

Received October 7, 2019, accepted November 21, 2019, date of publication December 5, 2019, date of current version December 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957793

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

CHARULEKA VARADHARAJAN¹, DEBORAH A. AGARWAL², WENDY BROWN^{3,6}, MADISON BURRUS¹, ROSEMARY W. H. CARROLL⁴, DANIELLE S. CHRISTIANSON², BAPTISTE DAFFLON¹, DIPANKAR DWIVEDI¹, BRIAN J. ENQUIST⁵, BORIS FAYBISHENKO¹, AMANDA HENDERSON⁶, MATTHEW HENDERSON², VALERIE C. HENDRIX², SUSAN S. HUBBARD¹, ZARINE KAKALIA¹, ALEXANDER NEWMAN⁶, BENJAMIN POTTER⁷, HEIDI STELTZER⁸, ROELOF VERSTEEG⁹, KENNETH H. WILLIAMS¹, CHELSEA WILMER⁶, AND YUXIN WU¹

¹Lawrence Berkeley National Laboratory, Earth and Environmental Sciences Area, Berkeley, CA 94720, USA

²Lawrence Berkeley National Laboratory, Computing Sciences Area, Berkeley, CA 94720, USA

³Bugs Unlimited, LLC Crested Butte, CO 81224, USA

⁴Desert Research Institute, Reno, NV 89512, USA

⁵Department of Ecology and Evolutionary Biology, The University of Arizona, Tucson, AZ 85721, USA

⁶Rocky Mountain Biological Laboratory, Crested Butte, CO 81224, USA

⁷Navarro Research and Engineering, Grand Junction, CO 81503, USA

⁸Fort Lewis College, Durango, CO 81301, USA

⁹Subsurface Insights, LLC, Hanover, NH 03755, USA

Corresponding author: Charuleka Varadharajan (cvaradharajan@lbl.gov)

This work was supported in part by the U.S. Department of Energy under Contract DE-AC02-05CH11231, in part by the WatershedFunction Scientific Focus Area funded by the U.S. Department of Energy, Office of Science, Office of Biological, and Environmental Research under Contract DE-AC02-05CH11231, in part by the National Energy Research Scientific Computing Center (NERSC), U.S. Department of Energy Office of Science User Facility operated under Contract DE-AC02-05CH11231, in part by the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) data repository operated under Contract DE-AC02-05CH11231, and in part by the Subsurface Insights for development of some of the core abilities used here under Award DE-SC0009732 and Award DE-SC0018447.

ABSTRACT The U.S. Department of Energy's Watershed Function Scientific Focus Area (SFA), centered in the East River, Colorado, generates diverse datasets including hydrological, geological, geochemical, geophysical, ecological, microbiological and remote sensing data. The project has deployed extensive field infrastructure involving hundreds of sensors that measure highly diverse phenomena (e.g. stream and groundwater hydrology, water quality, soil moisture, weather) across the watershed. Data from the sensor network are telemetered and automatically ingested into a queryable database. The data are subsequently quality checked, integrated with the United States Geological Survey's stream monitoring network using a custom data integration broker, and published to a portal with interactive visualizations. The resulting data products are used in a variety of scientific modeling and analytical efforts. This paper describes the SFA's end-to-end infrastructure and services that support the generation of integrated datasets from a watershed sensor network. The development and maintenance of this infrastructure, presents a suite of challenges from practical field logistics to complex data processing, which are addressed through various solutions. In particular, the SFA adopts a holistic view for data collection, assessment and integration, which dramatically improves the products generated, and enables a co-design approach wherein data collection is informed by model results and vice-versa.

INDEX TERMS Sensor systems and applications, sensors, geoscience, water resources, watershed, data management, data integration, data processing, co-design.

I. INTRODUCTION

Watersheds are fundamental units of the Earth, and are complex systems whose behavior and evolution are governed by coupled interactions between hydrological, biogeochemical, and ecological processes, as well as human factors such as land use [1]–[3]. It is becoming increasingly important to understand and predict how watersheds will respond to changing environmental conditions and anthropogenic perturbations to optimize management of water-energy resources. This requires access to a diverse set of observations at multiple spatial and temporal scales that mutually provide insight into the functioning of watershed subsystems, and guide the development and application of predictive modeling tools.

In particular, the use of sensors and environmental observatories for measuring watershed processes has exploded over the past few decades. Sensors can dramatically enhance data collection, for example by increasing the temporal resolution to sub-hourly or greater timescales, enabling data to be collected from remote locations or harsh periods where the field site can be inaccessible, alerting users of unusual or otherwise diagnostic conditions warranting further examination in a rapid response manner, and reducing the amount of manual effort and time needed to collect measurements. Thousands of sensors are now routinely being used in large networks including the United States Geological Survey (USGS) stream network [4], the National Ecological Observatory Network (NEON; <https://www.neonscience.org/>), and the United States Department of Energy's (DOE) Ameriflux network (<https://ameriflux.lbl.gov/>) that observe critical watershed parameters such as river discharge, surface and groundwater levels, water quality (temperature, pH, dissolved oxygen, conductivity), and fluxes of carbon, water and energy [5]. Emerging optical sensors for solutes such as nitrate, phosphate and dissolved organic matter (DOM) present an opportunity to measure and determine aquatic health and biogeochemical dynamics at unprecedented temporal and spatial resolution [6]–[8].

In turn, high-resolution data from sensors have led to novel discoveries in watershed hydrobiogeochemical behavior. For example, electrical conductivity and pH data at the Hore stream in Plynlimon, Wales, collected at hourly resolution revealed diurnal patterns that were not visible in the data collected at daily, weekly, and monthly intervals, thereby demonstrating the value of collecting data using automated sensors [9]. Similarly high-resolution nitrate and DOM concentrations in the Sleepers River in Vermont, US, were found to be influenced by peak snowmelt at seasonal time-scales, and in-stream biogeochemical processes at daily time-scales [10]. High-frequency data from monitoring networks have provided driving and validation data for a broad suite of watershed models such as the National Water Model [11], Amanzi-All Terrestrial Simulator [12],

PFLOTRAN [13], [14], SWAT [15] and are now leading to a new research paradigm where data mining and machine learning algorithms can be used to gain insights into Earth system science [16].

A number of cyberinfrastructure systems to acquire, process, and publish sensor network data have been developed for water monitoring sensor networks. The USGS maintains the National Water Information System (NWIS; <https://nwis.waterdata.usgs.gov/>), and the National Groundwater Monitoring Network (NGWMN; <https://cida.usgs.gov/ngwmn/>) to publish data collected from its monitoring networks [17]. The USGS data can be explored, subsampled and retrieved through user portals and the NWIS web services. The Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) has developed the Hydrologic Information System (HIS; <https://hiscentral.cuahsi.org/>). This system maintains a meta-data catalog of approximately 100 providers, from which data can be retrieved using web services.. The HIS enables users to access the data via the Hydroclient interactive portal and the WaterOneFlow web services. Additionally, CUAHSI maintains the Hydroshare system to share and publish data, and obtain digital object identifiers (DOIs) for citations and has developed the Observation Data Model (ODM2), an information model to store and manage diverse Earth observations [18]–[20]. The United States Environmental Protection Agency (EPA) maintains the Water Quality Exchange (WQX), which is a framework to enable data sharing across local, state, and federal agencies (<https://www.epa.gov/waterdata/water-quality-data-wqx>), and is developing the Interoperable Watershed Networks to handle continuous water quality data [21]. Collectively these systems contribute to the Open Water Data Initiative (<https://acwi.gov/spatial/owdi/>), and more recently the Internet of Water initiative [22], which seek to overcome challenges related to water data publication and discovery across the United States (US). The Ameriflux network has developed methods to standardize, QA/QC, gap fill, and process high-frequency eddy covariance and micrometeorological datasets [23].

Despite the immense progress made over the last two decades, significant gaps and challenges still remain for scientific researchers seeking to acquire, manage, and use data from watershed monitoring networks (e.g. [24]). While the need for, and advantages of using sensors and related cyberinfrastructure to advance watershed science are well documented, literature on challenges associated with sensor data collection, curation, processing, and use in models are relatively sparse (e.g., [25], [26]). There is an inherent assumption that the use of sensors for environmental observations will make data acquisition easier, and that data will become readily available for analysis once a sensor is installed in the field. In reality, the automation of data collection requires periodic maintenance and calibrations of the sensors, constant access to power, and the ability to download data at regular intervals, which for real-time streaming data requires internet

The associate editor coordinating the review of this manuscript and approving it for publication was Laxmisha Rai .

connectivity in the field. The data need to be formatted, documented, and organized in a manner that facilitates easy retrieval of large datasets. Data quality issues of various types including flatlining, sensor drift, and sensor malfunction need to be handled prior to data use. Often, data from various types of measurements need to be integrated across providers, which requires implementation of metadata and data standards, exchange protocols, and controlled vocabularies. Finally, the data need to be easily accessible through technologies such as APIs and web portals to enable their use in subsequent analyses, visualizations, and modeling.

This paper highlights the potential problems and solutions that can be involved in building the integrated field and data infrastructure necessary to make sensor data useful for scientific analysis and readily assimilable into predictive numerical models. First, we present the field- and cyber-infrastructure that we have developed to obtain an array of datasets across an emblematic watershed for the DOE's Watershed Function Scientific Focus Area (SFA) project. This infrastructure forms an end-to-end system that deals with many aspects of sensor-based measurements, including data acquisition, curation, quality analysis and quality control (QA/QC), integration with other data streams, visualization, and download. We then illuminate challenges faced in obtaining these data from both the field and data management perspectives, describe our solutions to some of these issues, and identify areas for improvement. Finally, we describe how such an end-to-end data pipeline can enable co-design strategies for situations in which data are used for model development, and models can help optimize data collection from sensor networks. Although we provide examples from the Watershed Function SFA, many of the challenges and solutions discussed are not unique to watershed science, and are broadly applicable to studies of other complex systems (e.g. other natural or urban environments, transportation infrastructure), where sensors are used to monitor the environment, generating heterogeneous data that are then used in predictive models.

II. FIELD INFRASTRUCTURE AND DATA COLLECTION

The Watershed Function SFA seeks to understand and predict the impacts of hydrologic perturbations such as drought and early snowmelt on the retention and release of water, carbon, nutrients, and metals in mountainous watersheds [27]. The SFA's primary field site is the East River (ER), a snow-dominated, headwater basin located in the Upper Colorado River Basin (UCRB) in the western US. The ER is considered to be representative of the snow-dominated headwater watersheds in the UCRB and is part of several testbeds in the US supported by the DOE's Biological and Environmental Research Subsurface Biogeochemistry Program (<https://doesbr.org>). The ER watershed spans 300 square km., and comprises the pristine East River and Washington Gulch drainages, as well as the mining-impacted Slate River and Coal Creek drainages. The watershed contains various life

zones, including alpine, subalpine, montane, and floodplain ecosystems.

The SFA has deployed extensive sensor instrumentation in the ER (see interactive map at <https://wfsfa-data.lbl.gov/watershed/>). The measurement network includes 17 stream-gauging stations paired with stream-water analyte measurements, 6 weather stations with soil moisture and temperature probes, 10 instrumented groundwater wells, ~40 digital repeat photography cameras (phenocams), and about 30 multilevel soil instruments.

The ER watershed also has significant infrastructure maintained by several federal, state, and local agencies that have different data systems. The USGS maintains gaging stations and collects water quality measurements in the East River, Coal Creek, Slate River and at Almont (confluence of the East and Taylor Rivers), and makes the data available through NWIS. Meteorological and snow data are available from the National Resources Conservation Service (NRCS) 'Butte' and 'Schofield' Snow Telemetry (SNOTEL) sites, Crested Butte Cooperative Observer Network (COOP) stations, and several Weather Underground stations. Additional water quality data are available from the EPA's STORET database and the National Water Quality Portal, which includes measurements by the Colorado Department of Public Health and Environment (CDPHE), and other local groups such as the Coal Creek Watershed Coalition and the Rivers of Colorado Water Watch. The EPA also has a Colorado Clean Air Status and Trends Network (CASTNET) station at Gothic, which collects data under the National Atmospheric Deposition Program (NADP). The Rocky Mountain Biological Laboratory (RMBL; www.rmbll.org) located near Crested Butte is a field station that has collected biological and ecohydrological measurements for 90 years. In addition, an eddy flux tower is maintained by the National Center for Atmospheric Research (NCAR).

For analysis and modeling efforts, the data from the SFA's sensor network need to be integrated with other diverse data collected by the SFA and its collaborators at the East River, as well as the data from the other agencies. Examples of additional data available include a variety of remote sensing measurements such as Light Detection and Ranging (LiDAR) [28], Airborne Snow Observatory [29], Airborne Electromagnetic, and hyperspectral surveys ground-based observations such as electrical resistance tomography, seismic data, time-lapse camera images of snow distribution and phenology, microbial community genome (metagenomic) investigations, soil and sediment biogeochemistry and mineralogy, river and porewater chemistry, and extensive measurements of snow density, snow water equivalent, snow chemistry and stable isotopes [30], [31].

III. SFA DATA MANAGEMENT FRAMEWORK

Environmental observatories such as the SFA's ER site typically require multiple components of an information system to curate, process, store, and publish data [26]. The SFA's Data Management Framework provides infrastructure and

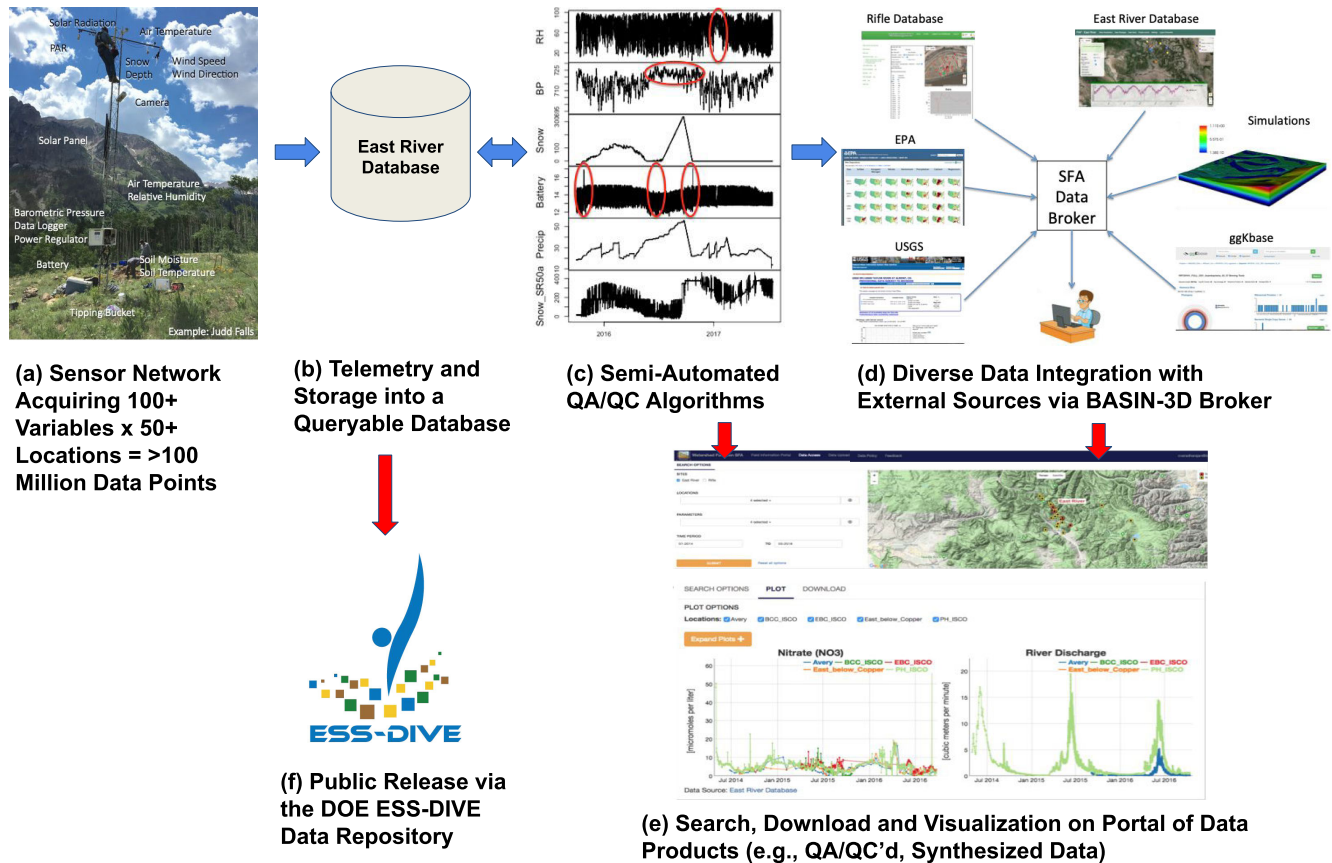


FIGURE 1. Overview of the Watershed SFA end-end Data Management Framework. The framework contains (a) a data collection and acquisition system involving a distributed sensor network across the watershed for diverse observations, (b) a queryable database with a workflow to telemeter and store data and associated data products with relevant metadata, (c) scripts for semi-automated QA/QC with cleaned data stored in the ERDB, (d) a data integration broker (BASIN-3D) to synthesize project data with external datasets for real-time queries, (e) an advanced data search and access portal for data discovery, exploratory analysis and download, and (f) periodic publication of data with DOIs in the DOE's ESS-DIVE repository.

services to support various aspects of the data lifecycle [32]. These include the ability to (1) manage, archive, and publicly release datasets as per the project's data policy, (2) enable the SFA team and the broader community to discover and access relevant datasets, (3) perform Quality Assurance and Quality Control (QA/QC) of priority datasets; and (4) enable efficient data collection, data integration, and product generation. We have developed specific tools for data management and preservation, QA/QC, data discovery and visualization (Figure 1), as described below.

A. EAST RIVER DATABASE

We have maintained a well-curated repository for the Watershed Function SFA's data and metadata [27] called the East River database (ERDB). It is a SQL database, which holds the metadata (such as locations) of monitoring stations and observational time-series, as well as data from hundreds of physical sensors that comprise over 100 million time point-value pairs (Table 1), and additional data such as water quality measurements from laboratory analysis of geochemical water samples. East River sensor measurements are telemetered and automatically ingested into the ERDB. An automated

data flow maps instrument output to an SFA-developed controlled vocabulary for standardization of variable names across instruments, and cross-comparison of data. The lab measurements of field samples are also standardized and integrated into the database regularly via Python scripts. Data from NRCS SNOTEL sites and Weather Underground are harvested from the sources and federated into the ERDB. Data can be retrieved through queries to the ERDB, and are optimized through the use of caching and materialized views. All data stored in the ERDB are accessible to the project team through an Application Programming Interface (API), and web portal interfaces (Section III.D).

B. QUALITY ASSURANCE AND QUALITY CONTROL (QA/QC) OF METEOROLOGICAL, HYDROLOGICAL, AND GEOCHEMICAL DATA

Quality Assurance(QA)/Quality Control(QC) is necessary to ensure the reliability of data and enable scientific accuracy. We perform QA/QC for the SFA's data using R (<https://www.r-project.org/>) and Python (<https://www.python.org/>) scripts, and document our

TABLE 1. The SFA maintains hundreds of sensors of various types in the ER watershed, which measure diverse datasets. Some sensors are placed at multiple sites or depths, and measure multiple variables; hence there typically are many data streams generated by a single instrument. The streaming data are automatically telemetered and ingested into a queryable database.

Sensor	Variables Measured	No. of sensors*	No. data streams
YSI EXO2 Multiparameter Sonde	Timestamp, Battery Voltage	1	2
YSI EXO Conductivity & Temperature Smart Sensor	Water Conductivity, Water Temperature	1	6
YSI EXO Optical Dissolved Oxygen Smart Sensor	Dissolved Oxygen	1	2
YSI EXO pH & ORP Smart Sensor	pH, Oxidation Reduction Potential (ORP)	1	3
YSI EXO fDOM Smart Sensor	Fluorescent Dissolved Organic Matter (fDOM)	1	2
YSI EXO Turbidity Smart Sensor	Turbidity	1	3
Campbell Scientific CS300 Pyranometer	Solar Radiation	6	9
R.M. Young 05103 Wind Monitor	Wind Speed, Wind Direction	7	14
Campbell Scientific HPM45C Temperature and Relative Humidity Probe	Air Temperature, Relative Humidity	7	14
Campbell Scientific CS105 Vaisala PTB101B Barometer	Barometric Pressure	6	6
Judd Ultrasonic Depth Sensor	Snow Thickness	5	5
Licor LI-190R Quantum Sensor	Photosynthetically Active Radiation (PAR)	5	10
Stevens Water HydraProbe	Soil Moisture, Soil Temperature	15	45
Campbell Scientific CS CNR1 Net Radiometer	Incoming Radiation, Reflected Radiation	6	24
Campbell Scientific Thermocouple	Temperature	6	12
Campbell Scientific SR50A-L Sonic Distance Sensor	Snow Thickness	6	6
Campbell Scientific CS650 Water Content Reflectometer	Water Content, Water Temperature	19	57
Decagon Devices MPS-6 Dielectric Water Potential Sensor	Water Potential, Water Temperature	54	108
Campbell Scientific 5TE Soil Moisture, Temperature, and Electrical Conductivity Sensor	Soil Moisture, Temperature, Electrical Conductivity	49	147
In-Situ Aqua TROLL 200 Data Logger	Water Level	2	12
Omega PX309-V050GI	Pressure	9	9
Campbell Scientific CS100 Barometric Pressure Sensor	Barometric Pressure	1	1
OTT HydroMet OTT Pluvio ² S	Precipitation	1	7
LabCell RT1 Soil Temperature Sensor	Temperature	15	15
Solinst Levellogger M5 Edge	Timestamp, Battery Voltage	10	20
HOBO Soil Moisture/Temp Sensor	Soil Moisture, Temperature	10	24
Campbell Scientific CR1000 Measurement and Control Data Logger	Timestamp, Battery Voltage	6	6
TOTAL		251	569

* This is a count of the number of sensors (as of June 2019) at individual point locations, where each location has a unique latitude, longitude and depth.

workflows using R Markdown (<https://rmarkdown.rstudio.com/>) and Jupyter Notebooks (<https://jupyter.org/>).

The QA/QC process for time-series data involves both the identification of problems in the data (QA), and subsequent data cleaning or processing (QC). An important component of the QA/QC procedure is gap filling of missing values. Our QA/QC workflow includes three major steps:

Step 1 (QA)—Cataloging data, determining the temporal frequency of sampling to assess data availability, and assessing data quality;

Step 2 (QC)—Processing and cleaning raw datasets, including custom procedures for formatting timestamps, detecting and removing duplicates, bad data and outliers, gap filling, and flagging QC-ed data; and

Step 3—Preparing time grids of the desired frequency, and exporting data into the desired format for modeling or further analysis.

The first step involves an automated data inventory workflow to catalog available meteorological, hydrological, and geochemical datasets combined with semi-automated methods to assess their quality (Step 1). A Python script implemented as a Jupyter notebook queries across all monitored stations and variables, and retrieves data using the ERDB API. The script creates high-level lookup tables of available data displaying the data types, locations, and temporal frequency of measurements, as well as the number of data points with other descriptive statistics such as the mean, min, and max values for each location. Cross-comparison of

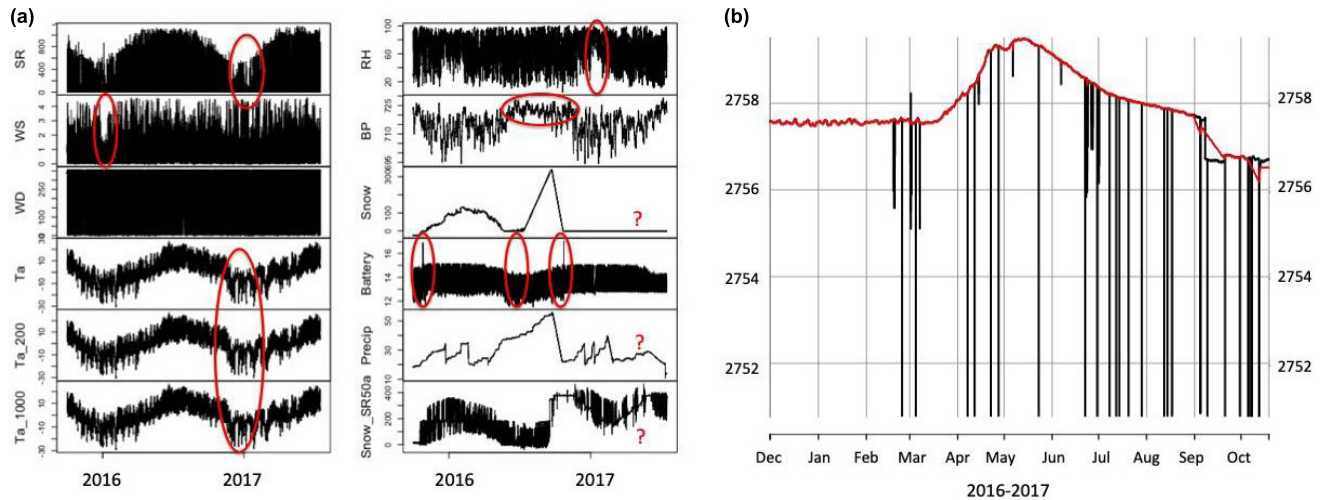


FIGURE 2. Examples of QA/QC for different types of sensor data collected at the East River, CO: (a) Time-series data from the Billy Barr meteorological station, with red ovals indicating the outliers and suspicious data; (b) Time-series plot groundwater level data obtained from a downhole pressure transducer in the monitoring well PLM-6; the black line shows the raw water level (absolute elevation in m), and the red line shows the QA/QC-ed, de-spiked, clean time-series of water level. Many of the spikes are caused due to the sensors being pulled out of the borehole for calibration or water sampling.

the datasets using their range, mean, and standard deviation values, and temporal sampling frequency enables quick identification of potential problems. Finally, the scripts plot the data inventory and time-series values for visual inspection. Thus, we collectively use the subroutines to perform real-time quality checks of data and determine the datasets that require subsequent application of the QA/QC procedures.

The next step involves processing and cleaning the raw time-series datasets through semi-automated scripts (Step 2). The workflow for cleaning time-series includes: (1) conversion of timestamps into a uniform ISO 8601 date/time format specified as YYYY-MM-DD HH:MM:SS [33]; (2) identification and removal of duplicates using the function “duplicated” in the “base” package of R; (3) removal of data points that are clearly outside the sensor measurement ranges; and (4) automatic detection and removal other bad data in time-series using the R package “tsoutliers” [34], which is a widely used method to detect outliers, level shifts and temporary changes [35]. Various criteria and metrics are applied in cleaning the datasets. For example, the cleaned data have to be consistent with the sampling resolutions and descriptive statistics from Step 1.

Meteorological datasets needed as driving inputs to climate models require additional processing. These datasets are gap filled using the Singular Spectrum Analysis method in the R package “Rssa” [36]. The final step involves aggregating the cleaned, gap-filled time-series into the desired temporal frequency, which is typically hourly or half-hourly time intervals for modeling needs, and exporting the data into files with csv or netCDF formats, for data analysis or input into simulation codes.

This QA/QC workflow was applied to diverse data from the UCRB, including data from 17 meteorological stations located in the East River watershed, and

additional meteorological data from other organizations, including North American Regional Reanalysis (NARR, produced by the National Centers for Environmental Prediction, NCEP), Western Regional Climate Center (WRCC) data, PRISM high-resolution spatial climate dataset (<http://www.prism.oregonstate.edu/>), NRCS SNOTEL data, and data from the USGS (Figure 2a). We also performed QA/QC of groundwater level data measured in monitoring wells (e.g. Figure 2b), river discharge measurements, and geochemical data.

The QA/QC-ed data products are used for a variety of purposes in the SFA including data analysis [27] and numerical modeling of hydrological and geochemical processes, including infiltration, evapotranspiration, and evolution of CO₂ concentrations, at the ER, using numerical codes such as the Community Land Model (CLM), Ecosys, PFLOTTRAN, and ParFlow [37]–[41].

C. DATA INTEGRATION

The SFA needs to obtain and integrate data from various providers for scientific analysis and modeling. Hence, we developed a data broker, BASIN-3D (Broker for Assimilation, Synthesis and Integration of eNvironmental Diverse, Distributed Datasets) that provides a generic framework to integrate diverse, multiscale data across a variety of additional data sources and environmental data types [27]. BASIN-3D is an extensible brokering software that enables creation of synthesized datasets for use in scientific analysis, modeling, and visualization. It is designed to enable researchers to search and visualize the latest data available from many providers, as easily as if it were an integrated dataset in local storage. BASIN-3D retrieves data from the providers in real-time, as-needed, which ensures that the latest version of the data is always served to the user. This is

particularly useful for sensor data, where the data stream is constantly changing with additional measurements and QA/QC processing.

BASIN-3D connects to data sources in real-time via web services, provides lists of the available data types and locations, retrieves subsets of data requested by the user, and transforms the data streams into uniform, abstracted formats to provide common nomenclature and formats required for an integrated dataset. The abstracted formats are based on the Open Geospatial Consortium (OGC) and ISO “Observations and Measurements” (OGC 10-004r3 / ISO 19156: 2013) and OGC “Timeseries Profile of Observations and Measurement” (OGC 15-043r3) data standards [42], [43]. The current implementation of BASIN-3D integrates time-series data; however the OGC/ISO framework also allows for future additions of diverse data types such as sample-based or remotely-sensed data. Notably, BASIN-3D is able to handle multiscale data, through the use of the OGC/ISO framework that allows for specification of hierarchies of spatial features (e.g., points, plots, sites, watersheds, basin). Thus, users can retrieve data for a specific site or river basin.

The BASIN-3D software is open source and available at <https://github.com/Watershed-Function-SFA/BASIN-3D>.

D. INTERACTIVE PORTALS FOR DATA ACCESS AND VISUALIZATION

The SFA data management framework also includes two Javascript-based interactive web portals to disseminate information on field activities to the public, and provide data discovery and access tools to the team. The first is a Field Information Portal, which is a publicly-accessible, interactive map providing information on the SFA’s field infrastructure and research activities (<https://wfsfa-data.lbl.gov/watershed/>). Detailed location pages provide information about each sampling location such as the measurements and instrumentation available there (Figure 3). Users are able to pick measurement themes (e.g. hydrogeology, biogeochemistry) or research themes (e.g. soil carbon) to filter sites of interest.

The second portal that builds on the BASIN-3D data integration broker (Section III.C), provides users an integrated, intuitive search and download of the data (Figure 4). The application of BASIN-3D for this portal requires an additional software component (the WFSFA data broker) to connect to the specific data sources of interest (i.e., the SFA’s East River and Rifle field databases, USGS NWIS), and map the data source vocabularies to the BASIN-3D synthesis models. This enables users to search across the SFA database and NWIS to retrieve variables of interest (e.g. discharge) in a uniform format for visualization and download.

The portal utilizes several third party Javascript libraries to enable various capabilities that make the user interface more intuitive. The features and associated libraries include search and selection of multiple inputs from dropdown lists (enabled by Bootstrap Multiselect; <https://github.com/davidstutz/bootstrap-multiselect>), slicing and dicing data (Crossfilter;

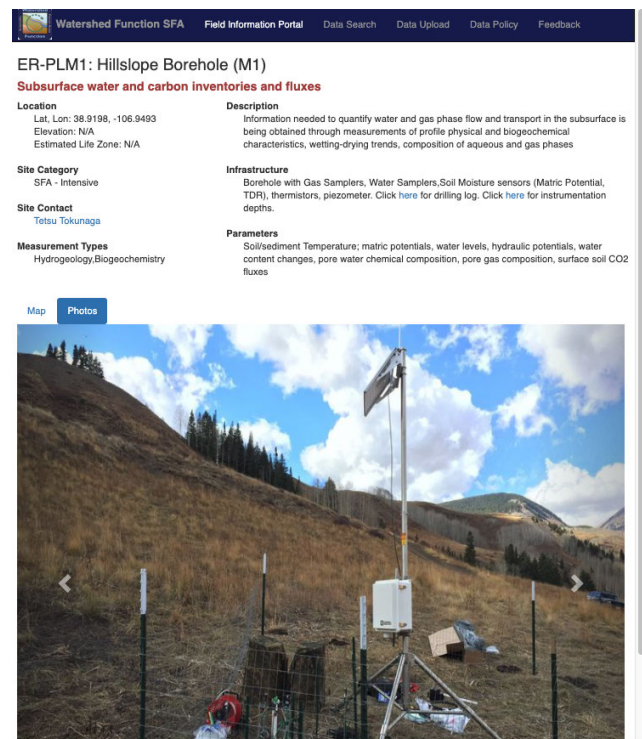


FIGURE 3. Screenshot from the Field Information Portal that contains detailed site pages describing the location and infrastructure available at each long-term sampling location.

<https://square.github.io/crossfilter/>), interactive maps (Google Maps; <https://cloud.google.com/maps-platform/> and Leaflet; <https://leafletjs.com/>), interactive tables (DataTables; <https://github.com/DataTables/DataTables>), and interactive visualizations for exploratory data analysis (Dygraphs charting library; <http://dygraphs.com> and other libraries utilizing D3.js; <https://d3js.org/>). Thus the portal allows users to seamlessly browse, integrate, sub-sample, download, and visualize integrated datasets across different databases without writing any code (see video in Supporting Information).

The portal is deployed on the Spin platform [44], a Docker-based system for hosting scalable web services provided by the DOE’s National Energy Research Scientific Computing Center (NERSC). The portal is deployed as a “microservices” architecture using Docker containers (<https://www.docker.com/>). Multiple instances are easy to “spin up” using isolated containers, which enables easier maintenance and deployment, and allows for scaling of resources to meet increasing user demand.

E. DATA PUBLICATION

The SFA sensor and other project data are made publicly available as per the project’s data management policy (available at <http://watershed.lbl.gov>). Data are first curated within the ERDB or an internal project data management portal. The data are then published periodically with relevant metadata through the DOE’s Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE)

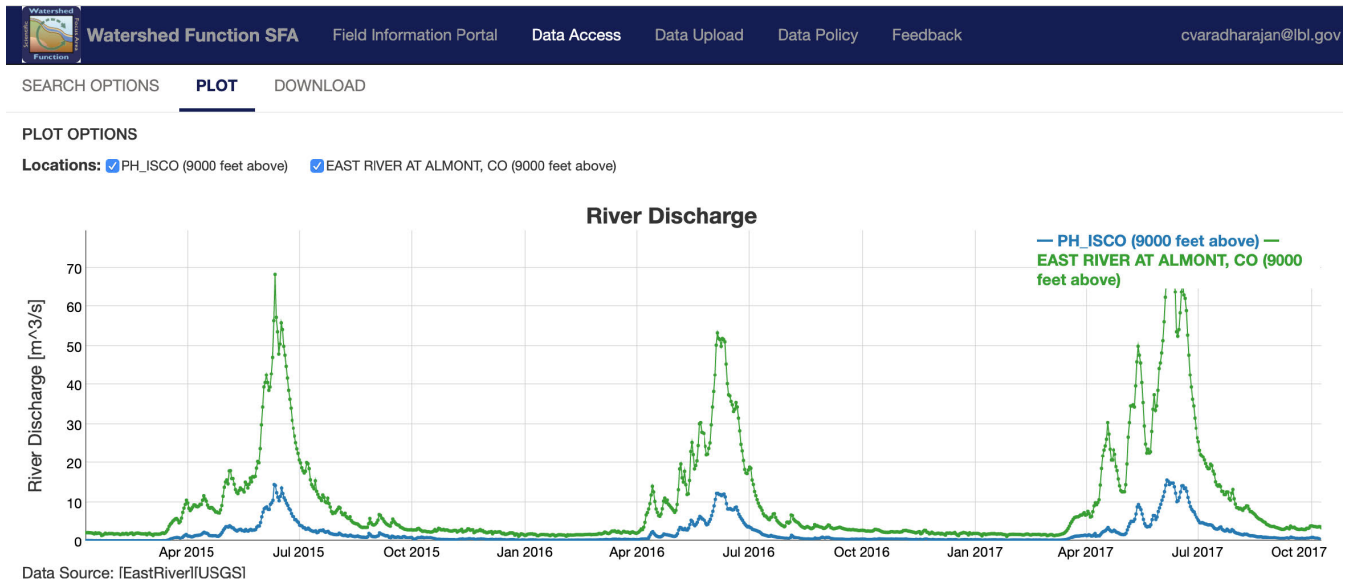


FIGURE 4. Screenshot from the SFA's data access portal showing the integrated visualization of multiple datasets provided by BASIN-3D using interactive visualizations. Here, discharge data from the SFA at its Pumphouse site (PH ISCO) and USGS at the downstream Almont gaging station (USGS ID 09112500) are shown over the same time period. A video providing a virtual tour of the SFA data portal is provided in the supplementary material.

data repository [45]. Sensor data are published in snapshots with a few years of data comprising a single data package (e.g. discharge data for water year 2015-2018 [46]).

IV. OVERCOMING CHALLENGES IN OBTAINING AND USING DATA FROM WATERSHED SENSOR NETWORKS

Organizations seeking to deploy sensor networks in watersheds will face an array of challenges that span practical field logistics to building cyberinfrastructure, data management support services, and processing pipelines for complex data. The issues are compounded by the fact that the technologies for obtaining and managing data are continuously evolving. Hence, the data acquired are becoming more complex over time, as new sensors and data acquisition hardware become available. Similarly, software and data processing capabilities, as well as community data/metadata standards, are rapidly changing. Projects need to anticipate some of these problems at the outset and dedicate staff time and resources to resolving them. In this section, we discuss various challenges encountered, and potential current or future solutions for building an end-to-end sensor data system, illustrated with practical examples from our sensor network in the ER.

A. DATA ACQUISITION

Many factors need to be considered for the deployment, maintenance, and curation of sensor measurements [47]. First, the network design needs to capture the spatial and temporal heterogeneity of the system given the research questions posed, stakeholder needs, and available resources. Secondly, data acquisition and maintenance protocols must be designed to obtain as many high-quality, continuous observations as possible, given the environmental conditions and logistics at the field site. Contingency plans must be made to handle

the inevitable, unexpected events that can lead to missing data or problems with data quality. Finally, all field installations, maintenance/calibrations, and other events must be well documented and made available for future QA/QC and data interpretation.

The challenges inherent to establishing and maintaining viable sensor networks in the field are compounded in snow-dominated, mountainous watersheds due to complex topography, animal/rodent activities, access, and safety considerations. In addition, harsh environmental conditions in the winter test power supplies, sensor applicability and survival, as well as limited telemetry options. Furthermore, site visits may not be possible during critical times of change in the environmental system. Data quality issues arise at all points in the data collection process.

As an example, the ER stream discharge network [31], [46] embodies many of these challenges. Discharge measurements require two types of data. First, one needs to measure river stage. In the ER, this is accomplished using a pressure transducer installed in the stream and corrected for barometric pressure. Second, instantaneous discharge observations are measured. For this, we rely on wading the stream with an Acoustic Doppler Velocimeter (ADV). From these data, one develops a depth-to-discharge relationship (rating curve) to establish near-continuous discharge values. However, discharge data quality is hampered in steep channels by turbulent conditions, which can result in high-frequency temporal variations that need to be smoothed in post-processing (Figure 5). Rating curves depend on high-flow observations when stream wading is often unsafe, and are invalidated by changes in channel cross-sectional geometry that can potentially occur during peak discharge. High-flows and woody debris in the streams can lead to loss of instrumentation and missing data.

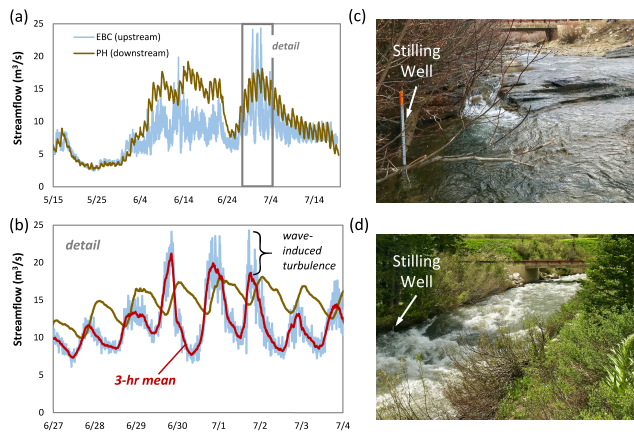


FIGURE 5. Discharge measurements collected at two sites in the East River to illustrate several challenges in data acquisition in mountainous watersheds. The upstream East Below Copper (EBC) location sees considerable turbulence during spring snowmelt in comparison to the downstream Pumphouse (PH) location. It is notable that the EBC transducer was washed away during peak discharge in 2016, 2017 and 2018. (a) A comparison of EBC and PH 10-minute hydrograph for runoff in Spring 2019 (b) the detail shows the 10-min data for June 27 to July 4, 2019. Wave-induced turbulence at EBC increases variance in the hydrograph to falsely produce more estimated streamflow. In contrast, the hydrograph for PH is smooth, indicating that no turbulence is observed at that location. The high-frequency noise from turbulence at EBC is removed using a simple 3-hour moving average (c) installation of EBC in October 23, 2018 at low discharge appears ideal with a large eddy and bedrock controls constraining the cross section, (d) EBC in June 21, 2019 experiences standing waves and turbulent conditions imposing on the transducer (inside the stilling well) and un-wadable conditions.

Winter iced-conditions can produce anomalously high pressure readings that need correction [48]. The spring freshet is a critical period of solute export that requires sensors be installed in the fall and survive the winter. However, sensors will freeze if temperatures drop below 0°C and can be swept away during peak runoff. With lack of telemetry, invalid or lost data are often not discovered until the summer when access to the transducers is safe. Subsequently, a critical annual event is missed.

Sensor networks deployed on land, such as along the ER hillslopes, face a unique set of challenges as well. A specific example is sensor and cable damage due to rodents or the grazing activities of cows. Temporal fencing during the active season, and other protective measures can mitigate damages to a certain degree, but complete damage avoidance is rare. Another problem that can occur are potential power outages during the winter season, when site access for battery replacement might be challenging or when the solar panel becomes buried under the snow. Even if the system's power consumption is configured to last through the winter, data transmission under harsh winter conditions can consume an abnormally high amount of power, resulting in faster drainage of the onboard batteries. Finally, unexpected events in the field can lead to missing or erroneous data. As a classic example, the solar panels used for CO₂ flux instrumentation at the East River were completely buried under a deep snowpack in the spring of 2019. This caused the top panel to later be

damaged by snow grooming for a ski race near our research site.

To address some of these challenges, we have designed our network to be robust by installing redundant sensors at all our stream sites. The network sustainability is reviewed annually to assess human resources, safety concerns, which data are critical to research goals, and if sensors should be removed, added, re-located or replaced to meet data quality standards. In addition, we are continually exploring new technologies and their applicability across a range of stream characteristics and ability to survive extreme environmental conditions. Specifically, we now use stand-alone bluetooth sensors to measure water levels (HOBO MX2001), which are proving capable of surviving winter conditions, and additionally are cost-effective for safe and rapid download of data without the need for a large power source, cell service, or satellite communication. We retrieve data weekly, which limits the potential for lost data even if the sensor is swept away during subsequent flows. We are also incorporating non-wading tracer techniques, and are investigating non-contact options such as hand-held microwave radar, scanning LiDAR, airborne GPR, and Interferometric stream radar (ISRad) through institutional support as well as collaborations with federal and university partners. Finally, we document details about our sensor installs (Section IV.B), and address some of the quality issues through post-processing and QA/QC (Section IV.C).

In the future, advances in low-cost/low-power sensors, wireless sensor networks, and related hardware (e.g. Arduino or Raspberry Pi boards) can make it possible to overcome many of these challenges, and particularly enable installation of several redundant sensors. Currently, low-cost sensors are not available for many of the measurements needed within the ER (e.g. water chemistry, soil moisture). Similarly, off-the-shelf options for ultra-low-power or energy harvesting devices are limited, particularly for remote monitoring in rugged, outdoor environments. Although, wireless sensor technologies are rapidly evolving with more developed instrumentation becoming commercially available, these still pose challenges (e.g. limited range, durability, cost) for practical field deployment at scale. Significant improvements in low-cost/ultra-low-power wireless sensor hardware and communication protocols (e.g. bluetooth, wireless mesh networks, LoRa) are needed to revolutionize monitoring of natural water systems.

B. DATA CURATION AND STORAGE

Considerations for storage and management of sensor data include the volume of data generated, heterogeneity of instrumentation, diversity of data types and formats, the need to federate external data or be interoperable with other databases, and requirements for data usage.

Data curation can be especially difficult when researchers from multiple institutions and scientific backgrounds collect data through a variety of sensors and data acquisition infrastructure, which typically occurs in many research projects including at our ER field site. The heterogeneity

in personnel, organizations, sampling preferences, equipment, and data types poses a number of challenges for data curation and archival. First, it becomes difficult to track additions or modifications to the sensor network, which occur for several reasons such as new sensor installations or instrument relocation. Second, maintaining adequate documentation and metadata about sensor characteristics (e.g. instrument types, observational footprints), maintenance (e.g. calibration times and procedures, replacements), and event details (e.g., install specifications, failure modes) becomes a time- and resource-intensive task. Moreover, the use of different sensor instrumentation and data acquisition hardware systems across research groups makes it challenging to create a common workflow for transmitting data into a database and comparing data across sensor types. For example, soil temperature at the ER is monitored at various locations with three types of sensors (RT1, 5TE, and MPS-6) that have inherently different characteristics such as support volume, orientation, and sensitivity (Table 1). Even for sensors from the same vendor, differences in installation (e.g., orientation angle for phenocams) can result in different responses. A variety of data acquisition systems are used in the ER, requiring data to be downloaded by exchanging a Secure Digital (SD) card, through a serial cable, over bluetooth, and through satellite or cellular modems. Each of these are subject to unique failure modes such as loss of power to a data logger or loss of wireless connectivity that can lead to gaps in data acquisition or transmission (Section IV.A). Additional challenges occur since data from other organizations need to be federated into the database. This requires that differences in metadata reporting, variable names, time zones, and data formats between different providers are reconciled.

We approach these challenges through a multi-pronged effort. First, we curate and store detailed metadata information on permanent infrastructure including sensor placement, sensor type and other details about the installation. In some cases, this information is not readily available, especially for public data obtained from external sources. Second, we identify locations using unique project-specific identifiers for each sampling site, and homogenize variable names by mapping to a controlled vocabulary that covers the range of data types used in the SFA. This helps mitigate issues that might arise due to inconsistent use of location or variable names. Third, we use Quantities, Units, Dimensions, and Types (QUDT), a NASA developed architecture to standardized unit conversion (<http://www.qudt.org/>). Then all data are stored in a queryable database following a modified version of the Observation Data Model [19], which supports storage of heterogeneous environmental data types. Finally, all data are made available through an API with token-based authorization.

Based on our experience, we recommend the use of community-accepted templates for metadata reporting and adoption of data standards as a solution to many of these challenges. Metadata templates would describe aspects such as sensor and acquisition systems, sensor location and

placement, calibration procedures, sample collection metadata, reporting units, timezone, owner and use restrictions, and sensor events such as installs/failures. Some of this metadata are typically not recorded, but provide valuable information needed for data interpretation to users. Examples of metadata templates that are relevant to watershed monitoring include the EPA's WQX continuous monitoring template (<https://www.epa.gov/waterdata/water-quality-exchange-web-template-files>) [7], FLUXNET's Biological, Ancillary, Disturbance and Metadata (BADM; <https://ameriflux.lbl.gov/data/badm-data-templates/>), and the FRAMES reporting templates for ecohydrological measurements [49]. In addition, adoption of standardized data reporting and formats, will greatly reduce the effort required to harmonize and store data, and enable interoperability of the database with other systems. Several standards are relevant for water data including the OGC/ISO Observations and Measurements [43] and related standards such as SensorML [50] and WaterML (<https://www.opengeospatial.org/standards/waterml>).

In large sensor networks, the data volumes generated from thousands of raw data streams, camera imagery, and all the different postprocessing versions will add up over time. In the future, solutions for compression, storage (e.g. fog and edge computing), and retrieval of data efficiently will be needed [51]–[53].

C. DATA ASSESSEMENT

The accuracy and uncertainty in climate, hydrological, and biogeochemical predictions depend on the representativeness and quality of the data collected in the field and laboratory. Commonly, a QA/QC process starts with a manual (i.e., subjective) inspection of data, followed by selection of a statistical algorithm to perform cleaning and gap filling that is most suitable for the particular dataset and investigation objectives.

QA/QC can pose several challenges in detection and cleaning of missing or erroneous data, corrections of sensor drift and periodic malfunctioning due to intermittent current surges and temperature shifts, as well as accounting for the effects of periodic calibration of sensors. First, QA/QC procedures are not generalizable, especially when the data have diverse structures, formats, and patterns. As an example, data cleaning requires the identification and removal of erroneous or inconsistent entries through statistical methods for anomaly/outlier detection. However, anomalies or outliers are not necessarily the outcomes of incorrect or inconsistent entries - they can infrequently result from extreme events, such as flood or drought periods in the ER. Therefore, any anomaly or outlier should be examined through the application of domain knowledge (e.g., in terms of ranges and other attributes, and with consideration of complementary datasets) before discarding the data. Similarly, effective handling of missing data, including gap filling, requires domain-specific knowledge for different types of meteorological, hydrological, and geochemical variables. Although a

variety of statistical methods and tools exist for estimating missing values or correcting inconsistencies, gap filling is not easy and can lead to erroneous data. Several sources can give rise to missing or inconsistent entries such as periodic malfunctioning of sensors, loss of power, sensor drift, technological glitches (i.e., hardware failure), signals that exceed the sensor measurement or calibration ranges, and human errors. It may be easy to fix technological or sensor related errors but not manual errors. Rigorous documentation of all actions conducted on sensors, particularly reinstallations and recalibrations, is required to correct for some of the manual errors in post-processing (Section IV.B).

To address some of these challenges, we created interactive scripts to allow users to access data using APIs and run cataloging, QA/QC or visualization subroutines in Jupyter Notebooks. Users can query for the data available at different stations, and group stations or datasets based on designated attributes (e.g., types of measurements). The modular workflow helps determine the datasets that need QA/QC, and additionally enables exploratory data analysis by helping researchers understand the data availability across the watershed to refine their science questions and determine data needs for modeling. Users can then run the subroutines that are most appropriate for the quality problems in the raw data streams.

We find that the use of interactive workflow tools such as Jupyter notebooks to be an effective method to provide transparency for data processing. The Jupyter tools also help address some of the common challenges in QA/QC [54] by generalizing across computing platforms and processing software (e.g., R and Python), and enabling users to tweak QA/QC procedures and parameters for their custom scientific purpose.

Further advancements in QA/QC can be achieved by using models (mechanistic or data driven), to identify and correct erroneous values, using a co-design approach (Section V). For example, we are working to implement mechanistic models (e.g., reactive transport models, such as PFLOTRAN and Crunch) in our QA/QC framework to provide sanity checks. Examples include flowing the data through models to confirm the charge balance in geochemical samples. In addition, the use of machine learning for anomaly detection, sequence predictions and signal reconstruction, also holds promise for fully automating a real-time QA/QC process. In the ideal situation, QA/QC algorithms are performed on real-time streaming data, with automated notifications being sent to the field team in the event of a sensor failure or suspicious data.

D. DATA INTEGRATION

Bringing data from a variety of sources into a unified framework is a difficult task but is increasingly needed to quantify and predict complex watershed behavior. Integrating data requires re-organizing data from their native formats to a generalized, abstracted format to create a synthesized product.

The primary challenge with integrating heterogeneous data is that they can have a variety of native formats or structures.

Science domains have different norms, and even within a domain, common data ontologies or organizational standards have often not been adopted. Therefore, data are organized and named in very different ways among data sources, requiring significant investment to learn the native format of each data source. Even then, the mapping to a generalized format may not be obvious. For example, the same term “site” can be used by different providers to represent a region containing multiple sampling activities or an individual sampling station. Differences in units also require additional processing, and in some cases may need domain experts to write conversion routines. Although unit ontologies and conversion libraries such as QUDT, UDUNITS (<https://www.unidata.ucar.edu/software/udunits/>), and Pint (<https://pint.readthedocs.io/en/0.9/>) now exist, no single library covers the entire range of physical, chemical, and biological data types that are collected at the ER site. Additionally, there is no consistency in the type of metadata provided by different sources. Some data sources include detailed metadata (e.g., methods, units, aggregation type), and others provide none. For example, data catalog and availability information, and sensor characteristics can be missing, making it difficult to create efficient queries to retrieve data.

One possible solution is for providers or data brokers to adopt internationally recognized data standards. For example, in the BASIN-3D data broker, we map data source terminologies onto the OGC/ISO data standards [43]. While data standards, such as those from the OGC/ISO and Federal Geospatial Data Committee (FGDC) are helpful in tackling some of the data diversity issues, they are not easy for domain scientists to adopt. This is because the standards are typically described at a high level and do not provide implementation guidance beyond simple examples for a few science domains. The standards also tend to contain specialized informatics terminology that is not easily interpretable by domain scientists. To overcome this problem in BASIN-3D, we used a hybrid approach for terminology choices by following the standards in most cases, yet modifying some opaque terminology to more common terms that can be easily understood by domain scientists (e.g., using `utc_offset` instead of `Zone-Offset`, `units_of_measurement` instead of `uom`).

Although BASIN-3D is the outcome of our initial efforts at synthesizing diverse multiscale data, much work remains to be done in the field of data integration for complex watershed datasets. For example, there is no single standard that encompasses all scientific data types that we use in the SFA, thus requiring the use of multiple standards. Moreover, it is difficult to track the various standards available, and their evolution over time. Besides the heterogeneity in metadata, data sources also differ in their QA/QC objectives and procedures, and thus comparing the quality of data available from different sources is an unresolved problem. Advances in data integration technologies are especially needed to achieve the vision of next-generation global sensor networks, wherein data from different sites can be synthesized to identify universal patterns of watershed behavior [47].

E. DATA ACCESS AND VISUALIZATION

Tools to discover and explore data must be intuitive and easily usable, to encourage adoption by domain scientists [47]. There are many challenges associated with designing an intuitive data retrieval and visualization platform for diverse, complex data types, and particularly for streaming sensor data. Search interfaces need to enable users to easily discover the parameters of interest from amongst hundreds of variables. High-resolution sensor data can impact the performance of data retrieval and visualization algorithms, and often require some type of aggregation or dimension reduction for display on small screens. Multiparameter charts can get complex very quickly, particularly when needing to display more than two co-located parameters with different units. Display of data from different time-zones requires conversion to a uniform time format. Heterogeneous data may need different representations - e.g. line plots and bar charts to display on the same chart, and there are an increasing number of visualization techniques and tools to choose from [55]. Users also need to view annotations on additional information such as QA/QC flags (e.g., duplicated data, outliers, corrected or gap-filled data), and sensor metadata for data interpretation.

While there is abundant literature on effective user interfaces for big and streaming sensor data, we have found two approaches to be particularly useful in designing such systems. The first is to employ user experience (UX) research methods in the design of web portal interfaces. UX is a well-established field of human-computer interaction research, and has been used effectively for about 30 years in designing products and web interfaces. This method develops an understanding of user needs through well-defined processes, such as interviews, development of use cases, and guidelines for developers, and can be an effective tool for improving usability. For the SFA, we employed UX research methodologies (e.g., interviews, use cases, mock-ups, and work process observation) that have been shown to work effectively for scientific data management and analyses [56]. Both the SFA's web portals were built after following a thorough UX process, and the time spent upfront in designing these interfaces resulted in products that have needed minimal modifications since their first release.

The second solution that has proved to be particularly effective for scientific data discovery is the use of interactive widgets and visualizations. We used a number of Javascript libraries to create intuitive search and display interfaces by customizing text boxes, mapping interfaces, tabular displays, and plots. Based on the feedback obtained through the UX approach, we eventually chose to create an interface that supported multiparameter charts to deal with the problem of displaying various types of data with different units. These charts are synchronized between all the variables and allow users to zoom in and out for exploratory data analysis (see video in Supporting Information). However, a potential pitfall of using third party libraries is their long-term maintenance or lack thereof, which can lead to incompatibilities between

software versions or security concerns. Thus the libraries need to be tracked over time to incorporate their modifications into the codebase. Additionally, these tools may not be performant with large datasets leading to increased latency in data retrieval or visualizations. We chose to aggregate sensor data to daily intervals, with the option of obtaining higher-resolution data for smaller periods to overcome the data-size limitations. For future improvements, parallelization of the data retrieval/visualization tasks can also help with performance [55].

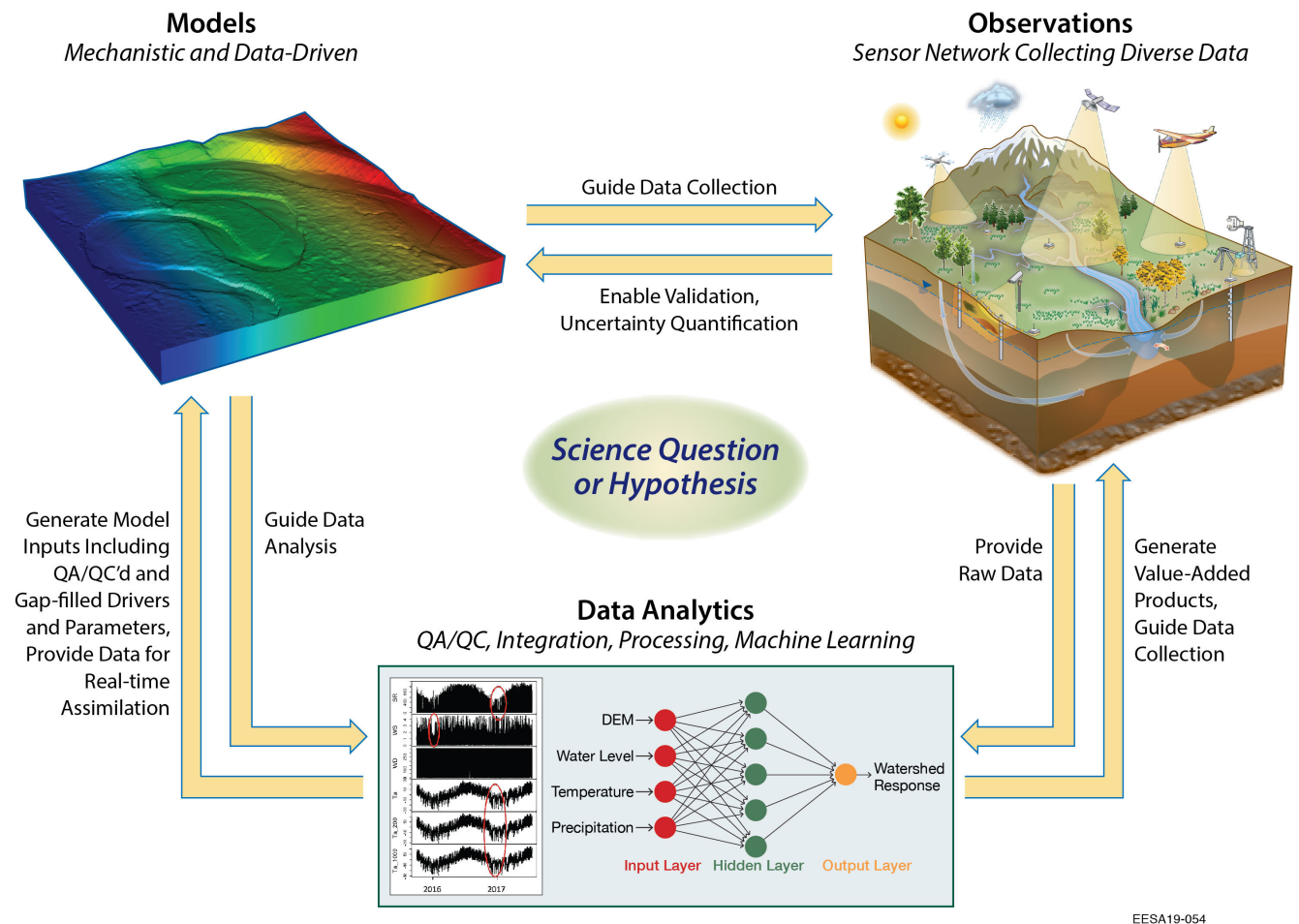
F. DATA PUBLICATION AND OWNERSHIP

Federally-funded research and monitoring programs are now required to publish their data in repositories to make their data findable, accessible, interoperable and reusable (FAIR) [57]. For most long-term sensor deployments, this entails periodic data archival with the appropriate metadata and usage policies that allow the public to understand and use the data.

A challenge faced by many sensor data collectors is the most effective means by which to distribute data publicly, given how repositories currently operate. Most data repositories, including ESS-DIVE, accept and publish "data packages", which are bundles of data files and metadata. The current recommendation from the scientific data community is that sensor data be published in snapshots, as periodic data collections with persistent identifiers in recognized repositories (http://wiki.esipfed.org/index.php/Sensor_Data_Archiving). This is the approach that we have chosen for publicly releasing data from the SFA's sensor network.

However this format for archiving and publishing data is not well-suited for streaming sensor data, and many challenges remain. For one, the dynamic nature of sensor data requires an ability to archive incremental changes (such as new data additions) and track different versions of the data, without necessarily assigning a new DOI. Data products from QA/QC or processing need to be linked to the original raw data streams. Data ownership tends to get complex, and may change for many reasons such as the addition of new technical staff or processing by a different person on the team. Thus assigning ownership on a constantly evolving data stream can be difficult, and currently requires the creation of many different data packages that credit the appropriate authors. However, this becomes confusing for data users, placing the burden on them to keep track of and cite all the different versions of the data accurately. The data package format also doesn't allow users to retrieve subsets of the data easily, and do not require users to adhere to data standards.

Better solutions for publishing streaming sensor data are urgently needed, with increasing amounts of data becoming publicly available due to requirements by funding agencies. An ideal approach for data providers would be the ability to publish a continuous data stream with persistent identifiers that allow identification of different versions of the data. From a user perspective, it'd be preferable to have the data organized such that it would be possible to retrieve



EESA19-054

FIGURE 6. Schematic illustration of co-design concepts that judiciously link models (both mechanistic and data-driven) and data across scales by involving assimilation, uncertainty quantification, and inverse strategies. Co-design strategies should be able to capture the highest uncertainty, capture outsized impacts of different processes, and predict larger scale system behavior. It should also be able to take advantage of increasingly-frequent autonomous datasets, seamlessly incorporate machine learning and other analytical tools for near real time exchange of information, and provide enhanced computational efficiency.

and visualize relevant subsets of the data using custom queries, and with the usage policies and citations clearly defined.

V. UTILIZING A CO-DESIGN APPROACH ENABLES BETTER MEASUREMENTS AND MODELING

Developing a predictive understanding of complex systems such as watersheds requires models to account for non-linear interactions between the intrinsic processes involved across various temporal and spatial scales, based on observations. However, the uncertainty associated with modeling and predicting watershed behavior has remained high despite the advent of advanced sensor networks and availability of high-resolution data for several reasons. First, even with sensors, data collection is expensive, and it is practically impossible to collect fine-resolution point measurements (e.g., soil moisture, groundwater level) everywhere at the watershed scale. Second, processes are linked across the range of scales and compartments (e.g., from the vegetative canopy to bedrock, from terrestrial hillslopes through

aquatic environments); therefore, intelligent networked sensing strategies are required for capturing the most valuable information needed to predict watershed function.

One solution to reduce uncertainties in model predictions, and simultaneously optimize sensor systems is to adopt the concept of co-design. Co-design refers to the intentional co-development of observation and modeling strategies across scales to enhance the two-way exchange of information. While developing observation strategies to inform models has been common in recent years, in practice there are often gaps in the datasets needed to inform the models, and large time delays between data collection and simulation. Few models currently take advantage of the increasingly frequent autonomous data streams available in near real-time. Strategies to systematically incorporate machine learning or other analytical tools to enhance the rapid exchange between observations and model predictions and to reduce computational efficiency are lacking. Models are not used routinely to optimize observation strategies such as the measurement type, frequency and placement.

Co-design strategies emphasize the *seamless and rapid exchange* of information and data analytics to address science questions or hypotheses about watershed behavior (Figure 6). Thus, co-design enables assimilation of integrated data streams into models to improve predictions and computational efficiency. It also enables models to be used to determine measurement approaches that reduce the uncertainty of system response to perturbation.

We are developing and testing formal co-design strategies within the SFA, to both improve our model predictions as well as our observational networks. As an example, we developed a three-dimensional flow and reactive transport model using PFLOTRAN [38], [58], which provided information about the system behavior before acquiring any field observations at two active meanders of the East River watershed. Subsequently, this model-derived understanding was used to frame science questions and collect geochemical samples along a two-dimensional transect for quantifying subsurface geochemical exports to the river, thereby demonstrating one aspect of the co-design strategy [13].

Adopting co-design strategies for development of field- and cyber-infrastructure can help address many of the challenges presented earlier. For example, sensor network configurations can be altered in response to model predictions, after accounting for practical logistical issues such as availability of power or site access. As model-informed autonomous data acquisition becomes more developed, cyber-infrastructure could play a role in translating model output to data acquisition control systems, and documenting the rationale for the adjustment in the data collection strategy in the database. Additionally, the use of co-design will trigger the collection of sufficient metadata about the sensor installation and maintenance to enable use of the data by models, including metadata that may not be traditionally recorded as part of sensor deployment [49]. This can include more detailed information on sensor placement and orientation, or other ancillary data needed to use the sensor outputs (e.g. intermediate manual measurements for calibration that can be used to validate gap filling procedures). Co-design will also offer an opportunity to coordinate and standardize data collection and reporting protocols, to minimize subsequent efforts needed to harmonize and convert diverse data into the formats required for use in the models. Finally, the incorporation of continual, real-time data QA/QC and processing in the co-design workflow will result in better products for modeling and analysis, as well as improved data collection. For example, we prepare some model input data by gap-filling at hourly (or higher) resolutions with all outliers removed, and transforming the data using detrending or normalization. The resulting products are not only used in models but are also used to improve and optimize data collection and sensor maintenance. Similarly, efforts are underway in the SFA to use models inform QA/QC and provide sanity checks (Section IV. C).

Thus, the use of co-design requires one to view the field data collection and subsequent treatment of the data as a

continuous, iterative process involving all aspects of the end-to-end field to data framework.

VI. SUMMARY

The Watershed Function SFA project generates and utilizes diverse datasets at its UCRB field site in East River, Colorado, which include hydrological, geochemical, geological, geophysical, ecological, microbiological, and remote sensing data. The project has developed extensive infrastructure and services to: (1) acquire data in the field, (2) manage the data with appropriate metadata; (3) enable the team and the broader community to discover, visualize and access relevant datasets; (4) perform Quality Assurance and Quality Control (QA/QC) of priority datasets; (5) enable efficient data integration, and product generation; and (6) publish data for use by the broader scientific community and public. Notably, this infrastructure includes tools for end-to-end data management and preservation, QA/QC, data discovery and download, and interactive visualizations. Such an extensive data pipeline is required to provide quality data products for scientific use and modeling.

However, the development and maintenance of such infrastructure and services is a complex endeavor, which requires addressing various problems that can arise with data acquisition, storage, curation, QA/QC, integration, and publication. Scientific research projects that deploy and utilize sensor networks in the natural environment need to anticipate and plan to overcome some of these challenges with an integrated view of the field-data system. Adopting a holistic, co-design approach will enable optimization of resources needed for solutions that include improving field logistics and sampling procedures, creating data processing and software tools, and ultimately methods to better integrate data with models. Efforts to simplify and scale such systems across observational platforms are needed to make it easier to use data from environmental sensor networks.

APPENDIX

A video of the SFA data portal is included as Supporting Information.

ACKNOWLEDGMENT

We also acknowledge the support of the Watershed SFA team, including project members who assisted with other aspects of field sampling and laboratory analysis, and those who were interviewed for the SFA data portal design. We thank Emily Robles for her assistance in preparing the video of the SFA data portal.

REFERENCES

- [1] K. A. Lohse, P. D. Brooks, J. C. McIntosh, T. Meixner, and T. E. Huxman, "Interactions between biogeochemistry and hydrologic systems," *Annu. Rev. Environ. Resour.*, vol. 34, no. 1, pp. 65–96, Oct. 2009.
- [2] J. J. McDonnell, M. Sivapalan, K. Vaché, S. Dunn, G. Grant, R. Haggerty, C. Hinz, R. Hooper, J. Kirchner, M. L. Roderick, J. Selker, and M. Weiler, "Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology," *Water Resour. Res.*, vol. 43, no. 7, Jul. 2007, Art. no. W07301.

- [3] M. Sivapalan, "Pattern, process and function: Elements of a unified theory of hydrology at the catchment scale," in *Encyclopedia of Hydrological Sciences*, M. G. Anderson and J. J. McDonnell, Eds. Hoboken, NJ, USA: Wiley, 2006. [Online]. Available: <https://onlinelibrary.wiley.com/doi/book/10.1002/0470848944>, doi: [10.1002/0470848944.hsa012](https://doi.org/10.1002/0470848944.hsa012).
- [4] S. M. Eberts, M. D. Woodside, M. N. Landers, and C. R. Wagner, "Monitoring the pulse of our Nation's rivers and streams—The U.S. geological survey streamgaging network," U.S. Geol. Surv. Fact Sheet, Reston, VA, USA, Tech. Rep. 2018-3081, p. 2, 2018, doi: [10.3133/fs20183081](https://doi.org/10.3133/fs20183081).
- [5] M. Rode, A. J. Wade, M. J. Cohen, R. T. Hensley, M. J. Bowes, J. W. Kirchner, G. B. Arhonditsis, P. Jordan, B. Kronvang, S. J. Halliday, R. A. Skeffington, J. C. Rozemeijer, A. H. Aubert, K. Rinke, and S. Jomaa, "Sensors in the stream: The high-frequency wave of the present," *Environ. Sci. Technol.*, vol. 50, no. 19, pp. 10297–10307, Oct. 2016.
- [6] P. J. Blaen, K. Khamis, C. Lloyd, C. Bradley, and S. Krause, "Real-time monitoring of nutrients and dissolved organic matter in rivers: Adaptive monitoring strategies, technological challenges and future directions," *Sci. Total Environ.*, vol. 569, pp. 647–660, Nov. 2016.
- [7] B. A. Pellerin, B. A. Stauffer, D. A. Young, D. J. Sullivan, S. B. Bricker, M. R. Walbridge, G. A. Clyde, Jr., and D. M. Shaw, "Emerging tools for continuous nutrient monitoring networks: Sensors advancing science and water resources protection," *J. Amer. Water Resour. Assoc.*, vol. 52, no. 4, pp. 993–1008, 2016.
- [8] J. T. Crawford, L. C. Loken, N. J. Casson, C. Smith, A. G. Stone, and L. A. Winslow, "High-speed limnology: Using advanced sensors to investigate spatial variability in biogeochemistry and hydrology," *Environ. Sci. Technol.*, vol. 49, no. 1, pp. 442–450, 2015.
- [9] J. W. Kirchner, X. Feng, C. Neal, and A. J. Robson, "The fine structure of water-quality dynamics: The (high-frequency) wave of the future," *Hydrol. Process.*, vol. 18, no. 7, pp. 1353–1359, 2004.
- [10] B. A. Pellerin, J. F. Saraceno, J. B. Shanley, S. D. Sebestyen, G. R. Aiken, W. M. Wollheim, and B. A. Bergamaschi, "Taking the pulse of snowmelt: *In situ* sensors reveal seasonal, event and diurnal patterns of nitrate and dissolved organic matter variability in an upland forest stream," *Biogeochemistry*, vol. 108, no. 1, pp. 183–198, 2012.
- [11] S. Sharma, R. Siddique, S. Reed, P. Ahnert, and A. Mejia, "Hydrological model diversity enhances streamflow forecast skill at short-to medium-range timescales," *Water Resour. Res.*, vol. 55, no. 2, pp. 1510–1530, 2019.
- [12] E. T. Coon, J. D. Moulton, and S. L. Painter, "Managing complexity in simulations of land surface and near-surface processes," *Environ. Model. Softw.*, vol. 78, pp. 134–149, Apr. 2016.
- [13] D. Dwivedi, C. I. Steefel, B. Arora, M. Newcomer, J. D. Moulton, B. Dafflon, B. Faybishenko, P. Fox, P. Nico, N. Spycher, R. Carroll, and K. H. Williams, "Geochemical exports to river from the intramander hyporheic zone under transient hydrologic conditions: East River Mountainous Watershed, Colorado," *Water Resour. Res.*, vol. 54, no. 10, pp. 8456–8477, Oct. 2018.
- [14] D. Dwivedi, B. Arora, C. I. Steefel, B. Dafflon, and R. Versteeg, "Hot spots and hot moments of nitrogen in a riparian corridor," *Water Resour. Res.*, vol. 54, no. 1, pp. 205–222, 2018.
- [15] A. Dagnew, D. Scavia, Y.-C. Wang, R. Muenich, C. Long, and M. Kalcic, "Modeling flow, nutrient, and sediment delivery from a large international watershed using a field-scale SWAT model," *J. Amer. Water Resour. Assoc.*, vol. 55, no. 5, pp. 1288–1305, 2019.
- [16] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and M. Prabhat, "Deep learning and process understanding for data-driven Earth system science," *Nature*, vol. 566, pp. 195–204, Feb. 2019.
- [17] D. Blodgett, J. Lucido, and J. Kreft, "Progress on water data integration and distribution: A summary of select U.S. Geological Survey data systems," *J. Hydroinform.*, vol. 18, no. 2, pp. 226–237, Jul. 2015.
- [18] J. S. Horsburgh, M. M. Morsy, A. M. Castronova, J. L. Goodall, T. Gan, H. Yi, M. J. Stealey, and D. G. Tarboton, "HydroShare: Sharing diverse environmental data types and models as social objects with application to the hydrology domain," *J. Amer. Water Resour. Assoc.*, vol. 52, no. 4, pp. 873–889, Aug. 2016.
- [19] A. S. Jones, J. S. Horsburgh, S. L. Reeder, M. Ramírez, and J. Caraballo, "A data management and publication workflow for a large-scale, heterogeneous sensor network," *Environ. Monit. Assessment*, vol. 187, no. 6, p. 348, 2015.
- [20] J. S. Horsburgh, J. Caraballo, M. Ramírez, A. K. Aufdenkampe, D. B. Arscott, and S. G. Damiano, "Low-cost, open-source, and low-power: But what to do with the data?" *Frontiers Earth Sci.*, vol. 7, no. 1, p. 67, 2019.
- [21] T. Slawewski, D. Young, B. Dean, B. Bergenroth, and K. Sparks, "Pilot implementation of the US EPA interoperable watershed network," *Open Geospatial Data, Softw. Standards*, vol. 2, no. 1, p. 13, 2017.
- [22] L. Patterson, M. Doyle, K. King, and D. Monsma, "Internet of water: Sharing and integrating water data for sustainability," Aspen Inst., Washington, DC, USA, Tech. Rep., 2017. [Online]. Available: https://assets.aspeninstitute.org/content/uploads/2017/05/Internet-of-Water-Report-May-2017.pdf?_ga=2.240874691.1535092405.1576651057-648714319.1576651057
- [23] G. Pastorello, D. Gunter, H. Chu, D. Christianson, C. Trotta, E. Canfora, B. Faybishenko, Y.-W. Cheah, N. Beekwilder, S. Chan, S. Dengel, T. Keenan, F. O'Brien, A. Elbashandy, C. Poindexter, M. Humphrey, D. Papale, and D. Agarwal, "Hunting data rogues at scale: Data quality control for observational data in research infrastructures," in *Proc. IEEE 13th Int. Conf. e-Sci. (e-Science)*, Oct. 2017, pp. 446–447.
- [24] L. Josset, M. Allaire, C. Hayek, J. Rising, C. Thomas, and U. Lall, "The U.S. water data gap—A survey of state-level water data platforms to inform the development of a national water portal," *Earth's Future*, vol. 7, no. 4, pp. 433–449, 2019.
- [25] A. S. Jones, Z. T. Aanderud, J. S. Horsburgh, D. P. Eiriksson, D. Dastrup, C. Cox, S. B. Jones, D. R. Bowling, J. Carlisle, G. T. Carling, and M. A. Baker, "Designing and implementing a network for sensing water quality and hydrology across mountain to urban transitions," *J. Amer. Water Resour. Assoc.*, vol. 53, no. 5, pp. 1095–1120, 2017.
- [26] J. S. Horsburgh, D. G. Tarboton, D. R. Maidment, and I. Zaslavsky, "Components of an environmental observatory information system," *Comput. Geosci.*, vol. 37, no. 2, pp. 207–218, 2011.
- [27] S. S. Hubbard, K. H. Williams, D. Agarwal, J. Banfield, H. Beller, N. Bouskill, E. Brodie, R. Carroll, B. Dafflon, D. Dwivedi, N. Falco, B. Faybishenko, R. Maxwell, P. Nico, C. Steefel, H. Steltzer, T. Tokunaga, P. A. Tran, H. Wainwright, and C. Varadharajan, "The East River, Colorado, Watershed: A mountainous community testbed for improving predictive understanding of multiscale hydrological–biogeochemical dynamics," *Vadose Zone J.*, vol. 17, no. 1, pp. 1–25, 2018.
- [28] H. Wainwright and K. Williams, "LiDAR collection in August 2015 over the East River Watershed, Colorado, USA," 2017. Distributed by Watershed Function SFA, doi: [10.21952/WTR/1412542](https://doi.org/10.21952/WTR/1412542).
- [29] T. H. Painter, D. F. Berisford, J. W. Boardman, K. J. Bormann, J. S. Deems, F. Gehrke, A. Hedrick, M. Joyce, R. Laidlaw, D. Marks, and C. Mattmann, "The Airborne Snow Observatory: Fusion of scanning lidar, imaging spectrometer, and physically-based modeling for mapping snow water equivalent and snow albedo," *Remote Sens. Environ.*, vol. 184, pp. 139–152, Oct. 2016.
- [30] Z. Fang, R. W. H. Carroll, R. Schumer, C. Harman, D. Wilusz, and K. H. Williams, "Streamflow partitioning and transit time distribution in snow-dominated basins as a function of climate," *J. Hydrol.*, vol. 570, pp. 726–738, Mar. 2019.
- [31] R. W. H. Carroll, L. A. Bearup, W. Brown, W. Dong, M. Bill, and K. H. Williams, "Factors controlling seasonal groundwater and solute flux from snow-dominated basins," *Hydrol. Processes*, vol. 32, no. 14, pp. 2187–2202, 2018.
- [32] W. K. Michener and M. B. Jones, "Ecoinformatics: Supporting ecology as a data-intensive science," *Trends Ecol. Evol.*, vol. 27, no. 2, pp. 85–93, 2012.
- [33] *Data Elements and Interchange Formats, Information Interchange, Representation of Dates and Times*, document ISO 8601:2004(E), 2004.
- [34] J. López-de-Lacalle, "tsoutliers R package for detection of outliers in time series," CRAN, R Package, 2016.
- [35] C. Chen and L.-M. Liu, "Joint estimation of model parameters and outlier effects in time series," *J. Amer. Stat. Assoc.*, vol. 88, no. 421, pp. 284–297, 1993.
- [36] A. Korobeynikov, A. Shlemov, K. Usevich, and N. Golyandina. (2015). *RSSA: A Collection of Methods for Singular Spectrum Analysis, R Package Version 0.13-1*. [Online]. Available: <http://CRAN.R-project.org/package=Rssa>
- [37] A. Tran, J. Rungee, B. Faybishenko, B. Dafflon, and S. Hubbard, "Assessment of spatiotemporal variability of evapotranspiration and its governing factors in a mountainous watershed," *Water*, vol. 11, no. 2, p. 243, 2019.
- [38] D. Dwivedi, I. C. Steefel, B. Arora, and G. Bisht, "Impact of intra-meander hyporheic flow on nitrogen cycling," *Procedia Earth Planet. Sci.*, vol. 17, pp. 404–407, Jan. 2017.

- [39] B. Arora, D. Dwivedi, S. S. Hubbard, I. C. Steefel, and K. H. Williams, "Identifying geochemical hot moments and their controls on a contaminated river floodplain system using wavelet and entropy approaches," *Environ. Model. Softw.*, vol. 85, pp. 27–41, Nov. 2016.
- [40] B. Arora, N. F. Spycher, C. I. Steefel, S. Molins, M. Bill, M. E. Conrad, W. Dong, B. Faybishenko, T. K. Tokunaga, J. Wan, and K. H. Williams, "Influence of hydrological, biogeochemical and temperature transients on subsurface carbon fluxes in a flood plain environment," *Biogeochemistry*, vol. 127, nos. 2–3, pp. 367–396, 2016.
- [41] L. M. Foster and R. M. Maxwell, "Sensitivity analysis of hydraulic conductivity and Manning's n parameters lead to new method to scale effective hydraulic conductivity across model resolutions," *Hydrol. Process.*, vol. 33, no. 3, pp. 332–349, 2019.
- [42] J. Tomkins and D. Lowe, *Timeseries Profile of Observations and Measurements*, OGC Standard 15-043r3, OGC Implementation Specification, 2016.
- [43] *Geographic Information—Observations and Measurements*, OGC document 10-004r1, ISO 19156:2011, V2.0, Open Geospatial Consortium, 2011.
- [44] C. Snively, G. Alvarez, V. Hendrix, S. Cholia, and S. Lasiewski, "Spin: A Docker-based platform for deploying science gateways at NERSC," in *Proc. Gateways*, 2018, doi: [10.6084/m9.figshare.7071770.v3](https://doi.org/10.6084/m9.figshare.7071770.v3).
- [45] C. Varadharajan, S. Cholia, C. Snively, V. Hendrix, C. Procopiou, D. Swantek, W. J. Riley, and D. A. Agarwal, "Launching an accessible archive of environmental data," *Eos*, vol. 100, 2019, doi: [10.1029/2019EO111263](https://doi.org/10.1029/2019EO111263).
- [46] R. Carroll and K. Williams, "Discharge data collected within the East River for the Lawrence Berkeley national laboratory watershed function science focus area (water years 2015–2018)," 2019. Distributed by Watershed Function SFA, doi: [10.21952/WTR/1495380](https://doi.org/10.21952/WTR/1495380).
- [47] M. Lehning, N. Dawes, M. Bavay, M. Parlange, S. Nath, and F. Zhao, "Instrumenting the earth: Next generation sensor networks in environmental science," in *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA, USA: Microsoft Corporation, 2009, pp. 45–51.
- [48] N. B. Melcher and J. F. Walker, "Evaluation of selected methods for determining streamflow during periods of ice effect," *U.S. Geol. Surv.*, Reston, VA, USA, Paper 2378, 1992. [Online]. Available: <https://pubs.usgs.gov/wsp/2378/report.pdf>
- [49] D. S. Christianson, C. Varadharajan, B. Christoffersen, M. Detto, B. Faybishenko, B. O. Gimenez, V. Hendrix, K. J. Jardine, R. Negron-Juarez, G. Z. Pastorello, T. L. Powell, M. Sandesh, J. M. Warren, B. T. Wolfe, J. Q. Chambers, L. M. Kueppers, N. G. McDowell, and D. A. Agarwal, "A metadata reporting framework (FRAMES) for synthesis of ecohydrological observations," *Ecol. Inform.*, vol. 42, pp. 148–158, Nov. 2017.
- [50] M. Botts, A. Robin, J. Greenwood, and D. Wesloh, "OGC sensorML: Model and XML encoding standard," Tech. Rep. OGC 12-000, 2014. [Online]. Available: <http://portal.opengeospatial.org/files/55939>
- [51] A. V. Dastjerdi and R. Buyya, "Fog computing: Helping the Internet of Things realize its potential," *Computer*, vol. 49, no. 8, pp. 112–116, Aug. 2016.
- [52] E. Baccarelli, P. G. V. Naranjo, M. Scarpiniti, M. Shojafar, and J. H. Abawajy, "Fog of everything: Energy-efficient networked computing architectures, research challenges, and a case study," *IEEE Access*, vol. 5, pp. 9882–9910, 2017.
- [53] Q. Yan and Y. Wang, "A kind of efficient data archiving method for historical sensor data," in *Proc. 3rd Int. Conf. Inf. Sci. Control Eng. (ICISCE)*, Jul. 2016, pp. 44–48.
- [54] J. S. Horsburgh, S. L. Reeder, A. S. Jones, and J. Meline, "Open source software for visualization and quality control of continuous hydrologic and water quality sensor data," *Environ. Model. Softw.*, vol. 70, pp. 32–44, Aug. 2015.
- [55] S. M. Ali, N. Gupta, G. K. Nayak, and R. K. Lenka, "Big data visualization: Tools and challenges," in *Proc. 2nd Int. Conf. Contemp. Comput. Inform. (IC3I)*, Dec. 2016, pp. 656–660.
- [56] L. Ramakrishnan, S. Poon, V. Hendrix, D. Gunter, G. Z. Pastorello, and D. Agarwal, "Experiences with user-centered design for the tigers workflow API," in *Proc. IEEE 10th Int. Conf. e-Sci. (e-Science)*, vol. 1, Oct. 2014, pp. 290–297.
- [57] M. D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," *Sci. Data*, vol. 3, Mar. 2016, Art. no. 160018.
- [58] G. E. Hammond, P. C. Lichtner, and R. T. Mills, "Evaluating the performance of parallel subsurface simulators: An illustrative example with PFLORAN," *Water Resour. Res.*, vol. 50, no. 1, pp. 208–228, 2014.



CHARULEKA VARADHARAJAN received the B.Tech. degree in civil engineering from IIT Madras, Chennai, India, in 2001, and the M.S. and Ph.D. degrees from the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA, in 2004 and 2009, respectively.

Since 2010, she has been with the Lawrence Berkeley National Laboratory, starting as a Post-doctoral Researcher, and currently a Research Scientist. She is also a Geochemist and a Data Scientist with the Earth and Environmental Sciences Area at Berkeley Lab, and also a Senior Fellow of the Berkeley Institute of Data Sciences. Her research focuses on the nexus of water, energy, and carbon for the purposes of sustainable water resource management. She is currently the Deputy Lead for the DOE's new environmental data archive, ESS-DIVE, and is developing software tools for data integration, analysis and visualization related to large DOE projects that study watersheds and tropical forests.



DEBORAH A. AGARWAL received the B.S. degree in mechanical engineering from Purdue University and the M.S. and Ph.D. degrees in computer engineering from the University of California at Santa Barbara, Santa Barbara. She is currently a Senior Scientist and the Data Science and Technology Department Head with the Lawrence Berkeley National Laboratory (LBNL). She is also an International Chair at Inria, France, and a Senior Fellow with the Berkeley Institute for

Data Science, University of California at Berkeley, Berkeley. Her current research focuses on developing computational tools to enable scientists to more effectively organize and use their data to address challenges. She has worked on projects involving watershed understanding, tropical forests, soil carbon, carbon capture, cosmology, particle accelerators, and satellite data.



WENDY BROWN received the B.S. degree in forest resource management (minoring in soils emphasis and geology) from the University of Montana, in 1984. Since 1983, she has been a Field Biologist, focused on aquatic entomology and a variety of other mountain ecological studies in the Colorado Rocky Mountains. In 1990, she formed a small business Bugs Unlimited, LLC, for contract work in field research. Since 2014, she has been working on the East River Scientific Focus Area

with the Lawrence Berkeley National Laboratory near the Rocky Mountain Biological Laboratory, Gothic, CO, USA.



MADISON BURRUS received the B.S. degree in environmental science (minoring in environmental engineering) from Stony Brook University, in 2017. Her studies in chemistry laboratories and nutrient-cycling lectures fostered an interest in water quality challenges. Since her graduation, she has been working with the Watershed Function Scientific Focus Area (Watershed SFA) with the Lawrence Berkeley National Laboratory on analyzing stream quality in the East River, CO, USA. She is currently working with the Watershed SFA Data Management Team on quality assurance and quality control processing of its diverse datasets.



basins, and mountainous watersheds. She is currently an Associate Research Professor and has been collaborating with the Watershed Function Science Focus Area, since 2014.



Energy and Resources Group, University of California at Berkeley, Berkeley.



cal, point-scale, and remote sensing measurements.



has extensive experience with machine learning techniques, and data assimilation approaches.



Area de Conservation, Guanacaste, and Costa Rica.

ROSEMARY W. H. CARROLL received the B.S. degree in physics from the Bates College, Maine, and the M.S. and Ph.D. degrees in hydrology from the University of Nevada, Reno, in 1999 and 2010, respectively, with a research emphasis on mercury transport and fate in river-reservoir systems. In 2000, she began working at the Desert Research Institute, Reno, NV, USA, on groundwater and surface water interactions across a range of environmental systems including wetlands, agricultural

DANIELLE S. CHRISTIANSON is currently a Computer Science Engineer with the Lawrence Berkeley National Laboratory, Data Science and Technology Department. As a member of the Usable Software Systems Group, she combines her background in mechanical engineering, ecological research of microclimate, and study of digital representations of natural systems to build data management tools for earth science research. She is an alumna of Texas A&M University and the

BAPTISTE DAFFLON received the Ph.D. degree in geophysics from the University of Lausanne, Switzerland. He is currently a Research Scientist with the Lawrence Berkeley National Laboratory, Geophysics Department. His current research interests include the understanding and quantification of subsurface hydrothermal—geochemical processes and their covariability (in space and time) with surface properties using advanced monitoring strategies including a variety of geophysical,

DIPANKAR DWIVEDI received the Ph.D. degree from Texas A&M University. He is currently a Research Scientist with the Lawrence Berkeley National Laboratory, Geochemistry Department. He is also working on advancing process understanding of coupled carbon and nitrogen cycling under transient hydrological conditions across scales. His research focuses on reactive transport modeling, soil carbon dynamics, redox processes, and organo-mineral interactions. He also

BRIAN J. ENQUIST received the B.A., M.S., and Ph.D. degrees in biology. He is currently a Professor of ecology and evolutionary biology with The University of Arizona. He is a broadly trained plant ecologist. His lab investigates how functional constraints at the level of the individual (anatomical and physiological) influence larger scale ecological and evolutionary patterns. His research also includes the monitoring of long-term dynamics of growth and change within a tropical forest in the



using deterministic and stochastic models.



BORIS FAYBISHENKO received the B.S. degree in hydrogeology and engineering geology, and the Ph.D. and D.Sc. degrees in hydrogeology. He is currently a Staff Scientist of the Lawrence Berkeley National Laboratory, Energy Geosciences Division. His research is related to environmental sciences, remediation, long-term monitoring, optimization of coupled energy-water systems under present-day and future climatic conditions, and analytical and numerical modeling

AMANDA HENDERSON received the B.A. degree in biology from the Kenyon College, OH, USA, and the M.S. degree in ecology and evolutionary biology from The University of Arizona. She is currently a Field Manager for the watershed function scientific focus area with the Rocky Mountain Biological Laboratory, Crested Butte, CO, USA.



MATTHEW HENDERSON received the B.S. degree in computer science from California Polytechnic University, Pomona. He has more than a decade of experience working on scientific computing projects and data. He was a Software Engineer with NASA JPL, where he worked on a number of Earth Science projects. He is currently a Computer Systems Engineer with the Lawrence Berkeley National Laboratory.



Engineer. She has more than 15 years of industry development experience. Her current work focuses on systems for preserving, expanding access to, and improving usability of critical data generated through DOE-sponsored research of terrestrial and subsurface ecosystems (ESS-DIVE, NGEE Tropics, and Watershed Function SFA).



using geophysical methods to explore how hydrological, geochemical, and biological processes interact to govern larger scale system behavior. She has been elected as a member of the American Academy of Arts and Sciences, and a Fellow of the American Geophysical Union (AGU) and of the Geological Society of America (GSA). She was a recipient of the Society of Exploration Geophysicists (SEG) Frank Frischknecht Award and the SEG Hal Mooney Award. She was a GSA Birdsall Dreiss Distinguished Lecturer. She is the PI of the Watershed Function Project.



include corporate sustainability and implementing environmental policies.



ALEXANDER NEWMAN received the B.S. degree in earth systems sciences from the University of California at Merced, Merced, and the M.S. degree in biogeochemistry from the Memorial University of Newfoundland. He is currently a Field Manager with the Rocky Mountain Biological Laboratory, where he is working with the Watershed Function Science Focus Area.



BENJAMIN POTTER received the B.S. degree in environmental sciences from Colorado Mesa University. He is currently an Experienced Scientist with a demonstrated history of working in the environmental and remote monitoring industry. His current work focuses on managing remote environmental monitoring stations at DOE Legacy Management sites. He is skilled in environmental process control and programming.



and Cryosphere in a Changing Climate.

HEIDI STELTZER received the B.S. degree in biology from Duke University and the Ph.D. degree in ecosystem science from the University of Colorado, Boulder. Since 2009, she has been a Professor with the Fort Lewis College. She has spent over two decades conducting field studies in CO, AK, and GL, USA, to understand the impacts of climate change. She is currently an Environmental Scientist and an Explorer. She is a Lead Author for the upcoming IPCC special report on the Ocean



ROELOF VERSTEEG received the Ph.D. degree in geophysics from Paris 7, in 1991. After working in academia and the DOE Idaho National Laboratory, he founded Subsurface Insights, in 2012. Subsurface Insights is a small business with the mission to develop software and hardware tools which provide actionable insights in ecosystem processes.

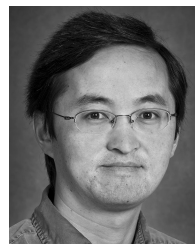


of biogeophysics, with a particular emphasis on the development of minimally invasive techniques for characterizing the subsurface activity of microbial communities and the products of their metabolism. His current research interests include seasonal variations in redox-sensitive elements, issues impacting the long-term mobility of contaminant metals in the subsurface, and the coupling of hydrological and biogeochemical processes that impact both. He serves as an Associate Editor for *JGR-Biogeosciences* and *Frontiers in Water*.

KENNETH H. WILLIAMS received the B.A. degree in geology, the M.S. degree in geophysics, and the Ph.D. degree in environmental science, policy and management from the University of California at Berkeley, Berkeley. He is currently a Staff Geological Scientist and a Program Lead of environmental remediation and water resources with the Lawrence Berkeley National Laboratory. He has authored or coauthored more than 140 peer-reviewed publications. He is a pioneer in the field



CHELSEA WILMER received the B.S. degree in environmental biology from the Fort Lewis College. She has more than three years of research experience in CO, USA. She is currently a Field Research Manager with the Rocky Mountain Biological Laboratory, Crested Butte, CO, USA. She currently focuses on climate science and snowmelt in the East River watershed.



YUXIN WU is currently a Staff Scientist with the Lawrence Berkeley National Lab, Earth and Environmental Sciences Area. He is trained in near surface and environmental geophysics, and has expertise in hydro-/bio- geophysics, and joint applications of geophysical, hydrological and biogeochemical methods for multi-disciplinary research. He has published more the 40 peer-reviewed articles. His research interests include geophysical monitoring of subsurface hydro-biogeochemical processes; characterization and monitoring of fracture flow related to energy production and waste storage; enhanced geophysical imaging approaches; and the development of novel imaging approaches for ecosystem interactions. He is a member of AGU and SEG.

• • •