hydrology

MDPI

Article

WaterDataHub 1.0: A Curated Discovery Resource for AI-Driven Water Research

Tea Dobovšek [1], Barbara Hribar-Lee [2], Miha Moškon [1], Andreja Žgajnar Gotvajn [2,3], Jane Tanner [4] and Zoran Bosnić [1,*]

[1] Faculty of Computer and Information Science, University of Ljubljana, 1000 Ljubljana, Slovenia; tea.dobovsek@gmail.com (T.D.); miha.moskon@fri.uni-lj.si (M.M.)
[2] Faculty of Chemistry and Chemical Technology, University of Ljubljana, 1000 Ljubljana, Slovenia; barbara.hribar@fkkt.uni-lj.si (B.H.-L.); andreja.zgajnar@fkkt.uni-lj.si (A.Ž.G.)
[3] Hugh Kelly Fellow, Rhodes University, Makhanda 6140, South Africa
[4] Institute for Water Research, Rhodes University, Makhanda 6140, South Africa; j.tanner@ru.ac.za
* Correspondence: zoran.bosnic@fri.uni-lj.si

Abstract

The increasing availability of publicly accessible data has created new opportunities for data-driven and AI-supported water research. Such data are particularly attractive due to their transparency, cost-effectiveness, and ability to support reproducible scientific studies. Moreover, publicly available datasets offer strong potential for methodological transferability across geographical locations, enabling comparative analyses, model generalization, and knowledge sharing across regions with diverse environmental and socio-economic conditions. In this work, we explore both the benefits and limitations of conducting AI-driven research using publicly available datasets, emphasizing their role in accelerating innovation and enabling broader participation in scientific inquiry. We specifically focus on water-related data, including environmental monitoring datasets that capture various aspects of water quality, availability, and ecosystem dynamics. These datasets typically rely on measurements such as chemical composition, temperature, flow rates, and pollutant levels, often collected through governmental or international monitoring initiatives. We introduce WaterDataHub 1.0, a curated discovery resource for publicly available data sources relevant to water research and artificial intelligence applications. The principal contribution of this work is WaterDataHub 1.0 itself: a centralized, cross-domain, and AI-oriented discovery resource that organizes fragmented water-related data sources within a common framework and facilitates their identification and reuse for data-driven research.

Keywords: water research; database; machine learning

Check for updates

Academic Editors: Vasilis Kanakoudis, Evangelos Keramaris and Francesco De Paola

Received: 22 October 2026
Revised: 8 November 2026
Accepted: 9 November 2026
Published: 11 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Water resource management is a critical global challenge driven by climate change, urbanization, and increasing demand for freshwater. These pressures have intensified the need for efficient monitoring, prediction, and management of water systems. Recent advances in artificial intelligence (AI) and machine learning (ML) provide powerful tools for addressing these challenges, enabling complex modeling and data-driven decision-making [1].

AI has been increasingly applied across water-related domains, including water quality assessment, hydrological forecasting, and infrastructure management. Systematic reviews indicate that AI-driven approaches significantly improve predictive accuracy and enable real-time decision support [2,3].

A key requirement for successful AI applications is access to high-quality and well-structured data. Publicly available datasets are essential for reproducibility, transparency, and scalability of research. However, public availability alone does not guarantee that water-related data can be readily discovered and reused. Water data are distributed across heterogeneous organizations, platforms, formats, and jurisdictions, creating challenges for data discovery and integration [4]. Existing infrastructures, such as the Water Quality Portal [5] and HydroShare [6], address parts of this challenge, while the FAIR principles emphasize the importance of findability, accessibility, interoperability, and reusability [7]. Nevertheless, facilitating the discovery and comparison of distributed resources across different water-research domains remains an important challenge.

The principal contribution of this work is WaterDataHub 1.0, a curated, cross-domain, and AI-oriented discovery resource for publicly available water-related datasets. Rather than introducing another repository for storing environmental observations, WaterDataHub provides a structured discovery layer over existing distributed resources. It consolidates data sources spanning six major water-research domains, characterizes them using common metadata relevant to their accessibility and potential use in AI and machine-learning workflows, and provides researchers with a unified entry point for identifying suitable data resources. In this way, WaterDataHub aims to facilitate data reuse, interdisciplinary research, reproducibility, and the development of data-driven approaches in water research [8].

The remainder of this paper is organized as follows. Section 2 describes the methodology used to develop WaterDataHub 1.0, including data-source identification, curation, and classification. Section 3 presents the collected resources, their characteristics, and potential AI/ML applications across the six research domains. Section 4 discusses the contributions and limitations of WaterDataHub and compares it with existing resources. Section 5 concludes the paper and outlines future directions.

2. Methods

The development of WaterDataHub 1.0 followed a structured methodology comprising three main stages: identification and curation of publicly available water-related data sources, characterization and classification of the identified resources using a common metadata framework, and organization of the resources into six thematic research domains. The following subsections describe these stages and the methodological decisions underlying the construction of WaterDataHub 1.0.

2.1. Data Source Identification and Curation

WaterDataHub 1.0 was developed through a structured process for identifying and curating publicly available water-related data sources, conducted in April and May 2025. Candidate sources were identified through:

- Review of peer-reviewed journal articles, predominantly published within the preceding five years, primarily using ScienceDirect and Google Scholar;
- Exploration of publicly accessible data repositories, governmental and international environmental databases, and research data infrastructures;
- Identification of data sources commonly used in recent water-related scientific and AI-based studies.

The search emphasized publications from the preceding five years to capture recent developments in water-related data resources and AI/ML applications, without excluding older publications describing relevant or established datasets. ScienceDirect and Google Scholar were used as complementary literature discovery tools, combining access to peer-reviewed journal publications with broader coverage across publishers and research disci-

plines. The identification process was not restricted to these tools, as it also included direct exploration of data repositories and research infrastructures.

The literature and repository searches used combinations of domain-specific water-related keywords and data-discovery terms, including "dataset", "database", "data repository", "open data", and "publicly available data". The searches covered the six research domains, using the following keywords and representative environmental variables or indicators to refine the identification of relevant data sources:

1. **Water quality monitoring:** "water quality" and "water quality monitoring", supplemented by parameters such as biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), dissolved oxygen, pH, conductivity, ammonia, nitrate, phosphate, heavy metals, *E. coli*, and chlorophyll-a.
2. **Hydrology and hydrodynamics:** "hydrology" and "hydrodynamics", supplemented by variables such as river discharge, water level, flow velocity, precipitation, snow depth, soil moisture, water temperature, waves, and ocean currents.
3. **Flood and drought forecasting:** "flood forecasting", "flood modeling", "flood hazard", "drought forecasting", "drought monitoring", and "drought index", supplemented by river discharge, water level, precipitation, soil saturation, soil moisture, and soil temperature.
4. **Aquatic ecosystems and biodiversity:** "aquatic ecosystems", "freshwater biodiversity", "ecological status", and "habitat quality", supplemented by biological and ecological indicators such as fish, macroinvertebrates, phytoplankton, zooplankton, macrophytes, amphibians, species richness, chlorophyll-a, cyanobacteria, harmful algal blooms (HABs), and trophic state.
5. **Water pollution and pollutant detection:** "water pollution", "water quality", "water monitoring", and "pollutant detection", supplemented by pollutant categories such as heavy metals, pesticides, microplastics, per- and polyfluoroalkyl substances (PFAS), pharmaceuticals, polycyclic aromatic hydrocarbons (PAHs), polychlorinated biphenyls (PCBs), and endocrine disruptors.
6. **Water treatment and wastewater management:** "wastewater", "water discharge", "water treatment", and "sewage", supplemented by treatment-related parameters such as BOD_5, COD, total suspended solids (TSS), ammonia nitrogen, total phosphorus, heavy metals, and nitrate.

Additional searches combined domain-specific terms with "machine learning", "data modeling", and "artificial intelligence" to identify data resources used in AI-based water research. Searches were conducted iteratively, with terminology adapted to individual research domains, relevant environmental variables and indicators, and additional resources identified during the review. To ensure broad coverage of established water-data infrastructures, the identification process considered major resources across water quality, hydrology, Earth observation, and environmental monitoring.

Candidate sources were screened according to the following criteria:

- Public accessibility of the data;
- Availability of data through direct download, APIs, or other documented access mechanisms;
- Relevance to at least one of the water-related research domains;
- Prior use in scientific or AI-based research, where documented;
- Availability of sufficient documentation to characterize the source within WaterDataHub 1.0.

Candidate sources were first checked for duplication and relevance to the defined research domains. Sources were then assessed for public data accessibility, availability of

documented access mechanisms, and sufficient metadata to support their characterization within WaterDataHub. Prior use in scientific or AI-based studies was considered as supporting evidence of relevance, but was not a mandatory inclusion requirement. Sources failing the accessibility, thematic relevance, or documentation requirements were excluded.

WaterDataHub 1.0 is intended as a curated rather than exhaustive catalog and should not be interpreted as providing complete coverage of all publicly available water-related data sources. The objective of the identification process was to achieve broad thematic coverage of publicly accessible data resources relevant to contemporary water research and potential AI/ML applications, rather than to obtain a statistically representative sample of all existing water-related datasets. Because these resources are distributed across heterogeneous and continuously evolving infrastructures, no comprehensive reference population was available against which the completeness of the selected collection could be quantitatively assessed. The resulting catalog therefore reflects the resources identified through the search procedures and inclusion criteria described above, without implying exhaustive coverage of any of the six research domains.

The initial search identified 180 candidate data sources. Following the screening process, nearly two-thirds of the identified sources were excluded: 35 duplicates, primarily resulting from different studies referring to the same public data sources; 68 sources without publicly accessible data; 4 sources outside the defined water-related scope; and 10 sources for which insufficient information was available for adequate characterization. The remaining 63 data sources were retained and included in WaterDataHub 1.0. The identification and screening process is summarized in Figure 1.

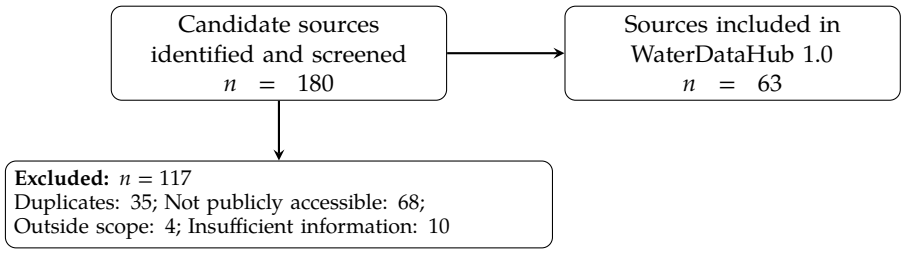


Figure 1. Data-source identification, screening, and inclusion workflow for WaterDataHub 1.0.

2.2. Classification Framework

To enable consistent organization and comparison of the identified data sources, we developed a classification framework based on their main characteristics and potential research applications. The collected data sources were characterized using the following metadata fields, informed by common practices in environmental data catalogs:

- Data source name (the official name of the data source or dataset, as provided by its publisher);
- Data provider (the organization or institution identified as responsible for maintaining or publishing the data) and access URL;
- Topic/Thematic domain (water quality, hydrology, floods and droughts, ecosystems, pollution, wastewater; assigned according to the variables provided and the source's stated purpose);
- Geographic coverage (global, continental/multinational, national, regional; assigned according to the geographical extent of the available data, rather than its spatial resolution);
- Typical variables (free text, e.g., pH, temperature, flow, water level, nutrients, biodiversity; representative variables identified in dataset descriptions and documentation);

- Data acquisition method (sampling, in-situ, crowdsourcing, modeling, remote sensing, administrative records, measurements; identified from the documented data collection or generation methods);
- Data format (tabular/text, spatial, image) assigned according to the data representations and products described or made available by the provider.

The metadata fields used in WaterDataHub are largely consistent with those commonly employed in environmental data catalogs and do not constitute a new AI-specific metadata standard. Their relevance to AI/ML research lies in supporting the preliminary identification of resources suitable for particular modeling tasks. For example, information about available variables helps identify potential predictors and target variables; geographic coverage supports the selection of resources for spatial modeling and cross-regional analyses; data acquisition methods and formats provide an initial indication of the data preparation and processing approaches that may be required; and thematic classification facilitates the identification of complementary resources across research domains. These characteristics support initial dataset discovery and shortlisting, while detailed assessment of spatial and temporal resolution, data quality, completeness, and suitability for a specific AI/ML application remains the responsibility of the researcher. This classification aligns with approaches used in recent reviews of hydrological and environmental datasets [3].

The six research domains are not mutually exclusive, and a data source may therefore be assigned to multiple domains when its content supports several types of water-related research. Classification is based primarily on the type and intended application of the provided data. In particular, *water quality monitoring* covers general physical, chemical, and biological indicators of water quality; *water pollution and pollutant detection* focuses specifically on contaminants and pollution events; and *water treatment and wastewater management* covers data related to treatment processes, wastewater systems, and treatment-plant operation. For example, a source containing both general water-quality parameters and pollutant measurements may be classified under both water quality monitoring and water pollution. This multi-label classification enables individual data sources to be represented in all relevant thematic domains, facilitating their identification across overlapping research areas.

Initial domain assignments were made by the first author according to the criteria described above, while ambiguous or overlapping cases were reviewed collaboratively and resolved through consensus among the authors. Classification consistency was therefore ensured through consensus review rather than through a formal inter-coder agreement analysis.

2.3. Thematic Classification of Sources by Research Area

The resources included in WaterDataHub 1.0 provide data relevant to various areas of water research. Based on the types of data offered by individual sources, six research areas were identified within the scope of this project:

- Water quality monitoring;
- Hydrology and hydrodynamics;
- Flood and drought forecasting;
- Aquatic ecosystems and biodiversity;
- Water pollution and pollutant detection;
- Water treatment and wastewater management.

The six-domain framework was developed to represent the major recurring application areas identified during the collection and characterization of water-related data resources. It is intended as a practical organizational framework for WaterDataHub rather than as a mutually exclusive or standardized taxonomy of water research.

Of the 63 data sources included in WaterDataHub 1.0, 30 (47.6%) were assigned to a single research domain, while 33 (52.4%) were assigned to multiple domains. Specifically, 19 sources were assigned to 2 domains, 9 to 3 domains, 1 to 4 domains, 2 to 5 domains, and 2 to all 6 domains. Overall, the classification resulted in 121 thematic assignments, reflecting the interdisciplinary nature of many water-related data resources.

3. Results

WaterDataHub 1.0 (http://bit.ly/water-data-sources, accessed on 1 November 2026) currently contains 63 curated publicly available data sources covering the six research domains defined in Section 2.3. The complete inventory and associated metadata are accessible through the WaterDataHub 1.0 catalog. The representative variables and indicators listed in Table 1 illustrate the types of environmental information associated with each domain, while data acquisition methods and formats are characterized separately in the WaterDataHub catalog. Because the domains are non-exclusive, individual resources may contribute to more than one category. The source-identification and screening procedure, including the reasons for excluding candidate sources, is documented in Section 2 and summarized in Figure 1.

Table 1. Quantitative overview of the data sources included in WaterDataHub 1.0 across the six research domains. As individual data sources may be assigned to multiple domains, the domain counts are not mutually exclusive.

Research Domain	No. of Sources	Geographical Coverage	Representative Variables and Indicators
Water quality monitoring	26	Global, continental, regional, national	pH, dissolved oxygen, nutrients, temperature
Hydrology and hydrodynamics	25	Global, continental, national	Precipitation, streamflow, groundwater level, water level
Flood and drought forecasting	18	Global, continental, regional, national	Precipitation, river discharge, soil moisture, drought indices
Aquatic ecosystems and biodiversity	18	Global, continental, regional, national	Species richness, biomass, chlorophyll-a, ecological indicators
Water pollution and pollutant detection	18	Global, continental, regional, national	Heavy metals, pesticides, microplastics, organic pollutants
Water treatment and wastewater management	14	Global, continental, regional, national	BOD_5, COD, suspended solids, nutrient concentrations
Total unique sources	**63**	–	–

Beyond their thematic distribution, the resources included in WaterDataHub 1.0 were quantitatively characterized with respect to data accessibility, temporal characteristics, and metadata completeness. These characteristics provide additional information on the practical usability of the identified resources and their potential suitability for data-driven research. Table 2 summarizes the main results.

The temporal characteristics in Table 2 distinguish between the documentation of temporal coverage and the availability of time-indexed observations. Documented observation dates or temporal coverage indicate that the resource specifies when observations were collected or the period covered by the data, without necessarily providing a time series. Historical time-indexed observations indicate that the resource contains measurements associated with different points in time, potentially supporting temporal analyses. These

characteristics are not mutually exclusive, and the availability of historical observations does not necessarily imply that the overall temporal coverage is explicitly documented.

Table 2. Quantitative characteristics of the data resources included in WaterDataHub 1.0. Accessibility categories are mutually exclusive, with each resource assigned to one category, whereas temporal characteristics are not mutually exclusive, as individual resources may exhibit multiple temporal properties.

Characteristic	No. of Resources	Percentage
Total resources	63	100%
Accessibility		
Direct access	41	65.1%
Registration, request, or other access mechanism required	13	20.6%
Both direct and gated routes	9	14.3%
Temporal characteristics		
Documented observation dates or temporal coverage	55	87.3%
Availability of historical time-indexed observations	56	88.9%
Recurring data updates documented	36	57.1%
Real- or near-real-time data documented	14	22.2%
Metadata completeness		
All principal metadata fields available	61	96.8%
Fields with complete metadata: name, provider, geographic area	63	100%
Fields with incomplete metadata: URL, topic, variables, method, format	62	98.4%

The quantitative characterization reveals variation among the identified resources in both accessibility and temporal properties. Most resources (79.4%) provide at least one documented direct access route, although some also require registration or permission for specific datasets. Historical time-indexed data are available in 88.9% of resources, and recurring updates are documented for 57.1%. Real- or near-real-time data are confirmed for a smaller proportion (22.2%). Temporal coverage varies across datasets and monitoring stations: documented examples include ICES phytoplankton observations from 1901 to 2025 and CHIRPS rainfall estimates from 1981 to near-present. A single overall temporal range could not be established for the heterogeneous collection of resources. The high completeness of catalog fields (96.8% of entries) supports resource discovery but does not, by itself, establish data quality, timeliness or suitability for a particular analysis.

The following sections characterize the types of publicly available data identified within each domain and, importantly, examine the AI-driven research opportunities they enable. This dual perspective reflects the central objective of WaterDataHub: not only to facilitate discovery of publicly available water-related data, but also to connect these resources with their potential use in data-driven and AI-supported research.

3.1. Water Quality Datasets

The data sources identified within the water quality domain provide physical, chemical, and biological measurements. Typical variables include temperature, pH, dissolved oxygen, and nutrient concentrations. The identified resources include both in-situ monitoring data and remotely sensed observations and vary considerably in their temporal and spatial coverage [9–17].

A review of more than 170 recent studies shows that ML methods are widely used for both single-variable and multi-variable prediction tasks, demonstrating high accuracy and flexibility [2]. From the perspective of AI-driven research, the characteristics of these publicly available data support several types of machine-learning tasks:

- **Regression and water-quality prediction**—prediction of continuous variables and indices, such as dissolved oxygen, nutrient concentrations, and water-quality indicators;
- **Classification**—assignment of observations or locations to predefined water-quality classes;
- **Time-series forecasting**—prediction of future water-quality conditions, trends, and seasonal variations;
- **Anomaly and event detection**—identification of unusual measurements, abrupt changes, and pollution events;
- **Spatial modeling and remote-sensing analysis**—estimation and mapping of water-quality characteristics across geographical areas.

These examples illustrate how the water-quality resources collected in WaterDataHub 1.0 can serve not only as sources of environmental observations but also as inputs for a range of AI-driven research tasks, depending on their variables, temporal resolution, spatial coverage, and data availability.

3.2. Hydrological Datasets

The data sources identified within the hydrology and hydrodynamics domain provide measurements and estimates of key hydrological variables, including precipitation, streamflow, groundwater levels, and evapotranspiration [18–23]. They include widely used resources such as CAMELS (Catchment Attributes and Meteorology for Large-sample Studies), GRDC (Global Runoff Data Centre), CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data), GLDAS (Global Land Data Assimilation System), and GRACE (Gravity Recovery and Climate Experiment), providing extensive spatial and temporal coverage and supporting the analysis of water availability, hydrological dynamics, and extreme events [3].

From the perspective of AI-driven research, these data support several types of machine-learning tasks:

- **Streamflow and runoff prediction**—estimation and forecasting of river discharge and catchment runoff;
- **Groundwater modeling**—prediction of groundwater levels and their temporal variation;
- **Hydrological time-series forecasting**—prediction of precipitation, streamflow, and other hydrological variables;
- **Water-availability modeling**—estimation of current and future water availability from multiple hydrological variables;
- **Spatiotemporal modeling**—analysis and prediction of hydrological processes across locations and time.

3.3. Flood and Drought Data

The data sources identified within the flood and drought domain combine meteorological, hydrological, and geospatial information, including precipitation, streamflow, soil and land characteristics, and remotely sensed observations [24–30]. Their integration provides the spatial and temporal information required for modeling extreme hydrological events and their development across different geographical areas.

From the perspective of AI-driven research, these data support several types of machine-learning tasks:

- **Flood forecasting**—prediction of flood occurrence, magnitude, and peak discharge;
- **Flood susceptibility assessment**—identification and mapping of areas with elevated flood risk;

- **Flood detection and mapping**—identification of inundated areas from satellite and remote-sensing data;
- **Drought prediction**—forecasting of drought occurrence, severity, and temporal evolution;
- **Spatiotemporal extreme-event modeling**—analysis and prediction of flood and drought patterns across space and time.

These applications particularly benefit from multi-source data integration, which combines complementary meteorological, hydrological, and Earth-observation information for modeling complex extreme events [3]. However, combining heterogeneous data sources also requires careful consideration of data quality and compatibility, including differences in spatial and temporal resolution, measurement methods, completeness, and consistency, as these factors may affect model reliability and the interpretation of results.

3.4. Aquatic Ecosystem and Biodiversity Data

The data sources identified within the aquatic ecosystems and biodiversity domain provide information on species distribution, biomass, ecological indicators, water conditions, and related environmental characteristics [31–34]. These data combine biological and environmental observations and can therefore support analyses of ecosystem status and its response to environmental change.

From the perspective of AI-driven research, these data support several types of machine-learning tasks:

- **Species and habitat modeling**—prediction of species distribution and habitat suitability;
- **Biodiversity modeling**—analysis and prediction of changes in biological diversity;
- **Ecosystem health assessment**—classification and prediction of ecological conditions from biological and environmental indicators;
- **Harmful algal bloom detection**—identification and monitoring of algal blooms using water-quality and remote-sensing data;
- **Environmental impact prediction**—assessment of ecosystem responses to environmental and climatic changes.

These applications demonstrate the importance of combining biological, environmental, and remote-sensing data for AI-supported monitoring and modeling of aquatic ecosystems.

3.5. Water Pollution and Pollutant Detection

The data sources identified within the water pollution domain provide measurements of contaminants such as heavy metals, pesticides, herbicides, industrial chemicals, pharmaceutical contaminants, petroleum derivatives, and microplastics [35–37]. Such data support the identification of contamination, analysis of pollution patterns, and assessment of changes in pollutant concentrations across space and time.

From the perspective of AI-driven research, these data support several types of machine-learning tasks:

- **Pollutant detection**—identification of the presence of contaminants from environmental and analytical measurements;
- **Pollutant classification**—classification of contaminants or pollution conditions based on measured characteristics;
- **Pollution prediction**—prediction of pollutant concentrations and their temporal evolution;
- **Pollutant dispersion modeling**—estimation of the spatial distribution and spread of contaminants;
- **Anomaly and pollution-event detection**—identification of unusual measurements and emerging contamination events.

AI-based analysis is particularly relevant in this domain because conventional analytical methods can generate large and complex datasets. Machine-learning methods can support automated and more efficient processing and interpretation of these measurements [38].

3.6. Water Treatment and Wastewater Management

The data sources identified within the water treatment and wastewater management domain primarily provide physical, chemical, and operational parameters related to wastewater generation, treatment processes, and treatment-plant operation [39–42]. Although closely related to pollution monitoring, this domain focuses specifically on the treatment and management of wastewater. Measurements collected at treatment-plant inlets and outlets can be used to assess treatment efficiency, while operational data provide information on process performance and plant conditions.

From the perspective of AI-driven research, these data support several types of machine-learning tasks:

- **Treated-water quality prediction**—prediction of effluent quality from influent and process parameters;
- **Treatment-process optimization**—optimization of treatment efficiency and operational conditions;
- **Fault and anomaly detection**—identification of equipment malfunctions, measurement errors, and abnormal process behavior;
- **Influent change detection**—detection of sudden changes in wastewater composition for timely process adaptation;
- **Infrastructure detection and monitoring**—identification and monitoring of wastewater infrastructure using geospatial and remote-sensing data.

Combining environmental and operational data can therefore support both predictive monitoring and data-driven management of wastewater treatment systems. Recent studies also demonstrate that integrating environmental and operational datasets improves predictive performance in wastewater systems [38].

3.7. Summary of AI Research Opportunities

Table 3 summarizes the main families of AI and machine-learning tasks supported by the data resources identified across the six thematic domains. The mapping reflects the types of applications discussed in Sections 3.1–3.6 and illustrates both the diversity of AI applications and the overlap between research domains.

Table 3. Illustrative mapping of AI/ML task families to the six thematic domains covered by WaterDataHub 1.0. Checkmarks indicate representative applications supported by the characteristics of the identified data resources. A dash indicates that an application was not specifically highlighted in this overview, rather than that it is inapplicable to the domain.

AI/ML Task Family	Water Quality	Hydrology	Flood & Drought	Ecosystems	Pollution	Wastewater
Prediction/regression	✓	✓	✓	✓	✓	✓
Classification	✓	✓	✓	✓	✓	✓
Time-series forecasting	✓	✓	✓	–	✓	✓
Anomaly/event detection	✓	✓	✓	✓	✓	✓
Spatial/geospatial modeling	✓	✓	✓	✓	✓	✓
Optimization/decision support	–	–	–	–	–	✓

The comparison shows that prediction and spatial modeling are applicable across all six thematic domains, whereas other AI task families are more domain-specific. It also demonstrates the interdisciplinary potential of the identified data resources, as similar AI methodologies can be applied to different types of water-related data and research questions.

3.8. Illustrative Use Case

To illustrate the practical use of WaterDataHub 1.0, consider a researcher aiming to develop a machine-learning model for predicting river water-quality conditions in the United States. The researcher requires water-quality measurements as potential prediction targets and complementary environmental variables that may improve model performance.

Using the water-quality category in WaterDataHub, the researcher identifies the Water Quality Portal (WQP), maintained by the U.S. Environmental Protection Agency (EPA) and the U.S. Geological Survey (USGS), as a relevant data source. Its catalog entry indicates the availability of measurements such as pH, nutrients, biochemical oxygen demand (BOD), dissolved oxygen (DO), temperature, and flow. The data are collected primarily through in-situ measurements and sampling and are available in tabular formats. Based on these characteristics and its coverage of the United States, WQP is shortlisted as a potential source of target variables and related predictors for the intended modeling task.

To identify complementary environmental information, the researcher examines resources classified under hydrology and identifies Copernicus Services as another candidate. According to its WaterDataHub entry, Copernicus provides remotely sensed information, including water temperature, turbidity, chlorophyll, water level, and other environmental characteristics, primarily in image-based formats. These variables may provide additional spatial and environmental context for water-quality prediction.

By comparing the two catalog entries, the researcher can identify their complementary characteristics: WQP provides primarily in-situ water-quality measurements in tabular format, whereas Copernicus offers remotely sensed environmental observations. Both are therefore shortlisted for further investigation. Their combination could potentially support the development of a machine-learning model integrating local water-quality measurements with broader environmental information.

This example demonstrates how WaterDataHub facilitates the identification and preliminary selection of complementary data sources across research domains using their geographic coverage, available variables, acquisition methods, and data formats. However, the catalog does not establish whether the shortlisted datasets provide matching locations, compatible observation periods, appropriate spatial and temporal resolutions, or sufficient data quality. These aspects must be verified through examination of the underlying datasets before integration and model development.

4. Discussion

The fragmented and heterogeneous nature of publicly available water data discussed above provides the principal motivation for WaterDataHub 1.0. The resource addresses this challenge by providing a centralized, curated discovery layer for publicly available datasets relevant to water research and AI-driven environmental applications.

Unlike conventional environmental data repositories, WaterDataHub 1.0 does not aim to replace existing data providers or duplicate raw datasets. Instead, it serves as a metadata-driven discovery platform that facilitates efficient identification of suitable datasets for specific research objectives. This distinction is important because the primary obstacle faced by many researchers is not necessarily data availability, but rather dataset discoverability and the ability to assess dataset suitability prior to data acquisition. By consolidating information from diverse sources into a unified framework, WaterDataHub is

designed to facilitate dataset discovery by providing a common entry point for identifying and comparing resources from different providers. Beyond discovery itself, the ability to compare relevant dataset characteristics within a common framework may improve research efficiency by supporting faster identification, preliminary assessment, and selection of data resources appropriate to a particular research question.

One of the major strengths of WaterDataHub 1.0 is its interdisciplinary coverage. The catalog brings together data sources from multiple domains, including water quality monitoring, hydrology, hydrodynamics, flood and drought forecasting, aquatic ecosystems, pollution monitoring, and wastewater management. Such breadth is particularly important for contemporary environmental research, where complex challenges increasingly require the integration of heterogeneous datasets originating from different scientific disciplines. WaterDataHub 1.0 therefore supports both domain-specific studies and broader interdisciplinary investigations.

Another important contribution of WaterDataHub 1.0 is its emphasis on artificial intelligence and machine learning applications. The rapid development of AI-based environmental analysis has increased demand for large, reliable, and well-characterized datasets. However, many environmental datasets were originally developed for monitoring or regulatory purposes rather than for machine learning applications. Consequently, researchers often face difficulties in determining whether a particular dataset possesses sufficient temporal coverage, spatial resolution, completeness, and accessibility for advanced analytical methods. By providing metadata relevant to data-driven applications, WaterDataHub provides information that can support researchers in assessing the suitability of individual resources for specific AI/ML tasks.

The database also reveals several important characteristics of the current water-data ecosystem. First, substantial differences exist in the availability and accessibility of datasets across geographical regions. Europe and North America account for a large proportion of publicly accessible and well-maintained environmental repositories, whereas significant data gaps remain in many developing regions. These disparities may influence the development and transferability of AI models, potentially limiting their applicability in areas where environmental challenges are often most severe. Second, considerable heterogeneity exists among datasets with respect to metadata standards, file formats, temporal resolution, and access mechanisms. Although recent initiatives promoting FAIR (Findable, Accessible, Interoperable, and Reusable) data principles have improved interoperability [7], further standardization remains necessary to facilitate large-scale integration of environmental information.

Several limitations of WaterDataHub 1.0 should also be acknowledged. WaterDataHub 1.0 currently relies on publicly accessible information and therefore reflects the availability and quality of existing metadata provided by data owners. Some potentially valuable data sources may have been omitted due to limited documentation, language barriers, restricted accessibility, or limitations of the search and identification process. Coverage may also be geographically biased toward regions and organizations providing well-documented resources in readily accessible languages and formats. WaterDataHub 1.0 should therefore be regarded as a curated and extensible resource rather than an exhaustive inventory of all publicly available water-related data sources. Furthermore, the rapidly evolving nature of environmental data infrastructures means that new repositories are continuously being established, while existing datasets may change in scope, format, or accessibility. Consequently, maintaining the relevance of WaterDataHub will require periodic updates, potentially supported by community contributions.

A further limitation of the present work is that WaterDataHub 1.0 has not yet undergone formal user-based validation. Consequently, its intended benefits in terms of dataset discov-

ery and research support should not be interpreted as empirically demonstrated outcomes. Future evaluation will include expert assessment of the coverage and characterization of included resources, user studies with researchers from relevant water-related domains, comparison of dataset-search efficiency with existing discovery approaches, and collection of structured user feedback. Such evaluation will provide evidence for both the practical utility of WaterDataHub and priorities for its further development.

Comparison with Existing Resources

Numerous repositories and data infrastructures provide access to water-related environmental data. Examples include Waterbase [43], GEMStat (Global Environment Monitoring System for Freshwater database) [44], EPA Water Quality Portal (WQP) [5,45], HydroSHEDS [46], NOAA OneStop [47], the Copernicus Data Space Ecosystem [48], HydroShare [49] and various national environmental information systems. These resources have become indispensable for water quality monitoring, hydrological modeling, environmental assessment, and climate-related research [50]. However, each platform is typically designed to support a specific data domain, geographical region, or organizational mission, resulting in a fragmented data landscape that can be challenging to navigate for researchers seeking comprehensive information resources.

Waterbase, maintained by the European Environment Agency (EEA), provides extensive environmental monitoring data across Europe, particularly for water quality assessment and implementation of the European Union Water Framework Directive [51]. Similarly, GEMStat, coordinated by the United Nations Environment Programme (UNEP), focuses on freshwater quality observations collected from monitoring stations worldwide. The Water Quality Portal integrates water quality measurements from multiple agencies within the United States and represents one of the largest publicly accessible repositories of freshwater observations. While these platforms provide valuable access to observational data, their primary purpose is the dissemination of specific datasets rather than facilitating cross-domain dataset discovery.

Hydrological repositories such as HydroSHEDS and HydroATLAS provide geospatial information related to river networks, watersheds, and hydrological characteristics. Satellite-based services, including Copernicus, Landsat, Sentinel, and Global Precipitation Measurement (GPM), offer large-scale Earth observation products that support environmental monitoring and modeling. Although these resources provide extensive spatial coverage and high-quality data products, researchers must often possess substantial domain knowledge to identify which datasets are most appropriate for a particular application.

WaterDataHub 1.0 differs from these resources primarily in its scope and intended function. While established platforms provide specialized capabilities for data storage, retrieval, analysis, or publication, WaterDataHub serves as a curated discovery layer connecting researchers to resources maintained by different providers. It does not duplicate the underlying datasets or replace their existing access mechanisms. Instead, it consolidates descriptive metadata across six water-research domains to facilitate the preliminary identification of relevant data sources, particularly for interdisciplinary and AI/ML-oriented research. Table 4 provides a comparison with five representative platforms, highlighting their different objectives, functionalities, and relationship to WaterDataHub.

The comparison in Table 4 illustrates the complementary roles of these platforms. WQP and GEMStat provide access to water-quality observations, HydroSHEDS offers specialized hydrographic products, Copernicus supports the discovery and processing of Earth-observation data, and HydroShare enables the sharing and publication of hydrological research resources. These platforms provide capabilities that WaterDataHub does not attempt to reproduce. WaterDataHub, in turn, provides a common entry point for

discovering resources across providers and research domains, including water quality, hydrology, floods and droughts, aquatic ecosystems, pollution, and wastewater management. Its principal added value lies in consolidating and organizing heterogeneous resources across thematic domains, rather than offering superior data access or processing capabilities within any individual domain.

Table 4. Comparison of WaterDataHub 1.0 with representative water-data repositories and discovery platforms.

Platform	Primary Scope	Data Access	Discovery	Main Distinction
Water Quality Portal (WQP)	US water-quality monitoring	Aggregated observations	Search and filtering	Detailed water-quality data
GEMStat	Global freshwater quality	Observations and indicators	Geographic exploration	Global water-quality focus
HydroSHEDS	Hydrographic and geospatial data	River and catchment products	Geospatial downloads	Specialized hydrographic data
Copernicus Data Space Ecosystem	Satellite Earth observation	Data, APIs, cloud processing	Advanced search and processing	Extensive satellite-data services
HydroShare	Hydrological research resources	Hosting and sharing	Search and collaboration	Research-data publication
WaterDataHub 1.0	Six water-research domains	Links to original providers	Cross-provider discovery	Curated, AI/ML-oriented catalog

For example, a researcher developing a water-quality prediction model can use WaterDataHub to identify both in-situ monitoring data from the Water Quality Portal and complementary remotely sensed environmental information from Copernicus Services within a single catalog, without independently searching multiple provider-specific platforms. This cross-provider discovery capability represents the principal added value of WaterDataHub, rather than replacing the specialized data access and processing functionalities offered by existing repositories.

The AI-oriented contribution of WaterDataHub lies not in introducing novel metadata fields, but in organizing existing descriptive metadata within a cross-domain discovery framework oriented toward potential AI/ML applications. Conventional environmental data catalogs also provide information on variables, geographic coverage, acquisition methods, and data formats. WaterDataHub brings these characteristics together across heterogeneous water-related resources and relates the six thematic domains to representative AI/ML task families, as illustrated in Table 3. This enables researchers to identify potentially relevant resources for tasks such as prediction, classification, anomaly detection, and spatial modeling. The framework therefore supports the initial discovery and comparison of candidate data sources rather than providing an automated or empirically validated assessment of their AI readiness.

The FAIR principles (Findable, Accessible, Interoperable, and Reusable) provide an important foundation for environmental data management and discovery [7]. WaterDataHub complements existing data infrastructures by providing a common descriptive overview of resources from different providers, including their thematic coverage, geographic scope, representative variables, acquisition methods, data formats, and access information. These characteristics can help researchers identify potentially relevant resources for AI/ML applications, but they do not establish data completeness, predictive

utility, or readiness for model development [52]. Such assessments require further examination of the underlying datasets.

Beyond its AI-oriented metadata framework, WaterDataHub serves as a centralized discovery resource that facilitates the identification and reuse of fragmented water-related data sources distributed across diverse platforms and repositories. Although many water-related datasets are publicly accessible, they are often difficult to locate due to inconsistent metadata standards, fragmented hosting infrastructures, and varying access mechanisms. By providing a unified entry point to distributed resources, WaterDataHub aims to facilitate the discovery of existing environmental information and help researchers identify available data access mechanisms. In this way, it may support the findability and accessibility dimensions of the FAIR principles, without modifying the underlying datasets or their metadata standards.

It is important to emphasize that WaterDataHub should be viewed as complementary to existing repositories rather than a competing infrastructure. The long-term value of the platform lies in its ability to connect researchers with appropriate data sources, facilitate interdisciplinary investigations, and support the development of reproducible data-driven workflows. As environmental data infrastructures continue to expand, the role of metadata-driven discovery platforms is expected to become increasingly important for maximizing the scientific utility of publicly available datasets.

5. Conclusions

The principal contribution of this study is WaterDataHub 1.0, a curated, metadata-driven discovery resource that brings together information on 63 publicly accessible water-related data sources across six research domains: water quality monitoring, hydrology and hydrodynamics, flood and drought forecasting, aquatic ecosystems and biodiversity, water pollution and pollutant detection, and water treatment and wastewater management. By consolidating descriptive metadata from heterogeneous providers, WaterDataHub facilitates the identification and preliminary comparison of resources relevant to interdisciplinary water research.

Unlike repositories that host or distribute specific datasets, WaterDataHub provides a complementary discovery layer that connects researchers to existing data infrastructures. Its application-oriented organization supports the initial identification of resources for AI and machine-learning tasks, including prediction, classification, anomaly detection, and spatial modeling. The framework does not introduce a new AI-specific metadata standard or independently assess dataset suitability for model development.

The study demonstrates the feasibility of organizing fragmented water-data resources within a common discovery framework. However, WaterDataHub 1.0 is a curated rather than exhaustive catalog, and its coverage depends on the sources identified and documented during the collection process. Furthermore, the platform's potential benefits for dataset discovery, research efficiency, and reproducibility have not yet been evaluated through formal user studies. These limitations should be considered when interpreting its contribution.

Overall, WaterDataHub 1.0 establishes a foundation for improving the findability and accessibility of existing water-related data resources. Its principal value lies in facilitating cross-domain discovery and helping researchers identify potentially relevant data sources for subsequent scientific analysis and AI-driven applications.

Future Directions

Future development of WaterDataHub will focus on three priorities: maintaining and expanding the catalog, improving discovery functionality, and evaluating its practical usefulness.

First, the catalog will be reviewed and updated approximately every two years to verify resource accessibility, correct outdated links, update metadata, and incorporate additional relevant data sources. Major changes to the content or structure will be documented through versioned releases. Community contributions will also be encouraged, with suggested additions and corrections reviewed by the WaterDataHub maintainers before inclusion.

Second, planned technical improvements include automated metadata harvesting and link checking, integration with external repositories through application programming interfaces (APIs), and more advanced search capabilities, including semantic search and resource recommendations. The incorporation of persistent identifiers and standardized dataset quality indicators will also be explored to improve resource characterization and support more informed dataset selection.

Third, future work will investigate the practical usefulness of WaterDataHub through user-based evaluations. Such studies could assess the effectiveness of dataset discovery, the clarity and usefulness of the metadata framework, and the extent to which the catalog supports researchers in identifying resources for interdisciplinary and AI/ML-oriented applications. The results would inform subsequent improvements to the platform and its maintenance procedures. In addition, the adoption and scientific impact of WaterDataHub will be monitored through available usage statistics, citations, and an analysis of subsequent studies that use the catalog to identify data resources, including the types of AI/ML applications supported by these resources.

Author Contributions: Conceptualization, T.D., B.H.-L. and Z.B.; methodology, T.D., B.H.-L. and Z.B.; software, T.D.; validation, T.D. and Z.B.; formal analysis, T.D. and Z.B.; investigation, T.D.; resources, Z.B., B.H.-L., A.Ž.G. and J.T.; data curation, T.D. and Z.B.; writing—original draft preparation, T.D. and M.M.; writing—review and editing, T.D., M.M. and Z.B.; visualization, T.D. and Z.B.; supervision, Z.B., B.H.-L., A.Ž.G. and J.T.; project administration, B.H.-L., M.M. and Z.B.; funding acquisition, B.H.-L., M.M., Z.B., A.Ž.G. and J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Slovenian Research and Innovation Agency (ARIS) through the research programmes P2-0209, P1-0201, P2-0359, and P2-0191, as well as the research project J5-70190. B.H.-L., M.M., A.Ž.G., and Z.B. acknowledge the support of the Water Research Center at the University of Ljubljana and the University of Ljubljana Development Fund (RSUL). This work was carried out within the framework of the Africa-Europe Clusters of Research Excellence (CoRE) project Water Resources Management for a Sustainable and Just Future, in collaboration with members of the CoRE Engineering for the Future and the CoRE Artificial Intelligence, Data Science, and Theoretical and Computational Thinking. The authors gratefully acknowledge the support of Rhodes University through the Hugh Kelly Fellowship.

Data Availability Statement: The WaterDataHub 1.0 catalog compiled and presented in this study is openly available at https://bit.ly/water-data-sources, accessed on 1 November 2026. The catalog provides metadata and links to the original data sources, which remain accessible through their respective providers.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Glossary of Data Resources and Acronyms

Glossary of major datasets, databases, and data infrastructures referenced in the manuscript.

Acronym	Full Name	Description
CAMELS	Catchment Attributes and Meteorology for Large-sample Studies	Catchment-scale hydrological, meteorological, and environmental data.
CHIRPS	Climate Hazards Group InfraRed Precipitation with Station Data	Precipitation estimates combining satellite observations and ground-based measurements.
GEMStat	Global Environment Monitoring System for Freshwater (GEMS/Water) database	Global database of inland water-quality observations.
GLDAS	Global Land Data Assimilation System	Global land-surface and hydrometeorological data products.
GRACE	Gravity Recovery and Climate Experiment	Satellite observations used to estimate changes in terrestrial water storage.
GRDC	Global Runoff Data Centre	International repository of river discharge observations.
Waterbase	Waterbase (European Environment Agency)	European datasets covering water quality, water resources, and related environmental indicators.
WQP	Water Quality Portal	Portal providing access to water-quality monitoring data from multiple US agencies.

References

1. Ed-Dehbi, W.; Ahlaqqach, M.; Benhra, J. Artificial Intelligence for Optimal Water Resource Management: A Literature Review. *Eng. Proc.* **2025**, *97*, 52. [CrossRef]
2. Yan, X.; Zhang, T.; Du, W.; Meng, Q.; Xu, X.; Zhao, X. A Comprehensive Review of Machine Learning for Water Quality Prediction over the Past Five Years. *J. Mar. Sci. Eng.* **2024**, *12*, 159. [CrossRef]
3. Hasan, F.; Medley, P.; Drake, J.; Chen, G. Advancing Hydrology through Machine Learning: Insights, Challenges, and Future Directions Using the CAMELS, Caravan, GRDC, CHIRPS, PERSIANN, NLDAS, GLDAS, and GRACE Datasets. *Water* **2024**, *16*, 1904. [CrossRef]
4. Colohan, P.; Onda, K. Water data for water science and management: Advancing an Internet of Water (IoW). *PLoS Water* **2022**, *1*, e0000017. [CrossRef]
5. Read, E.K.; Carr, L.; De Cicco, L.; Dugan, H.A.; Hanson, P.C.; Hart, J.A.; Kreft, J.; Read, J.S.; Winslow, L.A. Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resour. Res.* **2017**, *53*, 1735–1745. [CrossRef]
6. Horsburgh, J.S.; Morsy, M.M.; Castronova, A.M.; Goodall, J.L.; Gan, T.; Yi, H.; Stealey, M.J.; Tarboton, D.G. Hydroshare: Sharing diverse environmental data types and models as social objects with application to the hydrology domain. *J. Am. Water Resour. Assoc.* **2016**, *52*, 873–889. [CrossRef]
7. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef]
8. Dobovšek, T. A Systematic Review of Publicly Available Data Sources for Water Research. Bachelor's Thesis, University of Ljubljana, Ljubljana, Slovenia, 2026.
9. Ooi, K.S.; Chen, Z.; Poh, P.E.; Cui, J. BOD5 prediction using machine learning methods. *Water Supply* **2021**, *22*, 1168–1183. [CrossRef]
10. Lapucci, C.; Antonini, A.; Bohm, E.; Organelli, E.; Massi, L.; Ortolani, A.; Brandini, C.; Maselli, F. Use of Sentinel-3 OLCI Images and Machine Learning to Assess the Ecological Quality of Italian Coastal Waters. *Sensors* **2023**, *23*, 9258. [CrossRef]
11. Tso, C.H.M.; Magee, E.; Huxley, D.; Eastman, M.; Fry, M. River reach-level machine learning estimation of nutrient concentrations in Great Britain. *Front. Water* **2023**, *5*, 1244024. [CrossRef]
12. Byrne, A.; Lomeo, D.; Owoko, W.; Aura, C.M.; Nyakeya, K.; Odoli, C.; Mugo, J.; Barongo, C.; Kiplagat, J.; Mwirigi, N.; et al. LAQUA: A LAndsat water QUality retrieval tool for east African lakes. *Remote Sens.* **2024**, *16*, 2903. [CrossRef]
13. Baltodano, A.; Agramont, A.; Lekarkar, K.; Spyrakos, E.; Reusen, I.; van Griensven, A. Exploring global remote sensing products for water quality assessment: Lake Nicaragua case study. *Remote Sens. Appl. Soc. Environ.* **2024**, *36*, 101331. [CrossRef]
14. Ndou, N. Geostatistical inference of Sentinel-2 spectral reflectance patterns to water quality indicators in the Setumo dam, South Africa. *Remote Sens. Appl. Soc. Environ.* **2023**, *30*, 100945. [CrossRef]
15. Pahlevan, N.; Smith, B.; Alikas, K.; Anstee, J.; Barbosa, C.; Binding, C.; Bresciani, M.; Cremella, B.; Giardino, C.; Gurlin, D.; et al. Simultaneous retrieval of selected optical water quality indicators from Landsat-8, Sentinel-2, and Sentinel-3. *Remote Sens. Environ.* **2022**, *270*, 112860. [CrossRef]
16. Sheikholeslami, R.; Hall, J.W. Global patterns and key drivers of stream nitrogen concentration: A machine learning approach. *Sci. Total Environ.* **2023**, *868*, 161623. [CrossRef]

17. Arias-Rodriguez, L.F.; Tüzün, U.F.; Duan, Z.; Huang, J.; Tuo, Y.; Disse, M. Global Water Quality of Inland Waters with Harmonized Landsat-8 and Sentinel-2 Using Cloud-Computed Machine Learning. *Remote Sens.* **2023**, *15*, 1390. [CrossRef]
18. Khozani, Z.S.; Precht, E.; Ionita, M. Weekly streamflow forecasting of Rhine river based on machine learning approaches. *Nat. Hazards* **2024**, *121*, 4135–4153. [CrossRef]
19. Antonio, B.; McRae, A.T.T.; MacLeod, D.; Cooper, F.C.; Marsham, J.; Aitchison, L.; Palmer, T.N.; Watson, P.A.G. Postprocessing East African Rainfall Forecasts Using a Generative Machine Learning Model. *J. Adv. Model. Earth Syst.* **2025**, *17*, e2024MS004796. [CrossRef]
20. Pazola, A.; Shamsudduha, M.; French, J.; MacDonald, A.M.; Abiye, T.; Goni, I.B.; Taylor, R.G. High-resolution long-term average groundwater recharge in Africa estimated using random forest regression and residual interpolation. *Hydrol. Earth Syst. Sci.* **2024**, *28*, 2949–2967. [CrossRef]
21. Viviers, C.; van der Laan, M.; Gaffoor, Z.; Dippenaar, M. Downscaling and validating GLDAS groundwater storage anomalies by integrating precipitation for recharge and actual evapotranspiration for discharge. *J. Hydrol. Reg. Stud.* **2024**, *54*, 101879. [CrossRef]
22. Wang, Y.; Li, C.; Cui, Y.; Cui, Y.; Xu, Y.; Hora, T.; Zaveri, E.; Rodella, A.S.; Bai, L.; Long, D. Spatial downscaling of GRACE-derived groundwater storage changes across diverse climates and human interventions with Random Forests. *J. Hydrol.* **2024**, *640*, 131708. [CrossRef]
23. Dong, N.; Hao, H.; Yang, M.; Wei, J.; Xu, S.; Kunstmann, H. Deep-learning-based sub-seasonal precipitation and streamflow ensemble forecasting over the source region of the Yangtze River. *Hydrol. Earth Syst. Sci.* **2025**, *29*, 2023–2042. [CrossRef]
24. Iilonga, S.N.; Ajayi, O.G. Implementation of deep learning algorithms to model agricultural drought towards sustainable land management in Namibia's Omusati region. *Land Use Policy* **2025**, *156*, 107593. [CrossRef]
25. Rasheed, Z.; Aravamudan, A.; Zhang, X.; Anagnostopoulos, G.C.; Nikolopoulos, E.I. Combining global precipitation data and machine learning to predict flood peaks in ungauged areas with similar climate. *Adv. Water Resour.* **2024**, *192*, 104781. [CrossRef]
26. Nearing, G.S.; Cohen, D.; Dube, V.; Gauch, M.; Gilon, O.; Harrigan, S.; Hassidim, A.; Klotz, D.; Kratzert, F.; Metzger, A.; et al. Global prediction of extreme floods in ungauged watersheds. *Nature* **2024**, *627*, 559–563. [CrossRef]
27. Bojer, A.K.; Biru, B.H.; Al-Quraishi, A.M.F.; Debelee, T.G.; Negera, W.G.; Woldesillasie, F.F.; Esubalew, S.Z. Machine learning and remote sensing based time series analysis for drought risk prediction in Borena Zone, Southwest Ethiopia. *J. Arid Environ.* **2024**, *222*, 105160. [CrossRef]
28. Danso, S.Y.; Ma, Y.; Addo, I.Y. Delineating flood susceptibility zones using novel ensemble models—An application of evidential belief function, relative frequency ratio, and Shannon entropy. *Int. J. Appl. Earth Obs. Geoinf.* **2025**, *141*, 104669. [CrossRef]
29. Narin, O.G.; Sekertekin, A.; Bayik, C.; Bektas Balcik, F.; Arıkan, M.; Sanli, F.B.; Abdikan, S. Multi-Sensor Flood Mapping in Urban and Agricultural Landscapes of the Netherlands Using SAR and Optical Data with Random Forest Classifier. *Remote Sens.* **2025**, *17*, 2712. [CrossRef]
30. Alonso-Sarria, F.; Valdivieso-Ros, C.; Molina-Pérez, G. Detecting Flooded Areas Using Sentinel-1 SAR Imagery. *Remote Sens.* **2025**, *17*, 1368. [CrossRef]
31. Joshi, N.; Park, J.; Zhao, K.; Londo, A.; Khanal, S. Monitoring Harmful Algal Blooms and Water Quality Using Sentinel-3 OLCI Satellite Imagery with Machine Learning. *Remote Sens.* **2024**, *16*, 2444. [CrossRef]
32. Polazzo, F.; Domisch, S.; Flörke, M.; Rico, A. A modelling approach to assess climate change impacts on taxonomic and functional diversity of European stream macroinvertebrates: Implications for water quality monitoring. *Ecol. Indic.* **2024**, *166*, 112404. [CrossRef]
33. Effrosynidis, D.; Tsikliras, A.; Arampatzis, A.; Sylaios, G. Species Distribution Modelling via Feature Engineering and Machine Learning for Pelagic Fishes in the Mediterranean Sea. *Appl. Sci.* **2020**, *10*, 8900. [CrossRef]
34. Belayhun, M.; Chere, Z.; Abay, N.G.; Nicola, Y.; Asmamaw, A. Spatiotemporal pattern of water hyacinth (*Pontederia crassipes*) distribution in Lake Tana, Ethiopia, using a random forest machine learning model. *Front. Environ. Sci.* **2024**, *12*, 1476014. [CrossRef]
35. Rieger, N.; Olmedo, E.; Thiel, M.; Salvo, V.S.; Honorato, D.; Vasquez, N.; Turiel, A.; Piera, J. Seasonal hotspots of beach litter in the North-East Atlantic linked to aquaculture and river runoff. *Commun. Earth Environ.* **2024**, *5*, 751. [CrossRef]
36. Deng, R.; Zhu, T.; Zhou, W.; Liu, F.; Lin, X. Machine learning based water quality evolution and pollution identification in reservoir type rivers. *Environ. Pollut.* **2025**, *382*, 126668. [CrossRef]
37. Zhen, Y.; Wang, L.; Sun, H.; Liu, C. Prediction of microplastic abundance in surface water of the ocean and influencing factors based on ensemble learning. *Environ. Pollut.* **2023**, *331*, 121834. [CrossRef]
38. Frincu, R.M. Artificial intelligence in water quality monitoring: A review of water quality assessment applications. *Water Qual. Res. J.* **2024**, *60*, 164–176. [CrossRef]
39. Faria, B.; Mendes, R.; Lopes, C.; Picado, A.; Sousa, M.; Dias, J. Insights for Sea Outfall Turbid Plume Monitoring with High-Spatial-Resolution Satellite Imagery Application in Portugal. *Remote Sens.* **2023**, *15*, 3368. [CrossRef]

40. Jones, E.R.; van Vliet, M.T.H.; Qadir, M.; Bierkens, M.F.P. Country-level and gridded estimates of wastewater production, collection, treatment and reuse. *Earth Syst. Sci. Data* **2021**, *13*, 237–254. [CrossRef]
41. Wang, Y.; He, X.; Bai, Y.; Tan, Y.; Zhu, B.; Wang, D.; Ou, M.; Gong, F.; Zhu, Q.; Huang, H. Automatic detection of suspected sewage discharge from coastal outfalls based on Sentinel-2 imagery. *Sci. Total Environ.* **2022**, *853*, 158374. [CrossRef]
42. Li, H.; Zech, J.; Hong, D.; Ghamisi, P.; Schultz, M.; Zipf, A. Leveraging OpenStreetMap and Multimodal Remote Sensing Data with Joint Deep Learning for Wastewater Treatment Plants Detection. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *110*, 102804. [CrossRef]
43. European Environment Agency. Waterbase—Water Quality ICM, 2026. Available online: https://www.eea.europa.eu/en/datahub/datahubitem-view/fbf3717c-cd7b-4785-933a-d0cf510542e1 (accessed on 23 February 2026).
44. United Nations Environment Programme (UNEP); GEMS/Water Programme. GEMStat: The Global Freshwater Quality Database, 2026. Available online: https://gemstat.org/ (accessed on 23 February 2026).
45. EPA; USGS; NWQMC. Water Quality Portal. 2026. Available online: https://www.waterqualitydata.us/ (accessed on 23 February 2026).
46. HydroSHEDS Project. HydroSHEDS, 2026. Available online: https://www.hydrosheds.org/ (accessed on 23 February 2026).
47. National Oceanic and Atmospheric Administration (NOAA). NOAA OneStop Data Discovery and Access Platform. 2026. Available online: https://data.noaa.gov/onestop/ (accessed on 23 February 2026).
48. Copernicus Data Space Ecosystem. Copernicus Data Space Ecosystem. 2026. Available online: https://dataspace.copernicus.eu/ (accessed on 23 February 2026).
49. Tarboton, D.G.; Ames, D.P.; Horsburgh, J.S.; Goodall, J.L.; Couch, A.; Hooper, R.; Bales, J.; Wang, S.; Castronova, A.; Seul, M.; et al. HydroShare retrospective: Science and technology advances of a comprehensive data and model publication environment for the water science domain. *Environ. Model. Softw.* **2024**, *172*, 105902. [CrossRef]
50. Xu, D.; Xue, Z.; Hu, L.; Chen, H. From Multi-Scale Observations to Integrated Applications: A Review of Remote Sensing for Inland Water Quality Monitoring. *Front. Environ. Sci.* **2026**, *14*, 1832807. [CrossRef]
51. European Parliament; Council of the European Union. Directive 2000/60/EC establishing a framework for Community action in the field of water policy. *Off. J. Eur. Communities* **2000**, *L 327*, 1–73.
52. Jaywant, S.A.; Arif, K.M. Remote Sensing Techniques for Water Quality Monitoring: A Review. *Sensors* **2024**, *24*, 8041. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.