Environmental Modelling & Software 104 (2018) 94–101

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Characterizing hydrologic networks: Developing a tool to enable research of macroscale aquatic networks

Luke A. Winslow [a, *], Tobi H. Hahn [a], Sarah DeVaul Princiotta [b, c], Taylor H. Leach [a], Kevin C. Rose [a]

[a] Department of Biological Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180, USA
[b] Lacawac Sanctuary and Biological Field Station, Lake Ariel, PA 18436, USA
[c] Hancock Biological Station, Murray State University, Murray, KY 42071, USA

ARTICLE INFO

Article history:
Received 15 December 2017
Received in revised form
8 February 2018
Accepted 15 March 2018
Available online 26 March 2018

Keywords:
R package
Lakes
Streams
Hydrologic network
Limnology
Data integration
Macrosystem ecology

ABSTRACT

Addressing continental scale challenges affecting inland aquatic systems requires data at comparable scales. Critically, local in-situ observations for both lotic and lentic ecosystems are frequently fragmented across federal, state and local agencies, and nonprofit or academic organizations and must be linked to other geospatial data to be useful. To advance macro-scale aquatic ecosystem science, better tools are needed to facilitate dataset integration. Key to integration of aquatic data is the linking of spatial data to the hydrologic network. This integration step is challenging as hydrologic network data are large and cumbersome to manage. Here we develop a new R package, *hydrolinks*, to ease linking aquatic data to the hydrologic network. We use *hydrolinks* to evaluate the spatial data quality for all lake and stream sites available through the U.S. Water Quality Portal. We find that 76.5% of lake sites and 13.9% of stream sites do not correspond with mapped waterbodies.

Software availability

Name of software: *hydrolinks*
Developers: Tobi Hahn, Luke Winslow, Taylor Leach, Kevin Rose
Software Required: R
License: MIT license
Availability: Freely available through CRAN (cran.r-project.org), development version at github.com/lawinslow/hydrolinks
Archived version: https://doi.org/10.5281/zenodo.1169399

1. Introduction

Many of today's challenges in understanding aquatic systems operate at the regional to global scale (Heffernan et al., 2014). Addressing these macrosystem-scale challenges affecting inland aquatic ecosystems requires data at comparably large scales. From climate change impacts (e.g., Isaak et al., 2016; O'Reilly et al., 2015) to eutrophication and algal blooms (e.g., Chapra et al., 2017; Sinha et al., 2017), large regional to global scale insights depend on linking observations across diverse and spatially extensive lake and stream regions. Broad scale data is particularly important as both geographic and morphological heterogeneity of inland waters and diverse watershed characteristics drive how aquatic ecosystems respond to broad scale forcing (e.g., Read et al., 2015; Soranno et al., 2015; Winslow et al., 2015). To this end, researchers have built tools and data systems to automate integration and processing of aquatic data and features (e.g., Haag and Shokoufandeh, 2017; Horsburgh and Reeder, 2014; Read et al., 2011).

Some datasets relevant to aquatic research, such as land cover (Jin et al., 2013), elevation (Jarvis et al., 2008) and climate (Dee et al., 2011) are well curated and have been made publicly available at continental and global scales. But many potentially high-value, in-situ aquatic observations are fragmented across government, nonprofit, and academic organizations (Soranno et al., 2015). Further, available datasets are frequently described by latitude/longitude coordinates, often without clear, unambiguous reference to mapped surface water features. The lack of unique identifiers or unambiguous linking can make it difficult to connect aquatic data

* Corresponding author.
E-mail address: lawinslow@gmail.com (L.A. Winslow).

https://doi.org/10.1016/j.envsoft.2018.03.012
1364-8152/

to other *in-situ* datasets or important spatial context information, such as land use, hydrologic connectivity, or waterbody morphology (Hill et al., 2016) using alternative site identification schemes. While there have been recent efforts to collect and organize fragmented sources of aquatic data and link them to specific, mapped waterbodies, most current examples are limited in data-type, temporal span, or confined to a regional spatial extent (e.g., Oliver et al., 2017; Soranno et al., 2015). To advance macrosystem aquatic research, new tools are needed to accelerate the integration of fragmented data across both lentic and lotic ecosystems.

One powerful way to integrate fragmented aquatic datasets and connect those data with landscape features is to link *in-situ* aquatic data to mapped surface water features using unique identifiers. This allows for the unambiguous communication of the surface water feature from which a data point originates. Unfortunately, available tools to enable linking are only available on specialized geographic information system (GIS) platforms that are often foreign to ecologists and require special training. One such example is the U.S. Geological Survey's (USGS) Hydrologic Event Management tool (HEM; https://nhd.usgs.gov/tools.html), which is specific to only the USGS National Hydrography Dataset. Additionally, the HEM tool requires manual handling of the hydrologic network data, does not support lakes, and needs a local license for the ArcGIS software package. Bringing hydrologic network linking to general scientific computing platforms used by aquatic ecologists and generalizing the linking to include multiple commonly used, large-scale hydrologic network datasets would enable better, broad-scale data integration thus accelerating macrosystem hydrologic science.

Here we introduce *hydrolinks*, an R package to ease linking of aquatic data with the hydrologic network. *hydrolinks* automates access, retrieval, and local storage of large hydrologic network datasets, including the USGS National Hydrologic Dataset (NHD) high resolution (1:24k), NHD plus (currently 1:100k), and the global hydroLAKES datasets. The package includes a number of algorithms for linking data described by latitude and longitude to the hydrologic network, including variable-width buffer, centroid correspondence, and point-in-polygon linking. To provide example uses, we demonstrate the functioning of *hydrolinks* by linking a large, national-scale aquatic dataset to the U.S. hydrologic network. We then show how linking large datasets to the hydrologic network can enable integration of diverse data sources and highlight potential data quality issues in large, federal environmental databases.

2. Methods

2.1. Linking tool

The *hydrolinks* package is built to reduce the complexities of linking aquatic ecosystem data that is described by simple latitude/longitude geographic referencing to generally available hydrologic network geographic datasets. Based on the data and type of hydrologic linking requested, the package can automatically download required hydrologic network data to link latitude/longitude data points to streams or waterbodies (generally lakes or reservoirs). Here we provide an overview of the linking approaches used and how these large, complex data are stored and distributed.

2.1.1. Data access methods

One of the novel aspects of *hydrolinks* is to enable automatic, on-the-fly data access so the user does not need to download and manage the large hydrologic datasets for linking. The hydrologic datasets are split up into small, quick-to-download file sizes (~50 MB) that represent contiguous regions. Included in the package are data on the geographic bounding boxes associated with the sub-region files of each dataset. These bounding boxes are used when executing linking functions to determine which sub-region pieces of the hydrologic networks are needed. These data are downloaded using simple HTTP from an online source, verified complete using an MD5 hash, and then used in the data linking procedure. All downloaded files are locally cached to enable rapid re-linking of future data. The location of the local cache can be user-defined using the *local_path* function. This may be useful when an alternative cache drive or path is preferred over the default.

2.1.2. Current available datasets for linking

With the initial release of *hydrolinks*, we have included one global and two U.S. datasets. The hydroLAKES dataset (Messager et al., 2016) is a globally comprehensive dataset of lakes over 10 ha. For lakes and streams, both the NHD plus (v2; Moore and Dewald, 2016) and NHD high resolution (1:24k; Simley and Carswell, 2009) datasets are included. The NHD high resolution includes lakes and streams down to very small sizes (~0.1 ha for lakes; Winslow et al., 2014). The NHD plus is based on a lower resolution product (~1:100k) but includes a great deal of ancillary hydrologic data not covered by the base NHD high resolution dataset that may be useful to many users. We have opted not to include NHD medium resolution (1:100k) due to its similarity in resolution and coverage as the NHD plus. Further datasets that are freely available for redistribution may be added in future package updates and upon request.

2.1.3. Linking methods

The R package presented here, *hydrolinks*, implements several different hydrologic network linking procedures, with some differences based on linking to the network of moving water (discussed here as flowlines) or bodies of still water (here as waterbodies). Unless noted, geographic locations being linked to the network are described as geopoints (latitude and longitude pairs on the WGS84 datum). All methods return an R *data.frame* object with rows that correspond to linked flowlines and waterbodies with an additional column (titled *MATCH_ID*) corresponding to the matched geopoint supplied by the user.

For flowline linking, *hydrolinks* implements a single, flexible implementation. The majority of hydrologic network datasets represent streams and most rivers by lines (as opposed to polygons), so *hydrolinks* implements a snap-to-line linking procedure where a point is snapped to a stream flowline with the *link_to_flowlines* function. While some supplied geopoints may be geographically close to the snapped flowlines, other points may fall a large distance the nearest stream flowline. To differentiate between points that fall "close enough" to a flowline and those that may have erroneous geospatial information, hydrolinks differentiates geopoints that fall beyond a user-configurable distance from the hydrologic flowlines (default 100 m). Points which fall outside of that distance are not linked to the hydrologic network. In certain situations, the default may not be sufficient, so users can specify a custom maximum distance by specifying the *buffer* parameter.

Lakes have a different linking strategy than stream linking. Geopoints can be linked to lakes and reservoirs by a simple point-in-polygon approach. This process identifies which geopoints are contained within any given waterbody polygon. This linking procedure is implemented in *hydrolakes* in the included *link_to_waterbodies* function. Geopoints can also be specifically linked to the waterbody centroids of the chosen hydrologic dataset, with a point-to-centroid approach implemented by the *link_waterbody_centroids* function. However, data describing the location of lakes is sometimes inaccurate or taken at a point which does not overlap with the lake (Fig. 1). Two common issues are (1) lake

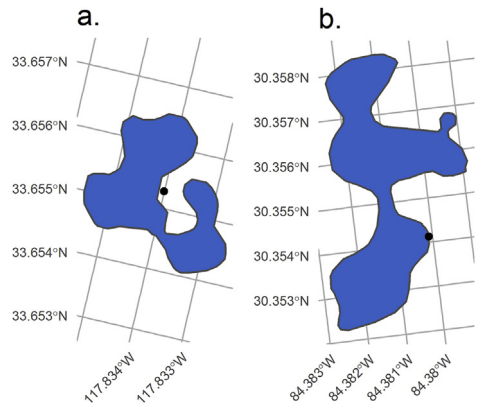

Fig. 1. - Two examples from the Water Quality Portal (WQP) of non-point-in-polygon matches for lake sites. Panel (a) is an example lake where the point was within 25 m of the lake polygon centroid, but not within 25 m of the shoreline and (b) shows a WQP point that did not fall on the lake polygon, but was within a 25 m buffer.

associated data being defined by the centroid of the lake polygon where the centroid lies outside the lake polygon itself (Fig. 1a). The non-overlapping centroid issue can occur on lakes with non-convex polygons (e.g., oxbow lakes). 2) Points may be taken very near to a lake (such as at a boat launch) but not geographically over the lake (Fig. 1b). These potential issues are handled in two ways. 1) *link_to_waterbodies* has an optional parameter, *buffer* (specified in meters), to match geopoints which fall within a buffer around each lake polygon. 2) The *link_waterbody_centroids* has a user defined maximum-distance cutoff, *buffer* (defaults to 25 m), which allows for a loosening of the restriction on matching geopoints to waterbody polygon centroids.

2.2. Dataset example

To show an example of large-scale insights enabled by *hydrolinks*, we examined the spatial quality of data available from one of the largest single sources of aquatic data available, the USGS Water Quality Portal (WQP; Read et al., 2017). The WQP is an interface to query several U.S. federal databases of aquatic data, including the U.S. Environmental Protection Agency (EPA) STOrage and RETrieval Dashboard (STORET) and the U.S. Geological Survey (USGS) National Water Information System (NWIS) database. We downloaded all national-scale lake and stream data available from the WQP using the USGS *dataRetrieval* R package (Hirsch and De Cicco, 2015). For stream and lake sites from the WQP, we used *hydrolinks* to link observations to mapped waterbodies using all potential methods. Stream sites were linked to the NHD using *link_to_flowlines* with a *buffer* of 100 m. Lake sites were linked in three ways. First, lake sites were linked using point-in-polygon (*link_to_waterbodies*) with no buffer; only lake sites directly over a waterbody polygon were matched. Lake sites were also linked using *link_to_waterbodies* with a buffer of 25 m. Finally, lake sites were linked using *link_waterbody_centroids* with a *buffer* of 25 m.

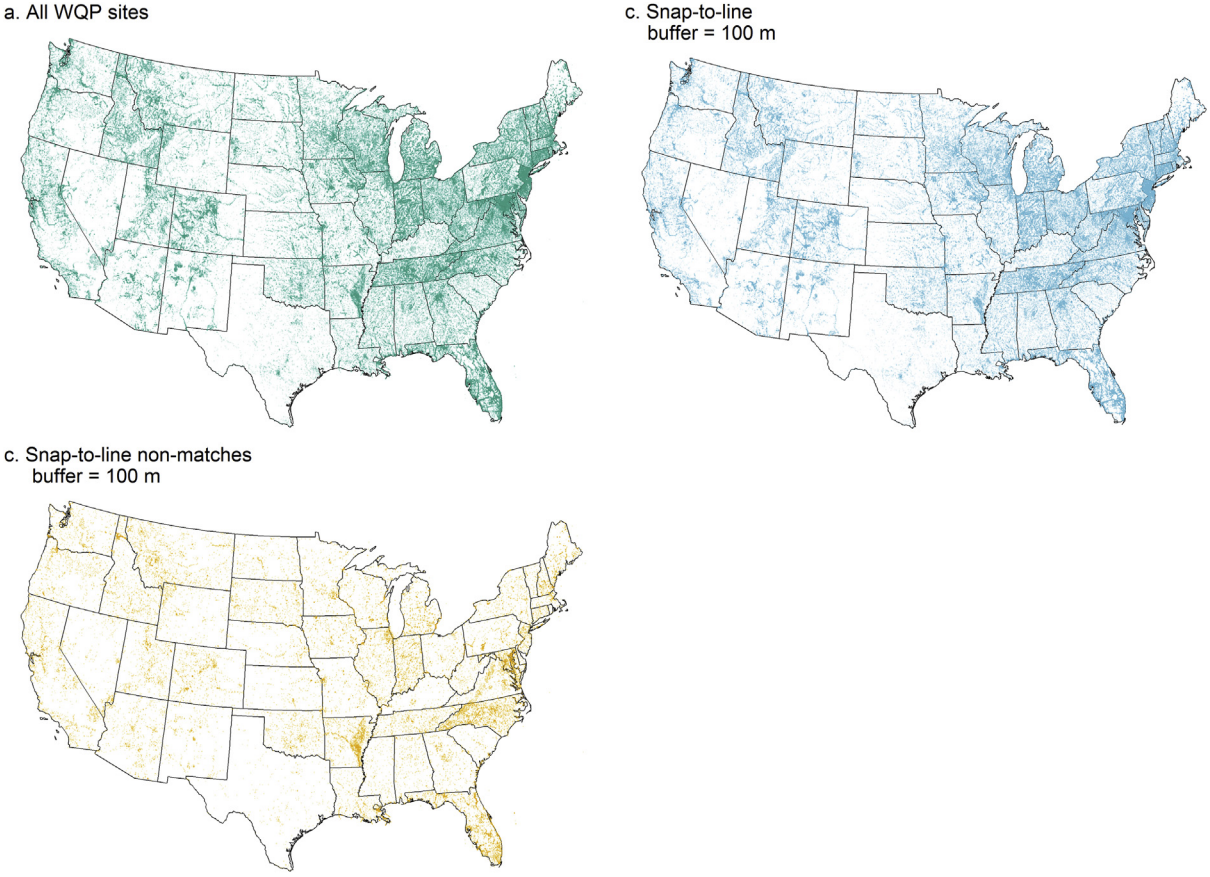

Fig. 2. Map of all lake sites in the Water Quality Portal (WQP) and their link with the National Hydrography Network high-resolution dataset. (a) All 149,905 WQP lake sites. (b) The 102,754 sites which linked directly with waterbodies using point-in-polygon with no buffer (*link_to_waterbodies* function, *buffer* = 0 m). (c) The 11,866 additional sites that matched lakes by point-in-polygon with a 25 m buffer (*link_to_waterbodies* function, *buffer* = 25 m). And (d) the 206 point-to-centroid matched sites with a maximum distance between the centroid and the point of 25 m (*link_waterbody_centroids* function, *buffer* = 25 m).

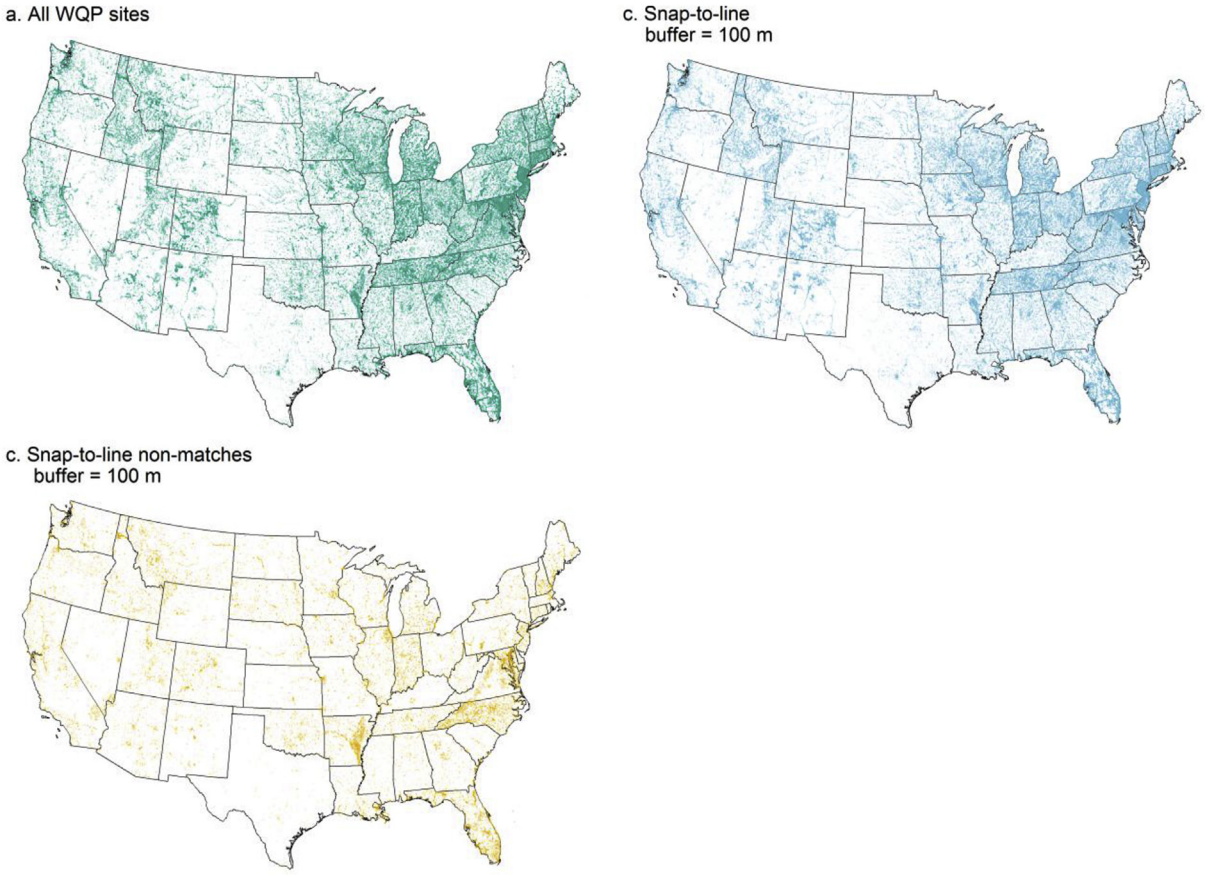

Fig. 3. Map of all stream sites in the Water Quality Portal (WQP) and their link to with the National Hydrography Network (NHD) high-resolution flowline dataset. A 100 m maximum distance cutoff between the WQP point and the flowline was used to define a linked stream site. Panels show (a) all 547,930 stream sites in the WQP, (b) all 471,668 snap-to-lines matches (*link_to_flowlines* function, *buffer* = 100 m), and (c) the remaining 76,262 WQP sites that were not matched to the NHD layer with a 100 m maximum distance cutoff between the point and the flowline.

Table 1
Number of Water Quality Portal (WQP) sites linked to NHD high-resolution hydrologic features using each of the linking procedures.

	Number of sites linked	
	Lakes	Streams
Total WQP sites	149,905	547,930
Point-in-polygon		
link_to_waterbodies, buffer = 0 m	102,754 (68.5%)	–
link_to_waterbodies, buffer = 25 m	11,866 (7.9%)	–
Point-to-centroid		
link_waterbody_centroids, buffer = 25 m	206 (0.1%)	–
Snap-to-line		
link_to_flowlines, buffer = 100 m	–	471,668 (86.1%)

3. Results

3.1. Water Quality Portal example

There are a large number of uniquely identified sample sites in the WQP (Figs. 2a and 3a). As of the date of access (August 1, 2017), there were 149,905 and 547,930 unique sites categorized as lakes and streams, respectively (Table 1). Of the reported WQP sites, 47,151 lake sites (31.5%) and 76,262 stream sites (13.9%) did not correspond with mapped hydrologic feature in the high-resolution NHD based on the linking procedure used here (point-in-polygon with no buffer for lakes, snap-to-line with 100 m buffer for streams). Using the point-to-polygon approach with a 25 m buffer increased the number of linked lake sites by 11,866 (7.9%) with an additional 206 lakes linked using the point-to-centroid approach with a 25 m buffer (Table 1). Using a larger buffer for lakes showed diminishing returns beyond 25 m (Supplemental Text 1).

The number of lake sites without matches to hydrologic features in the NHD varied spatially across U.S. states (Figs. 2 and 3). States that appear to have a high linked site density on the visualized map (Fig. 2) tended to have a combination of both a high abundance of lakes and a high numerical coverage percentage. For example, Wisconsin, Minnesota, and Michigan all have high numbers of natural lakes (Winslow et al., 2014) and had high percent lake coverage (Table 2). Some state's sampling is dominated by fewer large lakes, with a large divergence of percent covered by area and number (e.g., Utah, Maryland, and Tennessee all have high area but low numerical coverage). Other notable mentions include Florida, which has a high density of linked sites visually (Fig. 2a and b) and a relatively low percent of total observed lake area (Table 2). In general, the lake sites that did not match waterbodies were distributed similarly to the underlying whole population of lake sites, with a few exceptions. A few states dominated numerically the lake sites which did not match via point-in-polygon but did match well with a 25 m buffer (Fig. 2c; New Hampshire, Wisconsin, and Ohio). Wisconsin had the majority of lakes which matched via centroid (Fig. 2d), with only a few scattered among the other states.

For streams, the distribution of match and non-match sites had a different spatial distribution compared to linked lake sites (Fig. 3a–b; Table 2). Some states followed similar patterns with

Table 2
Statistics for sampling coverage of linked lake sites from the WQP summarized by state. Sites within Washington DC were included in Maryland and CONUS represents the contiguous United States.

State	Streams				Lakes			
	Observed by Length (%)	Observed by Number (%)	Median Length (m)	Total Flowlines in State	Observed by Area (%)	Observed by Number (%)	Median Area (m^2)	Total Lakes in State
CONUS	2.59	0.73	271	11,784,806,954	65.92	0.50	1802	6,500,264
Alabama	1.89	0.75	434	213,043,388	39.23	0.11	3265	109,064
Arkansas	2.78	1.07	471	221,279,838	37.83	0.10	1346	187,386
Arizona	1.21	0.49	522	485,341,545	48.13	0.44	1518	38,354
California	1.70	0.36	217	756,229,974	50.28	0.80	2181	102,217
Colorado	1.84	0.49	253	450,209,743	47.21	0.58	1296	115,142
Connecticut	8.46	2.00	131	22,183,795	20.09	0.71	3311	39,343
Delaware	5.40	1.32	172	9,206,802	9.56	0.80	2936	9690
Florida	7.26	1.64	205	170,497,690	31.09	0.80	5760	468,706
Georgia	2.88	0.89	296	190,787,240	23.13	0.08	6991	207,146
Iowa	3.80	1.45	599	183,864,337	40.58	0.27	1379	106,420
Idaho	4.26	1.89	588	291,338,659	70.77	0.37	1378	44,566
Illinois	4.60	1.49	442	193,962,896	36.02	0.39	1481	164,391
Indiana	0.32	0.13	141	440,438,573	31.49	0.65	2478	141,002
Kansas	1.52	0.53	579	296,009,992	47.90	0.20	1352	216,913
Kentucky	1.67	0.65	456	162,552,524	49.26	0.07	1031	193,614
Louisiana	1.93	0.41	288	221,695,088	52.90	0.08	3543	191,076
Massachusetts	5.36	1.07	127	29,830,639	20.10	1.16	4060	48,370
Maryland	14.06	3.98	210	40,678,338	91.72	0.32	2187	28,185
Maine	2.91	0.87	264	91,001,096	60.86	1.74	1726	32,336
Michigan	12.11	3.55	383	137,709,871	96.60	1.59	6355	93,947
Minnesota	9.01	2.88	500	169,615,578	82.45	4.17	1986	126,476
Missouri	1.37	0.58	474	296,199,009	57.01	0.16	976	273,790
Mississippi	1.51	0.43	304	261,573,841	21.72	0.10	1855	192,231
Montana	1.95	0.76	569	628,057,437	56.72	0.31	2379	148,321
North Carolina	1.91	0.60	229	231,633,366	16.13	0.20	2498	116,764
North Dakota	3.44	0.73	433	145,462,709	37.73	0.22	4217	248,855
Nebraska	2.32	0.54	526	205,722,710	24.45	0.26	2254	126,098
New Hampshire	6.93	2.23	227	29,839,188	52.39	3.14	3444	28,257
New Jersey	4.49	0.64	72	39,571,533	32.98	1.50	1811	25,068
New Mexico	1.57	0.58	588	382,954,284	37.74	0.27	1544	73,081
Nevada	0.72	0.27	531	507,130,141	49.90	0.63	1830	24,856
New York	7.06	2.10	295	162,307,061	47.80	1.09	1712	85,422
Ohio	8.08	3.25	572	148,279,208	49.28	1.45	1460	113,114
Oklahoma	2.29	0.72	337	269,141,699	69.90	0.09	1220	388,291
Oregon	0.91	0.20	208	509,513,074	41.51	0.53	1370	74,590
Pennsylvania	5.24	1.98	454	138,312,078	27.77	0.20	1471	94,896
Rhode Island	6.07	1.41	153	3,738,695	25.00	1.68	4128	8037
South Carolina	1.19	0.29	189	126,551,137	15.29	0.14	4587	91,049
South Dakota	1.86	0.52	442	264,054,783	60.66	0.32	2669	164,627
Tennessee	4.62	2.05	465	182,146,967	78.33	0.13	946	114,766
Texas	1.15	0.31	341	865,712,892	43.58	0.05	1180	1,008,001
Utah	2.29	1.19	737	301,631,042	82.82	0.80	1207	46,749
Virginia	4.39	1.60	292	177,019,466	43.86	0.29	2466	89,171
Vermont	4.59	1.33	231	42,431,861	93.38	1.06	965	29,696
Washington	1.16	0.27	214	391,801,056	38.63	0.99	2658	61,937
Wisconsin	12.13	4.35	537	142,610,324	83.15	5.74	1772	83,650
West Virginia	6.27	3.05	626	88,453,114	46.79	0.27	939	24,990
Wyoming	1.21	0.34	273	465,480,671	49.27	0.25	1703	99,613

Michigan, Wisconsin, Minnesota, and Maryland having top percent observed numbers for both lakes and streams. Some states showed large divergence between lake and stream coverage. While Connecticut was 5th in stream length coverage (8.46%) it ranked 44th for lake coverage by area (20.09%). Visually, these different linking densities were not as apparent in streams as in lakes, though mapping completion dates and resolution may be an important driver of this discrepancy in stream coverage stats.

Using *hydrolinks*, we examined how the distribution of observed lakes and streams compares to the full contiguous U.S. distribution (Fig. 4). The median area of observed lakes was 17.05 ha, compared to 0.14 ha for the complete distribution of contiguous U.S. lakes (Winslow et al., 2014, Fig. 4a). Stream sampling was similarly skewed towards larger, higher-order systems (except at the highest order systems; Fig. 5a), with a length-weighted sampled stream order of 2.44 versus the length-weighted whole U.S. distribution stream order of 0.22.

4. Discussion

hydrolinks automates the multiple, manual steps required to link coordinate-described data to the hydrologic network. Facilitating this process can help unify different limnological datasets as well as datasets that are best described in relation to the hydrologic network (e.g., watershed or buffer land use, lake morphology and stream flow and order characteristics). Linking these geopoint sites to the hydrologic network can improve data sharing, integration, and better describe the geographic and morphological

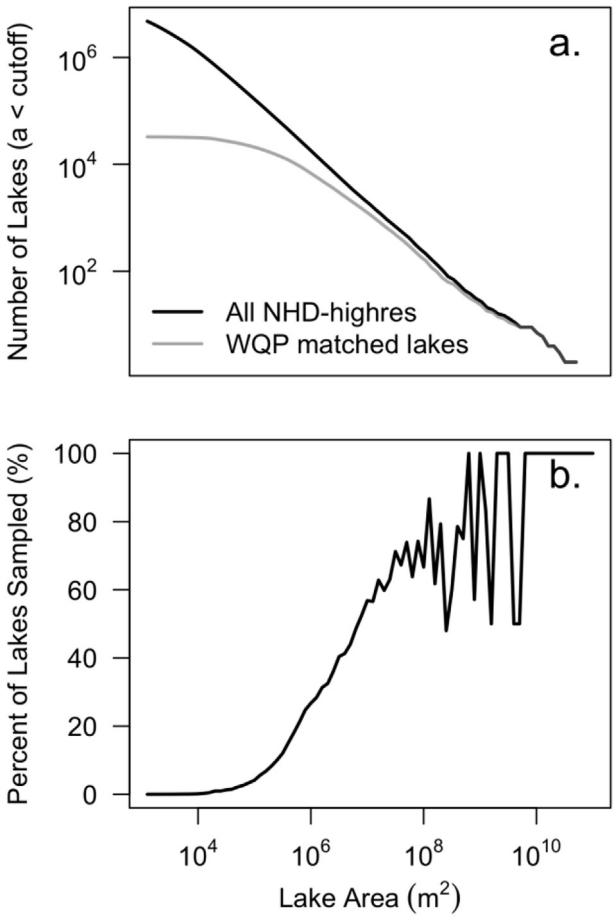

Fig. 4. The size-abundance distribution of lakes in the NHD high resolution layer and the matched Water Quality Portal (WQP) lakes. (a) Cumulative size-abundance distribution and (b) percentage of lakes. Lake observations have a large-lake bias with the majority of large lakes (over c. 10 km^2) versus the minority of small lakes (under c. 1 km^2) have archived observation data.

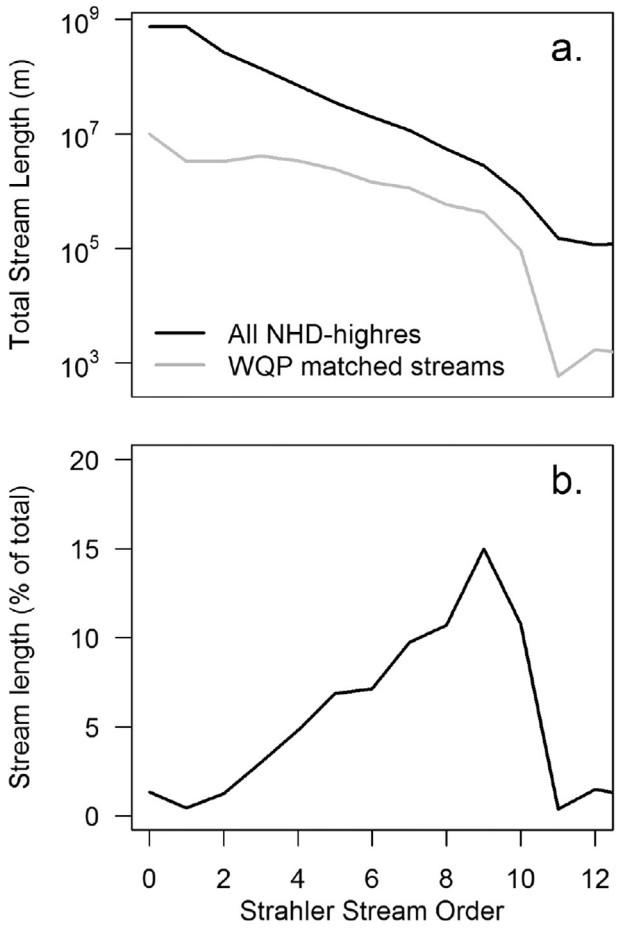

Fig. 5. Abundance distribution of monitored streams binned by stream order and summarized by total length. Panel (a) shows the total stream length connected with locations with associated Water Quality Portal (WQP) data contrasted with the total stream length for each stream order in the NHD high-resolution layer. Panel (b) shows the percent of each stream order segments with corresponding WQP observation site.

representation of inland waters within the observed data record.

The WQP example shown here highlights how the tools in *hydrolinks* can quickly link observations to value-added data products. Here, linking the WQP sites to the NHD allowed for the comparison of lake size and stream order for observed versus the full-network distribution of lakes and streams. Not surprisingly, observations are generally skewed towards larger aquatic systems (lakes and streams) as previously noted (e.g., Hanson et al., 2007), though some size-discrepancy may be due to higher required geolocation precision required for small lakes. It is difficult to make a simple comparison between the size-abundance distribution of lakes and streams. Using stream order normalized by stream length was generally effective in this situation. However, this strategy may not work well at stream orders over 10, which showed a drop in coverage that may not be representative, largely due to the limited number of these large rivers.

Further, tools for hydrologic network linking and verification for lakes and streams could be useful to both data providers and users of aquatic database data. In our example, we show a number of challenges in linking data from the WQP to the high-resolution NHD. While the specific reasons that WQP sites failed to link are unclear, these points warrant further investigation and possible correction by data providers. *hydrolinks* allows users to quickly verify the WQP latitude/longitude metadata, and slate non-linking site data for further examination or removal. For example, in the WQP data used here, the number of lake linking failures was reduced from 31.5% to 23.6% by including the 25 m buffer, indicating many sites are too-near the lake shoreline for simple point-in-polygon linking. Further, of the 21,600 lake sites that did not successfully link to a lake, 8000 can likely be explained as mislabelled stream sites (Supplemental Text 1). Using such techniques, *hydrolinks* can enable data input quality assurance procedures by providing a quick way to verify newly supplied data to the underlying WQP databases (e.g., STORET or NWIS) or identify and correct existing non-linking sites.

Different organizations frequently have their own, independent identification system for waterbodies. Further, different organizations often collect data on the same waterbodies or stream reaches but catalog and archive the data using organization-specific identification systems. For example, the U.S. Geological Survey frequently adopts the National Hydrography Network (http://nhd.usgs.gov) as a unifying identifier for aquatic data. But many state and local agencies use local identification schemes for waterbody and stream identification (e.g., Wisconsin Water Body Identification Codes, http://wi.dnr.gov). While it is unlikely that states and other local agencies will adopt a more national (or global) sample site ID procedure, *hydrolinks* can nonetheless help unify these datasets when integration is desired (e.g., Supplemental Text 2).

Without ID-based linking, name or latitude/longitude are often not sufficient or unambiguous enough to integrate hydrologic

datasets. Lake names are extremely ambiguous, with the most common lake names used in many locations (e.g., Long, Trout, Emerald, Clear, or Mud Lakes). Data taken from effectively the same lake or stream site can frequently have very different locations based on latitude and longitude. This is especially challenging when dealing with datasets from both large and small lakes. For example, in a large lake, data sampled from two locations 5 km distant can be thought of as describing the same aquatic system. On the other hand, in lake rich regions with many small lakes, a 5 km distance may cross many small, distinct, and heterogeneous lakes. Further, the loss of precision on latitude and longitude that can occur through data handling (e.g., through decimal truncation in Excel) may cause points to no longer geographically match lakes on the landscape.

Open datasets are more frequently being released using unique identifiers based on hydrologic network dataset identifiers. For example, recent lake datasets on lake depth (Oliver et al., 2016) and water temperature (Winslow et al., 2017) were released with both NHD unique identifiers and standard latitude and longitude geopoints. With *hydrolinks*, users of those datasets can quickly integrate their own and other datasets using the hydrologic network dataset identifiers. Furthermore, this tool will enable future datasets to be released using one or more broadly accessible unique identifiers to unambiguously link data to waterbodies and accelerate data integration across different published datasets.

hydrolinks joins a number of tools available to better integrate diverse aquatic data sets in scientific computing environments. Data management, especially of large and diverse data sets, can take outsized amount of time in scientific data analyses (Lohr, 2014). Tools that enable better data access and integration are at the forefront of accelerating environmental science and macrosystem ecological research. With the release of *hydrolinks*, we hope to enable better communication and integration of lake and stream datasets at large scales.

Acknowledgements

Authors thank Lesley Knoll for providing critical insight and early support that resulted in this project and the development this manuscript. S.D.P. and T.H.H. received financial assistance from the Pennsylvania Department of Conservation and Natural Resources - Wild Resource Conservation Program (grant # WRP-15533). T.H.H. also received support from New York State Energy Research and Development Authority (NYSERDA) Agreement No. 115876. L.A.W. and K.C.R. received support from the National Science Foundation (grant # MSB-1638704).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.envsoft.2018.03.012.

References

Chapra, S.C., Boehlert, B., Fant, C., Bierman, V.J., Henderson, J., Mills, D., Mas, D.M.L., Rennels, L., Jantarasami, L., Martinich, J., Strzepek, K.M., Paerl, H.W., 2017. Climate change impacts on harmful algal blooms in U.S. Freshwaters: a screening-level assessment. Environ. Sci. Technol. 51, 8933–8943. https://doi.org/10.1021/acs.est.7b01498.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Holm, E.V., Isaksen, L., Kallberg, P., Kohler, M., Matricardi, M., Mcnally, A.P., Monge-Sanz, B.M., Morcrette, J.J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J.N., Vitart, F., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137, 553–597. https://doi.org/10.1002/qj.828.

Haag, S., Shokoufandeh, A., 2017. Development of a data model to facilitate rapid watershed delineation. Environ. Model. Software 1–13. https://doi.org/10.1016/j.envsoft.2017.06.009.

Hanson, P.C., Carpenter, S.R., Cardille, J.A., Coe, M.T., Winslow, L.A., 2007. Small lakes dominate a random sample of regional lake characteristics. Freshw. Biol. 52, 814–822. https://doi.org/10.1111/j.1365-2427.2007.01730.x.

Heffernan, J.B., Soranno, P.A., Angilletta, M.J., Buckley, L.B., Gruner, D.S., Keitt, T.H., Kellner, J.R., Kominoski, J.S., Rocha, A.V., Xiao, J., Harms, T.K., Goring, S.J., Koenig, L.E., McDowell, W.H., Powell, H., Richardson, A.D., Stow, C.A., Vargas, R., Weathers, K.C., 2014. Macrosystems ecology: understanding ecological patterns and processes at continental scales. Front. Ecol. Environ. 12, 5–14. https://doi.org/10.1890/130017.

Hill, R.A., Weber, M.H., Leibowitz, S.G., Olsen, A.R., Thornbrugh, D.J., 2016. The stream-catchment (StreamCat) dataset: a database of watershed metrics for the conterminous United States. J. Am. Water Resour. Assoc. 52, 120–128. https://doi.org/10.1111/1752-1688.12372.

Hirsch, R.M., De Cicco, L.A., 2015. User guide to exploration and graphics for RivEr trends (EGRET) and dataRetrieval: R packages for hydrologic data. In: Hydrologic Analysis and Interpretation. US Geological Survey, Reston, VA, USA.

Horsburgh, J.S., Reeder, S.L., 2014. Data visualization and analysis within a Hydrologic Information System: integrating with the R statistical computing environment. Environ. Model. Software 52, 51–61. https://doi.org/10.1016/j.envsoft.2013.10.016.

Isaak, D.J., Young, M.K., Luce, C.H., Hostetler, S.W., Wenger, S.J., Peterson, E.E., Ver, J.M., Groce, M.C., Horan, D.L., Nagel, D.E., 2016. Slow climate velocities of mountain streams portend their role as refugia for cold-water biodiversity. PNAS 113, 1–6. https://doi.org/10.1073/pnas.1522429113.

Jarvis, A., Reuter, H., Nelson, A., Guevara, E., 2008. Hole-filled SRTM for the Globe Version 4, Available from the CGIAR-CSI SRTM 90m Databas [WWW Document].

Jin, S., Yang, L., Danielson, P., Homer, C., Fry, J., Xian, G., 2013. A comprehensive change detection method for updating the National Land Cover Database to circa 2011. Remote Sens. Environ. 132, 159–175. https://doi.org/10.1016/j.rse.2013.01.012.

Lohr, S., 2014. For Big-data Scientists,"janitor Work"is Key Hurdle to Insights. New York Times.

Messager, M.L., Lehner, B., Grill, G., Nedeva, I., Schmitt, O., 2016. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. Nat. Commun. 7, 13603. https://doi.org/10.1038/ncomms13603.

Moore, R.B., Dewald, T.G., 2016. The road to NHDPlus - advancements in digital stream networks and associated catchments. JAWRA J. Am. Water Resour. Assoc. https://doi.org/10.1111/1752-1688.12389, 20460, n/a-n/a.

O'Reilly, C.M., Sharma, S., Gray, D.K., Hampton, S.E., Read, J.S., Rowley, R.J., Schneider, P., Lenters, J.D., McIntyre, P.B., Kraemer, B.M., Weyhenmeyer, G.A., Straile, D., Dong, B., Adrian, R., Allan, M.G., Anneville, O., Arvola, L., Austin, J., Bailey, J.L., Baron, J.S., Brookes, J.D., Eyto, E. de, Dokulil, M.T., Hamilton, D.P., Havens, K., Hetherington, A.L., Higgins, S.N., Hook, S., Izmest'eva, L.R., Joehnk, K.D., Kangur, K., Kasprzak, P., Kumagai, M., Kuusisto, E., Leshkevich, G., Livingstone, D.M., MacIntyre, S., May, L., Melack, J.M., Mueller-Navarra, D.C., Naumenko, M., Noges, P., Noges, T., North, R.P., Plisnier, P.-D., Rigosi, A., Rimmer, A., Rogora, M., Rudstam, L.G., Rusak, J.A., Salmaso, N., Samal, N.R., Schindler, D.E., Schladow, S.G., Schmid, M., Schmidt, S.R., Silow, E., Soylu, M.E., Teubner, K., Verburg, P., Voutilainen, A., Watkinson, A., Williamson, C.E., Zhang, G., 2015. Rapid and highly variable warming of lake surface waters around the globe. Geophys. Res. Lett. 42, 1–9. https://doi.org/10.1002/2015GL066235.

Oliver, S.K., Collins, S.M., Soranno, P.A., Wagner, T., Stanley, E.H., Jones, J.R., Stow, C.A., Lottig, N.R., 2017. Unexpected stasis in a changing world: lake nutrient and chlorophyll trends since 1990. Global Change Biol. 1–13. https://doi.org/10.1111/gcb.13810.

Oliver, S.K., Soranno, P.A., Fergus, C.E., Wagner, T., Winslow, L.A., Scott, C.E., Webster, K.E., Downing, J.A., Stanley, E.H., 2016. Prediction of lake depth across a 17-state region in the U.S. Inl. Waters 6, 314–324. https://doi.org/10.5268/IW-6.3.957.

Read, E.K., Carr, L., Cicco, L. De, Dugan, H.A., Hart, J.A., Kreft, J., Read, J.S., Winslow, L.A., Hanson, P.C., 2017. Water quality data for national-scale aquatic research: the Water Quality Portal. Water Resour. Res. 53, 1–11. https://doi.org/10.1002/2016WR019993.

Read, E.K., Patil, V.P., Oliver, S.K., Hetherington, A.L., Brentrup, J.A., Zwart, J.A., Winters, K.M., Corman, J.R., Nodine, E.R., Woolway, R.I., Dugan, H.A., Jaimes, A., Santoso, A.B., Hong, G.S., Winslow, L.A., Hanson, P.C., Weathers, K.C., 2015. The importance of lake-specific characteristics for water quality across the continental United States. Ecol. Appl. 25, 943–955.

Read, J.S., Hamilton, D.P., Jones, I.D., Muraoka, K., Winslow, L.A., Kroiss, R., Wu, C.H., Gaiser, E., 2011. Derivation of lake mixing and stratification indices from high-resolution lake buoy data. Environ. Model. Software 26, 1325–1339. https://doi.org/10.1016/j.envsoft.2011.05.006.

Simley, J., Carswell, J., 2009. The National Map– Hydrography: U.S. Geological Survey Fact Sheet 2009–3054.

Sinha, E., Michalak, A.M., Balaji, V., 2017. Eutrophication will increase during the 21st century as a result of precipitation changes. Science vol. 357 (80), 405–408. https://doi.org/10.1126/science.aan2409.

Soranno, P.A., Cheruvelil, K.S., Elliott, K.C., Montgomery, G.M., 2015a. It's good to share: why environmental scientists' ethics are out of date. Bioscience 65, 69–73. https://doi.org/10.1093/biosci/biu169.

Soranno, Bissell, E.G., Cheruvelil, K.S., Christel, S.T., Collins, S.M., Fergus, C.E.,

Filstrup, C.T., Lapierre, J.-F., Lottig, N.R., Oliver, S.K., Scott, C.E., Smith, N.J., Stopyak, S., Yuan, S., Bremigan, M.T., Downing, J.A., Gries, C., Henry, E.N., Skaff, N.K., Stanley, E.H., Stow, C.A., Tan, P.-N., Wagner, T., Webster, K.E., 2015b. No TitleBuilding a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse. GigaScience 4, 1–15.

Winslow, L.A., Hansen, G.J.A., Read, J.S., Notaro, M., 2017. Data Descriptor: large-scale modeled contemporary and future water temperature estimates for 10774 Midwestern U.S. Lakes. Sci. Data 1–11. https://doi.org/10.1038/sdata.2017.53.

Winslow, L.A., Read, J.S., Hansen, G.J.A., Hanson, P.C., 2015. Small lakes show muted climate change signal in deepwater temperatures. Geophys. Res. Lett. 42, 355–361. https://doi.org/10.1002/2014GL062325.

Winslow, L.A., Read, J.S., Hanson, P.C., Stanley, E.H., 2014. Lake shoreline in the contiguous United States: quantity, distribution and sensitivity to observation resolution. Freshw. Biol. 59, 213–223. https://doi.org/10.1111/fwb.12258.