A could-based, open-souce tool and database for stream solute tracer studies

Submitted for consideration to be published in Environmental Modeling and Software

Tyler B. Balson[1] and Adam S. Ward[2,3]

1 - Luddy School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA
2 - O'Neill School of Public and Environmental Affairs, Indiana University, Bloomington, IN, USA
3 - Biological & Ecological Engineering, Oregon State University, Corvallis, OR, USA

Keywords: Open Hydrology, Reproducibility, Hydroshare, Jupyter, Data frameworks

Abstract

The growth in cloud-based computing platforms and open data repositories is enabling new workflows, standardization of analyses, and synthesis projects across many domains, including hydrologic science. Here, we established an open-source tool and database for Stream Solute Tracers, leveraging community resources including both HydroShare (for data storage, curation, and analysis) and HydroLearn (linking our tool to disciplinary educational resources). This framework provides users cloud access to a standardized toolbox for analysis of experimental data producing an standardized analysis for individual experiments. Individual experimental results are presented in the context of all studies in the database, providing context for interpretation of results. Both the input and output are aggregated within a PostgreSQL database providing a queryable, single database for stream tracer experiments. Standardization of the analysis of these experimental results will ensure reproducibility for future research, and this is built on a discipline-specific platform to reduce barriers to entry for users.

1.0 Introduction

The data deluge (Bell, Hey, and Szalay 2009) has driven a revolution in our approach to science, with data-driven analysis now complementing the traditional scientific method across a host of disciplines (e.g.,(A. Ward et al. 2022)). While a range of flexible, powerful data science techniques have emerged to augment traditional modes, the successful application of data science to disciplinary approaches ultimately relies upon high-quality, organized data sets for analysis. To unlock the potential gains from these approaches, datasets must be sufficiently large and with complete metadata such that a model can derive meaningful underlying relationships. In many areas of research, data synthesis is an insurmountable step because underlying data have not been organized nor reported with sufficient metadata to enable synthesis, ultimately hindering additional science from taking place. Efficient and effective data documentation, synthesis, and storage has long plagued scientists and researchers. Over the last 50 years we have transitioned from bound notebooks and paper records stored in individual offices to digital, open, cloud hosted data. This has created organization by individual projects and data types (e.g., StreamPULSE), research neworks (e.g., LTERNet, Ameriflux), non-profits (e.g., GLEON, CUAHSI), and agency-specific databases (e.g., ESS-DIVE). As the infrastructure for data collection has grown, researchers have been able to break away from a ‘one data set for one location’ paradigm and can now analyze data that span several locations, measurements, and scales (“NEON” 2022, “GLEON” 2022, “LTER” 2016, “AmeriFlux” 2022). With data collection networks established, researchers’ focus shifted to the next requisite steps: data discovery and interoperability (e.g., ESS-DIVE, and Hydroshare). The combination of established data collection networks and services for storage and discovery of these data enables synthesis efforts and scientific discovery that was heretofore impossible.

Data collection networks have expanded to serve increasingly narrow domains, enabled by the rapid rise in computing power and decrease in storage costs. For example, the StreamPulse project aimed to take the ‘pulse’ of a river or steam by collecting and analyzing in-stream dissolved oxygen data, providing a standardized approach to estimate whole-stream respiration and simultaneously building a database of more than 500 million data points spanning more than 700 sites (“StreamPULSE” 2022). This approach provides clear benefits to the community, with standardization of data collection and analysis enabling intercomparisons. Moreover, the aggregated data in StreamPULSE enables synthesis, which is already generating insights not possible without the underlying database and analytical tools (Ulseth et al. 2019; Savoy et al. 2019; Bernhardt et al. 2018; Appling et al. 2018). Despite being a relatively niche application (whole stream respiration), the use of shared protocols for data collection and analysis has clear scientific benefits. With widespread access to computing power, similar applications could be implemented to standardize, organize, and enable systems of specific data types or experimental approaches. Indeed, this approach is being actively employed in several synthesis efforts including the 100 Plastic Rivers project (“100 Plastic Rivers - a Global Investigation” 2022), the 1000 intermittent rivers project (Thibault Datry et al. 2016; T. Datry et al. 2018; Shumilova et al. 2019; Schiller et al. 2019), the Cellulose Decomposition Experimental (CELLDEX; (Tiegs et al. 2019), WHONDRS (Stegen and Goldman 2018)), and countless similar initiatives. One similar area that is lacking a standardized approach for data storage and analysis is stream solute tracer studies.

Stream solute tracers have been a popular tool to study transport and transformation of solutes through streams and their connected river corridors for decades (e.g., (Thackston Edward L. and Schnelle Karl B. 1970; Fischer et al. 1979; Stream Solute Workshop 1990; Bencala and Walters 1983). While a host of approaches exist to collect and analyze these data, the community presently lacks any standardization, making intercomparison challenging. From an initial focus on travel times (e.g., “time of passage” studies) and advection-dispersion modeling (e.g., (Fischer et al. 1979)), the toolkit for interpretation of these data has grown substantially. Current, popular techniques include advection-dispersion modeling [ibid], interpretation of holdback (Danckwerts 1953), temporal moments (Harvey and Gorelick 1995; Gupta and Cvetkovic 2000), channel water balance (Payn et al. 2009) , separation of mass involved in transient storage (Wlostowski et al. 2017) , and StorAge Selection frameworks (Ciaran J. Harman 2015; C. J. Harman, Ward, and Ball 2016; A. S. Ward, Kurz, et al. 2019). Moreover, several models exist to interpret findings using an inverse modeling approach, including the popular Transient Storage Model (Runkel, McKnight, and Andrews 1998; Bencala and Walters 1983), STAMMT-L [Haggerty and Reeves, 2002], and several additional model formulations e.g., (Wörman et al. 2002; Boano et al. 2007). Here, we develop and demonstrate a standardized workflow for data analysis, integrating the approach with a grassroots training effort that provides support to researchers (“Stream Solute Tracers” n.d.). Providing this framework for the analysis will ensure reproducibility, which is currently a well documented challenge in scientific research (Fidler and Wilcox 2021). The establishment and function of a database for stream solute tracers that automates and standardizes analyses for individual researchers, aggregates a sufficiently large and complete data set to enable future synthesis, and we demonstrate one application of this approach.

To support an open-source, accessible tool and maximize accessibility, we design our tool on a cloud-based platform. Cloud platforms have become a staple across the scientific community with many domain specific fields creating designated platforms. In the hydrologic sciences, HydroShare ("Find, Analyze and Share Water Data" 2022) provides a platform for cloud based data-analysis and storage. Cloud based tools are popular because they standardize the data management required by funding agencies, enable repeatable analysis of data sets, and provide universal access for discovery and analysis of data sets. In short, such tools underpin FAIR data and software, a cornerstone of modern scientific best practices (Stall et al. 2019; Enabling FAIR Data Community 2018). Furthermore, hosting solute tracer analysis on HydroShare will also enable the aggregation and storage of experimental data across suites of sites and the accompanied analysis. Making these data openly available will promote additional science and collaboration. Just as the StreamPulse network is informing our understanding of whole-stream respiration, so too will a mature stream solute tracer data set support scientific inquiry and synthesis spanning a variety of temporal and spatial stream scales.

Our objective in this study is to document the development and deployment of an open-source, cloud-based framework to standardize the processing and interpretation of stream solute tracer data. Our deployment includes standardization of metadata and building of an open, accessible database to enable aggregation of data sets for future synthesis. This tool will promote reproducible and sharable science openly available. With reproducibility and openly available data we can develop and provide training that is transferable for use across a variety of courses. We are motivated by a recent HydroLearn module detailing experimental and analysis methods, integrating our database and analysis tool with a scaffolding of educational material. Standardizing the workflow for data storage and analysis to follow state-of-the-science instruction will enable researchers to confidently and consistently execute and interpret solute tracer studies, and build toward future synthesis.

2.0 Methods

2.1 Field Methods & Time Series Analysis

This tool is built upon the established reach-scale solute tracer studies presented in the Stream Solute Tracers HydroLearn Module ("Stream Solute Tracers" 2020). These methods follow those implemented in several studies (e.g., (A. S. Ward, Wondzell, et al. 2019; A. S. Ward, Zarnetske, et al. 2019; Payn et al. 2009; A. S. Ward et al. 2013)). Briefly, the approach uses a series of two conservative solute tracer tracer releases. The first release is at the downstream end of the study reach and used for dilution gauging. The second release is at the upstream end of the study reach, used for dilution gauging at the upstream end of the study reach and also observed at the downstream location to assess short- and long-term storage within the reach. The key data generated from this approach are three time-series (s) of background-corrected solute tracer concentration (g/m^3), which are the key input data to subsequent analysis. Additional input data for each experiment, which support analyses and assembly of a comprehensive database, including time of year, stream order, and geospatial location. Input file examples and templates are provided for users to easily format their own data sets in the HydroShare entry associated with this tool(“Data_example_resource” 2021).

Analysis of solute tracer data follows exactly the methods outlined in the associated HydroShare course. These are standard analyses in the field, and are well-documented in the studies cited below. Briefly, time series are analyzed to calculate mass recovery and channel water balance (Payn et al. 2009), holdback (Danckwerts 1953), temporal moments (Schmid 1995) , StoraAge Selection (C. J. Harman, Ward, and Ball 2016), and the partitioning of recovered mass between transient storage and advection-dispersion (Wlostowski et al. 2017; Runkel 2002). Implementation of each analysis is detailed in the Jupyter Notebook documented here (“SoluteTracerTool” 2021), and explained in detail in the associated HydroLearn course(“Stream Solute Tracers” 2020).

2.2 Implementation on HydroShare

HydroShare provides the conduit for researchers that perform stream tracer solute experiments to access tools for their analysis and storage of their results. Instead of needing their own code, we have provided standardized script for the user to perform their own analyses. Our code is deployed using an interactive, python-based tool that is accessible through HydroShare and usable as a jupyter notebook. The notebook format allows line by line execution of code on the users’ data, providing two key benefits. First, the line by line execution enables a workflow for researchers to check each step while analyzing their own data. Second, this allows for integration of rich text and links to the associated HydroLearn modules, providing a scaffolding for users that documents what is happening in each section of the code.

This framework was built with the assumption that the experiment was a slug injection and the user has the values associated with this procedure (e.g., tracer masses), consistent with several recent studies (A. S. Ward, Wondzell, et al. 2019; A. S. Ward, Zarnetske, et al. 2019; Payn et al. 2009) and following the protocols prescribed in the associated HydroLearn course. This tool requires three time-series of concentrations associated with the upstream and downstream injection sites. The users are required to have these three files post-processed and formatted in two column vectors per file representing the time in seconds from injection in the first column and concentration of tracer above background in (g/m^3) in the second column. These three files are uploaded to HydroShare by users with an accessible resource ID, which is supplied to the tool. The resource ID can be used within the existing *hsclient* package on HydroShare, enabling rapid information passing using RESTful services to pass the user data to the tool. The results of the analysis are passed back to the location of that resource ID and stored in a collection as part of an ongoing compilation of results.

Using *hsclient* and the users resource ID, the users files are sent to the container running the jupyter notebook. The files are imported as *.zip files and with the module *shutil* are unzipped within the container. The data from the files is then extracted and formatted using pandas. The tool then plots and displays the time-series of the users data. The tool then proceeds to analyze the data in a series of steps, which include:

1. Estimating advective timescale and the modal velocity of the solute tracer and by estimating longitudinal dispersion using two methods
2. Discharges at the up- and downstream end of the study reach are estimated by dilution gauging and using trapezoidal integration via numpy. Mass recoveries are also calculated for each tracer release, as well as net and gross changes in discharge along the study reach.
3. Short term storage calculations are computed using the scipy.integrate method with the second downstream injection concentration and time data (datafile_name). From this we can determine the maximum timescale of the observation (e.g., t99 (Mason, McGlynn, and Poole 2012)). This eliminates some subjectivity when determining the time at which the tracer is no longer present, since trailing can be challenging to observe with field probes (Schmadel 2011). The temporal moments, coefficient of variation, skewness, holdback function and apparent dispertivity and dispersion are computed using the timescale determined from the short term storage calculations. Using the scipy class interpolate.interp1d linear interpolations for 5,10,25,50, 75, 90 and 95 time percentiles are computed in addition to t99.
4. Given a stream reach (known volume) the ranked storage selection function (rSAS) provides a way to account for the age of infinitesimal segments of water within the volume provided the discharges are known. The rSAS theory can provide a way to characterize the hydrochemistry of varying types of flow systems relevant to hydrology and biogeochemistry (Rinaldo et al. 2015).

With computations complete, the results of the analysis are stored in a database that will aggregate and make accessible solute tracer data. We constructed a relational database for a body of previously published and novel observations using a popular, open source database PostgreSQL (“PostgreSQL” 2022) selected for its reliability and robustness. Using a database will provide features not possible though traditional file storage like querying data based on the metadata for each experiment. Additionally the build out of this database with new data will be more efficient than compiling and storing excess csv or similar file types.

2.3 Enabling contextual interpretation of results

2.3.1 Displaying results in context

Upon completion of the analysis and storage of the results, we next display the results in the context of a larger database of stream solute tracer studies that is aggregated from users of the tool, (FIG 5). The database itself is housed within a collection on HydroShare and openly available (Balson 2022). After the user supplies their resource ID, their files are automatically transferred to a community-accessible collection for incorporation to a stream-tracer database hosted on HydroShare. This transfer and updating of a database enables a contextual component to this analysis previously not possible. The user now has the capability to see their experimental results in the context of all prior solute tracer studies that have been analyzed. For example, code is provided to display the experiment being analyzed in context of the database using boxplots and cumulative distributions. This enables users to assess - for example - if their site has low, average, or high turnover or skewness, which may frame their interpretation of results. Additionally this provides the user additional data checks. One example of a data check is that users can determine if their experiment appears an outlier relative to existing data. This may cause users to critically reflect on whether this is expected or indicate that data processing leading up to HydroShare may warrant further scrutiny, providing a 'soft' QA/QC to the researcher.

2.3.2 Building a database for future analysis

In addition to providing context for individual users, the database itself will grow and persist to enable future synthesis efforts. Future use cases might include analysis of studies from a particular region of interest, based on stream order, as a function of discharge, or any host of other combinations of factors. As data are added to this database, future work in filtering and classifying results is supported by the standard database-agnostic querying supported by PostgreSQL. In addition to simple searches of the database, the tool might also enable more advanced analyses. For example supervised learning techniques for classification (e.g., K nearest neighbor) could be used in order to evaluate if streams can be classified from their stream-tracer results and available metadata. Likewise, unsupervised learning techniques could be deployed on the data to see how the data can be separated based only on the characteristics within the dataset. These provide useful information whether they can or cannot be effectively classified because the underlying why can still be evaluated further, and perhaps to assess when and where new studies are needed to resolve trends discovered via such analyses.

3.0 Case Study: Stream solute tracers at the H.J. Andrews Experimental Forest

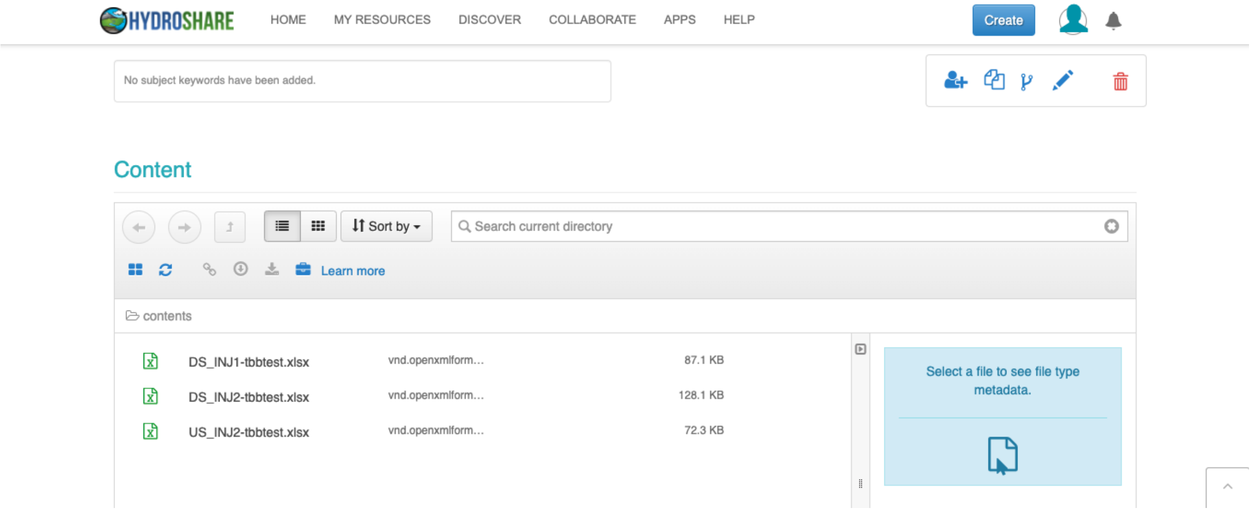
3.1 Experimental Information

We populated the database with 120 solute tracer experiments presented in (A. S. Ward, Zarnetske, et al. 2019), each of which was collected in accordance with the protocols outlined in ("Stream Solute Tracers" 2020). Tracer studies span 1st to 5th order reaches in the H.J. Andrews Experimental Forest (Western Cascades, Oregon, USA). All injections used NaCl as a conservative tracer, recording specific conductance and using a site-specific calibration curve to convert the time series into background-corrected concentrations. We used this previously published data set to confirm that our implementation here was consistent with individual analyses of the timeseries, and because it provides a large set of data for analysis.

In addition to the pre-populated database, we selected one additional solute tracer study to act as a demonstration of how a user can analyze new data and compare it to the pre-existing database. For this, we used the first injection documented in (A. S. Ward et al. 2013), conducted in the lower 50-m of WS01 at the HJ Andrews Experimental Forest. The release and documentation followed the same protocols as the synoptic data set and our standard methodology ("Stream Solute Tracers" 2020). While previously published, this experiment has never been compared to those in our database, so comparing this result to the basin-wide synoptic study is a useful exercise in assessing the repeatability of solute tracer studies.

3.2 Results

3.2.1 Analysis of the individual BTC

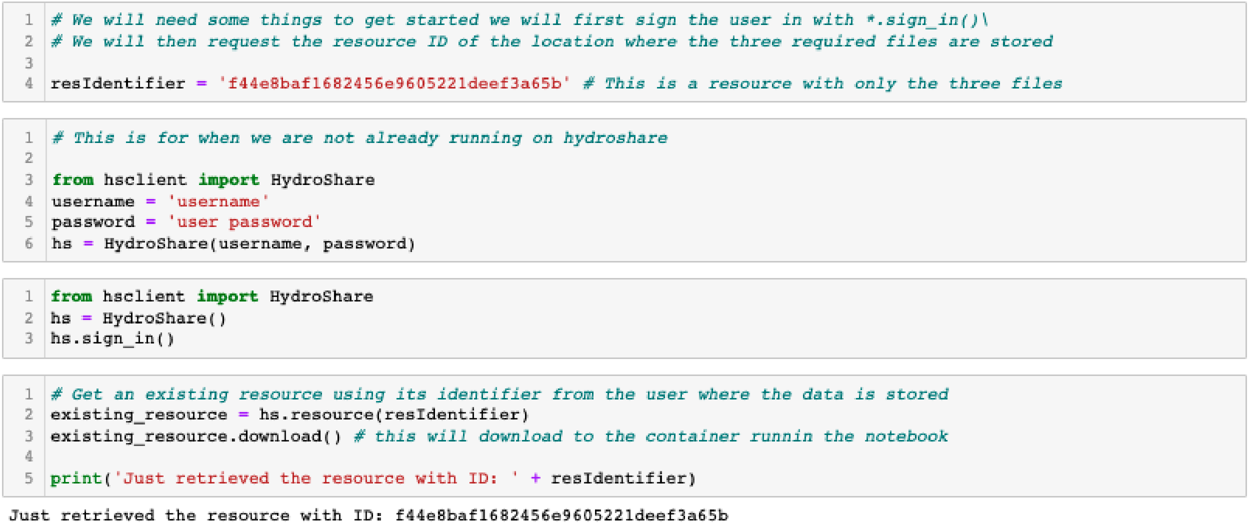


No subject keywords have been added.

DS_INJ1-tbbtest.xlsx	vnd.openxmlform...	87.1 KB
DS_INJ2-tbbtest.xlsx	vnd.openxmlform...	128.1 KB
US_INJ2-tbbtest.xlsx	vnd.openxmlform...	72.3 KB

Fig 1: File structure in Hydroshare. Users upload three background-corrected concentration timeseries as an input to the workflow documented here.

The three required input files are prepared in a format required (column 1 as time elapsed in seconds, column 2 as tracer concentration in g/m^3 above background) and uploaded to the users HydroShare space as a resource (see, Fig. 1), a process that will create a unique resource ID. The default naming scheme of the files are DS_INJ1, US_INJ2 and DS_INJ2, with both .csv and .xlsx formats accepted and templates provided. The resource ID associated with the three files is entered on line 4 of the ‘User Input’ section along with the users HydroShare credentials (see Fig. 2). This section creates a connection between the user's HydroShare space (i.e., the location of the uploaded files) to the notebook using the *hsclient* python package. If the user has specific naming conventions tied to their files that need to be retained, the user does have the ability to change the names of their files within the ‘File Names’ cell on lines 13,16 and 19 of the notebook (see Fig. 3).



```
# We will need some things to get started we will first sign the user in with *.sign_in()\
# We will then request the resource ID of the location where the three required files are stored

resIdentifier = 'f44e8baf1682456e9605221deef3a65b' # This is a resource with only the three files
```

```
# This is for when we are not already running on hydroshare

from hsclient import HydroShare
username = 'username'
password = 'user password'
hs = HydroShare(username, password)
```

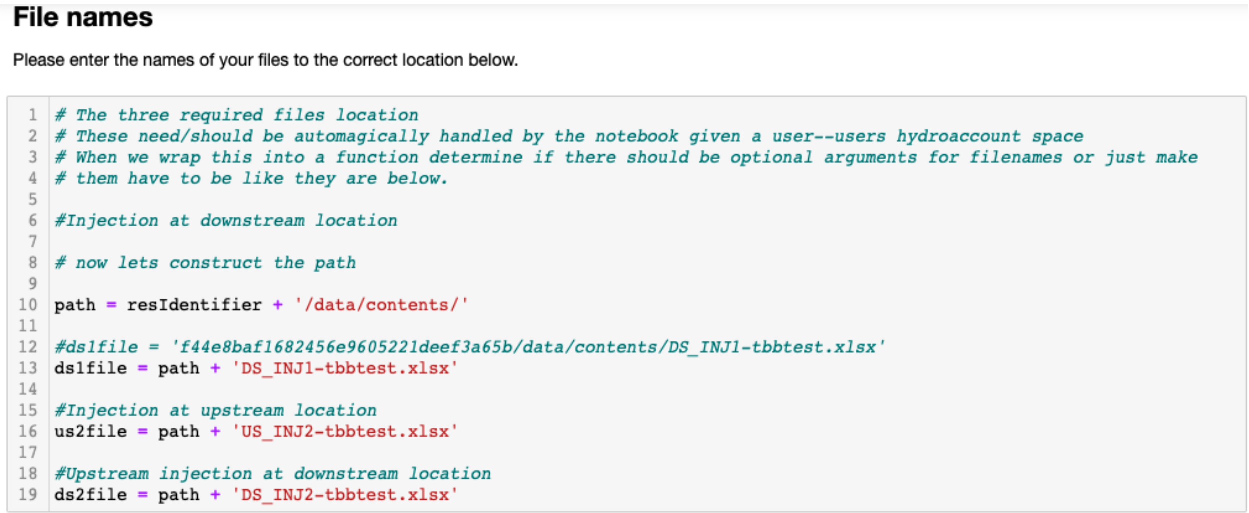
```
from hsclient import HydroShare
hs = HydroShare()
hs.sign_in()
```

```
# Get an existing resource using its identifier from the user where the data is stored
existing_resource = hs.resource(resIdentifier)
existing_resource.download() # this will download to the container runnin the notebook

print('Just retrieved the resource with ID: ' + resIdentifier)
```

```
Just retrieved the resource with ID: f44e8baf1682456e9605221deef3a65b
```

Fig 2: User credentials of resource ID.



File names

Please enter the names of your files to the correct location below.

```
# The three required files location
# These need/should be automagically handled by the notebook given a user--users hydroaccount space
# When we wrap this into a function determine if there should be optional arguments for filenames or just make
# them have to be like they are below.

#Injection at downstream location

# now lets construct the path

path = resIdentifier + '/data/contents/'

#ds1file = 'f44e8baf1682456e9605221deef3a65b/data/contents/DS_INJ1-tbbtest.xlsx'
ds1file = path + 'DS_INJ1-tbbtest.xlsx'

#Injection at upstream location
us2file = path + 'US_INJ2-tbbtest.xlsx'

#Upstream injection at downstream location
ds2file = path + 'DS_INJ2-tbbtest.xlsx'
```

Fig 3: File naming conventions can be customized, as in the above with '-tbbtest' appended to the default filenames.

After the user has entered their resource ID, credentials and ensured the proper naming conventions were used the breakthrough curves from the experimental data are plotted. This provides a first opportunity for the user to confirm their data quality and visually inspect the inputs being analyzed. These plots are then automatically saved to the user's HydroShare resource alongside the uploaded files. Once the user is satisfied their raw data appears as expected the remainder of the analysis can proceed. A results summary file is automatically generated as a .csv file and stored in the same HydroShare resource as the three files and the breakthrough curves (see Tab. 1, Fig. 4).

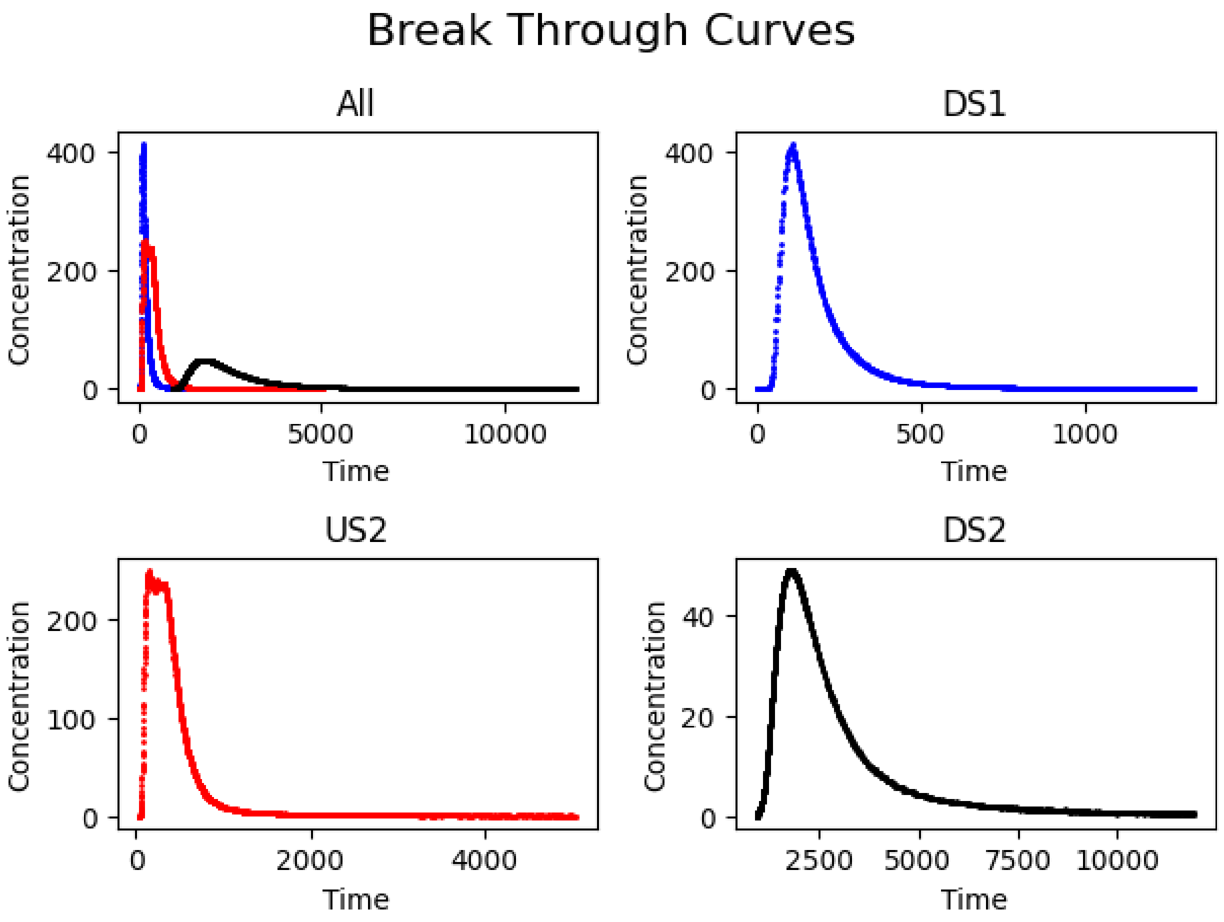


Fig 4: Break through curves

The length L in meters is	53
Start date and time	2016-08-01 14:55:00
End date and time	2016-08-01 15:13:00
mass1 in grams	622
mass2 in grams	1221.4
Q_us (m3/s)	0.011
Q_ds (m3/s)	0.011
dQ (m3/s)	0.000
Qlossmin (m3/s)	-0.002
Qgainmin (m3/s)	0.002
Qlossmax (m3/s)	-0.003
Qgainmax (m3/s)	0.002

mloss (g)	-235.769
mrec (g)	985.631
V (m/s)	0.033
M1	256591237.099
M1 Normalized (hr)	0.773
M2	6.07E+21
M2 Normalized (hr^2)	0.172
M3	-1.56E+30
M3 Normalized (hr^3)	0.137
skewness (unitless)	-0.003
skewness norm (unitless)	1.906
CV (m)	0.537
appdispersivity (m2/hr)	4.571
appdispersion	313.241
Holdback	0.654
t05 (hr)	1366.139
t10 (hr)	1495.491
t25 (hr)	1791.489
t50 (hr)	2305.189
t75 (hr)	1366.139
t90 (hr)	4737.588
t95 (hr)	6057.90
t05 norm (hr)	0.379
t10 norm (hr)	1495.491
t25 norm (hr)	1791.489
t50 norm (hr)	2305.189
t75 norm (hr)	1366.139
t90 norm (hr)	4737.587
t95 norm (hr)	6057.9
tpeak (hr)	1770
cpeak (g/m3)	49.065

cpeak Normalized	1.916

Table 1: Output results from a representative experiment using the Stream Tracer Tool

3.2.2 Analysis of the BTC in context of the pre-existing database

Once the user has the results of their own experiment, we proceed to display the result in context of the database. For example, while calculation for individual data would confirm our analysis that the coefficient of variation (CV; a measure of symmetrical spreading associated with longitudinal dispersion) of our test case is 0.54, this provides little context. To provide additional understanding, results for the analyzed data set are displayed for the user alongside histograms built from the aggregated database (Fig. 5). Visual inspection confirms the CV value of our test case appears to be central to the distribution of the CV's stored in the database. While we have discussed CV, plots are automatically generated in the notebook to visualize results of individual studies across a host of parameter values. For example, the temporal and central moments (first row Fig. 5), our test value is in the highest frequency bin, which is the leftmost bar. Values for the holdback functions, CV and skewness are central to the distributions (bottom row, Fig. X5).

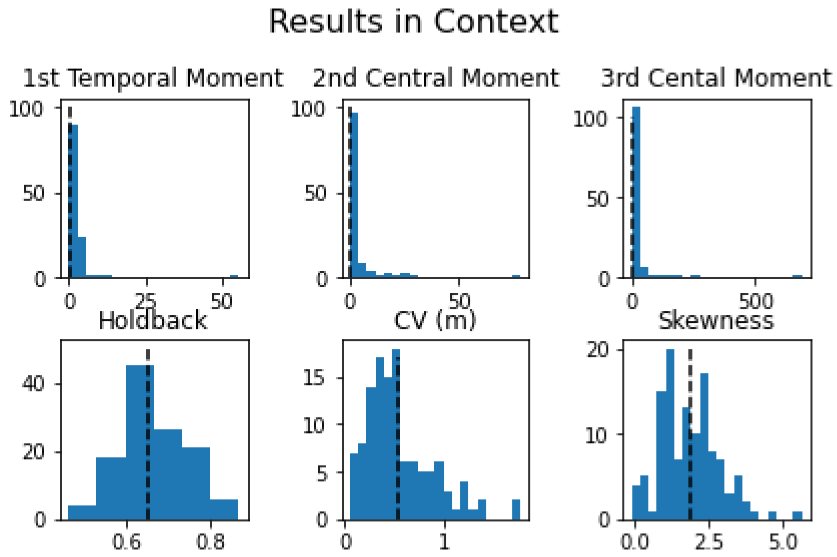
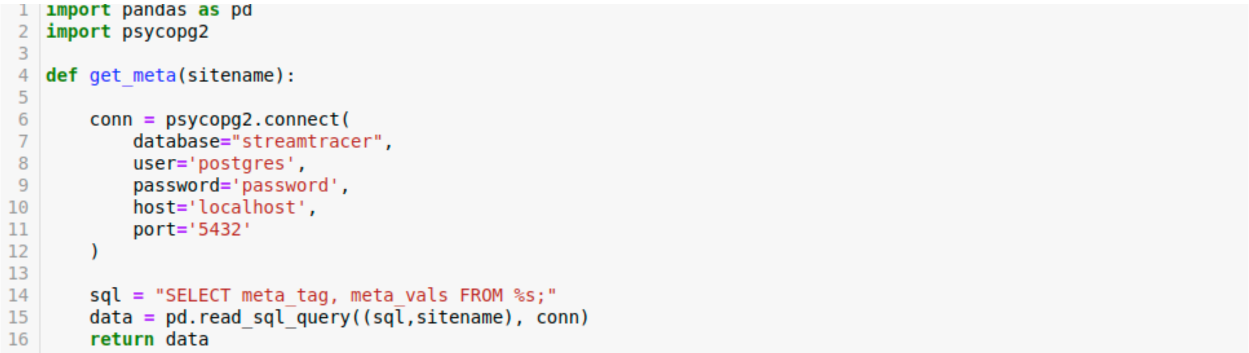


Fig 5: Results for the new data set being analyzed (vertical dashed line in each panel) compared to distributions built from the aggregated database of results. Note this is one of several figures that can be produced in the notebook to enable rapid inspection of results in comparison to those previously analyzed.

3.3 Searchable Database

As tracer studies are aggregated and the database expands, they are aggregated in an SQL-style database, providing the additional capability of query searches. Researchers accessing the database have the ability to extract all information on a given set of experiments. For example if a user is interested in evaluating all streams with an experimental reach length between 50 and 100 meters, these data can easily be extracted with a query. Examples of common SQL searches will be available to the user via the documentation in HydroShare, the python function (get_meta) built to extract the metadata is shown as one example (Fig. 6). Additionally the extendability of SQL will allow users to customize their queries and tailor them for their exact needs. This flexibility will promote the inclusion of additional analysis to be performed ad-hoc to the current analysis.



```
import pandas as pd
import psycopg2

def get_meta(sitename):

    conn = psycopg2.connect(
        database="streamtracer",
        user='postgres',
        password='password',
        host='localhost',
        port='5432'
    )

    sql = "SELECT meta_tag, meta_vals FROM %s;"
    data = pd.read_sql_query((sql,sitename), conn)
    return data
```

Fig 6: Example of common SQL searches for the database

3.4 Classification Techniques

To demonstrate the usefulness of this framework we incorporated a non-supervised classification technique (k-means clustering). This analysis extends the SQL function of the database (Section 3.3) for use within a function. This function can extract the entire database or subsets of data from a search criteria. As an example we performed k-means clustering on the data. We followed standard techniques to reduce the number of features and determine the appropriate number of clusters for the k-means model. Application of unsupervised learning techniques as a characterization method has become a useful approach across a host of applications (Ayustyana, Wibisono, and Sihombing 2021; Ishitsuka et al. 2022; Yang et al. 2019).

A first step in model building is data preprocessing and visualization, we used a pairplot from the seaborn package to visualize the dataset (Fig 7). From inspection, we identify what appear to be separable data (e.g., velocity and skewness columns). For demonstrative purposes, we built a model with a reduced feature set to showcase all model steps. Before reducing the features the data was first standardized using the built in StandardScaler function. To reduce the number of features principal component analysis (PCA) was used. A common strategy is to select the number of features that preserves 80-90% of the variance. The results, presented here by the cumulative explained variance (Fig. 7), show that using five (5) features accounts for 86.8% of the explained variance.

Once the number of components was determined we feed the dimensionally reduced input data to our k-means algorithm. To determine the number of clusters to use with the k-means algorithm we applied the elbow method to the within-cluster sum of squares (WCSS, Fig. 7). The selection of two (2) or six (6) clusters would be reasonable (Fig. 7), and to complete the demonstration we selected six clusters. To visualize the potential for these data to be separable the second and third components were plotted against each other. The data appear to have significant overlap using six clusters (Fig. 7) and plotting the same two components against each other using two clusters shows better separation. Again, we underscore our objective here is to spark ideas about what these data enable for users, not draw conclusions from this sample analysis.

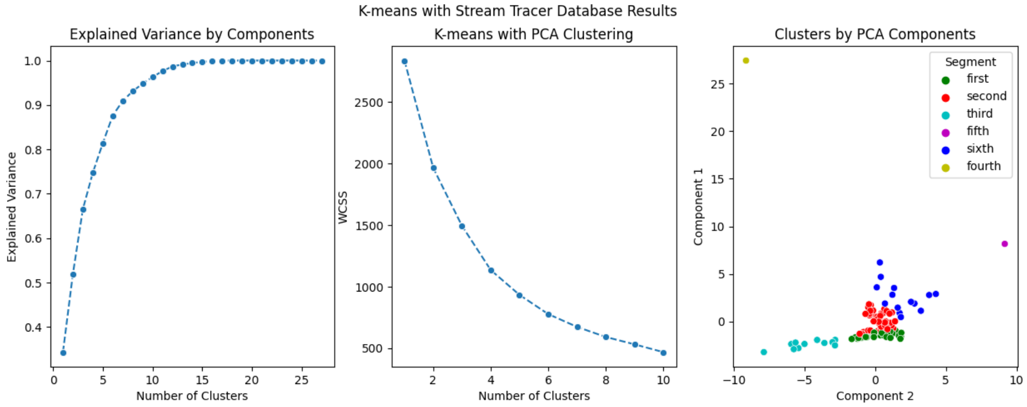


Fig 7: Results of the classification using unsupervised learning.

4.0 Conclusion

We deployed a cloud based tool that performs analysis of a user's experimental stream tracer data. These data are then added to a database that we have initially populated with one large study set (A. S. Ward, Wondzell, et al. 2019) to establish a community-accessible database that is openly available and will grow with each unique user dataset. In addition to being openly accessible, this database and the associated workflow have associated DOIs, making them increasingly discoverable and citable. The analysis performed creates standardized outputs that facilitate data sharing and easy comparison across experiments, including repeatable analyses for this common stream study technique. In addition to the database, the user data and results are automatically stored under a resource ID within HydroShare that allows these data to be citable and discoverable in individual researchers' entries. The existence of this tool will allow researchers to focus on experimental design, and the results of their study versus what analysis should be performed, where data should be stored and how to cite their own data.

One important contribution of this framework is that researchers can now evaluate their data against all other experimental data that have been aggregated. This contextual component was not possible without an openly available datasource. Having these data organized and accessible will enable comparisons across sites, approaches, and synthesis that has heretofore not been possible. We anticipate that compiling these data and enabling intercomparison will inevitably lead to new hypotheses to be tested, which could be compared against previous experiments on the same stream segments.

Additionally, this database readily enables data science approaches that may prove fruitful. For example, researchers may use these data to build supervised classification models based on reported metadata (e.g., stream order), or extend the unsupervised classification methods we presented on subsets of data (e.g., stratified by stream order, latitude, geologic setting, or a host of other bases). Enabling data to be used across suites of analysis from domain specific (e.g., hydrology) to the forefront of deep learning will provide new perspectives on traditional results previously not possible.

Lastly the development of this tool follows precisely the structure taught through the HydroLearn modules(“Stream Solute Tracers” 2020) . As a result, consistency in training new researchers to conduct and analyze results parallels the tools we provide. Previously, researchers were left to create workflows for analyses, limiting repeatability by others and allowing the potential for different implementations across the field. . Now there is one streamlined tool for analysis and training. This ensures consistent training which will enable these analysis and data to be leveraged for decades.

Acknowledgements

This research has been supported by the U.S. Department of Energy (DE-SC0019377), the National Science Foundation (grant nos. EAR-1360276 and EAR-1652293), the 2021 CUAHSI Hydroinformatics Innovation Fellowship, the University of Birmingham (Institute of Advanced Studies) and the Fulbright—University of Birmingham Scholar program, and with resources from the home institutions of the authors including the Burnell and Barbara Fischer Fellowship.

References

"100 Plastic Rivers - a Global Investigation." 2022. University of Birmingham. June 29, 2022. https://www.birmingham.ac.uk/research/water-sciences/projects/plastic-rivers.aspx.

"AmeriFlux." 2022. AmeriFlux. 2022. https://ameriflux.lbl.gov/.

Appling, Alison P., Robert O. Hall, Charles B. Yackulic, and Maite Arroita. 2018. "Overcoming Equifinality: Leveraging Long Time Series for Stream Metabolism Estimation." *Journal of Geophysical Research: Biogeosciences*. https://doi.org/10.1002/2017jg004140.

Ayustyana, E., S. A. Wibisono, and F. M. H. Sihombing. 2021. "Coal Characterization of South Sumatera Basin Using the Unsupervised Machine Learning Method." *IOP Conference Series: Earth and Environmental Science* 830 (1): 012043.

Balson, Tyler. 2022. "StreamTracerDataBase." https://www.hydroshare.org/resource/cdf1f32fec304f339a7a1dfeabd5f253/.

Bell, Gordon, Tony Hey, and Alex Szalay. 2009. "Computer Science. Beyond the Data Deluge." *Science* 323 (5919): 1297–98.

Bencala, Kenneth E., and Roy A. Walters. 1983. "Simulation of Solute Transport in a Mountain Pool-and-Riffle Stream: A Transient Storage Model." *Water Resources Research*. https://doi.org/10.1029/wr019i003p00718.

Bernhardt, E. S., J. B. Heffernan, N. B. Grimm, E. H. Stanley, J. W. Harvey, M. Arroita, A. P. Appling, et al. 2018. "The Metabolic Regimes of Flowing Waters." *Limnology and Oceanography*. https://doi.org/10.1002/lno.10726.

Boano, F., A. I. Packman, A. Cortis, R. Revelli, and L. Ridolfi. 2007. "A Continuous Time Random Walk Approach to the Stream Transport of Solutes." *Water Resources Research* 43 (10). https://doi.org/10.1029/2007wr006062.

Danckwerts, P. V. 1953. "Continuous Flow Systems: Distribution of Residence Times." *Chemical Engineering Science* 2 (1): 1–13.

"Data_example_resource." 2021. https://www.hydroshare.org/resource/f44e8baf1682456e9605221deef3a65b/.

Datry, T., A. Foulquier, R. Corti, D. von Schiller, K. Tockner, C. Mendoza-Lera, J. C. Clément, et al. 2018. "A Global Analysis of Terrestrial Plant Litter Dynamics in Non-Perennial Waterways." *Nature Geoscience* 11 (7): 497–503.

Datry, Thibault, R. Corti, A. Foulquier, D. von Schiller, and Klement Tockner. 2016. "One for All, All for One: A Global River Research Network." *Eos*. https://doi.org/10.1029/2016eo053587.

Enabling FAIR Data Community. 2018. *Commitment Statement to Enabling FAIR Data in the Earth, Space, and Environmental Sciences*. https://doi.org/10.5281/zenodo.1451971.

Fidler, Fiona, and John Wilcox. 2021. "Reproducibility of Scientific Results." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2021. Metaphysics

Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/scientific-reproducibility/.

“Find, Analyze and Share Water Data.” 2022. Hydroshare. 2022. https://www.hydroshare.org/.

Fischer, Hugo B., John E. List, C. Robert Koh, Jorg Imberger, and Norman H. Brooks. 1979. *Mixing in Inland and Coastal Waters*. Academic Press.

“GLEON.” 2022. 2022. https://gleon.org/.

Gupta, A., and V. Cvetkovic. 2000. “Temporal Moment Analysis of Tracer Discharge in Streams: Combined Effect of Physicochemical Mass Transfer and Morphology.” *Water Resources Research* 36 (10): 2985–97.

Harman, Ciaran J. 2015. “Time-Variable Transit Time Distributions and Transport: Theory and Application to Storage-Dependent Transport of Chloride in a Watershed.” *Water Resources Research* 51 (1): 1–30.

Harman, C. J., A. S. Ward, and A. Ball. 2016. “How Does Reach-scale Stream-hyporheic Transport Vary with Discharge? Insights from rSAS Analysis of Sequential Tracer Injections in a Headwater Mountain Stream.” *Water Resources Research* 52 (9): 7130–50.

Harvey, Charles F., and Steven M. Gorelick. 1995. “Temporal Moment-Generating Equations: Modeling Transport and Mass Transfer in Heterogeneous Aquifers.” *Water Resources Research*. https://doi.org/10.1029/95wr01231.

Ishitsuka, Kazuya, Hiroki Ojima, Toru Mogi, Tatsuya Kajiwara, Takeshi Sugimoto, and Hiroshi Asanuma. 2022. “Characterization of Hydrothermal Alteration along Geothermal Wells Using Unsupervised Machine-Learning Analysis of X-Ray Powder Diffraction Data.” *Earth Science Informatics* 15 (1): 73–87.

“LTER.” 2016. LTER. LTER Network. August 10, 2016. https://lternet.edu/.

Mason, Seth J. K., Brian L. McGlynn, and Geoffrey C. Poole. 2012. “Hydrologic Response to Channel Reconfiguration on Silver Bow Creek, Montana.” *Journal of Hydrology*. https://doi.org/10.1016/j.jhydrol.2012.03.010.

“NEON.” 2022. 2022. https://www.neonscience.org/.

Payn, R. A., M. N. Gooseff, B. L. McGlynn, K. E. Bencala, and S. M. Wondzell. 2009. “Channel Water Balance and Exchange with Subsurface Flow along a Mountain Headwater Stream in Montana, United States.” *Water Resources Research*. https://doi.org/10.1029/2008wr007644.

“PostgreSQL.” 2022. PostgreSQL. July 8, 2022. https://www.postgresql.org/.

Rinaldo, Andrea, Paolo Benettin, Ciaran J. Harman, Markus Hrachowitz, Kevin J. McGuire, Ype van der Velde, Enrico Bertuzzo, and Gianluca Botter. 2015. “Storage Selection Functions: A Coherent Framework for Quantifying How Catchments Store and Release Water and Solutes.” *Water Resources Research*. https://doi.org/10.1002/2015wr017273.

Runkel, Robert L. 2002. “A New Metric for Determining the Importance of Transient Storage.” *Journal of the North American Benthological Society* 21 (4): 529–43.

Runkel, Robert L., Diane M. McKnight, and Edmund D. Andrews. 1998. “Analysis of Transient Storage Subject to Unsteady Flow: Diel Flow Variation in an Antarctic Stream.” *Journal of the North American Benthological Society* 17 (2): 143–54.

Savoy, Philip, Alison P. Appling, James B. Heffernan, Edward G. Stets, Jordan S. Read, Judson W. Harvey, and Emily S. Bernhardt. 2019. “Metabolic Rhythms in Flowing Waters: An

Approach for Classifying River Productivity Regimes." *Limnology and Oceanography*. https://doi.org/10.1002/lno.11154.

Schiller, D., T. Datry, R. Corti, A. Foulquier, K. Tockner, R. Marcé, G. García-Baquero, et al. 2019. "Sediment Respiration Pulses in Intermittent Rivers and Ephemeral Streams." *Global Biogeochemical Cycles* 33 (10): 1251–63.

Schmid, Bernhard H. 1995. "On the Transient Storage Equations for Longitudinal Solute Transport in Open Channels: Temporal Moments Accounting for the Effects of First-Order Decay." *Journal of Hydraulic Research* 33 (5): 595–610.

Shumilova, Oleksandra, Dominik Zak, Thibault Datry, Daniel von Schiller, Roland Corti, Arnaud Foulquier, Biel Obrador, et al. 2019. "Simulating Rewetting Events in Intermittent Rivers and Ephemeral Streams: A Global Analysis of Leached Nutrients and Organic Matter." *Global Change Biology* 25 (5): 1591–1611.

"SoluteTracerTool." 2021. https://www.hydroshare.org/resource/6f3eee8ca110402ca432ebee52ab286f/.

Stall, Shelley, Lynn Yarmey, Joel Cutcher-Gershenfeld, Brooks Hanson, Kerstin Lehnert, Brian Nosek, Mark Parsons, Erin Robinson, and Lesley Wyborn. 2019. "Make Scientific Data FAIR." *Nature*. https://doi.org/10.1038/d41586-019-01720-7.

Stegen, James C., and Amy E. Goldman. 2018. "WHONDRS: A Community Resource for Studying Dynamic River Corridors." *mSystems* 3 (5). https://doi.org/10.1128/mSystems.00151-18.

"StreamPULSE." 2022. 2022. https://data.streampulse.org/.

"Stream Solute Tracers." 2020. 2020. https://edx.hydrolearn.org/courses/course-v1:IndianaUniversity+E710+2020_Fall/about.

"Stream Solute Tracers." n.d. Accessed June 29, 2022. https://edx.hydrolearn.org/courses/course-v1:IndianaUniversity+E710+2020_Fall/about.

Stream Solute Workshop. 1990. "Concepts and Methods for Assessing Solute Dynamics in Stream Ecosystems." *Journal of the North American Benthological Society* 9 (2): 95–119.

Thackston Edward L., and Schnelle Karl B. 1970. "Predicting Effects of Dead Zones on Stream Mixing." *Journal of the Sanitary Engineering Division* 96 (2): 319–31.

Tiegs, Scott D., David M. Costello, Mark W. Isken, Guy Woodward, Peter B. McIntyre, Mark O. Gessner, Eric Chauvet, et al. 2019. "Global Patterns and Drivers of Ecosystem Functioning in Rivers and Riparian Zones." *Science Advances* 5 (1): eaav0486.

Ulseth, Amber J., Robert O. Hall, Marta Boix Canadell, Hilary L. Madinger, Amin Niayifar, and Tom J. Battin. 2019. "Distinct Air–water Gas Exchange Regimes in Low- and High-Energy Streams." *Nature Geoscience*. https://doi.org/10.1038/s41561-019-0324-8.

Ward, Adam, Jennifer Drummond, Angang Li, Anna Lupon, Marie Kurz, Jay Zarnetske, James Stegen, et al. 2022. "Advancing River Corridor Science beyond Disciplinary Boundaries with an Inductive Approach to Catalyse Hypothesis Generation." https://doi.org/10.31223/x54w44.

Ward, Adam S., Michael N. Gooseff, Thomas J. Voltz, Michael Fitzgerald, Kamini Singha, and Jay P. Zarnetske. 2013. "How Does Rapidly Changing Discharge during Storm Events Affect Transient Storage and Channel Water Balance in a Headwater Mountain Stream?" *Water Resources Research* 49 (9): 5473–86.

Ward, Adam S., Marie J. Kurz, Noah M. Schmadel, Julia L. A. Knapp, Phillip J. Blaen, Ciaran J.

Harman, Jennifer D. Drummond, et al. 2019. “Solute Transport and Transformation in an Intermittent, Headwater Mountain Stream with Diurnal Discharge Fluctuations.” *WATER* 11 (11): 2208.

Ward, Adam S., Steven M. Wondzell, Noah M. Schmadel, Skuyler Herzog, Jay P. Zarnetske, Viktor Baranov, Phillip J. Blaen, et al. 2019. “Spatial and Temporal Variation in River Corridor Exchange across a 5th-Order Mountain Stream Network.” *Hydrology and Earth System Sciences* 23 (12): 5199–5225.

Ward, Adam S., Jay P. Zarnetske, Viktor Baranov, Phillip J. Blaen, Nicolai Brekenfeld, Rosalie Chu, Romain Derelle, et al. 2019. “Co-Located Contemporaneous Mapping of Morphological, Hydrological, Chemical, and Biological Conditions in a 5th-Order Mountain Stream Network, Oregon, USA.” *Earth System Science Data* 11 (4): 1567–81.

Wlostowski, Adam N., Michael N. Gooseff, William B. Bowden, and Wilfred M. Wollheim. 2017. “Stream Tracer Breakthrough Curve Decomposition into Mass Fractions: A Simple Framework to Analyze and Compare Conservative Solute Transport Processes.” *Limnology and Oceanography, Methods / ASLO* 15 (2): 140–53.

Wörman, Anders, Aaron I. Packman, Håkan Johansson, and Karin Jonsson. 2002. “Effect of Flow-Induced Exchange in Hyporheic Zones on Longitudinal Transport of Solutes in Streams and Rivers.” *Water Resources Research* 38 (1): 2–1 – 2–15.

Yang, Xuezhi, Xian Liu, Aiqian Zhang, Dawei Lu, Gang Li, Qinghua Zhang, Qian Liu, and Guibin Jiang. 2019. “Distinguishing the Sources of Silica Nanoparticles by Dual Isotopic Fingerprinting and Machine Learning.” *Nature Communications* 10 (1): 1620.