Report on recommendations for the storage of data

Project Identification	
Project Full Title	Water scenarios for Copernicus Exploitation
Project Acronym	Water-ForCE
Grant Agreement	101004186
Starting date	01.01.2021
Duration	36 months

Document Identification	
Deliverable number	D4.3
Deliverable Title	Report on recommendations for the storage of data
Type of Deliverable	Report
Dissemination Level	Public (PU)
Work Package	WP4
Leading Partner	USTIR/FVB-IGB

History of Changes		
Date	**Version**	**Comments**

04.08.2022	V0	First Concept (Evangelos Spyrakos, USTIR; Igor Igashawara, IGB)
02.09.2022	V0.1	Skeleton Draft for feedback from WP4 members
05.12.2022	V0.9	First complete Draft (Harriet Wilson, USTIR; Josh Millard, USTIR)
03.01.2023	V1.1	Full Draft for internal feedback ()

Water-ForCE is a CSA that has received funding form the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 101004186.

List of Acronyms

AB	Advisory Board
AGA	Annotated Grant Agreement
CA	Consortium Agreement
CSA	Coordination and Support Action
CT	Coordination Team
DoA	Description of Action
DMP	Data Management Plan
EB	Executive Board
EC	European Commission
IPR	Intellectual Property Rights
FA	Funding Authority
GA	Grant Agreement
GAs	General Assembly
GPDR	General Data Protection Regulation
PIP	Project Implementation Plan
PM	Person-months
PO	Project Officer
SDGs	Sustainable Development Goals
TL	Task Leader
WG	Working Group
WP	Work Package

Table of Contents

1. Introduction 7

1.1 Water-ForCE 7

1.2 Purpose of the document 8

1.3 Scope of the report 8

1.4 Definitions 9

1.4.1 In situ data 9

1.4.2 Database and data storage 9

2. Optimal data organisation and storage 11

3. Review of in situ data repositories 14

3.1 Purpose 16

3.1.1 Monitoring scheme repositories 16

3.1.2 Project-based repositories 16

3.1.3 Open data repositories 17

3.1.4 Open managed data repositories 17

3.1.5 Metadata repositories 18

3.1.6 Project-based databases 19

3.2 Data scope 19

3.3 Intended users 20

3.4 Funding 20

3.5 Data providers 21

3.6 Data provider interface 22

3.7 Repository content 23

3.7.1 Variables 23

3.7.2 Water-bodies considered 23

3.7.3 Spatial scope 23

3.7.4 Temporal scope 24

3.7.5 Database types 24

3.7.6 Versions and updates 24

3.7.7 Unique identifiers 25

3.7.8 Metadata 25

3.7.9 Storage 28

3.8 End-user interface 28

3.8.1 Findability 28

3.8.2 Downloading data & data permission 29

3.8.3 Help support 29

3.8.4 Machine-readability 29

3.9 End-users 30

4. Future needs and recommendations 32

4.1 Barriers to the uptake/use of in situ data repositories by satellite-EO end-users 32

4.2 Challenges faced by in situ data providers, gatherers and database managers 35

4.4 Recommendations 36

4.4 Future possibilities for storing in situ data 39

4.4.1 Metadata repositories for in-situ data 39

4.4.2 Extend an existing database 40

4.4.3 A built-for-purpose repository 40

4. Conclusion 41

References 42

Executive summary

This report summarises the optimal data information and framework for coupling in situ data and satellite-EO products. Through a review of existing in situ databases and interviews with data managers, needs for the future organisation and storage of in situ data for use with satellite-EO products were identified. Actions that may facilitate the wider use of in situ data within satellite-EO projects were emphasised.

The review has been structured in the following sections:

1. Description of existing in situ databases of water quality, water quantity and water-leaving reflectance data from inland and coastal waters,
2. Review of the characteristics of the databases highlighting advantages and challenges for coupling with satellite-EO projects, including the data organisation, data format, metadata description and data quality,
3. Summary of the needs and recommendations of in situ databases for improved use with satellite-EO.

1. Introduction

1.1 Water-ForCE

The Horizon-2020 project Water-ForCE (Water scenarios For Copernicus Exploitation) will develop a Roadmap for Copernicus Inland Water Services.
The Roadmap will contain:

- Analysis of user communities' landscape
- Analysis on how Copernicus water-related services can support policy development and monitoring of their implementation
- Gap analysis of the Copernicus water-related service portfolio
- Identification of future potential higher-level biogeochemical products
- Technical requirements for future Copernicus sensors to improve the water-related service portfolio
- Proposal for organising *in situ* measurement networks to validate Copernicus remote sensing and modelling products and to provide complementary data not collected by remote sensing
- Proposal on how to define relationships between Core Services and Downstream services
- Scenarios of the most optimal delivery of water services to different user communities.

The Water-ForCE project is coordinated by the University of Tartu (Estonia) with 20 participating organisations from all over Europe. It connects experts in water quality and quantity, in policy, research, engineering and service sectors.

This report is part of Work Package 4 (WP4) "Aligning *in situ* and satellite Earth observation activities" which is trying to establish clear links between *in situ* and satellite observation networks to ensure that they can mutually benefit from data collection and sharing

1.2 Purpose of the document

It is widely recognised that combined approach of in situ and satellite data can deliver a powerful combination to observe and verify change at frequency need to respond to hydrological events and provide early warning. In situ data offer ample opportunities to calibrate and validate Earth observation data and products. However, there are some disconnects between remote sensing and in situ observation research. Foer example, the COINS report entitled "Lake water quality in-situ data requirements and availability" (Carvalho et al., 2021), highlighted the data gaps in existing water quality in situ data for use with satellite-EO data. They call for greater consideration on the design of in situ sampling programs and protocols for satellite-EO users (D4.2) and the further investigation into existing repositories (this report).

1.3 Scope of the report

Task 4.3 Data integration within and between observation networks aims to assess the current landscape and provide best practices on methods and frameworks for combining in situ data with Earth Observation data towards improved inland water monitoring. The communities of in situ data providers (identified in T4.1) and remote sensing experts (identified in T2.1) will be consulted to evaluate the optimal data information and framework needed for coupling these two communities. It is expected to synthesize a framework for data organization,

which could include data format, metadata description, data quality and other important information which will be highlighted from these two groups. The idea is to summarize the best methods for archiving and sharing data which will allow an easy interpretation and usage of the dataset.

1.4 Definitions

1.4.1 In situ data

In this report, **in situ data** refers to measurements collected directly from the location of interest (e.g. soil moisture collected using a probe in the soil) usually represented as numbers (or text and multimedia). Variables of particular interest include water quality, water quantity and water leaving reflectance (See Section 3 D4.2). Water leaving reflectance was selected as it is key to matching in situ data to satellite data, and hence was included in addition to the two broader categories of water quality and quantity. In situ data can be collected at different frequencies for varied time periods and represent different spatial areas. In situ data may be collected for different reasons, for example as part of a project sampling campaign or a long-term monitoring program. In situ data may be compiled in a dataset associated with a unique body of work. In situ data from inland and coastal waters may be coupled or matched with satellite-EO products for various applications, including:

- Cal/Val of satellite data.
- To fill temporal and spatial observational gaps within in situ monitoring schemes (or vice versa to fill gaps in satellite time series due to cloud cover for example).
- To train and test machine learning models
- To develop satellite algorithms (e.g., to detect cyanobacteria).

1.4.2 Database and data storage

The following terminology is used in this document.

- **A data table** is a tabular representation of data, e.g., a single excel spreadsheet.
- A **dataset** is one or more DataTables (e.g, an excel file with 1 or more sheets).
- A **database** is an organised collection of datasets that are interlinked or inter-referenced between the datasets (e.g. a project on GitHub or a data package on Zenodo). This is sometimes described as a data package by certain repositories.
- A **repository**, barring any domain specific usage, is a term to describe a location for storing data (e.g. Github or Zenodo).
- A **metadata repository** is a term for a data aggregator or indexer, which does not store the actual datasets on their servers but they do store the actual metadata on their servers (e.g., GEOSS portal).

The organisation and storage of data is important for the future use and re-use of the data and has a direct impact on the end-users of the data. Repositories are typically built with the intention of being easily accessed, managed, and updated. Key existing sources for both abstract and practical guidelines on the best-practice of data repositories, especially within the scientific research or satellite-EO community, include;

- The FAIR Data Principles; Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016).
- TRUST principles for digital repositories; Transparency, Responsibility, User Focus, Sustainability, Technology (Lin et al., 2020).
- Open Geospatial Consortium Standards and Resources (https://www.ogc.org/).
- Re3 Registry of research data repositories (https://www.re3data.org/).

2. Optimal data organisation and storage

Data repositories can be evaluated in terms of 1) the intended purpose of the repository and the funding set-up, 2) the data providers/producers, 3) the interface between the data providers and the repository, 4) the content and structure of the repository, 5) the interface between the repository and the end-users, and 6) the intended end-users (Figure 1).

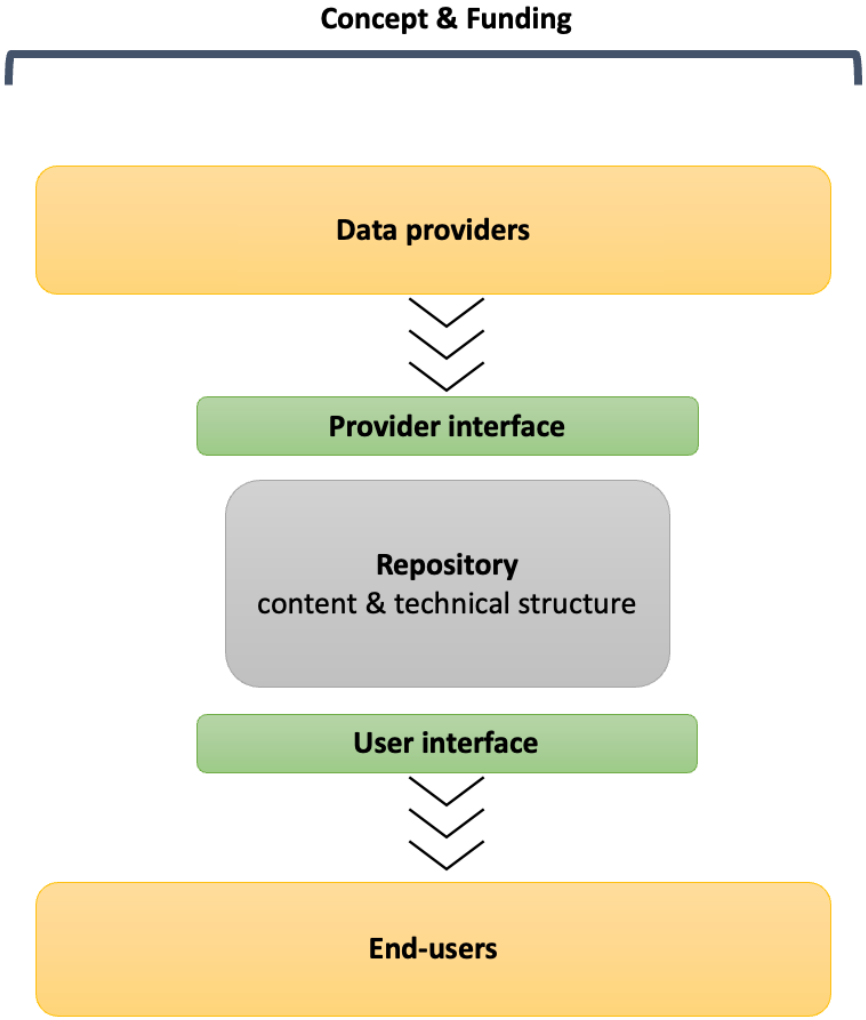

Figure 1. *Schematic showing the database characteristics considered in this report and the process from data providers to the end data users.*

The optimum organisation and storage of in situ data for use with satellite-EO depends on the exact application and the rapid development of remote sensing practices. However, an optimum system may include some of the following conditions;

- **Purpose**: Repository is well-known, and clearly contains inland/coastal in situ data for use with satellite-EO data
- **Funding**: Repository has sustainably funding (e.g., +10 years guaranteed)
- **Data producers/providers:** Data is provided by single or multiple producers who are motivated to keep their data in the repository and are aligned/harmonised with the use of the data. Data is high-quality and producers are trustable.
- **Content**:
    - All key in situ reference data is available including water-leaving reflectance, water quality and water quantity.
    - Temporal/spatial representation to match with satellite data
    - Good metadata and unique identifiers.
    - Standardised formatting.
    - International/Transboundary datasets.
    - In situ data can be matched-up with satellite overpasses, or datasets are already matched-up
    - Data is published in a timely manner
    - Data can be added or updated
- **End users:**
    - Repository is open-access and limits the barriers to new users
    - Allows download
    - Data is well-known or a metabase facilitates the discoverability

- Machine-readable

- **Technical:**
  - Data is stored in a sustainable way and at a reasonable cost
  - Repository is structurally good

3. Review of in situ data repositories

A list of existing in situ data repositories and databases were compiled by the consortium group (Table 1). When discussed collectively we will call them in situ data repositories. All include at least some observed water quality, water quality or water leaving reflectance data from inland or coastal water bodies. These in situ data are either currently, or could in the future be used in satellite-EO data processing. Repositories which were conceived and managed within Europe are of particular focus, however some particularly well-known databases from outside of the EU were included.

Table 1. *Characteristics of in situ repositories including water quality, water quantity, water-leaving reflectance data from inland and coastal areas..*

			Concept								Provider								Content						Access	
	Data focus	Intended community	Metadata repository	Data Repository	Database	Continous	Soft	EU-based	Open upload*	Project	Monitoring program	Mixed (open)	Mixed (managed)	Includes EU data	Water quality	Water quantity	Water leaving reflectance	Lakes	Rivers	Coastal / Oceans	National	Regional	Global	Download*	Includes an API
AERONET-OC	Ocean	Sat-EO		●		●			(●)				●	●	●		●	●	●	●			●	●	●
CUAHSI Hydroshare	Environmental	Reserachers		●		●			(●)			●		●	●	●		●	●	●		●		●	●
Danube Basin Water Quality (ICPDR)	Water quality	Management		●		●		●			●			●	●	●			●			●		(●)	
EMSO ERIC Data Portal	Ocean	Sat-EO		●			●	●						●	●					●		●		(●)	●
Environmental Data Initiative	Environmental	Reserachers		●		●			(●)			●		●				●	●	●			●	●	●
GEMStat	Water quality	Public/researchers		●		●		●	(●)				●	●	●			●	●				●	●	●
GEOSS Portal	Environmental	Sat-EO	●			●		●						●				●	●	●		●		●	
Global Runoff Data Portal	Runoff	General (?)		●		●		●					●	●		●		●	●	●			●	●	
HydroATLAS	Hydrology	General (?)		●			●			●				●		●		●	●	●			●	●	
ISIMIP (Input data)	Temp and met	Modellers		●			●	●		●				●	●	●		●	●	●			●	●	●
KNB	Environmental	Reserachers		●		●			(●)			●		●	●			●	●	●			●	●	●
LIMNADES	Inland and coastal	Sat-EO		●			●	●		●				●	●		●	●	●	●			●	(●)	●
MONOCLE	Inland and coastal	Sat-EO / in-situ		●			●	●		●				●	●		●	●	●	●			●	(●)	
NETLAKE	Lake	In-situ	●				●							●				●					●		
Observatory on LAkes (OLA)	Water quality	Management		●		●		●			●			●	●	●		●				●		(●)	
PANGAEA	Environmental	Reserachers		●		●		●	(●)				●	●	●	●	●	●	●	●			●	●	●
SeaBASS	Ocean	Sat-EO		●		●		●	(●)		●	●		●			●	●		●			●	●	●
Vannmiljø (Norway)	Water quality	Management		●		●		●			●			●	●	●		●	●	●	●			(●)	
VISS (Sweden)	Water quality	Management		●		●		●			●			●	●	●		●	●	●	●			(●)	●
Water Quality Archive (England)	Water quality	Management		●		●		●			●			●	●	●		●	●	●	●			●	●
Waterbase (EEA)	Water quality	Management		●		●		●			●			●	●	●		●	●	●		●		●	
Waterinfo.be (Belgium)	Water quality	Management		●		●		●			●			●	●	●		●	●	●	●			●	●
WISA (Austria)	Water quality	Management		●		●		●			●			●	●	●		●	●		●			(●)	
Zenodo	Environmental	Reserachers		●					(●)			●		●				●	●	●			●	●	●

* Brackets indicate that email or login is required to access open data

3.1 Purpose

Four key groups were identified for data repositories, in terms of the intended purpose. These were 1) monitoring schemes, 2) project-based, 3) open repositories with mixed data sources, and 4) open repositories with managed or harmonised data sources. Two additional relevant groups were identified, which were metadata repositories and project-based *databases* rather than repositories. These groups are summarised in the section below.

3.1.1 Monitoring scheme repositories

Monitoring schemes, such as national or regional water body monitoring schemes, require data storage. Data may be collected for the European Environment Framework or other legislation. This report included 9 national and regional monitoring schemes, including the European Environment Agencies Waterbase repository. Repositories designed to store data from national or regional monitoring schemes have some specific requirements. In particular, they are designed for upload only by designated parties within the scheme. Data within a monitoring scheme may be more harmonised, in terms of collection methods and metadata etc, although this is not necessarily always the case. An additional specificity is that they may be designed to have regular data uploads or updates. Many national or regional monitoring repositories do not all public download, or are only available on email request and do not offer a clear catalogue of the available data. Funding for monitoring scheme repositories is often continuous and a dedicated person or team may be responsible for curating and managing datasets and the repository.

3.1.2 Project-based repositories

Projects often require a designated space to store the input and output data of the project. We include 6 examples of project-based repositories, including MONOCLE and LIMNADES, which are two projects that were fully or partially funded by Horizon2020 and were concerned with enhancing the use of in situ inland and

coastal data within EO-based water quality services There are various common issues related to project-based repositories. These include the funding structure, which is typically not continuous (i.e. stops when the project ends) or data storage funding, which may not be included in the proposal at all. The management of project-based repositories may vary including persons within the project or database management team from the institute.

3.1.3 Open data repositories

Open data repositories are those which are entirely open for upload and download, usually within a certain thematic topic. These repositories include Zenodo, Environmental Data Initiative (hereafter "EDI") and CUASHI Hydroshare (hereafter "Hydroshare"). These repositories can differ in the degree of openness. For example, there are typically limitations on how much data you can upload as an individual or a single project. Additionally, EDI in particular requires you to contact them via email to determine if your data is appropriate for the site, and they prioritize work from within the related scientific community. In contrast, Zenodo and Hydroshare can be done just through their online portal. These repositories typically have continuous funding. One clear benefit is that there is no need to manage data use or ownership, since by publishing the data, the authors specify the terms of use. However, they store massive amounts of data, either on a dedicated server or cloud, which has substantial cost. Moreover, they require large investment into software development and maintenance of the website itself. Existing open-access international repositories included in our analysis were funded by research and innovation related government bodies. These may have thematic requirements, for example, data may be only environmental research data. These repositories are also suitable for small projects to store their data and make it available for re-use, by allowing it to be published.

3.1.4 Open managed data repositories

Another important class of repository is open repositories but which have a degree of data management and harmonisation. The repositories included in this group are AERONET-OC, GEMstat, Global Runoff Data Portal, and PANGAEA. These databases are different to open repositories as the result is a more curated data product, although there is some crossover with monitoring schemes (Section 2.1) and open data repositories (Section 2.3). For example, PANGAEA allows open upload of data from different providers, similar to an open data repository but has a data editorial team which "maintain the integrity and authenticity of the data" (https://www.pangaea.de/). Similarly, GEMstat requires the voluntary participation of monitoring schemes from around the world, but attempts to harmonise the data post-collection to provide a map-based water quality tool (https://gemstat.org/). These repositories offer open-access download in most cases, but some datasets may require additional requests. Such repositories require continuous funding and data management resources.

3.1.5 Metadata repositories

Metadata repositories are repositories that store metadata and can be used for data discovery. The metadata repositories included in this report were EMSO ERIC data portal to direct users to EMSO observatory data based on location, the GEOSS portal designed with EO-users in mind, and NETLAKE which is a metadata repository for high-frequency lake monitoring data. Metadata repositories vary in their data scope and intended user. They also vary in their funding and whether they are one-off projects or if they are actively maintained. While metadata repositories do not have to store the actual data, they do have to provide relevant metadata and links to the datasets themselves, in order to be useful. Maintenance can be automated or manual or a hybrid. For example, NETLAKE was a manual database put together by the lake monitoring community, while the GEOSS portal links end-users to EO and in situ data while using automated updating, where links are checked and some metadata is automatically populated from data websites (See D4.5 for more on

GEOSS portal).

3.1.6 Project-based databases

Finally, some relevant in situ datasets are stored in databases. Smaller projects, for example, might require a single database that can be uploaded on an open repository such as EDI or Hydroshare. There are many such databases, which can be found through open data repositories or metadata repositories. Two examples included here are the European Multi Lake Survey dataset containing biogeochemical parameters collected from +100 lakes over 2018 (https://portal.edirepository.org/nis/mapbrowse?packageid=edi.176.5; Mantzouki et al., 2018) and the Global data set of long-term summertime vertical temperature profiles from 153 lakes (https://portal.edirepository.org/nis/mapbrowse?scope=edi&identifier=705; Pilla et al., 2022) which are both stored on EDI as single Data Packages. Another example is The Surface Water Chemistry (SWatCH) database which comprises data from various sources but has been standardized into one high-quality transboundary database (Rotteveel et al., 2022) stored on Zenodo (https://doi.org/10.5281/zenodo.6484939).

3.2 Data scope

Only two of the repositories were designed specifically to store inland or coastal data, and they were MONOCLE and LIMNADES repositories. Three repositories (AERONET-OC, EMSO ERIC, and SeaBASS) were ocean focused but included a small number of lake or coastal monitoring sites. A few repositories were very wide in scope, and were classified as concerned with environmental data at large, which includes some inland and coastal data (Hydroshare, EDI, KNB, Zenodo). The largest proportion of the databases have a water-body monitoring focus in a specific region. These included OLA, Danube basin water quality portal, Vannmiljø (Norway), VISS

(Sweden), Water Quality Archive (England), Waterbase (EEA), Waterinfo (Belgium), WISA (Austria).

3.3 Intended users

A few of the repositories were designed specifically for the EO-data user community; PANGAEA, MONOCLE, LIMNADES and the GEOSS portal. A greater number were designed for research in general, or the in situ data user community. Most spatially oriented repositories are assumed to be designed for the management of water-bodies and environmental monitoring of that region. This could include EO-users but not primarily.

3.4 Funding

Funding is particularly important when considering the scope of the viability of a repository and the longevity of it. Funding can include the expense involved in the development, curation, maintenance of the repository. It may also include hosting, storage and server costs, a designated team and help-contact.

All the repositories were largely supported by public funding. In general, detailed information on the amount and source of funding received by each repository was limited, however two basic groups were identified. This was whether the repository had continuous or soft funding. Of the repositories, most (14 repositories) received funding for the foreseeable future, which was considered continuous (Table 1). A minority of 6 repositories (Table 1), were soft-funded, meaning they received short-term funding, for example from a 3-year project, or were no longer receiving any formal funding.

Monitoring schemes, open data and managed open data repositories were mostly continuously funded. Regional and national monitoring schemes are largely funded by government-based bodies, such as the European Environment Agency (EEA), or in the UK, the Environment Agency (EA). Open data and managed open data

repositories are typically funded by Research and Innovation oriented international groups. For example Zenodo which is funded by the EU OpenAire projects, or GEMstat which was funded through the United Nations Environment Program (UNEP). Different levels of visibility are associated with each repository, and it is beyond the scope of this document to fully evaluate this. Some databases, for example GEMstat, have expressed 'insufficient funding' as a problem (GEMS water strategy 2020-2024).

In contrast, project-based repositories were typically soft-funded. Discussion with data managers and experts highlighted various scenarios related to soft-funded repositories. For example, once the project funding is suspended the database or repository can be abandoned regardless of how widely used it is. Alternatively, as in the case of the LIMNADES repository, the maintenance may be taken up by the residing institution from goodwill, or the time of unpaid former project participants. In such cases, there may be little investment in the continued growth of the repository. Soft-funded repositories may also become unviable if the software becomes out of data and there is insufficient funding to update it.

Notably, in academic contexts, data maintenance is not always covered by research funding. For example, in German institutions database management and repository curation is typically the responsibility of the academic institution and is not funded by third-party funds.

3.5 Data providers

Open access repositories typically allow data to be freely uploaded and hence the data provider could be anyone. There were 8 repositories included which allowed open upload. All these still require some regulation on uploads, for example, the data provider must make an account for identification, provide some specific information about the data. Some data repositories have limitations on the storage and the

length of time that data will be stored. Some repositories also require a manual screening process, such as EDI, where data providers are requested to email first and certain research affiliations are prioritised. When uploading data, it is possible to stipulate the terms of use, but anyone can access the data and there is less control for the data provider.

For other repositories, the collection, upload and repository management are conducted within the same operation. This can be the case for some national monitoring schemes (e.g. Water Quality Archive gov.uk). Some project-based repositories also have cohesive collection, upload and storage (e.g., MONOCLE), or if the data was not collected by the repository operation itself, it may have been actively gathered, processed and uploaded by members of the project (e.g., ISIMIP, LIMNADES). In some cases, data providers want to protect or limit the access to their data, for example, to ensure the providers have the first opportunity of publication or to prohibit overlapping projects.

3.6 Data provider interface

Data providers can vary in expertise, resources and intentions. Hence, their needs from a repository can differ. The interface between the data provider and the repository determines the process of data upload, and the level of regulation and support during this process. Requirements for providing data can include metadata and interoperability standards (discussed in the later section and in Deliverable D4.2), QAQC standards and a level of completeness before publishing. Support offered to data providers can include; a help email or chat link, manual data handling and completeness checks, automated data handling tools and a designated management person/team to oversee the data upload and documentation.

Typically, open data repositories, such as Zenodo and Hydroshare, just provide the storage space and recommendation on metadata. In such cases, the data quality

and documentation are the responsibility of the provider. Both repositories provide a technical help desk. Repositories such as EDI, GEMstat or PANGAEA can provide an additional level of support in terms of data harmonisation, metadata completeness checks or even manually check the data. In project based databases, such as ISIMIP (lake sector) and LIMNADES, data was gathered in a one-off data call and they requested specific metadata to be provided with the observed data. In some cases, the project would also facilitate this process, or QAQC the data. It is not easy to assess data repositories where there is a closed upload process.

3.7 Repository content

3.7.1 Variables

Key in situ variables for coupling with satellite-EO based data are water quantity, water quality and water-leaving reflectance measurements. None of the databases considered in this report included all three types of data. Most databases or repositories included water quantity (17) and water quality (13). Only 5 repositories included water-leaving reflectance data, which is essential for matching data with optical satellite data. Notably, in-situ observations are not typically matched-up to the same time and geographical location as remote sensing data.

3.7.2 Water-bodies considered

Coverage of inland waters is very important, including lakes, rivers and coastal areas. 20 of the repositories include data from lakes. However, notably the range is high from just 3 lakes in the AERONET-OC repository, to more than 1000 lakes in the GEMstat database. Rivers were also well represented within the data repositories, with 19 repositories containing data. Coastal areas were also included in 19 of the data repositories.

3.7.3 Spatial scope

All repositories included data from European water bodies, as a criterion for the

original search. Further to this, 4 repositories included national data (Norway, Sweden, England, Belgium and Austria), 7 repositories included regional data and 13 repositories consider a global scope. Global data repositories do not necessarily include an even distribution globally, but have no spatial constraints.

3.7.4 Temporal scope

Most repositories included various temporal ranges depending on the dataset. The greatest temporal scope was the water quantity data provided by the GRDC which includes data from 1806. OLA, EDI, and ISIMIP all store long-term datasets including around 50-60 years of monitoring data. Water-leaving reflectance datasets were shorter in temporal scope. LIMNADES contains data from 1990 to 2021. AERONET-OC and SeaBASS contain ~20 years of data.

Another consideration in temporal scope is the timeliness of data upload, following its collection. This largely depends on the condition in which the data is uploaded (e.g., is raw data uploaded), whether there is any embargo on the data (e.g., a period where the data is not open-access), and the speed/resources attributed to data processing and upload.

3.7.5 Database types

The typically preferred data format is data table data in TAB-delimited or excel files, which are presented in a relational database. NetCDF files were also possible and some Shape files (e.g. HydroATLAS).

3.7.6 Versions and updates

Repositories use different practices for data versioning and updating. Some repositories require each dataset to have multiple timestamps, e.g., MONOCLE include multiple timestamps with each dataset, for the collection of the data, the

processing of the data, the ingestion of the data. This allows raw data to be uploaded, and different levels of processing. Some repositories just upload one version of the data. Once published repositories may not allow data to be edited to updated.

3.7.7 Unique identifiers

Unique identifiers allow data to be uniquely named and reduce the possibility of duplication. Various different unique identifiers exist including ARK, DOI, PURL, URL, UUID. All of the in situ databases explored in this study use unique identifiers, typically DOI's. This creates some transparency and traceability in terms of what the data is and how it can be found again. Open repositories such as Hydroshare or Zenodo etc assign a DOI automatically. Repositories themselves can also have unique identifiers if they meet sufficient standards, for example, the Re3 standard which includes Pangaea, zenodo, EDI and Hydroshare.

3.7.8 Metadata

Metadata is data about the data and is required to interpret the data. Various independent sources provide well defined protocols for metadata in general, and for scientific variables. Some specific examples of metadata standards include ISO 19115, CSDGM, ANZLIC, GEODCAT, Ckan. There is no specific formalised metadata scheme for incorporating in situ data into satellite-EO products. AERONET-OC and SeaBASS provide guides for recording data with metadata for match-up with satellite data. Additionally, LIMNADES metadata were conceived by the GEO AquaWatch group of satellite-EO specialists with particular focus on future use with inland EO water quality products. The product of this collaboration was the LIMNADES metadata template which includes metadata for stationary measuring stations in general (Table 3) and for the different variables ranging from biogeochemical variables such as phytoplankton abundance, to water leaving radiance (Table 4). It is clear that spectral data, shown in Table 4., requires a substantially higher number of metadata and information on the measuring process,

such as the equipment which was used to acquire the data, geometry of the measuring process employed by the equipment, and production process used available.

Table 3. *List of metadata required for monitoring stations in the LIMNADES repository.*

Essential	**Desirable**
Unique code to identify your station	Individuals/contact involved in data collection
The primary email contact address	Institution samples were collected on behalf of
First time recorded at station	name of cruise data were collected on
Initial latitude position	name of project data were collected for
Initial longitude position	Funding body or institution
	Last time recorded at station
	Final latitude position
	Final longitude position
	Description of weather and measured wind speed, Air temperature
	Observed wave height, Observed cloud cover, Air pressure,
	Measured water depth
	Water temperature at surface
	Water clarity with secchi depth
	Name of inland or coastal water body

Table 4. *Metadata required for data submission in the LIMNADES database, described for each variable.*

		Pigments	Dissolved element fraction	Suspended matter	Water temperature	Salinity	Turbidity	Absorption coefficients	Scattering coefficients	Radiometric quantity	Derived radiometric quantity	Particle size distribution	Aerosol optical thickness	Phytoplankton abundance
ID	Identification code	•	•	•	•	•	•	•	•	•	•	•	•	•
Methodology information	Link to extraction protocol or paper	•	•					•						
	Analytical method used	•	•	•	•	•	•	•	•	•	•	•	•	•
	Calculation computation method									•	•			
	Calibration history									•	•			
	Full name of analytical instrument	•	•	•	•	•	•	•	•	•	•	•	•	•
	Link to calibration or assurance report								•					
	Processing protocol								•					
	List any correction details							•						
	List any processing methods							•						
	Serial number of instrument	•	•	•	•	•	•	•	•	•	•	•	•	•
	Length of time it took before storage	•	•					•						
	Temperature the same was stored at	•	•					•						
	Was correction for straylight applied?									•	•			
	Was self shading correction applied?									•	•			
	Was sun zenith angle taken into account?									•	•			
	Was the sample fixed or fresh?							•						
	Was tilt sensor calibration performed?									•	•			
	Were dark measurements performed?									•	•			
	Were Immersion coefficients taken into account?									•	•			
	What was used as a reference value							•						
Instrument metadata	Average latitude for all measurements										•			
	Average longitude for all measurements										•			
	Depth at which the sample was taken	•	•	•	•	•	•	•	•	•	•	•	•	•
	Depth of measuremnt							•		•	•			
	Instrument timestamp for each measurement									•				
Field information	Depth at which the sample was taken	•	•	•	•	•	•	•	•	•	•	•	•	•
	Method used to sample	•	•	•	•	•	•	•	•	•	•	•	•	•
Data	Number of spectra used in aggregation										•			
	Pigment name	•												
	Standard deviation for aggregated measurements									•	•	•	•	
	The absolute uncertainty associated with the value	•	•	•	•	•	•		•					
	The value recorded by the instrument	•	•	•	•	•	•							
	The value recorded by the instrument at a specific bin							•						
	The wavelength bins recorded by instrument							•	•	•	•	•	•	•
	Type of measurement		x	x										
	Type of uncertainty assoctaed with value	•	•	•	•	•	•	•	•	•	•	•	•	•
	Units of bins							•	•	•	•	•	•	
	Units of measurement	•	•	•	•	•	•	•	•	•	•	•	•	•
Data Sharing	The chosen data policy	•	•	•	•	•	•	•	•	•	•	•	•	•
	The date for licenced data (C-F) to be made open access	•	•	•	•	•	•	•	•	•	•	•	•	•

3.7.9 Storage

Data storage can be very expensive and consume a lot of energy. The two main storage options are cloud-based and dedicated servers. Typically, dedicated servers are more appropriate for larger repositories, although cloud-based storage can be more accessible and reduce the needs of a data manager, including back-ups and security. There is a very large range in storage capacity between the repositories included in this report. For example, LIMNADES repository is a total of 50 GB while Zenodo offers 50GB per dataset and has an unlimited number of datasets. It also has 100 petabytes of space (https://help.zenodo.org/).

3.8 End-user interface

3.8.1 Findability

Findability refers to how easily a potential user can find or discover the data and whether there is sufficient information to interpret the data products. This can determine whether the data will be found and used correctly. Findability is one of the key variables of the FAIR protocol, which stipulates that data should be easy to find by both computers and humans. Key considerations in evaluating this include the discovery metadata standards used, whether there is a catalogue explaining the data services, key words to facilitate searching and retrieval, and the visibility of the repository; e.g. in Google search, in publications.

There were large differences between the findability of the data repositories explored. The open repositories, such as Zenodo and Hydroshare are very easy to find in general, and have their own search engines that use the metadata and keywords of each data package to search for relevant data. Hence, the findability depends on how comprehensive and relevant the metadata and keywords are. Moreover, the GEOSS portal typically links up. Individual databases/packages can also be found using Google search domain.

More harmonised data repositories, offer map-based browsers for data exploration prior to download. These include MONOCLE (H2O20) GEMStat and EMSO ERIC have open data maps, which allow some degree of data exploration and catalogue. Some of the repositories require an account to be made prior to the catalogue of data can be explored. Examples of this include Danube Basin Water Quality portal and LIMNADES.

3.8.2 Downloading data & data permission

The procedure to download data varied between the repositories. All repositories (excluding the metadata repositories) allow open-access download. Of the repositories, 14 were fully open, meaning there was no requirement of a login to download at least some of the data, but may still be terms and conditions. The other 8 data repositories require some identification or additional regulation. For some of the repositories this is just a login account but it may also mean that data providers need to approve each request for download, which is the case for LIMNADES repository. In some cases there is data which is not open-access, for example PANGAEA. Download limitations are also typical but depend on the repository.

3.8.3 Help support

Not all databases provide clear instructions on how to download data. Even where data is available for download, it can be inaccessible due to the complexity of the process and lack of educational resources. Potential users may require help in the download process, through a live or email-based help desk. Furthermore, where there is insufficient metadata (Section 3.4.5) the data may not be confidently used.

3.8.4 Download and machine-readability

Some databases support the retrieval of their content via API (Application programming interface) which enables the data to be machine-readable. Ways to do this include a simple open search implementation or interface such as RESTful

(REpresentational State Transfer). Machine-readable data repositories allow in situ data to be retrieved regularly or to check for updated data (i.e. a new version of data) which would be necessary if in situ data is to be assimilated into water quality satellite-EO data services . More than half of the data repositories have an API, but a substantial proportion did not have an API or had insufficient information on the website to determine this. Metadata repositories such as GEOSS were not machine-readable but it is not so necessary as they do not store the desired data themselves. Other data repositories are unable to provide an API due to legal reasons. For example the GRDC are prohibited to offer an API access point, since they require identified access and explicit (human) acceptance of the Terms of Use and Data Protection REgulations following the WMO Congress (https://www.bafg.de/GRDC/EN/02_srvcs/21_tmsrs/210_prtl/faqs.html).

Another consideration is whether databases allow the download of specific pieces of data (e.g. hence benefit of cubes), while others require download of large chunks of data

3.9 End-users

End-users are a diverse group, similarly to the data providers, with different levels of expertise and needs. Satellite-EO based end-users include; university students, researchers with different levels of expertise, and water managers. Different repositories have different end-users in mind, which was discussed in Section 2.3. Some are more aimed towards the public, while others are serving the needs of a specific group of the scientific community or the national environmental framework. The number of end-users is not available for most data repositories. For those that were available, there is a very large range, for example LIMNADES had 170 registered users in total, while open repositories such as GEMstat have around 400 per year (Table 5). Open repositories overall have very large numbers of users but for the relevant databases or data packages the numbers are lower. The number of citations largely depends on the acknowledgement requirements for different data, which

is stipulated by the repository and the data providers. Pangaea, Zenodo and Hydroshare give usage details per package.

Table 5. *Estimated number of users for repositories where possible and corresponding number of citations retrieved using Google Scholar in Jan 2023.* (Table to be continued)

	Est. number of users	**Est. number of citations (Google Scholar)**
LIMNADES	~170 registered users in total	
GRDC	~700 per month	~150 per year
GEMstat	~400 per year	
ISIMIP (lake sector)	~200 total	

4.Future needs and recommendations

The organisation and storage of in situ data is an important influencing factor in the compatibility of in situ data with satellite-EO data. In this section, the focus will be to 1) consider the best practises for organising and storing in situ data for different applications with satellite-EO data, 2) identify the barriers in providing this, and 3) offer some future scenarios for improving the storage of in situ data for uptake in satellite-EO products.

4.1 Barriers to the uptake/use of in situ data repositories by satellite-EO end-users

The review of existing in situ data repositories highlighted some substantial challenges and barriers to the uptake of in situ data by the satellite-EO data user community. Here, some of the key overall challenges are highlighted, but it is important to note that the list of repositories included in this report was not exhaustive, and all the data with the repositories was not inspected.

The key issues were as follows;

- Inland and coastal in situ data is scattered across many different repositories, in many different formats and with different levels of accessibility. A lack of transboundary, standardised datasets / many inconsistent datasets from different sources. This is a barrier for new or potential users looking for specific variables or spatio-temporal data. This is also a barrier for bigger scale attempts to gather and standardise relevant data to facilitate the use within the satellite-EO community
- Repositories (and monitoring schemes or sampling campaigns), are often not designed or focused on inland and coastal communities (e.g. limnologists, oceanographers), nor for EO communities, which is hence reflected by;

    - A lack of radiometric data
    - A lack of data from both inland and coastal waters
    - A lack of in situ data that matches up with satellite overpasses
    - A lack of data already formatted to be matched up with EO data
- Repositories, particularly project-based repositories tend to be poorly maintained - often 'frozen' or 'abandoned' after the project end-date
- Repositories are often difficult to be accessed by new users (e.g., require a login before clarity on what datasets are available)
- Lack of machine-readability prohibits automated data gathering which would be necessary for many applications

Coupling satellite-EO data with in situ data can be done with various applications and have different barriers (Table 6).

Table 6. *Barriers to the use of in situ data repositories for different applications coupling satellite-EO and in situ inland/coastal data.*

Application	**Specific needs from in situ data**
Calibration of optical data	• Lack of water-leaving reflectance data or other radiometric quantities
Validation - measuring uncertainty in sat-EO data	• Water-leaving reflectance data (in situ reference data • large datasets to be representative of spatio-temporal resolution of satellite data and the variability of inland/coastal water bodies • Secured-buoy data for precision georeferencing • Matched-up in situ and satellite-EO data
Algorithm development and validation	• lake water reflectance, Inherent Optical Properties and water constituents
Train AI models / machine learning	• Requires large amount of data •
Data fusion & downscaling	• Requires large amount of data
Reconstruction of missing data (e.g. dead lines, gaps, cloud cover)	• Requires large amount of data with good temporal/spatial coverage
Real-time use	• Timeliness of data publishing • automated machine-readable access

4.2 Challenges faced by in situ data providers, gatherers, and database managers

In situ database providers face various challenges in ensuring this data is effectively used. We identify some of the key barriers. These include;

- **Funding barriers**; the most obvious challenge to providing a repository is funding. All aspects of data collection, storage and organisation are influenced by the amount of funding and the resources provided to curating and maintaining the data repository. Some key examples of challenges which have funding barriers are:
    - Project funding does not cover the storage of data
    - Project funding does not cover the maintenance of database after project end-date
    - Time for data management is not included in the project funding
    - Insufficient funding for the desired harmonisation, quality control etc of data
    - Longevity of funding is not confirmed
- **Complexity/skill barriers**: Many different aspects of in situ data storage and organisation requires specific skills and experience which can be a barrier. Examples of challenges that have complexity barriers are below;
    - Building a data repository (especially in academic context)
    - Formatting spectral data in matched-up with satellite data
    - Metadata for spectral/radiometric data can be complicated and requires experienced persons (see Table 3; metadata requirements from LIMNADES)
    - Data ownership / hybrid requires automated or manual request and permission process
    - Data ownership requires legal understanding and many times it depends on laws set by different countries

    - Effective automated quality check or data processing
- **Time/effort barriers:** Even where the funding is available, some of the requirements are time consuming and arduous. These include;
    - Harmonising and standardising data from different sources
    - Need for manual quality check and assurance prior to publishing data
    - Formatting data, especially matched-up data can be highly time consuming (Pavlehan et al., 2021).
- **Social/organisational barriers** - data repositories may not be optimised by data providers or users for various reasons including;
    - Reluctance by potential users to use the repository even if the criteria for uptake are met
    - Lack of incentive for data providers to submit data or to meet the necessary quality and metadata requirements
    - Lack of attribution for data gatherers to sustain repositories

4.4 Recommendations

Alignment between data collection, data gatherers and end-users

As noted in other deliverables (e.g., D4.2, In situ workshop), there is a strong need for greater alignment between data producers, providers, gatherers, managers, funders and end-users. It was also apparent from this review, that inland and coastal data users, and satellite-EO community, are not a priority by in situ data producers or repository curators. When evaluating, planning and executing any future actions to improve the uptake of in situ data by the satellite-EO community, it will be necessary to get input from all stakeholders. For example, data producers and end-users need to communicate to determine how data producers can be incentivised to produce high-quality data and how end-users can get relevant and sufficient data for their applications, with particular focus on enhancing the match-up with satellite overpasses (Carvalho et al., 2021) . Aligning may also require agencies and service

providers of satellite data products to become more involved in the collection and long-term monitoring of necessary in situ measurements (Loew et al., 2017). Another aspect may include the standardisation of practises for in situ observation networks (D4.2), or alternative observational data (D4.4). Notably, however, standardisation will require a 'to-and-fro' process, since both data producers and data users need to have their needs met to make a sustainable alliance. In addition, it is worth noting that the goal is not always to have a perfectly uniform data collection regime, particularly in areas where the methodologies are still evolving. As noted by Pahlevan et al., (2021) when collecting radiometric in situ data to match with satellite data for the ACIX-Aqua, "creating a large pool of data using a combination of various methods likely minimizes any systematic errors in our performance assessments".

Centralised coordination

There is also a real need to capture and review the scope and efforts of different repositories and initiatives across the EU, and globally. Following this report, a systematic overview is suggested to identify the gaps for different scientific communities. One of the criticisms of scientific data is that repositories or databases are developed without thought, either for a short-term project or in response to the specific needs of a group of scientists, and may not be useful, accessible or compatible with the needs of the wider scientific community. Hence, careful analysis is necessary to optimise the use of a repository. Analysis or overview could be useful in discerning where needs are being fully met, where re-organisation is needed, whether an existing repository can be expanded or where a complete gap exists. This might include identifying all current repositories, substantial monitoring schemes, and even grasp the amount of less formal data, including that form low-cost sensors or citizen science. A following dissemination activity might be how to maximizing the data that is already available. For this we can investigate projects such as the MONOCLE project, where data was harmonise from various sources including low-cost sensors or citizen science sources. Maximising the use of existing

data may also involve increasing the profile of existing data and accessibility of data, while also providing capacity building activities to prohibit barriers to uptake.

Promoting open-access data

Open-access oriented science could also save a lot of problems, but there is a need to compensate or incentivise data producers. When data is open-access it removes the need for repositories to manage the release of the data. It also allows data to be uploaded efficiently so it can be used rapidly after collection, which is essential for applications such as near real-time monitoring or early warning systems (e.g., cyanobacteria alert).

Greater focus on data management

Data management in general appears to be overlooked within academic projects that propose a database or repository, funding agents and at the institutional level. For example, including data management planning within a project proposal would help to ensure data is made accessible before the end of a project and completion of funding. Funding agents may also consider including data management as part of the funding criteria. Funders may also require information on what will happen to the repository or database after the project. At the institutional level there could be greater investment in permanent data repository curators and managers, who have the expertise and time to work on the maintenance and consistency of datasets and improve the usability of data that institutions produce. Where there is a lack of focus on data management, there is a loss of quality and a decline in usability of the data.

Greater focus on new or potential users

There are very large barriers faced by new or potential users. Whether this be due to 1) insufficient discovery metadata, 2) barriers to accessing the data, 3) lack of educational/instructional guide on how to use the data, 4) lack of metadata and interpretability, 5) lack of trust in the data. Although repositories mostly meet FAIR

requirements (see D4.5), there is a substantial way to go before the data is fully accessible to a member of the general public, an undergraduate student or a water manager, and is fully maximised in it's possible use. - could use the in situ survey results.

Future use of repositories

Repositories in general could improve. For example, focus from FAIR to TRUSTworthy repositories (Lin et al., 2020). Copernicus Water Services: Inland and coastal in situ data is still a long way off from being used directly in Copernicus Water Services? There is also the added consideration of whether data is being stored efficiently and sustainably.

4.4 Future possibilities for storing in situ data

There are various future options for storing in situ data for the purpose of coupling with satellite-EO data. Some of these possible options are considered here.

4.4.1 Metadata repositories for in-situ data

One option is a metadata repository that facilitates the discovery of in situ data for the use with satellite-EO data. Currently there is no designated metadata repository that offers a central access point for different inland and coastal in situ data repositories. GEOSS and NETLAKE, however, both include in situ data that could be used with satellite-EO data. The main benefit of metadata repositories is likely to be the discovery of data, connecting end-users to data providers. A metadata repository could also provide an overview of data options, for example, with a map-based browser to guide users to relevant data providers. This may improve clarity on where the data gaps are and also improve communication between data providers and satellite-EO based end-users. A metadata repository removes the risk of copying data, and may have lower curation and maintenance costs than a data hosting repository, although if it is to be useful it would need to be sustainably funded. Notably, however, a metadata repository does not necessarily meet many of the

requirements of many of the satellite-EO data applications, but may be a good feasible option, in the right direction. Another option would be to have a Re3 type central body which coordinates all EU-based data repositories, which could require certain metadata etc standards to be met.

4.4.2 Extend an existing database

Data repositories that are already storing in situ data for inland and coastal waters could be extended, with sufficient funding and resources, to further meet the needs of the satellite-EO community. For example, SeaBASS and LIMNADES have are both vital in providing match-up data between in situ reference data and satellite overpasses. SeaBASS for example, could be expanded to include more inland water data. LIMNADES could also be refunded, upgraded, and expanded to include more match-up data or the accessibility and findability could be enhanced to ensure the data is more widely used. Other repositories, such as PANGAEA, may also already have the framework, capacity and flexibility to house the existing or new data.

4.4.3 A built-for-purpose repository

A built-for-purpose data repository could be designed with the intention and purpose of storing and organising data for different applications with satellite-EO data. A model for this might be adopted from existing managed open data repositories included in this report (e.g. GEMsWater), where data is consolidated from many sources and can be well presented for new end-users. A designated in situ data repository would need to standardise and reformat data from various sources, into a centralised data system. Lessons on how to do this efficiently might also be learnt from the MONOCLE repository and also the SWatCH and ACIX-Aqua projects. There are many challenges in building a specific repository, and preparation would require considering and recruiting the advice of all stakeholders (i.e., data providers, different end-users), to ensure the repository is well-received and used. Moreover, a substantial amount of effort and funding would be required and

hence the need for the repository would need to be confirmed. This may include determination of the exact scope (.e.g, inland and coastal data only, for validation or all coupling with in situ data). Flexibility would also be advised, considering how rapidly the methods of the satellite-EO community evolve. Similarly, data repository curation and commonly accepted data formats also change over time and would require specialist knowledge to ensure longevity. Hence, this undertaking requires.

Funding and organisation structure could take vicarious forms. One example is that storage space is made available to EU-based national or long-term monitoring schemes on the EOS cloud, alleviating costs for data storage by institutions. In return, collection protocol may be encouraged, promoting harmonisation across data producers and the uptake of practices relevant for satellite-EO applications. Alternatively, the liaison, management and harmonisation of data could be carried out by a data harmonising agent, which could potentially be a new space for a business innovation.

4. Conclusion

The organisation and storage of in situ data is important in the coupling of in situ and satellite-EO data. A number of existing inland and coastal data repositories were reviewed in this report. The review highlighted various unmet needs from the satellite-EO community as well as from data curators and managers responsible for gathering, organising and storing data. This report can be used as a starting place for future efforts to align the different stakeholders (e.g., data producers, gatherers and end-users) to improve the uptake of in situ data in satellite-EO based applications.

References

Carvalho, L., Ruescas, A., Stelzer, K., Brockmann, C., Taylor, P., Dobel, A., Nash, G. and Fry, M., 2021. Lake water quality in-situ data requirements and availability. Copenhagen, European Environment Agency, 70pp. (Unpublished)

Golub, M., Thiery, W., Marcé, R., Pierson, D., Vanderkelen, I., Mercado-Bettin, D., Woolway, R.I., Grant, L., Jennings, E., Kraemer, B.M. and Schewe, J., 2022. A framework for ensemble modelling of climate change impacts on lakes worldwide: the ISIMIP Lake Sector. *Geoscientific Model Development*, *15*(11), pp.4597-4623.

Lehner, B., Messager, M.L., Korver, M.C., Linke, S. (2022). Global hydro-environmental lake characteristics at high spatial resolution. Scientific Data 9: 351. doi: https://doi.org/10.1038/s41597-022-01425-z

Lin, D., Crabtree, J., Dillo, I., Downs, R.R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R. and Khodiyar, V., 2020. The TRUST Principles for digital repositories. *Scientific Data*, *7*(1), pp.1-5.

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., Thieme, M. (2019). Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. Scientific Data 6: 283. doi: https://doi.org/10.1038/s41597-019-0300-6

Loew, A., Bell, W., Brocca, L., Bulgin, C.E., Burdanowitz, J., Calbet, X., Donner, R.V., Ghent, D., Gruber, A., Kaminski, T. and Kinzel, J., 2017. Validation practices for satellite-based

Earth observation data across communities. *Reviews of Geophysics*, *55*(3), pp.779-817.

Pahlevan, N., Mangin, A., Balasubramanian, S.V., Smith, B., Alikas, K., Arai, K., Barbosa, C., Bélanger, S., Binding, C., Bresciani, M. and Giardino, C., 2021. ACIX-Aqua: A global assessment of atmospheric correction methods for Landsat-8 and Sentinel-2 over lakes, rivers, and coastal waters. *Remote Sensing of Environment*, *258*, p.112366.

Rimet, F., Anneville, O., Barbet, D., Chardon, C., Crepin, L., Domaizon, I., Dorioz, J.M., Espinat, L., Frossard, V., Guillard, J. and Goulon, C., 2020. The Observatory on LAkes (OLA) database: Sixty years of environmental data accessible to the public. *Journal of Limnology*, *79*(2), pp.164-178.

Rotteveel, L., Heubach, F. and Sterling, S.M., 2022. The Surface Water Chemistry (SWatCh) database: A standardized global database of water chemistry to facilitate large-sample hydrological research. *Earth System Science Data*, *14*(10), pp.4667-4680.

Warren, M.A., Simis, S.G., Martinez-Vicente, V., Poser, K., Bresciani, M., Alikas, K., Spyrakos, E., Giardino, C. and Ansper, A., 2019. Assessment of atmospheric correction algorithms for the Sentinel-2A MultiSpectral Imager over coastal and inland waters. *Remote sensing of environment*, *225*, pp.267-289.