The NETLAKE Metadatabase—A Tool to Support Automatic Monitoring on Lakes in Europe and Beyond

Eleanor Jennings, Elvira de Eyto, Alo Laas, Don Pierson, Georgina Mircheva, Andreja Naumoski, Andrew Clarke, Michael Healy, Kateřina Šumberová and Daniel Langenhaun

Abstract

Sharing data is a keystone of collaborative science. A fundamental barrier, however, can be a lack of knowledge on what is being collected, where, and by whom. The aim of NETLAKE (COST Action ES1201) was to build a network of sites and individuals to support development and deployment of automatic sensor-based systems on lakes and reservoirs in Europe. To support this, NETLAKE developed a metadatabase which could provide answers to questions on where lakes were monitored, details on the frequency and duration of monitoring, contact details, and which sensors were being used. Development included challenges related to time and resources, and indeed to communication between lake scientists and database experts. In total, metadata for 71 European lakes were captured. The resulting data revealed interesting facts; for example, seven sites had archives that spanned over a decade, only seven of these lakes were used as drinking water sources, and one was a large fish pond. GLEON, the Global Lake Ecological Observatory Network, and two pan-American projects are now adding their metadata and the metadatabase is developing into a tool for the global community which can promote high frequency monitoring and facilitate network science.

Introduction

Sharing data is one of the cornerstones of collaborative science. Over the last few decades, the increased use of automated sensors for lake monitoring (for reviews see Marcé et al. 2016; Meinson et al. 2015), and the availability of more and more "Big Data" from these systems, has fuelled a parallel increase in collaborative network science amongst limnologists, for example, through the Global Lake Ecological Observatory Network, GLEON (www.gleon.org; Weathers et al. 2013). A fundamental barrier to collaborative science, however, can be answering fairly basic questions such as: "Is there anybody out there?" or "What data are being collected, where, and by whom?" Automatic systems have been used to monitor lakes since the 1970s, but the technology became more widely used once more sophisticated control and communication

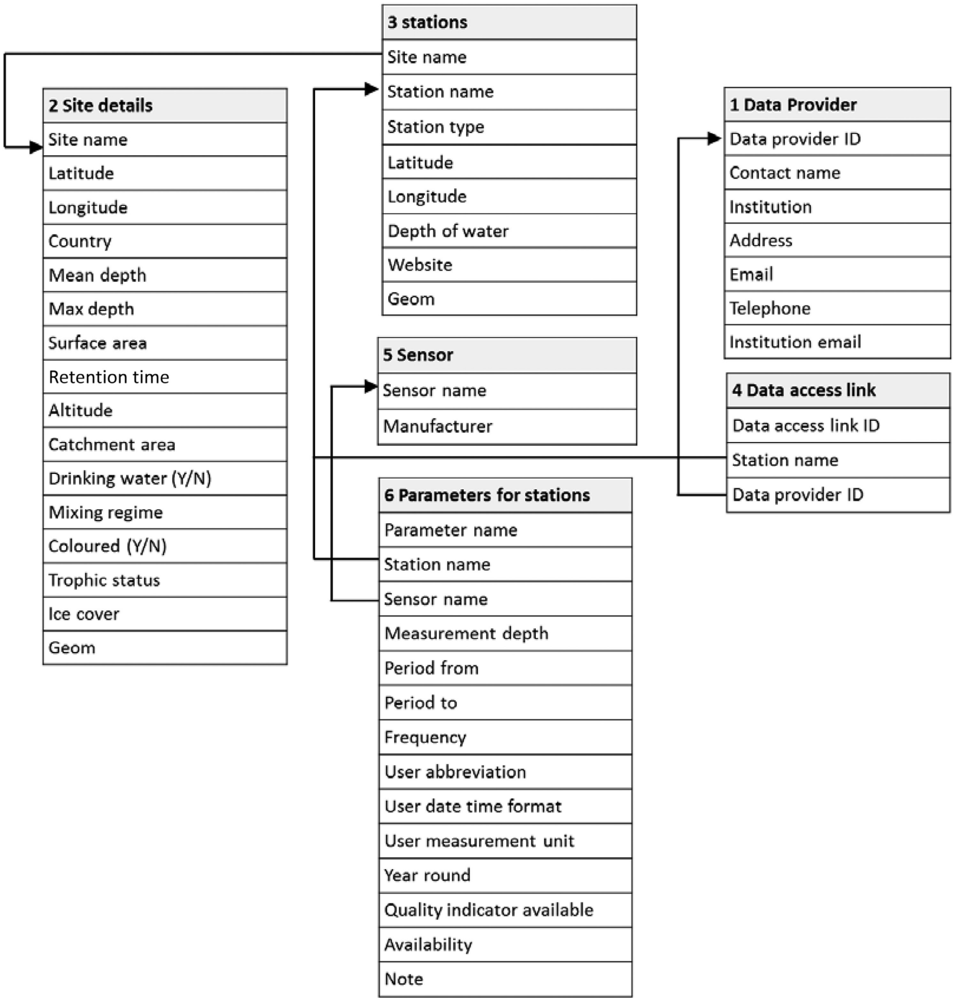

FIG. 1. Schema showing the metadatabase tables and relationships.

systems had been developed in the 1990s. Some of the most sophisticated systems ever produced in Europe were those developed using European Union (EU) LIFE funding in the late 1990s (Rouen et al. 2000). These were later used to support two larger EU projects on the effects of climate change on lakes (George 2010) and were designed to record a full suite of meteorological variables as well as a number of water quality attributes. Most systems have been deployed in natural lakes, but they are also now being used in the management of water supply reservoirs (Marcé et al. 2016).

At the time of the inception of the NETLAKE COST Action (Networking Lake Observatories in Europe, ES1201), a number of these original European systems were still in use as part of national projects, and many of the scientists involved in their management were members of GLEON, which was founded in 2005. However, it was recognised by the NETLAKE Action proponents that to undertake collaborative network science on lakes in Europe and beyond, there was a need to link both data recorded by these existing systems and their data providers to the measurements acquired in other lakes where newer monitoring platforms were being deployed. A common database for lake monitoring systems had long been an aim of GLEON, but that network had also recognised the numerous challenges that existed to making such a database a reality, not least being institutional and in some cases national restrictions on the sharing of data. Capturing information on the lake monitoring systems themselves in a metadatabase that could be used by the global lake community was seen by the NETLAKE participants as an equally important objective and was one of the key aims of NETLAKE, which ran from 2012–2016. Here, we describe how we developed this tool and the types of metadata that were captured in it, in particular that for European lakes (www.dkit.ie/netlake/netlake-resources/netlake-metadatabase).

Challenges identified and met

The NETLAKE Action had a four year timeframe within which to meet its objectives. Mindful of this time limit, it identified a number of key challenges to development of the metadatabase in initial meetings and how these might be best addressed.

Restricted resources: Limitations on personnel time and budget can be key obstacles in getting any project completed, and that included this metadatabase. EU COST Actions provide funding only for networking activities between researchers (i.e., workshops, meetings, and inter-partner visits), but not budget, for example, for personnel or IT resources. With this in mind, the NETLAKE working group on Data Acquisition and Management prioritized the metadatabase design as a task that needed to be tackled early in the four year Action in a structured manner, if it was to be brought to completion. The group also decided to use an open source database system and database administration software to ensure that the metadatabase was more sustainable after the Action lifetime.

Metadatabase design: The metadatabase design was tackled in the first year of development, recognizing that having an appropriate physical design that worked for the researchers who would use it was essential for its effectiveness and efficiency. The design was formulated during three expert opinion workshops which included both lake scientists and participants with expertise in metadatabase design. It is important to emphasize the degree of collaboration this required between the experts from the different disciplines, especially at the early stages when differing groups each used their own language and buzzwords. As a first step in design, it was important to define the metadatabase entities (tables) as described in the final schema (Fig. 1). This was followed by a second step, where the expert panel had to discuss and define the primary and foreign key constraints for each table. A third and final step then required the lake experts to again work in tandem with the database experts to define the relationships between these tables. Resolving issues that arose here had important consequences for the final database design complexity and functionality. Questions like: "How many parameters are measured for each station?" and "How many stations contain the same sensor?" needed discussion, understanding, consensus, and agreement. In this way, the working group defined the relationships between the entities in the metadatabase (one:one (1:1), one:many (1:m) and many:many (m:m)), thus laying the groundwork for the database programmer to finalize the design. The final schema was written in the open source database management system PostgreSQL (www.postgresql.org) before being again subject to scrutiny and testing by participants from both disciplines.

Metadata input: The next step, the database population, was a shared task undertaken using the adminer open source software, which also provided an interface for database management and querying (www.adminer.org). Data input was organized remotely using a moderator who issued passwords to designated data providers for each site. Using site-based personnel also ensured that the workload was shared and that the person who knew most about the system was entering the data, thereby reducing the chance for errors. An instruction manual with step-by-step instructions for data input in adminer was distributed. A second manual for querying the metadatabase both with adminer and also using Quantum GIS, the open source GIS platform (QGIS Development Team, 2017), was also developed. The online access for querying was by a separate "read only" password, which allowed the user to query and output metadata from the database, but not to edit it.

Encouraging data providers: An advantage of developing the metadatabase as part of an EU COST Action was that all participants had officially signed a Memorandum of Understanding, as national representatives, to support the Action objectives. This, together with a timetable that recognised the already heavy workloads of researchers, meant that metadata input for the European sites was largely complete by the end of the Action in October 2016. Data input at a global level has been slower, but is being integrated through GLEON. It was also recently boosted by participation of all sites included in the pan-American SAFER (Sensing the Americas' freshwater ecosystem risk from climate change (CRN 3038)) and Pampa 2 projects.

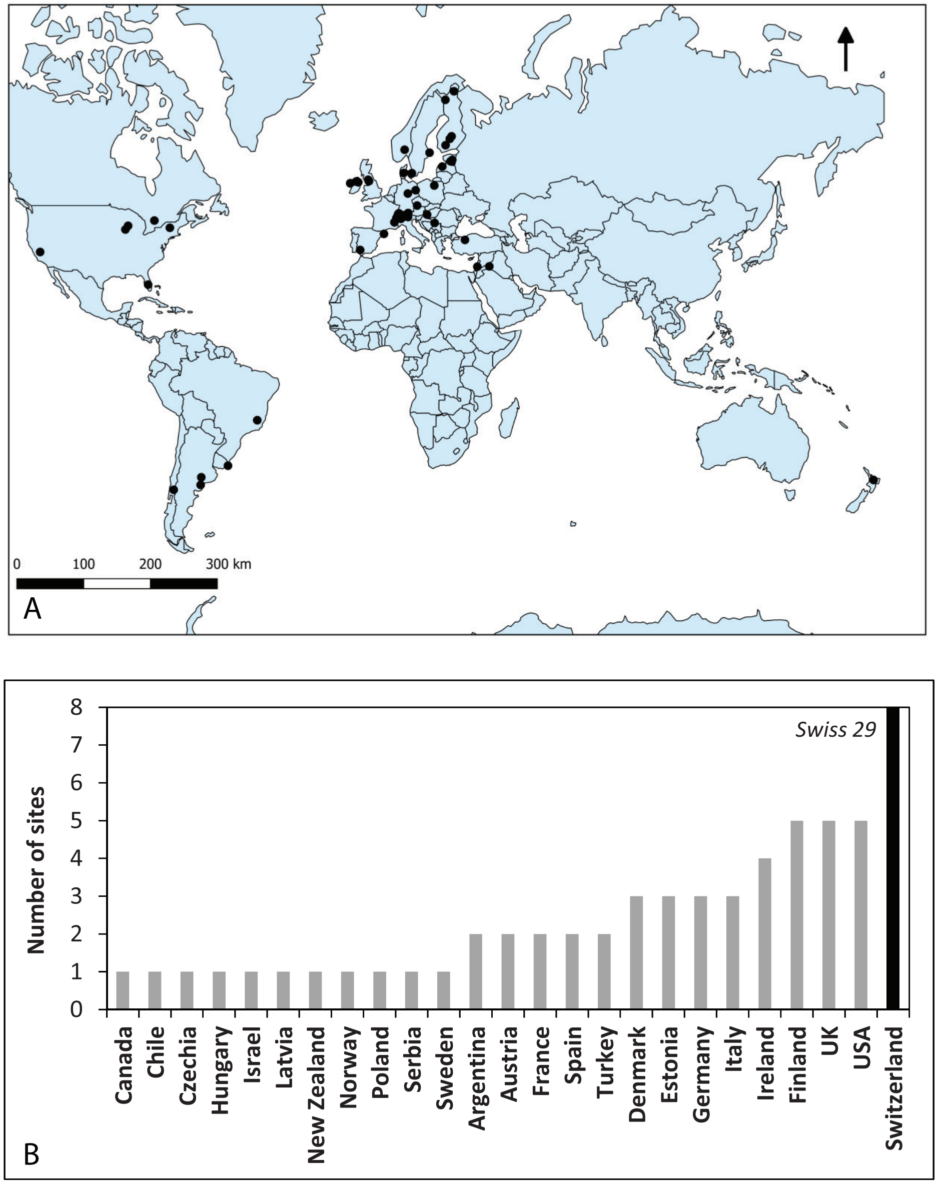

FIG. 2. A) Map showing the global distribution of metadatabase lake sites; B) number of lake sites in each of 25 countries.

Data types

After the metadatabase was organised with the tables and relationships between them defined (Fig. 1), the ecological and database experts agreed on the type of different parameters (attributes) that were measured for each station. Moreover, the type (e.g., string, numeric, or text) and length were defined for each attribute in the tables. They included details on basic lake physical characteristics. Any lake ("site" in the metadatabase) could have more than one station, including open water and littoral stations, associated land-based stations on the lakeshore (e.g. meteorological stations), and stations on the inflow and outflow. As well as general lake physical characteristics, the site metadata included lake trophic status, mixing regime, and presence of ice cover; whether it was classed as a coloured or clear lake; and whether it was used as a drinking water source. The metadata for individual stations were linked to the providers using a simple data access link. The main metadata for the actual high frequency monitoring data were contained in the "parameters-for-stations" table. Importantly, polling of the high frequency monitoring community during the database design phase identified the need for a controlled vocabulary, as many different abbreviations were being used for the same parameter by different data providers. This was due to differences in local common usage, but also to differences in names used in proprietary software provided by various sensor and logger manufacturers. Of the stations at which water temperature was measured (defined as Water_Temperature in the controlled vocabulary), for example, 19 different "user abbreviations" were recorded, including Water_temp, Sonde_Temperature, Temperature, wt, and Temp. For this reason, metadata were collected using both an agreed controlled vocabulary (based on one developed by GLEON; Winslow et al. 2008), and the user abbreviation. This table also collated metadata on the time period for which data were available, the general data frequency (i.e., sub-daily, hourly, and sub-hourly), the user time and date format, and an indicator of data access status (open or restricted access). Another entry for that station was used to show where a new sensor had been deployed at any site for a given parameter.

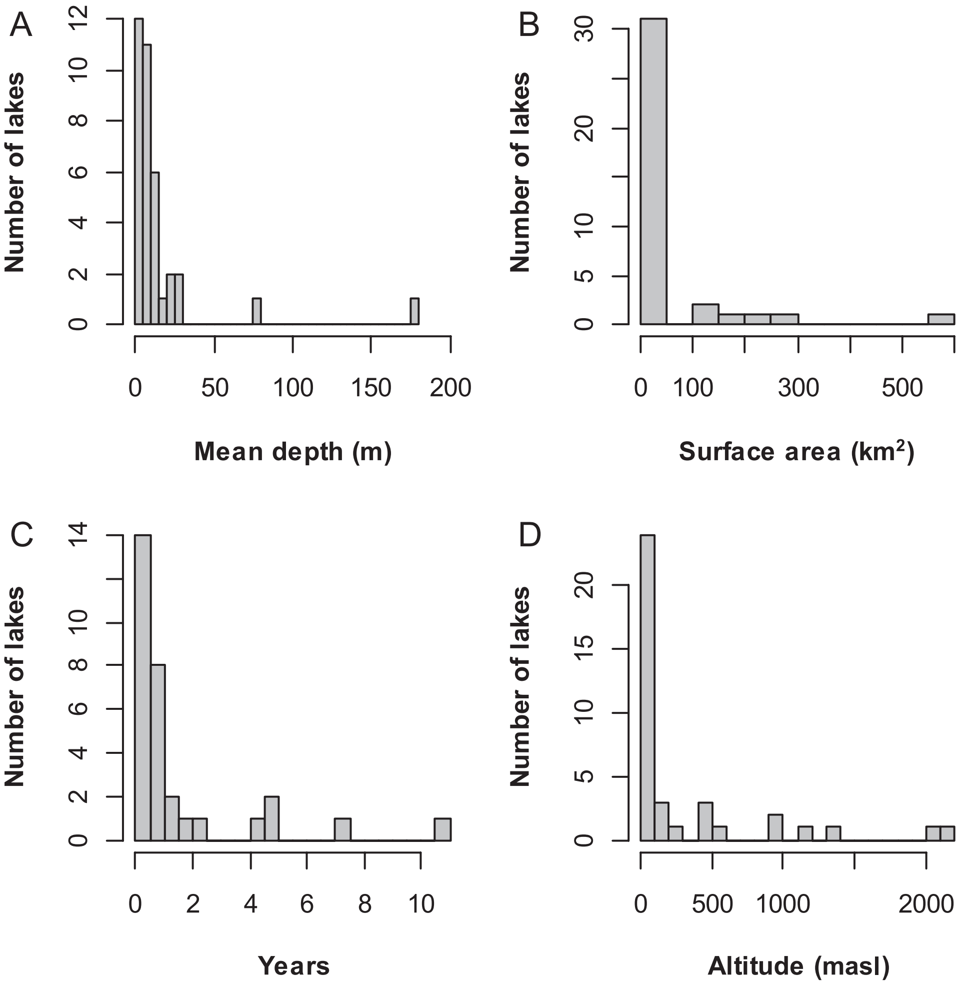

FIG. 3. Lake characteristics for the 42 European lakes (excluding the 29 Swiss water temperature monitoring sites): A) mean depth; B) surface area; C) retention time; D) altitude.

Metadatabase content

The database contained metadata for 83 lakes and 104 stations as of June 2017, provided by 46 different data providers. The data providers were all universities or research centres, and these metadata therefore captured the use of high frequency monitoring for research purposes rather than for management. The sites came from 25 countries (Fig. 2), mostly within Europe, reflecting participation in NETLAKE. Of the 83 lakes, 12 were outside of Europe and were part of the more recent and on-going expansion of the original metadatabase to include other GLEON sites and sites from the pan-American SAFER and Pampa 2 projects. Of the 104 stations, 56 were classed as open water sites, 33 as littoral, nine as land based (these were mainly associated meteorological stations), with three inflow sites and four outflow sites. Within Europe, 29 sites were part of a long-term project assessing water temperature only in Swiss lakes, 28 of which were littoral, and many of which were at higher altitudes.

With the exception of the set of Swiss sites, metadata were available for 42 other European sites. These tended to be shallower lakes, have a smaller surface area, and be situated at lower altitudes (Fig. 3). More than half of these sites (26 lakes) had a mean depth <10 m. The deepest lakes were the two large Italian lakes: Orta (80 m mean depth) and Maggiore (177 m mean depth). Only seven of these 42 lakes were at altitudes greater than 500 m above sea level (masl) (Tovel, Orta, and Maggiore (Italy); Kinneret (Israel); Langjern (Norway); Eymir (Turkey); and Piburgersee (Austria)), and 26 in total were at altitudes below 100 masl. Of the six lakes with the largest surface areas (> 100 km^2), it was of note that three were also very shallow. These were Balaton (Hungary), the lake with the largest surface area of 592 km^2 and a mean depth of 3.2 m; Võrtsjärv (Estonia; mean depth of 2.8 m and surface area of 270 km^2); and Vanajanselka (Finland; mean depth of 7 m and surface area of 103 km^2). Lakes with retention times of less than one yr dominated the European sites (30 lakes), and of these, 14 had retention times of less than six months, with eight of these in the wetter regions of western Europe in Norway, Ireland, and the United Kingdom. The 42 European sites included dimictic (19), monomictic (10), and polymictic (9) lakes, and one meromictic lake (Furnace, Ireland). They also included nine lakes classed as coloured (mean annual value > 30 mg PtCo L^{-1}). Eutrophic (16), mesotrophic (11), and oligotrophic (14) sites were all equally represented.

Measured parameters

There were 140 different sensor types (manufacturer and model) captured in the metadatabase. These details, linked to the contact details of the data providers, have proved to be a particularly useful resource for the lake monitoring community. Most sensors were deployed at fixed depths, and generally in the surface waters. As of June 2017, there were only seven lakes in the total metadatabase with winch-operated, variable depth systems. These were Saadjärv (Estonia), Jyväsjärvi (Finland), Müeggelsee (Germany), Furnace (Ireland), Kinneret (Israel), Erken (Sweden), and George (USA). All stations in the overall metadatabase had water temperature sensors, which included both lower cost, stand-alone temperature loggers and chains of platinum resistance thermometers (PRTs). Other commonly measured parameters included dissolved oxygen concentration (37 stations), conductivity (27), chlorophyll fluorescence (25), and pH (20). Some variables were measured at fewer sites. Only nine lakes had sensors for the phytoplankton pigment phycocyanin (giving an estimate of phycocyanin-containing phytoplankton biomass), while six lakes had data available for coloured dissolved organic matter. The latter lakes were mostly from regions with high organic matter soils, and hence highly coloured water.

Sensors were also commonly used to measure meteorological parameters, including a range of different light measurements. Six lake platforms had surface PAR (photosynthetically active radiation) sensors. Surprisingly, underwater PAR data, an important variable for assessing lake biota, were available from only eight sites. Balaton had a delayed fluorescence system, which also provided data on light photosynthesis curves. Only a small group of seven sites (Muggelsee (Germany); Balaton (Hungary); Feeagh (Ireland); Windermere, Esthwaite Water, and Blelham (UK); and Erken (Sweden)) had data for multiple parameters for more than 10 yrs, being sites that had all participated in the original EU projects (George 2010). However, data for multiple parameters for at least 5 yrs were also available from Kinneret (Israel), Langjern (Norway), Furnace (Ireland), Bassenthwaite (UK), and Anterne (France), as well as Harp Lake (Canada). Monitoring frequency was generally recorded as "sub-hourly," however typical time steps included measurements from 2 min to 5 min (Marcé et al. 2016). Interestingly, of 430 separate parameter datasets in the full metadatabase, 246 were listed as having "Restricted access," 140 as "Open access," while for 39 data streams this field was left blank. Restricted access meant that data providers or their institutes needed to give permission before data use.

How can this metadatabase help the monitoring community?

The metadatabase is supporting network science at three different levels. 1) At a **regional and global level**, it has captured, for the first time, metadata on lake monitoring for the European research community. Now, as the metadatabase expands, it has the potential to become a truly global resource. 2) At a **project level**, it has allowed working group and project leaders to archive details for their project sites in one location (e.g., SAFER, and the Swiss long-term water temperature projects). Such metadata can be required throughout a project's lifetime, for example for reports and for final publications. 3) At the level of the **individual scientist or water manager**, the metadatabase can be used to see what others are doing and potentially to identify collaborators for new projects or publications which need to include high frequency datasets from multiple lakes or for multiple parameters. It also provides a wealth of information for the new user, those literally dipping their toes into the world of high frequency lake monitoring, by allowing them access to the contact details of scientists who are already measuring parameters that they are interested in and are willing to provide advice. Indeed this aspect, combined with excellent new guidance in the "NETLAKE Guidelines for Automatic Monitoring Station Development" (Laas et al. 2016) and "NETLAKE Toolbox for the Analysis of Data Analysis" (Obrador et al. 2016), two other Action outputs, ensure that new users can more rapidly get up to speed with best practice for *in-situ* lake monitoring technology.

Future development

The concept of the NETLAKE metadatabase grew from the need of the lake high frequency monitoring community in Europe to know who was measuring what, where, and when, in order to bring collaborative research efforts to fruition. Realization of this vision took dedicated time and effort from many people, including the data providers, and while the metadatabase is now a reality, its development is on-going. There are still requirements in terms of a more user friendly and intuitive front-end, especially for users not familiar with database querying. On-going metadata input will follow the process developed during the design, but updating of the content will likely require volunteer moderators perhaps at a national level. As it currently stands, however, it is now a useful tool for the global community, acting as a one-stop-shop to answer the question of who is doing what, where, and when.

Acknowledgment

This article is based upon work from the NETLAKE COST Action ES1201, supported by COST (European Cooperation in Science and Technology). K.Š. was also supported by the grant RVO 67985939. The authors thank the many Action members and metadata providers for all their work on this project.

References

George, G. 2010. The Impact of Climate Change on European Lakes. In: G. George [ed.], The Impact of Climate Change on European Lakes. Springer Science and Business Media B.V., pp. 1–13.

Laas, A., E. de Eyto, D. Pierson, and E. Jennings [eds.], 2016. NETLAKE Guidelines for automatic monitoring station development. Technical report. NETLAKE COST Action ES1201. 58pp. http://eprints.dkit.ie/id/eprint/524

Marcé, R., G. George, P. Buscarinu, M. Deidda, J. Dunalska, E. de Eyto, G. Flaim, H-P. Grossart, V. Istvanovics, E. Moreno-Ostos, et al. 2016. Automatic high frequency monitoring for improved lake and reservoir management. Environmental Science & Technology. 50: 10780–10794.

Meinson, P., A. Idrizaj, P. Nõges, T. Nõges, and A. Laas. 2015. Continuous and high-frequency measurements in limnology: history, applications, and future challenges. Environ. Rev. 24: 52–62. doi: 10.1139/er-2015-0030.

Obrador, B., I. D. Jones, and E. Jennings [eds.], 2016. NETLAKE toolbox for the analysis of high-frequency data from lakes. Technical report. NETLAKE COST Action ES1201. 60pp. http://eprints.dkit.ie/id/eprint/530

QGIS Development Team, 2017. QGIS Geographic Information System. Open Source Geospatial Foundation. URL http://qgis.osgeo.org.

Rouen, M. A., G. G. George and D. P Hewitt. 2000. Using an automatic monitoring station to assess the impact of episodic mixing on the seasonal succession of phytoplankton. Verh. Internat. Verein. Limnol. 27: 2972–2976.

Weathers, K., P. C. Hanson, P. Arzberger, J. Brentrup, J. Brookes, C. C. Carey, E. Gaiser, D. P. Hamilton, G. S. Hong, B. W. Ibelings, V. Istvanovics, E. Jennings, K. Bomchul, T. Kratz, F-P. Lin, K. Muraoka, C. O'Reilly, C. Piccolo,

E. Ryder, and G. Zhum. 2013. The Global Lake Ecological Observatory Network (GLEON): the evolution of grassroots network. Limnology and Oceanography Bulletin 22: 71–73.

Winslow, L. A., B. J. Benson, K. E. Chiu, P. C. Hanson and T. K. Kratz. 2008. Vega: a flexible data model for environmental time series data. In: C. Gries and M. B. Jones [eds.]. Proceedings of the Environmental Information Management Conference 2008; 10–11 Sep 2008; Albuquerque, NM. https://conference.ecoinformatics.org/public/conferences/1/eim-2008-proceedings.pdf.

Eleanor Jennings, Dundalk Institute of Technology, Dublin Road, Ireland; eleanor.jennings@dkit.ie

Elvira de Eyto, Marine Institute, Newport, Ireland

Alo Laas, Estonian University of Life Sciences, Institute of Agricultural and Environmental Sciences, Centre for Limnology, Tartu, Estonia

Don Pierson, Uppsala University, Sweden

Georgina Mircheva, Ss. Cyril and Methodius University in Skopje, Macedonia

Andreja Naumoski, Ss. Cyril and Methodius University in Skopje, Macedonia

Andrew Clarke, Dundalk Institute of Technology, Dublin Road, Ireland

Michael Healy, Limerick County Council, Ireland

Kateřina Šumberová, Institute of Botany of the Czech Academy of Sciences, Department of Vegetation Ecology, Brno, Czech Republic

Daniel Langenhaun, Daniel Langenhaun, IGB Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Müggelseedamm, Berlin, Germany