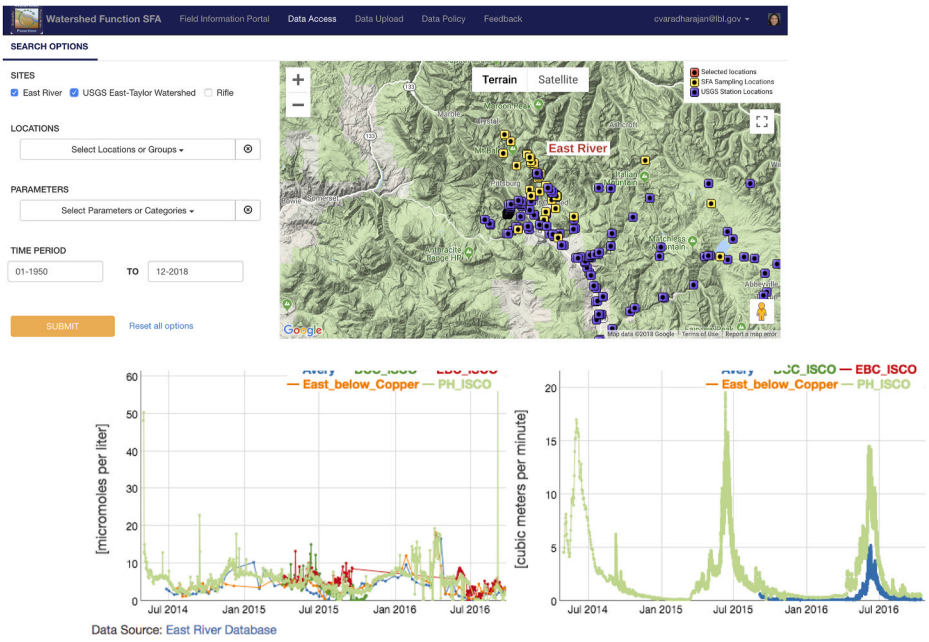

Fig. 6. WFSFA Data Portal provides an interactive data-driven user interface for exploring watershed data across USGS sites in the East Taylor Watershed and the Watershed Function SFA's Rifle and East River field sites.

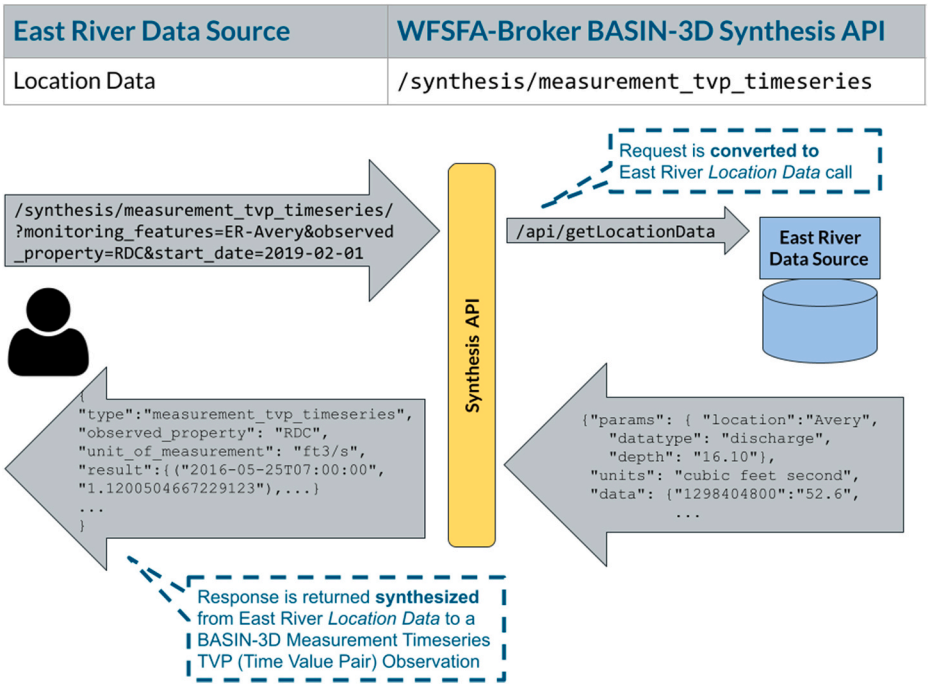

Fig. 7. Example workflow describing the functionality behind a Synthesis API request. A user requesting river discharge data (RDC) for a given location (Avery station at the SFA's East River field site) makes a request to the WFSFA, which retrieves the required data from the East River data source on demand and transforms the returned object into the generalized BASIN-3D synthesis models.

the data sources. It also requires support for JSON, as a concise and easily parsable format used in modern software applications. With BASIN-3D, we integrated data from NWIS (discharge, water temperature), and Daymet (meteorological parameters) for stations in all regions of the CONUS into a single data frame with harmonized timestamps and units using the default APIs provided by the data sources. This data frame was also stored in the HDF5 file format for offline access by the analytical algorithms. BASIN-3D allows users to create their own plugins to flexibly connect to additional data sources of interest.

2. **Connection to private databases** - A critical need for one of our use cases was to integrate data from a project's multiple private databases with data from public sources such as the USGS or EPA (Section 3.2). We note that even within a single project, there may be a need to synthesize data from multiple sources, some of which may not be available from a public provider and do not provide data conforming to existing standards. BASIN-3D also supports both token and OAuth

authentication making it convenient to connect to private databases that use these two common authentication mechanisms.

3. **Supporting diverse data types** - Although the current version of BASIN-3D only handles time-series data, it is extensible to represent more general observations including remotely-sensed and simulation data because of its use of the OGC concepts. For example, to incorporate remote sensing data, an *Earth Observation Result* type, or more specifically an *Optical Earth Observation Result,* can be defined to handle a standard gridded data format, such as that described as *eop: Earth Observation in the* OGC Earth Observation Metadata Profile for Observation and Measurements (OGC 10-157r4; see Appendix 1b). A rich set of additional parameters such as image resolution and processing information can be selectively included to fully describe the *Observation.* Similarly, for simulated data a model-specific *Result* type can be defined if an existing *Result* type (e.g., gridded or time-series) is not suitable. An *Observation* parameter can be defined to indicate that the data are simulated. Currently, we have implemented a lean data model that is based on our use cases and can be extended as needed to support additional data types.

   BASIN-3D also provides a mechanism for maintaining and amending a controlled list of *Observed Properties* for a variety of measurements. It comes with a default vocabulary (Appendix 2), created to support the diversity of measurements in our applications. However this list can be replaced with any preferred terminology or ontology by updating the *Observed Property* names in the BASIN-3D catalog, updating the mappings of the new terms to the data source terms in a plugin mapping *csv* file, and registering the updated plugin with the BASIN-3D Synthesis models. Following this, the new controlled vocabulary will work without requiring any changes to the plugin code. The controlled vocabulary focuses on the *Observed Property* alone and does not distinguish between units, aggregation type or sampling method, which are considered characteristics of the *Observation*. Separation of the *Observed Property* from the abstracted data representation *Observation Results* provides the flexibility to handle diverse observation data types in a similar manner. For example, it is easy to synthesize continuous sensor-based hydrological time-series data with manual geochemical measurements using the *Observation Results* of type *Measurement Timeseries TVP*. Thus, BASIN-3D clients (e.g., web portals, visualization, analysis software) only need to use the *Measurement Timeseries TVP* to work with heterogeneous, time-series data across sources.
4. **Supporting different spatial representations** - BASIN-3D can be expanded to support a variety of features and groupings by abstracting *Spatial Sampling Features* to the OGC's four geometries (point, curve, surface, solid). By specifying *Monitoring Features* with parent-child relationships, observations can be specified and synthesized at any defined spatial scale, as well as aggregated across scales. This provides flexibility with querying and integrating data across different scientific representations of sampling designs and aggregation (e.g. by a particular field site or across a river basin).

We also note that the challenges with data integration are not unique to geosciences and are relevant to other fields such as intelligent transportation systems, autonomous vehicles, smart cities, and industrial applications utilizing the Internet of Things (IoT) where similar or alternate solutions to BASIN-3D may exist. One example is a European open source IoT framework FIWARE (Cirillo et al., 2019), where adhoc integration modules and FIWARE IoT agents translate data provided by cities to the Next Generation Service Interface standard data format (OMA, 2012). These "adapters" are similar to our concept of a plugin, which generate data in a uniform format and also generate metadata following the FIWARE and Open and Agile Smart Cities (OASC) data models for downstream applications. More broadly there have been decades of research and applications based on semantic data integration to link data from heterogeneous sources onto a reference ontological model (e.g. Charpenay et al., 2018 and references therein). The semantic tools provide more sophisticated approaches to data harmonization in comparison to BASIN-3D, but do not handle aspects of spatial data integration and authentication necessary for our use cases. Further work is needed to explore and adopt concepts from the considerable body of literature on data integration beyond the geosciences.

4.2. Advantages and limitations of using BASIN-3D to address data integration challenges

We encountered several data integration challenges across our two use case applications. The first is to reconcile differences in concepts, formats, data models and semantics that can vary widely across data sources (Varadharajan et al., 2019). For example location terms such as "basin", "site" and "point" are used differently by both data providers and consumers. Sampling hierarchies can be organized using any combination of geometries (e.g. a transect containing wells, both *curves* as per the OGC), which makes it difficult to define a rigorous hierarchical description of sampling feature types. Similarly, data sources use different terms for variables and only a few use well-described ontologies. For example, NWIS has over 20,000 parameter codes distinguishing between sampling methods (e.g. filter size), units and aggregation (e.g. daily versus instantaneous). Additional transformations such as unit conversions, quality checks or other processing may be needed prior to synthesis. For example, NWIS has different parameter codes for discharge in ft^3/sec (00060) and m^3/sec (30208); and queries using the two codes return different results, requiring unit transformations to synthesize discharge data across one source. Some data may need to pass QA/QC checks before they are ready to be used, which is challenging since a myriad of data quality flags and methods are used across sources.

Data sources also differ in how they organize and provide access to data. In some cases, sources provide a collection of web services instead of a single end point. For example, NWIS has multiple web services for retrieving daily and instantaneous values for surface water, groundwater and water quality data. Sometimes sources do not provide key metadata and queries for discovery and synthesis. For example, we found critical information was not provided or easily accessed, such as the depths of sensors and reference datums in queries retrieving borehole information, the position of the timestamp in queries retrieving time-series data (i.e., whether the time stamp reflected the start, middle or end of the observation), descriptions of sampling protocols and data quality measures, and a list of all the observed properties measured at a given location. Finally, authentication requirements for data systems pose an additional barrier to users.

We addressed many of these challenges using a plugin model that provides flexibility to the critical tasks of mapping and transforming data source objects and terminologies to the generalized BASIN-3D synthesis model (Section 2.3). Additional data transformations for unit conversions and quality checks can also be specified in the plugin. By default, both the units and quality flags provided by the source are passed on to the user in the BASIN-3D *Observation Results* (except for NWIS river discharge, which is harmonized to m^3/sec). BASIN-3D also provides for three simple QA/QC flags - "checked", "unchecked", and "null". The plugin design is flexible and any number of endpoints can be specified within a single plugin to accommodate for differences in how sources organize and provide data. BASIN-3D plugins incorporate two widely-used authentication protocols to lower the bar to access systems that require authentication.

The BASIN-3D brokering approach to connecting with data sources, although versatile, also has its limitations. Currently, an understanding of both the data source and the BASIN-3D synthesis models, along with Python familiarity, are needed to build a custom data access plugin. Because of the diversity of data sources, a separate plugin for each data source is typically required. However as efforts to standardize data sources advance, we anticipate that one plugin may be able to support multiple data sources. Ideally the plugins should be designed so they are

easy to create or update as sources make structural changes such as adding new data categories or modifying terminologies. However, creation and maintenance of plugins is currently time-consuming especially if data source structures and semantics are difficult to map onto BASIN-3D's schema. Typically it involves the judgement of domain experts, who determine how to map the various terms and formats to the Synthesis models. Plugin development also requires expertise in software engineering, which may not be available to all scientific groups. The Python library is bundled with a plugin to the public USGS National Water Information System (NWIS; https://nwis.waterdata.usgs.gov) that serves as an example for developers to build custom plugins to new data sources. However, a more general template and testing tools that facilitate plugin creation and maintenance would lower barriers to this approach.

The current version of BASIN-3D specifies default preferred units (Appendix 2), but does not include default unit conversion libraries that would accommodate a variety of units provided by sources. Creating a generalized unit conversion tool is difficult, given the wide variety of physical, chemical and biological variables. Although existing unit libraries such as the Udunits-2 (https://www.unidata.ucar.edu/software/udunits/) or Pint (https://pint.readthedocs.io/en/stable/) can be utilized, a generalized broker will need custom unit conversions for non-physical units (e.g. for geochemical data).

Major feature enhancements to BASIN-3D are needed to scale to additional sources and handle more diverse and big data. First, the synthesis models must be extended to handle additional data types such as geospatial or remote sensing products and model data (Section 3.2). Then, BASIN-3D needs the ability to retrieve large datasets using protocols such as OpenDAP (https://www.opendap.org/) and OGC Web Coverage Service (https://www.unidata.ucar.edu/software/tds/current/reference/WCS.html). For large time series data, paging returned results in the JSON is also an option. Retrieving big data can be a challenge to implement if sources do not support large data transfer protocols or provide basic metadata such as the size of the results or number of data points. BASIN-3D will also need the ability to cache data retrieved from prior queries as it retrieves data on-demand for each query and currently does not have a mechanism to store returned results. Caching data can enable efficient data retrieval, improve performance and the user experience particularly when retrieving large datasets, and will enable access to sources during outages (Blodgett et al., 2015). Caching would also enable users to track versions of data over time and potential impacts of change in the data to analyses or simulations (Ghoshal et al., 2018).

Finally, a generalized data integration framework should provide detailed provenance information along with data usage rules and citations. This can be difficult to implement in a broker, particularly when there are various versions of the data and if sources do not track or provide the information. BASIN-3D currently enables identification of the source name and URL along with returned results but does not have support for additional citation information within its synthesis models. Creating dynamic data citations for on-demand queries is an important area for improvement to credit the sources and provide users of the synthesized data an ability to precisely cite versions of the data retrieved.

4.3. Using standards to enable a generalizable, extensible brokering framework

BASIN-3D has been designed to harmonize, integrate and query diverse datasets that result from a range of field investigations, monitoring networks and model simulations. In particular, use of the OGC and FGDC standards provides a means to support flexible synthesis of diverse measurement configurations and data types using abstracted data structures (Section 4.1). These standards are a suitable choice as they have been developed over several years to enable interoperability across data systems and have achieved consensus across and adoption by various organizations.

We encountered some challenges implementing these data standards as the underlying construct for integration. First, it was not easy to use the standards partly because they are specified at a high level and do not provide implementation guidance beyond some simple, limited examples. For example, the OGC standards do not specify implementation of *Monitoring Features* or its parent features for different shapes or resolve how collections of spatial hierarchies should be organized. Standards also use specialized terminologies that domain scientists may not be familiar with. For BASIN-3D, we had to balance constraints of following the OGC standard using the specified terminology (e.g., using observed_property and measurement_tvp_timeseries), while making the concepts and Synthesis API calls logical to domain scientists. Thus in a few cases we deviated from OGC definitions or terminologies to improve the usability of BASIN-3D for scientific researchers or for other practical reasons. For example, while the OGC standards differentiate between the feature being observed and the representative sampling feature upon which the actual observation is made, BASIN-3D does not make this distinction because most data sources only include information on the Sampling Feature and a 'Feature' in one case may be a 'Sampling Feature' in another. Thus all spatial entities are *Monitoring Features* in BASIN-3D; however, the data model is implemented as hierarchical classes which enables expansion to support any OGC Feature entity. Similarly, all Feature entities use *Feature Types* instead of specific, entity-based types (e.g., Spatial Sampling Feature Type) for practical implementation.

It is also difficult to identify and track multiple, evolving standards potentially applicable to a situation. There are many standards that are highly suitable to a particular data type and sampling design, but few that are generalizable across the broad suite of diverse measurement types and sampling hierarchies that may be relevant to an interdisciplinary watershed study. As an example, we considered the widely-used GeoJSON specification (http://geojson.org) as the coordinate representation for *Monitoring Features*. GeoJSON has become a de facto "standard" for geospatial data proposed by the Internet Engineering Task Force but is yet to be approved (https://tools.ietf.org/html/rfc7946). However, GeoJSON has several limitations including the inability to extend geometries beyond simple types, such as representation of circles or meshes (OGC, 2017). GeoJSON only represents geographic coordinates, using the World Geodetic System 1984 (WGS 84) datum, with longitude-latitude pairs of a relevant shape (e.g. Polygon), which does not accommodate for location information using other reference datums (e.g. a plot's center point with lengths of the four sides). Thus in many implementations, GeoJSON has to be extended in an arbitrary manner by adding attributes into a non-standardized 'properties' field. As another example, the WaterML 2.0 standard is specifically tailored for water time-series observations (https://www.opengeospatial.org/standards/waterml). However, WaterML cannot be used to represent LiDAR, hyperspectral or snow pit observations from the East River. Additionally, JSON representations are more common in modern applications, as they are less verbose than XML-based standards, more easily parsed and better suited for big data, and also easier for domain scientists to use. Hence, we chose the higher-level OGC standard as the default model for data integration since its use enables BASIN-3D to be extensible to diverse data types without using multiple standards, and also allows for JSON representations.

5. Summary and conclusions

There is a critical need to integrate heterogeneous, multiscale data to address complex environmental challenges. Often, data is distributed across many sources that do not share common structures and formats. Generalized data management frameworks that integrate diverse, distributed data can facilitate their use in analysis and modeling. We developed BASIN-3D, a data-brokering integration framework that retrieves subsets of data from different sources for on-demand queries and

integrates the data into a unified view. BASIN-3D applies concepts from OGC and FGDC standards to create an extensible framework that allows creation of custom plugins to map data sources to a common synthesis model. It can be used as a Django application to support web services and portals, as well as a simple Python library for analysis and modeling. We present implementations of BASIN-3D to integrate diverse time-series data for two DOE projects that study the impacts of hydrological perturbations at watershed to regional scales.

We encountered several challenges in building this framework, despite having just two applications. It was difficult to harmonize spatial elements, variable names, data quality terms and units across sources, and it required a combination of domain expertise and software engineering to create software mappings to a common data-model schema. Data sources typically do not provide data with a user's view, and sometimes information needed to synthesize data are unavailable or not easily accessible. Although existing standards provide much value in representing diverse data using a generalizable abstracted approach, they are difficult to use and provide minimal guidance on the details of their implementation. Some of these challenges can be addressed using BASIN-3D's generalized brokering framework. The data integration needs of our applications are broadly applicable to a large number of environmental studies of complex systems. For example, the Critical Zone Observatories have similar interdisciplinary data that require integration across platforms (Zaslavsky et al., 2011). BASIN-3D also applies to policy initiatives that seek to integrate data across sources, such as the federal Open water data initiative (https://acwi.gov/spatial/owdi/) and California's Open and Transparent Water Data Act (https://water.ca.gov/ab1755). The BASIN-3D synthesis constructs are generalizable and can be extended to integrate data types beyond time-series for a wide range of environmental field and modeling investigations.

Authorship statement

Charuleka Varadharajan was responsible for the conceptual development of BASIN-3D and implementation of its first version. Valerie Hendrix and Danielle Christianson have made equally significant contributions as the lead author and are the primary software developers on BASIN-3D. Madison Burrus has contributed to testing the software deployment for the WFSFA broker and portal. Catherine Wong is responsible for documentation and testing of the open source Github repository. Susan Hubbard is a senior author who is the lead principal investigator on the Watershed Function SFA project and was responsible for developing the vision and ideas for the project's data integration. Deborah Agarwal is a senior author who was responsible for supervising the development of BASIN-3D and implementation for the Watershed Function SFA project.

Computer code availability

BASIN-3D (basin3d at version 0.2.0 and django-basin3d at version 0.0.4 at time of publication) is available open source on Github https://github.com/BASIN-3D with Berkeley Lab's modified BSD License (https://github.com/BASIN-3D/basin3d/blob/main/LICENSE). This version contains most of the software concepts and implementation described here. Some newer components, such as the support for HDF5 file formats and the Daymet integration (Section 3.1) are not in the current version, but will be available in future releases. Documentation for basin3d is available at https://basin3d.readthedocs.io/. Documentation for the web version is available at https://django-basin3d.readthedocs.io/.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is supported as part of the Watershed Function Scientific Focus Area, the iNAIADS DOE Early Career Project, and the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award no. DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. We acknowledge the support of the Watershed SFA team who provided feedback for the scientist-centered design exercises. We also acknowledge the anonymous reviewers whose comments helped improve the manuscript significantly.

Appendix. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cageo.2021.105024.

References

BERAC, 2013. BER Virtual Laboratory: Innovative Framework for Biological and Environmental Grand Challenges; A Report from the Biological and Environmental Research Advisory Committee. DOE/SC-0156.

Blodgett, D., Lucido, J., Kreft, J., 2015. Progress on water data integration and distribution: a summary of select US Geological Survey data systems. J. Hydroinf. 18, 226–237. https://doi.org/10.2166/hydro.2015.067.

Christianson, D.S., Varadharajan, C., Christoffersen, B., Detto, M., Faybishenko, B., Gimenez, B.O., Hendrix, V., Jardine, K.J., Negron-Juarez, R., Pastorello, G.Z., Powell, T.L., Sandesh, M., Warren, J.M., Wolfe, B.T., Chambers, J.Q., Kueppers, L.M., McDowell, N.G., Agarwal, D.A., 2017. A metadata reporting framework (FRAMES) for synthesis of ecohydrological observations. Ecol. Inf. 42 https://doi.org/10.1016/j.ecoinf.2017.06.002.

Charpenay, V., Käbisch, S., Kosch, H., 2018. Semantic data integration on the web of things. In: Proceedings of the 8th International Conference on the Internet of Things (IOT '18). Association for Computing Machinery, New York, NY, USA, pp. 1–8. https://doi.org/10.1145/3277593.3277609. Article 3.

Cirillo, F., Solmaz, G., Berz, E.L., Bauer, M., Cheng, B., Kovacs, E., 2019. A standard-based open source IoT platform: FIWARE. In: IEEE Internet of Things Magazine, vol. 2, pp. 12–18. https://doi.org/10.1109/IOTM.0001.1800022, 3.

Cox, S., 2011. Observations and Measurements v2.0 OGC Document 10-004r1 (Also Published as ISO 19156:2011 - Geographic Information – Observations and Measurements).

FGDC, 1998. Federal Geographic Data Committee. FGDC-STD-001-1998. Content Standard for Digital Geospatial Metadata (Revised June 1998) (Washington, D.C).

Genesereth, M., 2010. Data integration: the relational logic approach. In: Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, pp. 1–97. https://doi.org/10.2200/S00226ED1V01Y200911AIM008.

Ghoshal, D., Ramakrishnan, L., Agarwal, D., 2018. Dac-man: data change management for scientific datasets on HPC systems. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, SC '18. IEEE Press, Piscataway, NJ, USA, 72:1–72:13.

Haas, L.M., Lin, E.T., Roth, M.A., 2002. Data integration through database federation. IBM Syst. J. 41, 578–596. https://doi.org/10.1147/sj.414.0578.

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. Software 24, 879–888. https://doi.org/10.1016/j.envsoft.2009.01.002.

Horsburgh, J.S., Aufdenkampe, A.K., Mayorga, E., Lehnert, K.A., Hsu, L., Song, L., Spackman Jones, A., Damiano, S.G., Tarboton, D.G., Valentine, D., Zaslavsky, I., Whitenack, T., 2016. Observations Data Model 2: a community information model for spatially discrete Earth observations. Environ. Model. Software 79, 55–74. https://doi.org/10.1016/j.envsoft.2016.01.010.

Hubbard, S.S., Varadharajan, C., Wu, Y., Wainwright, H., Dwivedi, D., 2020. Emerging technologies and radical collaboration to advance predictive understanding of watershed hydrobiogeochemistry. Hydrol. Process. 34, 3175–3182. https://doi.org/10.1002/hyp.13807.

Hubbard, S.S., Williams, K.H., Agarwal, D., Banfield, J., Beller, H., Bouskill, N., Brodie, E., Carroll, R., Dafflon, B., Dwivedi, D., Falco, N., Faybishenko, B., Maxwell, R., Nico, P., Steefel, C., Steltzer, H., Tokunaga, T., Tran, P.A., Wainwright, H., Varadharajan, C., 2018. the East river, Colorado, watershed: a mountainous community testbed for improving predictive understanding of multiscale hydrological–biogeochemical dynamics. Vadose Zone J. 17 https://doi.org/10.2136/vzj2018.03.0061.

Kakalia, Z., Varadharajan, C., Alper, E., Brodie, E.L., Burrus, M., Carroll, R.W.H., Christianson, D., Dong, W., Hendrix, V., Henderson, M., Hubbard, S., Johnson, D., Versteeg, R., Williams, K.H., Agarwal, D.A., 2021. The Colorado East River

community observatory data collection. Hydrol. Process. 35 (6), e14243. https://doi.org/10.1002/hyp.14243.
Khalsa, S.J.S., 2017. Data and metadata brokering – theory and practice from the BCube project. Data Sci. J. 16, 1. https://doi.org/10.5334/dsj-2017-001.
Krysnova, V., Arnold, J.G., 2008. Advances in ecohydrological modelling with SWAT—a review. Hydrol. Sci. J. 53, 939–947. https://doi.org/10.1623/hysj.53.5.939.
Maxwell, R.M., Condon, L.E., Kollet, S.J., 2015. A high-resolution simulation of groundwater and surface water over most of the continental US with the integrated hydrologic model ParFlow v3. Geosci. Model Dev. (GMD) 8, 923–937. https://doi.org/10.5194/gmd-8-923-2015.
Nativi, S., Craglia, M., Pearlman, J., 2013. Earth science infrastructures interoperability: the brokering approach. IEEE J. Sel. Top. Appl. Earth Obs. Rem. Sens. 6 (3), 1118–1129. https://doi.org/10.1109/JSTARS.2013.2243113.
Nativi, S., Mazzetti, P., Craglia, M., Pirrone, N., 2014. The GEOSS solution for enabling data interoperability and integrative research. Environ. Sci. Pollut. Res. 21, 4177–4192. https://doi.org/10.1007/s11356-013-2264-y.
OGC, 2017. Testbed-12 JSON and GeoJSON user guide [WWW Document]. URL. http://docs.opengeospatial.org/guides/16-122r1.html#_ogc_needs_that_geojson_does_not_cover.
OMA, May 2012. "NGSI Context Management," tech. rep., approved version 1.0. http://www.openmobilealliance.org/release/NGSI/V1_0-20120529-A/OMA-TS-NGSI_Context_Management-V1_0-20120529-A.pdf.
Pastorello, G., Papale, D., Chu, H., Trotta, C., Agarwal, D., Canfora, E., Baldocchi, D., Torn, M., 2017. A new data set to keep a sharper eye on land-air exchanges. Eos, Trans. Am. Geophys. Union 98. https://doi.org/10.1029/2017EO071597.
Ramakrishnan, L., Poon, S., Hendrix, V., Gunter, D., Pastorello, G.Z., Agarwal, D., 2014. Experiences with user-centered design for the tigres workflow API. E-science (e-Science). In: 2014 IEEE 10th Int. Conf. https://doi.org/10.1109/eScience.2014.56.
Rode, M., Wade, A.J., Cohen, M.J., Hensley, R.T., Bowes, M.J., Kirchner, J.W., Arhonditsis, G.B., Jordan, P., Kronvang, B., Halliday, S.J., Skeffington, R.A., Rozemeijer, J.C., Aubert, A.H., Rinke, K., Jomaa, S., 2016. Sensors in the stream: the high-frequency wave of the present. Environ. Sci. Technol. 50, 10297–10307. https://doi.org/10.1021/acs.est.6b02155.
Sprague, L.A., Oelsner, G.P., Argue, D.M., 2017. Challenges with secondary use of multi-source water-quality data in the United States. Water Res. 110, 252–261. https://doi.org/10.1016/j.watres.2016.12.024.
Tomkins, J., Lowe, D., 2016. Timeseries profile of observations and measurements v1.0 OGC document 15-043r3l. http://docs.opengeospatial.org/is/15-043r3/15-043r3.html.
USGS Water Resources, 2020. Integrated water availability assessments (IWAAs) [WWW Document]. URL. https://www.usgs.gov/mission-areas/water-resources/science/integrated-water-availability-assessments-iwaas?qt-science_center_objects=0#qt-science_center_objects.
Varadharajan, C., Agarwal, D.A., Brown, W., Burrus, M., Carroll, R.W.H., Christianson, D. S., Dafflon, B., Dwivedi, D., Enquist, B.J., Faybishenko, B., Henderson, A., Henderson, M., Hendrix, V.C., Hubbard, S.S., Kakalia, Z., Newman, A., Potter, B., Steltzer, H., Versteeg, R., Williams, K.H., Wilmer, C., Wu, Y., 2019. Challenges in building an end-to-end system for acquisition, management, and integration of diverse data from sensor networks in watersheds: lessons from a mountainous community observatory in East river, Colorado. IEEE Access 7, 182796–182813. https://doi.org/10.1109/ACCESS.2019.2957793.
Zaslavsky, I., Whitenack, T., Williams, M., Tarboton, D.G., Schreuders, K., Aufdenkampe, A., 2011. Proceedings of the Environmental Information Management Conference, 28–29 September. EIM'2011, Santa Barbara, CA, pp. 145–150.