Getting freshwater spatiotemporal data on track

Sami Domisch[1],Giuseppe Amatulli[2], Vanessa Bremerich[1], Luc De Meester[1], Mark Gessner[1], Hans-Peter Grossart[1], Rita Adrian[1]

[1]Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Müggelseedamm 301, 12587 Berlin, Germany
[2]Yale University, Centre for Research Computing, New Haven, CT, 06511, USA

Date of submission: July 24th 2020

Abstract

Spatiotemporal freshwater-related earth system data are currently poorly organized and its full potential for research or management is rarely exploited, due to data disparity and its missing interoperability given the different data standards and formats. It is especially the spatial structure of water bodies, i.e. the river network and lakes with time legacy effects that require a specialized workflow for earth system data integration into freshwater research. Engaging different freshwater-related research disciplines such as hydrology, chemistry, geography, remote sensing, climatology and ecology effectively within a single framework poses an urgent prerequisite for an effective FAIR data management of environmental and biodiversity data, especially in the light of climate and land use changes and feedback mechanisms between earth systems.
Here we propose a full start-to-finish pilot project that optimizes the integration of earth system data into freshwater research and for earth system science in general by accounting for the crucial, yet often neglected connectivity within freshwater water bodies themselves and their terrestrial catchments. The central tool will be the new "GeoFRESH" online platform that provides the integration, processing, management and visualization of various standardized spatiotemporal freshwater-related earth system data. The platform allows exploiting the full potential of environmental, physical and biodiversity freshwater data along the hydrographical network, and will deliver a critical tool for stakeholders from research, sustainable water management and for monitoring freshwater ecosystem services alike. In addition, the data interoperability will allow informing NFDI$_4$Earth by enhancing the freshwater-specific data exchange among realms. The proposed pilot project represents a long-overdue component in freshwater-related FAIR data management, and once established, will prove itself as an essential key element for NFDI$_4$Earth given its high potential for advancing a seamless freshwater data integration across earth system disciplines.

I. Introduction

Freshwater ecosystems are characterized by the connectivity of water bodies within the river network, as well as their high degree of fragmentation where isolated lakes and ponds can be considered as islands analogous to the terrestrial realm[1,2]. Integrating earth system data into freshwater research and management and vice versa represents currently a major challenge given this (i) specific spatial structure and (ii) typical time legacy effects, with (iii) system-specific differences in these relationships between flowing (rivers) and standing (lakes, ponds and wetlands) systems. More specifically, the spatial structure derives from a spatial nestedness in catchments combined with the dendritic network structure of rivers, where legacy effects in time

are due to flowing water and retention time[3,4]. Because of these freshwater-specific characteristics, any simple data overlay of e.g. “terrestrial” land cover, climate or dams at a given lake or stream reach position is likely to provide only a fraction of the potential dependencies between these drivers and the freshwater environment[5]. Instead, a large-scale framework is required that routes the environmental information through the hydrographical network taking the time legacy effects into account. This is critical because the effects of changes in (i) the environment, such as high nutrient loadings in upstream croplands, (ii) connectivity via dams or weirs, or (iii) extreme events such as floods, droughts, or forest fires within the upstream catchments, cascade into water bodies located downstream, impacting the habitat conditions and the freshwater communities. Here, data that has been “freshwaterized”, i.e., tailored towards freshwater ecosystems substantially improves research and scenario development in freshwater systems[6]. By better integrating the specificities of the spatial structuring and temporal dependencies in freshwater systems, the improved data integration and scenarios for freshwater systems will also boost NFDI$_4$Earth research in climate (cf. climate feedbacks) and terrestrial systems (cf. interdependencies of terrestrial and inland systems). Currently, however, freshwater-specific spatiotemporal data is rarely harmonized but scattered across different repositories. The kitchen sink of data sources, types and their varying degree of operationality and interoperability poses a major challenge in their usability. Hence a large fraction of its full potential is not used.
Here we propose a fully integrated online “GeoFRESH” platform that provides the integration, processing, management and visualization of various standardized spatiotemporal freshwater-related earth system data. GeoFRESH will ensure that a rich suite of data relevant for freshwater ecosystems will be available and interoperable for researchers and stakeholders from other NFDI$_4$Earth systems by maximizing the FAIR[7] principles.

II. Pilot description

The proposed pilot project seeks to overcome the above outlined challenges by (i) preprocessing and tailoring data towards freshwater systems, and (ii) by building the new open-source GeoFRESH platform for hosting the data and enabling FAIR data management and interoperability among NFDI$_4$Earth systems. Within the pilot we demonstrate the effectiveness of this approach for two case studies: the prediction of freshwater cyanobacterial (blue-green) algal blooms in lakes globally (long-term data: Geisha-database; https://www.geisha-stormblitz.fr/), and predicting freshwater fish species habitat suitability[8] in South America.

For the pre-processing we will employ a suite of open-source geo-computational programming tools and rely on the highly versatile GRASS-GIS[9], GDAL/OGR[10], and pktools[11] software which are characterized by their high degree of interoperability, reproducibility, and fast and scalable processing capabilities. Data will be harmonized towards (i) the data standard (i.e., GeoPackage for vector, GeoTiff and NetCDF for raster), (ii) the extent and spatial resolution to a common standard, and (iii) by adding the respective water body identifier as an additional attribute. Here, we will capitalize on a novel, global hydrographical network at 90 m spatial resolution that is currently being developed in collaboration between IGB and Yale University (demo available at http://spatial-ecology.net/hydrography-demo/). Each water body, i.e., sub-catchment, stream reach, lake and reservoir, is treated as a discrete unit with a unique identifier. By employing graph theory, the routing and tracing of information while traversing through the network can be linked to each single data set, allowing the retrieval of information related to each water body

unit, taking longitudinal connectivity and time legacy effects into account. Where needed, data interpolation of point measurements along the river network will be accomplished by the open-STARS[12] and the SSN[13] tools in GRASS-GIS and R[14]. Building on (i) latest data products and methods developed in-house at IGB enables a very flexible handling of data and routines, thus highlighting the high feasibility of the planned work. Moreover, the pilot project will (ii) build on the IGB's long-term ecological research programme based on existing long-term data of lakes and rivers (Freshwater Research and Environmental Database, FRED; https://fred.igb-berlin.de/), (iii) extends the already established spatial and research data management infrastructure at IGB, and (iv) takes advantage of the strong expertise at IGB in state-of-the-art statistical modelling in space and time.

The central tool of the pilot project will be the new GeoFRESH platform which will be built around GeoNode[15] (https://geonode.org/) (available at IGB at: https://geo.igb-berlin.de/), which is an open-source framework designed to build stable and scalable spatial data infrastructures (SDI). It offers a consistent and easy-to-use interface, allowing also non-specialized users to find, process, share or download data and easily create thematic interactive maps from the available spatial layers. It contains an open geospatial data catalogue, which exposes metadata using an Open Geospatial Consortium (OGC) compliant Catalogue Service for the Web (CSW) to provide search capabilities and ensure interoperability within larger SDI networks across NFDI$_4$Earth. GeoNode relies on PostgreSQL with a PostGIS (https://postgis.net/) extension as a spatial database for vector data storage and GeoServer (http://geoserver.org/) as a spatial data server. They support all major vector and raster formats and supply geospatial information using standard OGC protocols such as Web Map Service (WMS), Web Feature Service (WFS) or Web Coverage Service (WCS), thereby providing broad interoperability and the possibility to directly connect to the data with a local GIS client to perform GIS analyses or create advanced cartographic representations. Most importantly for NFDI$_4$Earth, both GeoServer and PostGIS offer a multitude of powerful and flexible geospatial operations, which will allow performing additional processing steps on the uploaded data. For instance, users will be able to retrieve data tailored to freshwater ecosystems, intersect points or areas of interest and thus including upstream areas and taking longitudinal connectivity and time legacy into account within their specific analyses.

The case studies will tap on available species data at IGB (FRED, GEISHA; https://www.geisha-stormblitz.fr/), and we will employ machine-learning (ML) algorithms given their capacity to harness massive data, where traditional techniques would succumb due to computational constraints[16,17]. Moreover, ML techniques have proven to be viable solutions regarding the envisaged case study analyses[18,19]. All code and helper scripts for users will be deposited in the open-source GitHub repository. Taken together, **the innovation of the proposed pilot project builds upon initiating a unique freshwater data management and processing platform** to maximize the reusability and interoperability of freshwater-specific data among Earth Systems.

III. Relevance for the NFDI4Earth

The GeoFRESH pilot project will have far-reaching impacts on future freshwater research as well as research on Earth Systems in general, and has a high uptake potential for a variety of stakeholders. We identified scientists, data curators, and university teachers as direct potential

users who would interact with the platform, public authorities and decision-makers to make use of the flexible visualization and dissemination tools, and IGB as the infrastructure provider.

The streamlining of the data and workflows has the great potential to enhance the research data life cycle. After a successful pilot implementation phase, we envision the GeoFRESH platform to be a strong contribution to NFDI$_4$Earth by providing data interoperability for a variety of long-term research questions. This includes, for instance, (i) an improved understanding of the effects of climate and land use change on freshwater habitats and organisms[20], (ii) by examining substances entering freshwaters[21], (iii) assessing sediment transportation and the effects of altered connectivity due to dam construction[22], and (iv) how changes in water flow (droughts, floods) affect biodiversity[22]. More broadly, (v) we envision GeoFRESH to be a key for enabling large-scale scenario projections regarding ecosystem services in collaboration with the ARIES platform (http://aries.integratedmodelling.org/)[23]. By capitalizing on the interoperability among NFDI$_4$Earth disciplines, (vi) we expect GeoFRESH to strongly contribute towards an improved understanding of the global carbon cycle and fluxes in greenhouse gas emissions given the larger than expected outgassing of carbon dioxide and also methane from freshwaters to the atmosphere[24,25]. Finally, (vii) we envision that GeoFRESH has the high potential to contribute to the ISIMIP initiative (https://www.isimip.org/) via global freshwater-specific data and model outputs.

The pilot project tackles all elements of the FAIR criteria[7]. The data will be (i) *findable* on GeoFRESH (or if no publishing allowed, will be properly referenced), (ii) *accessible* through (batch) download, (iii) *interoperable* given the SDI among databases as well as the support of all data standards and geospatial computational tools, and (iv) *reusable* with a rich description of relevant metadata attributes and clear and accessible data usage licenses. Wherever possible, all data will be hosted on IGB servers, adhering to the open-access regulations of the data.

IV. Deliverables

The pilot project will yield four deliverables: (i) the GeoFRESH platform including the open-source geo-computational routines, (ii) technical guidelines and "how-to" vignettes for processing data, (iii) a roadmap document entitled "*Getting freshwater spatiotemporal data on track – towards the interoperability of freshwater-specific data*", and (iv) initiating a peer-refereed publication for dissemination purposes, exemplifying the effectiveness of the platform.

V. Work Plan & Requested funding

The tasks of the one-year pilot project (Table 1) will be undertaken by a data steward specialized in web GIS (100% E11 / level 4 position; €75.718 salary) and a postdoctoral researcher experienced in spatiotemporal modelling (50% E13 / level 4 position; €42,425 salary). The requested funding therefore amounts to €118,143. The IGB holds all necessary infrastructures, i.e., compute and storage servers within its domain, and contributes an additional €50,000 for personnel during the initial development phase of GeoFRESH.

Work phases of the pilot project	**Q1**			**Q2**			**Q3**			**Q4**		
	1	2	3	4	5	6	7	8	9	10	11	12
Data mobilization and harmonization												
Routing of environmental data												
Batch processing for GeoFRESH												
Technical guide documents												
Data integration												
Case study analyses												
Roadmap document												
Publication												
Interoperability of NFDI4Earth Pilots												

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

VI. References

1 Fagan, W. F. Connectivity, fragmentation, and extinction risk in dendritic metapopulations. *Ecology* **83**, 3243-3249 (2002).

2 Mims, M. C., Olden, J. D., Shattuck, Z. R. & Poff, N. L. Life history trait diversity of native freshwater fishes in North America. *Ecol Freshw Fish* **19**, 390-400 (2010).

3 Lehner, B., Verdin, K. & Jarvis, A. New global hydrography derived from spaceborne elevation data. *Trans Amer Geophys Union* **89**, 93–94 (2008).

4 Messager, M. L., Lehner, B., Grill, G., Nedeva, I. & Schmitt, O. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nat Commun* **7**, 13603 (2016).

5 Domisch, S., Jähnig, S. C., Simaika, J. P., Kuemmerlen, M. & Stoll, S. Application of species distribution models in stream ecosystems: the challenges of spatial and temporal scale, environmental predictors and species occurrence data. *Fundam Appl Limnol* **186**, 45–61 (2015).

6 Domisch, S., Amatulli, G. & Jetz, W. Near-global freshwater-specific environmental variables for biodiversity analyses in 1 km resolution. *Sci Data* **2**, 150073 (2015).

7 Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).

8 Elith, J. & Leathwick, J. R. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annu Rev Ecol Evol S* **40**, 677-697 (2009).

9 Neteler, M., Bowman, M. H., Landa, M. & Metz, M. GRASS GIS: A multi-purpose open source GIS. *Environ Modell Softw* **31**, 124-130 (2012).

10 GDAL/OGR Geospatial Data Abstraction Library. Open Source Geospatial Foundation. http://gdal.osgeo.org (2020).

11 McInerney, D. & Kempeneers, P. *Open source geospatial tools. pktools is available at http://pktools.nongnu.org*, (Springer, 2014).

12 Kattwinkel, M. & Szöcs, E. openSTARS: An Open Source Implementation of the 'ArcGIS' Toolbox 'STARS'. R package version 1.1.0. https://CRAN.R-project.org/package=openSTARS (2018).

13 Hoef, J. M. V., Peterson, E. E., Clifford, D. & Shah, R. SSN: An RPackage for Spatial Statistical Modeling on Stream Networks. *J Stat Softw* **56** (2014).

14 R Development Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org (2020).

15 Corti, P. *et al.* GeoNode: an open source framework to build spatial data infrastructures. *PeerJ Preprints* **7**, e27534v27531 (2019).

16 James, G., Witten, D., Hastie, T. & Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*. Vol. 103 (Springer, 2017).

17 Al-Jarrah, O. Y., Yoo, P. D., Muhaidat, S., Karagiannidis, G. K. & Taha, K. Efficient Machine Learning for Big Data: A Review. *Big Data Res* **2**, 87-93 (2015).

18 Olden, J. *et al.* Machine Learning Methods Without Tears: A Primer for Ecologists. *Q Rev Biol* **83**, 171-193 (2008).

19 Thomas, M. K., Fontana, S., Reyes, M., Kehoe, M. & Pomati, F. The predictability of a lake phytoplankton community, over time-scales of hours to years. *Ecol Lett* **21**, 619-628 (2018).

20 Sala, O. E. *et al.* Global biodiversity scenarios for the year 2100. *Science* **287**, 1770-1774 (2000).

21 Shen, L. Q., Amatulli, G., Sethi, T., Raymond, P. & Domisch, S. Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. *Sci Data* **7**, 161 (2020).

22 Zarfl, C., Lumsdon, A. E., Berlekamp, J., Tydecks, L. & Tockner, K. A global boom in hydropower dam construction. *Aquat Sci* **77**, 161-170 (2014).

23 Villa, F. *et al.* A methodology for adaptable and robust ecosystem services assessment. *PLoS One* **9**, e91001 (2014).

24 Allen, G. H. & Pavelsky, T. M. Global extent of rivers and streams. *Science* (2018).

25 Bodmer, P., Wilkinson, J. & Lorke, A. Sediment Properties Drive Spatial Variability of Potential Methane Production and Oxidation in Small Streams. *J Geophys Res Biogeosciences* **125** (2020).