Connecting rivers and lakes FAIRly

Sami Domisch[1], Vanessa Bremerich[1], Afroditi Grigoropoulou[1], Maria M. Üblacker[1], Jaime Garcia Marquez[1], Yusdiel Torres[1], Thomas Tomiczek[1], Giuseppe Amatulli[1], Igor Ogashawara[1], Hans-Peter Grossart[1], Martin Friedrichs-Manthey[2]

[1]Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Department of Ecosystem Research, Müggelseedamm 310, 12587 Berlin, Germany
[2]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstraße 4, 04103 Leipzig, Germany

Date of submission: 31.03.2023

Abstract

Freshwater research is characterized by a particular challenge: the different freshwater habitat types, such as rivers and lakes, are treated largely separated although being embedded within the freshwater continuum. This challenge impacts an array of research disciplines, all above understanding the impact of climate change and water scarcity for water management, biodiversity and society. The problem can be traced back to the geospatial delineation of rivers and lakes, stemming from different algorithms and processing pipelines. The proposed pilot project aims to address this challenge by (i) connecting latest high-resolution river and lake geospatial datasets globally by delineating the individual drainage basins contributing to lakes, and by (ii) pro-actively deploying functions that are interoperable with the upcoming NASA SWOT high-resolution water body data. Moreover, (iii) the pilot project will tap on a recently published national fish dataset to demonstrate the potential and feasibility of the approach for describing species distributions. All newly-developed data and functions will be FAIRly onboarded onto the existing GeoFRESH online platform and the *hydrographr* R-package, maximising the uptake by the research community who can integrate these easily into their analysis workflows. The pilot project has the high potential to provide a significant stepping stone towards building a seamless freshwater continuum by connecting data and tools in terms of river and lake research, and to strengthen the interlinkage between the NFDI4Earth and NFDI4Biodiversity consortia.

I. Introduction

Freshwater ecosystems are characterized by the unique lateral and longitudinal connectivity among habitats that drives a multitude of processes such as discharge, sediment transportation, nutrient and pollution dynamics, or species dispersal[1]. Addressing the connectivity comprises a crucial cornerstone for analysing and interpreting patterns and processes in Earth System Science (ESS), water management and freshwater biodiversity, yet its integration into analysis workflows – with the exception of hydrology – has been sluggish. This slow development can be attributed to the computationally intense data preparation and analyses that account for connectivity and spatial autocorrelation, requiring advanced knowledge in GIS programming[2]. Recent efforts to address the

lack of easy-to-use data and tools regarding the seamless connectivity at unprecedented high spatial resolution have resulted, for instance, in the Hydrography90m dataset[3] that provides a global stream network at 90m spatial resolution, which, in conjunction with the *hydographr* R-package, provides the basis to perform scalable and standardized high-resolution freshwater-related geospatial analyses without the need to learn a variety of GIS programming tools.

Lakes make up ca. 3% to the Earth's surface (opposed to the <1% share by streams and rivers) and play a crucial role in freshwater ecosystems. They contribute towards biodiversity, water quality, discharge dynamics, groundwater replenishment, and act as sediment and nutrient traps. Rivers and lakes consist of fundamentally different habitat types, and from a data and research perspective, they are typically treated separated. Yet they are not isolated, but embedded within the larger freshwater continuum. This separation can be traced back to the different algorithms for delineating a river network (flow routing) and the lake outline (remote sensing of water bodies). Though these data types overlap spatially, they are not linked and the water flow can not be routed seamlessly.

Considering climate change and water scarcity with its cascading impacts, it is obvious that this long-standing separation from a spatial connectivity perspective needs to be overcome. A first step towards this challenge is to geospatially connect rivers and lakes. When delineating river networks across lakes, they usually show a fishbone, or striped pattern (Fig. 1) while still portraying the lake inflow (contributing drainages) and pour point (outflow) correctly. When overlaying remotely-sensed (and thus spatially-explicit) lake boundary information, these two data types, i.e. network and lake boundary, can be dissolved to create a seamless hydrographical network integrating lakes, which is the objective of the pilot project.

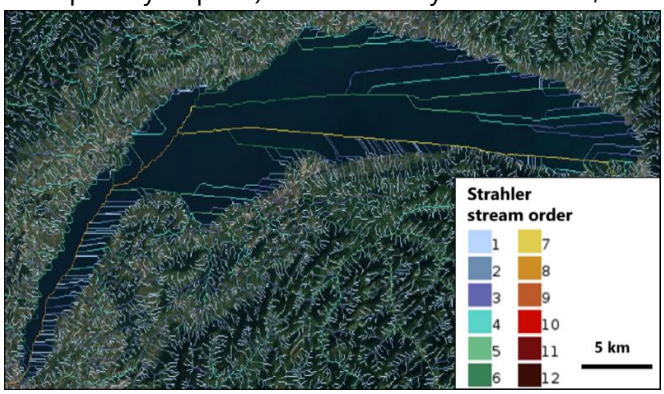

Figure 1 | The Hydrography90m stream network plotted on Lake Geneva in Switzerland. Note the fishbone pattern of the network across the lake surface. Online visualization available at geo.igb-berlin.de/maps/351/view.

Current best practices are given by the HydroLAKES[4], HydroRIVERS[5] and the LakeATLAS datasets[6] that provide valuable hydro-environmental features across 1.4M lakes globally. While undoubtably useful, small streams are not adequately depicted in the underlying network, although they contribute to the total stream length by 70%[ref.7]. The recently-developed Hydrography90m stream network[3] addresses this challenge by providing a high-resolution network delineation, such that small water bodies and the environmental as well as biodiversity features can be linked to specific headwater stream reaches.

Our vision is to support researchers in the Earth System and biodiversity sciences by developing geospatial data processing tools that connect river and lake research. The pilot project has the high potential to support multiple research disciplines simultaneously, advance the interoperability and to strengthen the ties between the NFDI4Earth and NFDI4Biodiversity consortia.

I. Pilot description

Our proposed 1-year pilot project aims to provide a seamless freshwater continuum across rivers and lakes. Moreover, a case study will employ a newly-collated freshwater fish dataset and hence demonstrate the potential of the geospatial tools towards freshwater sciences across disciplines. To this end, we have performed important, initial explorative analyses that demonstrate the high potential and feasibility of the approach (see online visualisation at https://geo.igb-berlin.de/maps/244/view, showing the overall as well as top-contributing drainage basins of Lake Titikaka in South America).

Regarding the methodological workflow we will follow a two-pronged approach. We will first finalize our preliminary processing pipeline and apply our workflow using the HydroLAKES dataset[4] to integrate the 1,4M lakes with a size of >10ha globally onto the Hydrography90m network. Second, we will provide our tools as generic functions to be onboarded on the GeoFRESH platform and the *hydrographr* R-package such that users can integrate custom lake / water body data into the network. This approach has the benefit in yielding (i) ready-to-use and FAIR data across the seamless river-lake continuum globally that can be instantly used in a given analysis workflow, and (ii) contribute towards Open Science by providing the code that can be further tailored to specific needs. Given that the NASA Surface Water and Ocean Topography (SWOT) mission will release remotely-sensed information of standing water bodies >6ha in size in 2023/2024 (the satellite was launched in December 2022), our aim is to be pro-active: we expect the SWOT data to be a game-changer in freshwater research, as it will allow to monitor global water storage changes every 21 days, contributing significantly to a high-resolution representation of the Earth's surface freshwater resources.

The technical backbone of the pilot project is provided by the openly available Hydrography90m stream network[3], the HydroLAKES[4] dataset and an open-source toolbox of geo-computational database and programming software (PostgreSQL, Postgis and pgRouting, GeoNode[8], GRASS-GIS[9], GDAL/OGR[10], and GuidosToolbox Workbench[11]), which are all characterized by their high degree of interoperability, reproducibility, and fast and scalable processing capabilities within the Linux environment, providing the essential building blocks towards Open Science and FAIR data[12]. The computation will be done in collaboration with the Yale HPC. We will start with lakes across Germany prior upscaling the analysis to the global scale, as to demonstrate the potential and feasibility of the tools towards describing the distributions of the German freshwater fish fauna as a case study. This harmonised freshwater fish occurrence and abundance data for 12 federal states in Germany[13] consists of 174897 occurrence records for 72 fish species in rivers as well as lakes, and hence provides an ideal foundation for a case study providing public analysis workflows to the research community, and highlighting the linkage between the two NFDI consortia.

The pilot project focuses on surface water-fed lakes, and we highlight the potential of this initiative as a "door opener" towards future research avenues to account for e.g. groundwater-fed lakes, further develop the processing workflow to address the particularities of deep versus shallow lakes, or extend residence time analyses from lakes to the entire freshwater continuum.

II. Relevance for the NFDI4Earth

The pilot project addresses a long-overdue component in geospatial freshwater-related Earth System and biodiversity sciences. The resulting output, i.e. the data as well as the functions to be used on custom data, maximizes the uptake potential by the research community given that both users as well as developers can capitalize on the output. We expect that scientists, data curators, university teachers to be stakeholders that can directly benefit from the pilot project.

The proposed pilot project builds on the GeoFRESH (NFDI4Earth) and the *hydrographr* R-package (NFDI4Biodiversity). Opposed to merely adding incremental steps, this pilot project will widen the horizon significantly by integrating further research disciplines and allowing researchers from ESS and biodiversity to benefit alike. Moreover, the pilot project will create valuable synergies with the European Open Science Cloud (EOSC) and the ongoing AquaINFRA project (https://aquainfra.eu/) that aims to create virtual research environments tailored towards aquatic sciences.

The uptake will be supported by integrating the data and tools into the GeoFRESH platform and the *hydrographr*-package. Both provide scalable analysis workflows that harness powerful open-source software that remain, however, out of scope to most researchers given the required advanced programming skills. Users of GeoFRESH and the *hydrographr*-package can therefore benefit from the software in the backend without having to interact with it directly. The pilot project will follow this trajectory and provide data and easy-to-use functions to be directly integrated into analysis workflows. The pilot project tackles all elements of the FAIR criteria[12], as all data and code will be publicly stored in the IGB's Freshwater Research and Environmental Database, FRED, and on GitHub, respectively.

III. Deliverables

The pilot project will yield four deliverables: (i) the technical operability of the pilot in terms of the global lake catchment data, (ii) integrating the analysis tools into the GeoFRESH platform and the *hydrographr* R-package, (iii) a vignette and tutorial exemplifying the case study workflow, and (iv) a roadmap document entitled "*Towards a seamless geospatial representation of freshwater habitats*".

IV. Work Plan & Requested funding

The tasks of the one-year pilot project (Table 1) will be undertaken by a postdoctoral researcher specialized in freshwater geospatial analyses. We hence request the pre-defined budget of €71.163,36.

Work phases of the pilot project	Q1			Q2			Q3			Q4		
	1	2	3	4	5	6	7	8	9	10	11	12
Finalizing analysis workflow												
Global lake catchments												
Case study analyses												
Functions for GeoFRESH & hydrographr												
Documentation / vignettes												
Roadmap document												

Table 1 | Planned work phases of the pilot project for the postdoctoral researcher (shaded in blue). Q refers to quarters; numbers refer to months.

References: [**1**] Wohl, E. (2017) *Prog Phys Geogr* 41, 345-362 [**2**] Peterson, E.E. et al. (2013) *Ecol. Lett.* 16, 707-719 [**3**] Amatulli, G. et al. (2022) *Earth Syst. Sci. Data* 14, 4525-4550 [**4**] Messager, M.L., et al. (2016) *Nat Commun* 7, 13603 [**5**] Lehner, B. & Grill, G. (2013) *Hydrol Process* 27, 2171-2186 [**6**] Lehner, B., et al. (2022) *Sci Data* 9, 351 [**7**] Lowe, W.H. & Likens, G.E. (2005) *BioScience* 55(3), 196-197 [**8**] Corti, P. et al. (2019) *PeerJ Preprints* 7, e27534v27531 [**9**] Neteler, M., et al. (2012) *Environ Modell Softw* 31, 124-130 [**10**] GDAL/OGR Geospatial Data Abstraction Library (2022) [**11**] Vogt, P. et al. (2022) *Ecography* e05864 [**12**] Wilkinson, M.D. et al. (2016) *Sci Data* 3, 160018 [**13**] Leibniz Institute for the Analysis of Biodiversity Change (LIB) (2022) Occurrence dataset https://doi.org/10.15468/c75fky accessed via GBIF.org on 2023-03-26.