Received: 30 September 2021 | Accepted: 8 September 2022

DOI: 10.1111/2041-210X.13992

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Machine learning applications in river research: Trends, opportunities and challenges

Long Ho | Peter Goethals

Department of Animal Sciences and Aquatic Ecology, Ghent University, Ghent, Belgium

Correspondence
Long Ho
Email: long.tuanho@ugent.be

Funding information
Fonds Wetenschappelijk Onderzoek, Grant/Award Number: 1253921N

Handling Editor: Aaron Ellison

Abstract

1. As one of the earth's key ecosystems, rivers have been intensively studied and modelled through the application of machine learning (ML). With the amount of large data available, these computer algorithms are ever increasing in numerous fields, although there is ongoing scepticism and scholars still question the actual impact and deliverables of algorithms.
2. This study aims to provide a systematic review of the state-of-the-art ML-based techniques, trends, opportunities and challenges in river research by applying text mining and automated content analysis.
3. Unsupervised and supervised learning have dominated river research while neural networks and deep learning have also gradually gained popularity. Matrix factorisation and linear models have been the most popular ML algorithms, with around 1300 and 800 publications on these topics in 2020 respectively. In contrast, river researchers have had few applications in multiclass and multilabel algorithm, associate rule and Naïve Bayes.
4. The current article proposes an end-to-end workflow of ML applications in river research in order to tackle major ML challenges, including four steps: (1) data collection and preparation; (2) model evaluation and selection; (3) model application; and (4) feedback loops. Within this workflow, river modellers have to balance numerous trade-offs related to model traits, such as complexity, accuracy, interpretability, bias, data privacy and accessibility and spatial and temporal scales. Any choices made when balancing the trade-offs can lead to different model outcomes affecting the final applications. Hence, it is necessary to carefully consider and specify modelling goals, understand the data collected and maintain feedback loops in order to continuously improve model performance and eventually reach the research objectives. Moreover, it remains crucial to address the users' needs and demands that often entail additional elements, such as computational cost, development time and the quantity, quality and compatibility of data. Furthermore, river researchers should account for new technologies and regulations in data collection and protection that are transforming the development and applications of ML, most notably data warehouse and

information management with multiple-cycles that are becoming a cornerstone of the integration of ML in decision-making in river and ecosystem management.

KEYWORDS
artificial intelligence, machine learning, remote sensing, river research

1 | INTRODUCTION

It has been found that over 50% of the world's population lives closer than 3 km to a surface body of fresh water (Kummu et al., 2011). Because of this ubiquitous pattern of settlement, riverine landscapes have been modified immensely, which often destroys and fragments habitats, causes the loss of biodiversity, concentrates contaminants and alters flow regimes (Cooper et al., 2013). A notable example is intensive agriculture that has increased erosion and sediment loads in many places, and their runoff has discharged nutrients and chemicals to streams and rivers, making eutrophication one of the main problems of many freshwater ecosystems (Foley et al., 2015). In addition to being diffuse source for pollutants, urbanisation has also introduced a vast amount of contaminants to rivers via its discharge points as more than 80% of the wastewater from human activities is still discharged directly into rivers and the sea (United Nations, 2015). It was estimated that over 1000 rivers—above all urban rivers—are responsible for transporting 0.8–2.7 million metric tons of plastic waste into the ocean per year, accounting for 80% of global plastic emissions (Meijer et al., 2021). It is also worth noting that 50,000 hydropower dams (higher than 15 m), capable of accumulating from 7000 to 8300 km^3 of freshwater, have been built on global rivers (Lehner, Liermann, Revenga, Vorosmarty, et al., 2011; Lehner, Liermann, Revenga, Vörösmarty, et al., 2011). These dams create severe impacts on river hydromorphology, flow discharge, thermal dynamics and inter-basin water transfers, consequently leading to ecological degradation in respect of river habitats and species (Hauer et al., 2017).

Equally important is the fact that rivers are essential in biogeochemical cycles of carbon (C), and nitrogen (N) because of their vital role in regulating the global hydrological cycle that is a key component of the cycling of these biogeochemicals (Schimel, 1995). Besides acting as a natural source of greenhouse gases (GHGs), rivers also serve as conduits for GHGs released into the atmosphere from soil pore water, groundwater and sediments as a result of substantial terrestrial-to-aquatic C flux of 5.1 $Pg\,C\,year^{-1}$ (Drake et al., 2018; Hotchkiss et al., 2015). The global emissions of CO_2 from streams and rivers were estimated around 3.9 $Pg\,C\,year^{-1}$ (Drake et al., 2018) while the figures for inland water CH_4 and N_2O evasions were 159 $Tg\,C\,year^{-1}$ and 1.26 $Tg\,N\,year^{-1}$ respectively (Beaulieu et al., 2011; Kroeze et al., 2005; Saunois et al., 2020; Stanley et al., 2016). These bodies of water are strongly affected by consequences of climate change such as sea level rise, warming water and flow modification (Guneralp et al., 2015). For example, great temperature increases and flow decreases are estimated for rivers in the south-eastern United States, Europe, eastern China, southern Africa and southern Australia (van Vliet et al., 2013).

To properly address the abovementioned issues, it is necessary to understand the mechanisms of rivers and their interactions with other environmental components, and machine learning (ML) appears a highly promising method for gaining a better understanding of river systems and improving decision-making through better use of data and quantitative evidence. Note that ML is a branch of artificial intelligence (AI) that allows computer systems to perform tasks linked to the behaviour of intelligent beings (Dobbelaere et al., 2021). By enabling computers to perform specific tasks, ML models can carry out complex processes by learning from data without a need for explicit pre-programmed systems (Royal Society, 2017). Figure 1 shows the main goals of ML applications which can be categorised into four groups, that is, data analysis, feature detection, prediction and forecasting and system learning (Witten & Frank, 2005). To this end, ML applications are based on both statistical and heuristic methods, such as linear regression or ensemble models, to automatically construct models and discover patterns from field data. In recent years, there have been numerous successful applications of ML as a result of increasing computational power and massive data availability. In fact, the successes of ML, together with the advent of Big Data, the Internet of Things and the fourth industrial revolution, have inspired researchers and engineers to apply and optimise them in many fields, including wastewater engineering (Zhao et al., 2020), public health (dos Santos et al., 2019), environmental engineering (Gibert et al., 2018) and ecology (Gobeyn & Goethals, 2019).

There are numerous opportunities for applying ML in river and ecosystem management, given the development of many new technologies, and infrastructure for collecting, distributing and analysing data. However, the ongoing excitement surrounding ML applications may just be a recurrence of the previous hypes around neural networks applications that were then followed by disappointment and criticism due to over-ambitious promises from developers and unnaturally high expectations from end-users (Linden & Fenn, 2003). This disappointment was also the result of a lack of insights into the proper selection and use of the methods, often combined with unclear standards and a lack of good practices starting from data collection to decision-making (Goethals et al., 2007). Building on these points of view, this review aims to address the following questions in the context of ML applications in river research: (1) What are the major ML-based algorithms and main research topics that use these tools in river research? and (2) How have these ML-based algorithms

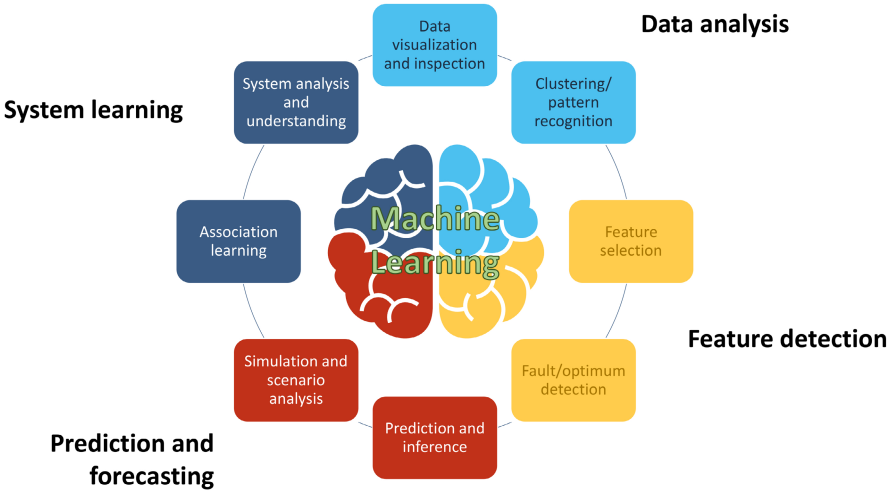

FIGURE 1 Main goals of machine learning applications

and research topics evolved in recent decades? Moreover, during the research project we also proposed the first end-to-end ML workflow in the context of river research, in which important and emerging issues were thoroughly discussed and numerous trade-offs were also deliberated. To this end, we applied text mining and automated content analysis (ACA) to review a large amount of river literature in a data-driven and impartial manner, taking advantage of 'big bibliographic data' in river research (Nunez-Mir et al., 2016; Vugteveen et al., 2014).

2 | MATERIALS AND METHODS

2.1 | Data collection

We collected bibliographic records from the Scopus website, covering publications from the timeframe 1950–2020, 1950 being the year of the first record we found on the selected topic. The Scopus database contains the largest international abstract and citation collection of peer-reviewed scientific literature (Scopus Elsevier, 2016). We identified and collected publications using the ML algorithms in the field of river research in a two-step procedure whereby the first step focused on ML categories while the second step aimed to classify ML algorithms.

In the first step, we categorised river research that applied ML algorithms into six categories, including supervised learning, unsupervised learning, neural networks and deep learning, semi-supervised learning, reinforcement learning, and self-supervised learning. These six categories were categorised based on the classification of Ayodele (2010) and Nguyen et al. (2019). For this, we applied a search query with a common template: TITLE-ABS-KEY (river AND *) in which * was filled with the name and abbreviation of each category. Note that supervised learning, reinforcement learning, and unsupervised learning are different by design. In particular, supervised learning aims to build a function that maps links between input(s) and labelled output(s), reinforcement learning is a family of algorithms that learns an optimal policy maximising return given an observation of the world, and unsupervised learning aims to find patterns in unlabelled data (Ayodele, 2010). In contrast, semi-supervised learning, self-supervised learning, neural networks and deep learning cannot be distinguished completely from the other ML categories (Ayodele, 2010; Doersch et al., 2015; Kostopoulos et al., 2018; Schmidhuber, 2015). Semi-supervised learning falls between unsupervised learning and supervised learning as this ML category learns from both labelled and unlabelled data (Kostopoulos et al., 2018); self-supervised learning can be categorised as an intermediate algorithm between supervised and unsupervised learning where the label scarcity can solved by automatically generating labels from the data itself (Doersch et al., 2015); while neural networks and deep learning can be both supervised and unsupervised (Schmidhuber, 2015).

In the second step, we focused solely on ML algorithms in supervised learning and unsupervised learning because of their substantially higher number of applications compared to semi-supervised learning, reinforcement learning and self-supervised learning while details of ML algorithms in neural networks and deep learning in river research merit another review. In particular, we used the categorisation in scikit-learn (2019) as a reference for the ML algorithms since scikit-learn (2019) is the most comprehensive and open-sourced ML library in Python, a popular language among data scientists and software developers (Hao & Ho, 2019). Subsequently, we filled * in the common template of the search queries with the name and abbreviation of ML methods in each algorithm which were found in the scikit-learn (2019). Especially noteworthy is that there are hybrid ML models that are a combination of multiple ML algorithms (Kotsiantis et al., 2006) which are largely omitted in the list of ML algorithms in scikit-learn (2019). For example, Wang et al. (2004) presented a hybrid Flexible NBTree model: a decision tree consisting of leaf nodes that contain General Naive Bayes algorithm, a variant of the standard Naive Bayesian classifier. More recently, including the integration of linear regression-deep neural network models, a hybrid HybPAS model was proposed by Albalawi et al. (2019). Hence, it is important to note that this review focuses mostly on large-scale trends of individual ML categories and algorithms, while details of the specific

hybrid models merit another review. Definitions of all ML algorithms and categories can be found in the **Glossary** while details of their search query can be found in **Supplementary Material A**.

It is important to note that river researchers could use more than one ML algorithm in their studies that, together with the indistinctness among ML categories/algorithms, created shared publications among ML categories/algorithms in the collected bibliographic records. Hence, to analyse and compare different ML categories/algorithms in river research, we collected separated bibliographic records for each ML category/algorithm. Then, when we analysed the global trends of all ML applications in river research, we filled * with all ML algorithms in the common template of the search query to eliminate duplicated records. We calculated shared publications among the ML categories and algorithms and this can be found in the **Supplementary Material E**. It is important to note that in order to assess and compare the publication performance of countries, we added a term (LIMIT-TO[AFFILCOUNTRY, **]) in which ** was filled by the name of countries in the search query. Regarding the publication performance of the EU 28 region, this term was filled with the name of all 27 current EU member states and the United Kingdom.

2.2 | Trends in river research topics

2.2.1 | Categorisation of research topics

To categorise the research topics within river research, we used text mining to search for the corresponding terms of each topic in authors' keywords of the publications. This categorisation is based on co-word analysis whose principle is that scientific fields can be characterised and analysed based on the patterns of keywords in publications (Callon et al., 1991). Specifically, when novel concepts and methods are applied in scientific fields, authors use keywords or their combinations to represent those elements. Hence, by identifying the keywords that appear together commonly in the literature, it is possible to recognise research topics based on the concepts and methods within scientific fields (Neff & Corley, 2009). Initially, we defined topics and their corresponding terms based on those in hydropower reservoirs research (Jiang et al., 2016) and in lakes and reservoirs research (Ho & Goethals, 2020). In these publications, these topics were defined by using Term Frequency-Inverse Document Frequencies Transformation (TF–IDF) and Latent Dirichlet Allocation (LDA) models to determine similarities with respect to terms and contents of the publications. More specifically, TF–IDF is used to evaluate the importance of a word in a document in a collection or corpus which increases proportionally in relation to the number of times a word appears in the document but is offset by the frequency of the word in the corpus (Robertson, 2004). LDA is a topic model that can be used to classify text in a document to a particular topic by using Bayesian methods to build a topic-per-document model and words-per-topic model, modelled as Dirichlet distributions (Blei et al., 2003). Taking these topics as a point of reference, we ultimately selected 25 research topics for this study by adding relevant yet absent topics while removing those we judged to be irrelevant. Specifically, while we retained topics on environmental and socio-economic issues, we removed topics on construction and operation and added topics on groundwater, estuaries, land use and movement. Subsequently, we selected publications within each topic by searching for corresponding terms in the authors' keywords. The details of the chosen topics and their corresponding terms can be found in **Supplementary Material B**. The publications that were not classified into any of the 25 research topics were added into the Misc category. As with the classification of the ML algorithms, there exist studies that have investigated more than one research topic, which led to shared publications among different research topics in river research. The shared publications among the research topics can be found in the **Supplementary Material E**.

2.2.2 | Temporal trends

To understand the trends in research topics over time, we ranked the topics by decade, starting from the 1980s while the pre-1980s publications were agglomerated. Note that the early publications were listed incompletely in Scopus databases; hence numerous subsequent papers excluded these from their research (Ho & Goethals, 2020; Jiang et al., 2016; McCallen et al., 2019; Qian et al., 2015). Subsequently, we applied the Mann–Kendall trend test to examine increasing or decreasing trends within the 25 research topics using the KENDALL package (McLeod & McLeod, 2015) in R (R Core Team, 2014). Despite being under the data assumption of independent and identically distributed (i.i.d.), the Mann–Kendall test can deal with non-normally distributed data, outliers and nonlinear trends (Mann, 1945). More details of the Mann–Kendall trend test can be found in **Supplementary Material C**.

2.2.3 | Thematic clusters

To identify thematic clusters within the research topics, we used data-driven hierarchical clustering following Ward's minimum variance method (Ward, 1963) for estimating the (dis)similarity among the 25 research topics via the Euclidean distances. Together with factor analysis and principal component analysis, hierarchical clustering is one of the commonly used multivariate statistical techniques for the identification of research themes (Neff & Corley, 2009). Illustrated by a dendrogram, the Euclidean distances representing the (dis)similarities between research clusters were measured in relation to publication level, which was based on the frequency of research topics appearing in the same publication (Jiang et al., 2016). More specifically, when research topics frequently appear in the same publications, represented by a close distance in the dendrogram, it is suggested that there is an interdisciplinary field in river research based on these topics. The hierarchical clustering analysis was implemented via STATS (R Core

Team & Worldwide Contributors, 2002) and clusterR (Mouselimis et al., 2019) packages in R. The flowchart and explanation of the hierarchical clustering using Ward's method can be found in **Supplementary Material C** while its mathematical details can be found in the original paper (Ward, 1963).

2.3 | Dashboard

An equally important point is that to facilitate the data accessibility, analysis and visualisation for the readers, we developed an interactive application using R Shiny package (Chang et al., 2015). This application allows the application's user interface to be customised to provide an elegant environment for displaying user-input controls and simulation output (Wojciechowski et al., 2015). The output can be simultaneously updated with changing input. Thanks to this option, the users can access, analyse and visualise the collected publications in a quick, flexible and informative way. Our Shiny application is available online at https://env-research.shinyapps.io/ML_river/.

3 | TRENDS IN MACHINE LEARNING APPLICATIONS IN RIVER RESEARCH

3.1 | Global trends of machine learning applications

Overall, there was an unprecedented increase in ML applications in river research field from 2000 to 2020. In particularly, the number of publications on the topic increased more than 10-fold, from 310 publications in 2000 to 3444 in 2020, leading to a total of 30,962 publications collected by 21 September 2020. Compared to the 6% overall percentage increase in river research during the 2010s, ML publications in this field have increased at a faster rate of around 11% on average. Similarly, the number of countries which have contributed more than 10 publications increased threefold, from 33 in the 1990s to 100 in the 2010s. During this period, we observed an increase in studies with authors from Chinese institutions from only 12 publications in 2000 to 1413 in 2020, making China the current world leader regarding the number of publications in this field (Figure 2). River research from China accounts for around 26% of all ML-related river research, surpassing the EU 28 and the United States (responsible for 24% and 20% respectively). This may be the result of massive Chinese programmes of investment in research and development, especially in the AI field, as in 2017, China became the second-largest spender after the United States for research and development with around 370.6 billion dollars (Lee, 2018). Following China, the EU 28, and the United States are India, Canada and Brazil, each contributing around 5%. These leading countries together account for almost 90% of the total publications that apply ML in integrated river management. More details of the global trends of ML applications in integrated river management can be found in Figure D.1 in SI.

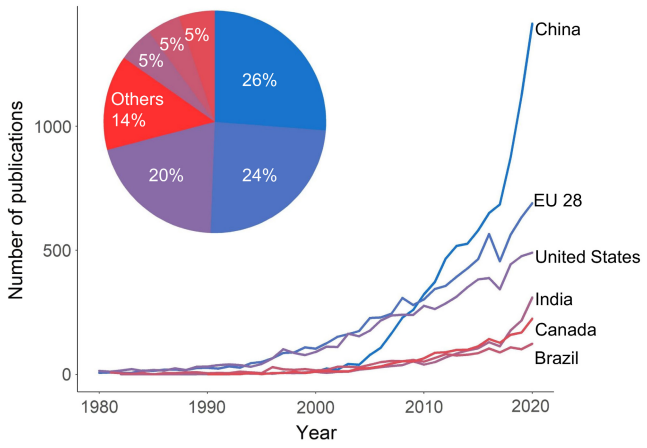

FIGURE 2 Total number and proportion of publications applying machine learning in river research in the most productive countries and regions are illustrated in line and pie charts respectively. The colours in the pie chart represent the same country as they represent in the line chart.

3.2 | Trends in machine learning categories

Figure 3 indicates the yearly evolution of the applications in each ML category and the percentages for each of them over the last decades in river research. It appears that unsupervised learning and supervised learning have dominated the field of river research while the applications of semi-supervised, self-supervised and reinforcement learning have remained very limited (lower than 0.3% of the total publications). In particularly, before the 1990s, around 98% of the ML applications in river research were either unsupervised learning or supervised learning. After that, the proportion of unsupervised learning or supervised learning decreased as river modellers increasingly applied neural networks and deep learning to around 11% of the total publications during 1990s and up to almost 21% during the 2000s. Thereafter, the proportion of neural network and deep learning applications decreased gradually to 15% in 2020 despite their sharp increase to almost 600 publications in river research from 2018 to 2020. Regarding supervised learning, despite its ML applications featuring in almost two-thirds of the total publications related to river management before the 1980s, the usages of supervised ML algorithms by river researchers declined over subsequent decades. Only around 30% of the total publications applied the ML category during the 2000s. After that, the proportion slightly increased to 40% during the 2010s and to 46% in 2020. In contrast, starting from around 30%, the percentage of publications applying unsupervised ML in river management had grown to more than half of the total publications during the 1980s and 1990s. However, this dominance shrank to around 45% of the total publications between 2000 and 2020, similar to the number of studies featuring supervised learning. As unsupervised algorithms are frequently applied in exploratory settings, the early dominance of papers featuring unsupervised learning perhaps indicates the fact that river researchers were focusing on better understanding river systems rather than predicting their

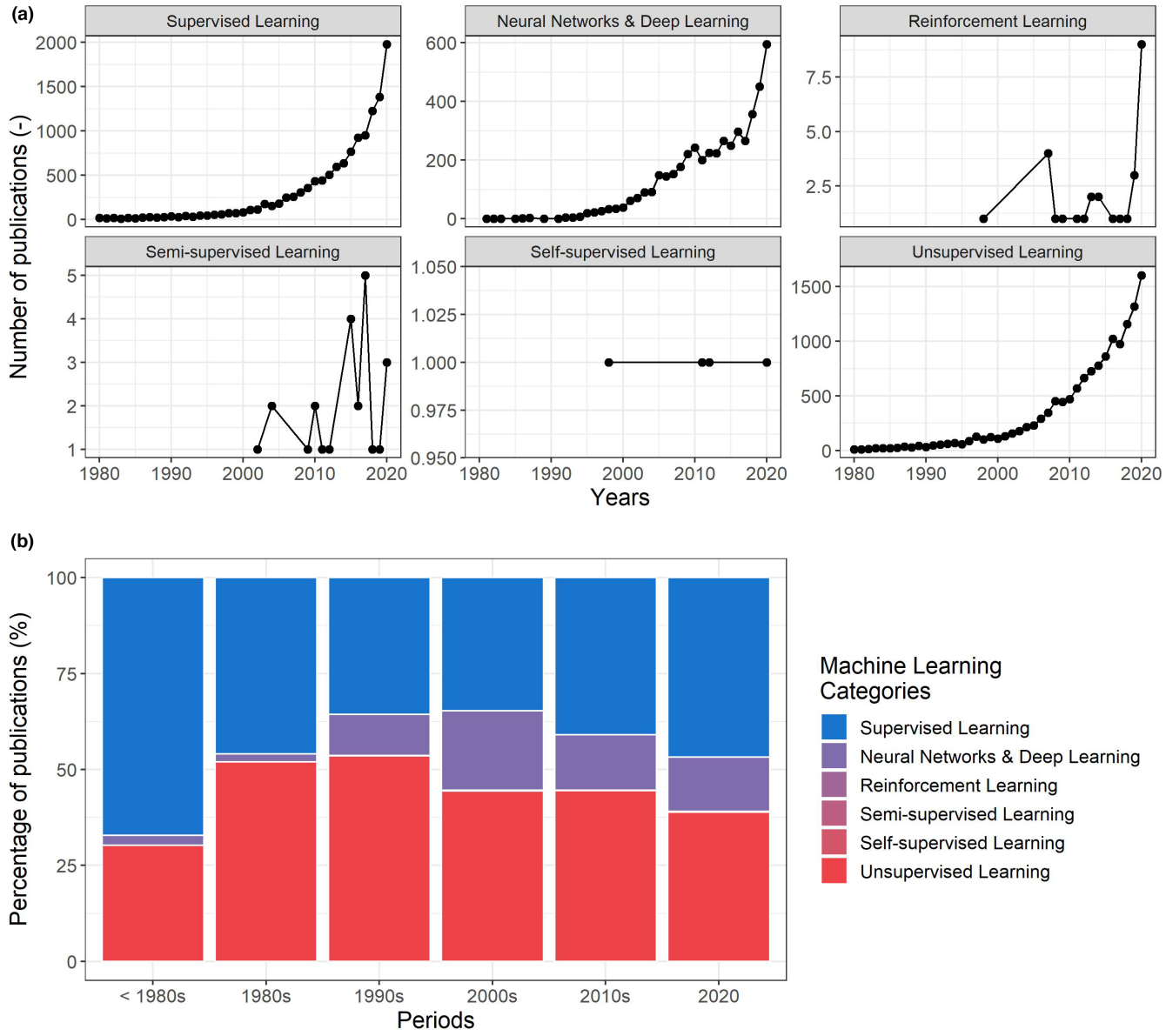

FIGURE 3 Applications of six machine learning categories in river research. (a) Yearly evolution of the machine learning categories in river research since 1980. (b) Percentage of publications applying the machine learning categories in river research. Note that because of low percentage of applications of reinforcement learning, semi-supervised learning and self-supervised learning (<0.3%) in river research, only supervised learning, unsupervised and neural networks and deep learning appear in (b).

behaviours (Müller & Guido, 2016). From 2010 to 2020, attention turned towards predictive models which can be more useful for estimating and forecasting river states from certain inputs (Provost & Fawcett, 2013). Note that the limited number of publications applying reinforcement learning, semi-supervised learning and self-supervised learning reveals an opportunity to exploit their application in river research in the future. Also noteworthy is the fact that there have been few shared publications among the ML categories—less than 5% of the total publications in each ML category—which suggests a limited number of publications applying more than one ML category in river research (**Supplementary Material E**).

3.3 | Trends in machine learning algorithms

Several ML algorithms have long been applied in the field, such as clustering, ensemble methods, linear models, manifold learning and matrix factorisation, while many have only been implemented from 2000 to 2020 (Figure 4). The newcomers include stochastic gradient descent, Naïve Bayes, associate rule, multiclass and multilabel algorithms, and support vector machines. Despite being new in the field, support vector machines have already gained popularity with 110 applications in 2020, meaning they rank sixth among the 14 algorithm groups. Matrix factorisation and linear models have been the two most ML algorithms with more than

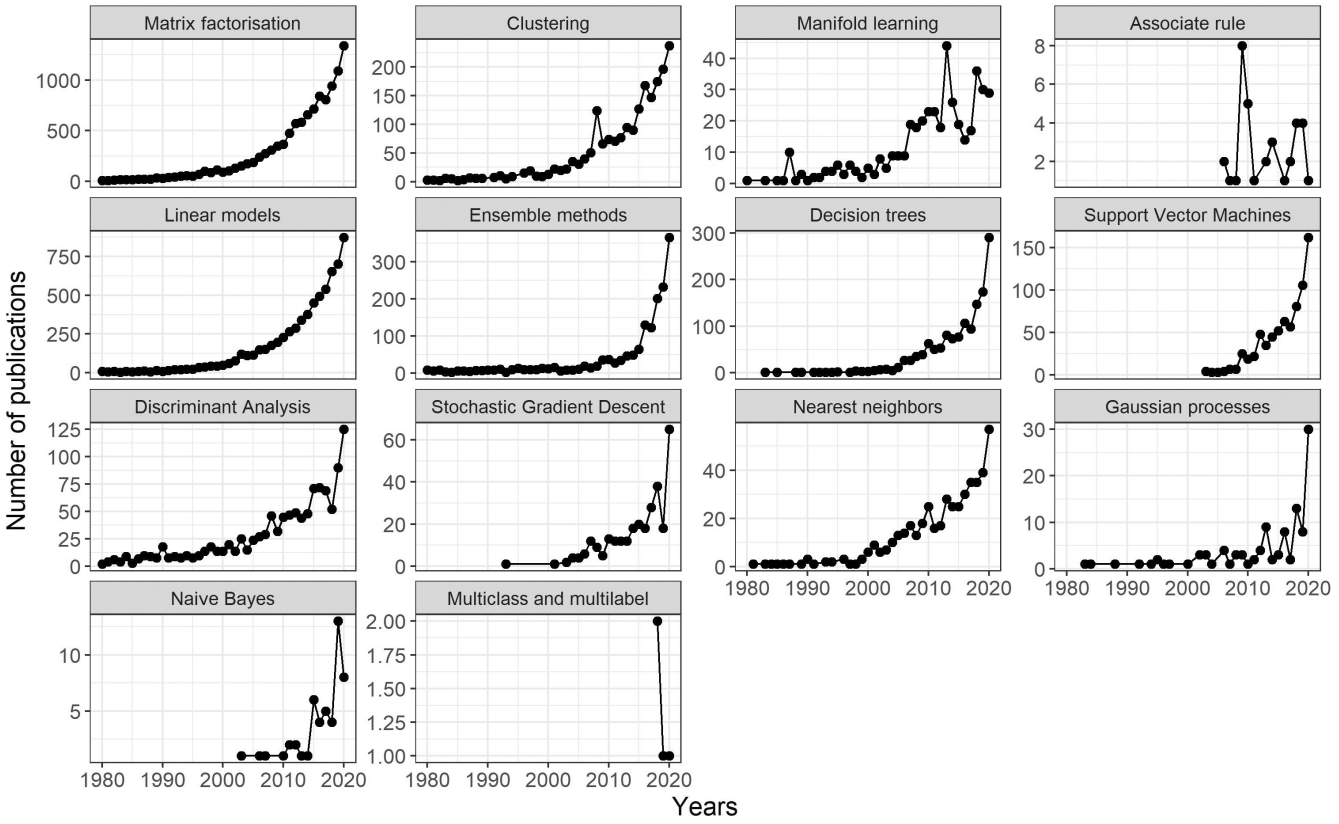

FIGURE 4 Evolution of the major supervised and unsupervised machine learning applications in river research over the past four decades. Note that matrix factorisation, clustering, manifold learning and associate rule are unsupervised while the other algorithms are supervised.

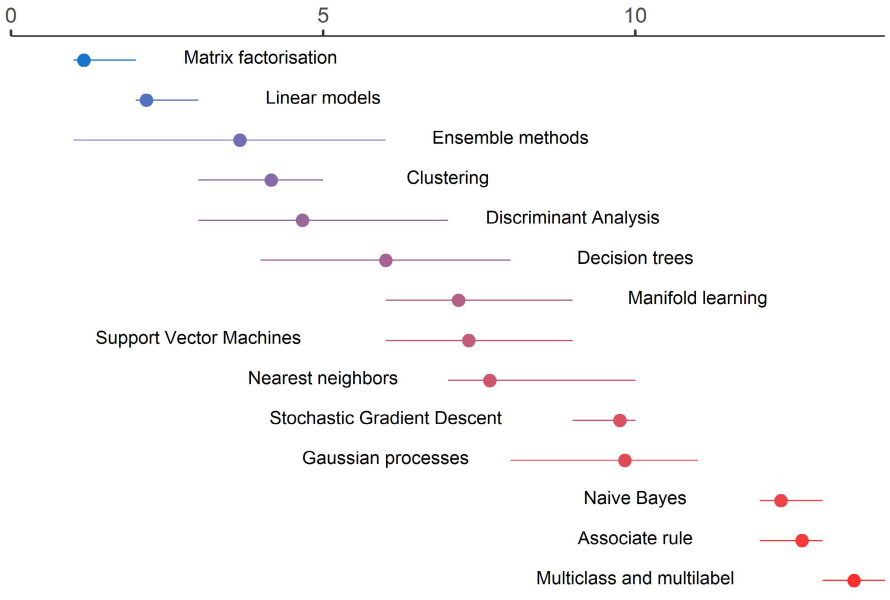

FIGURE 5 Ranking of the major supervised and unsupervised machine learning algorithms in integrated river management over the four decades (1980s, 1990s, 2000s and 2010s) and in 2020. The dots represent the median rank of each machine learning algorithm over the study period while the bars represent the maximum and minimum ranks of each machine learning algorithm over the study period. Note that matrix factorisation, clustering, manifold learning and associate rule are unsupervised learning while the other algorithms are supervised learning.

1300 and 800 publications in 2020, respectively. Conversely, river researchers have had few applications in multiclass and multilabel algorithm, associate rule and Naïve Bayes, which each have less than 15 publications per year.

Figure 5 indicates the median, maximum and minimum ranks of the 14 MI algorithms in integrated river management over the four decades (1980s, 1990s, 2000s and 2010s) and in 2020. Based on the median ranks over the study period, matrix factorisation, linear models, clustering and ensemble methods have been applied the most in river research. Interestingly, of the four most popular ML algorithms, two were supervised and two were unsupervised. This comparable proportion between the two main ML categories

suggested that within the aims of river modelling research there has been a balance between research for explanatory (via unsupervised learning) and predictive (via supervised learning). Note that matrix factorisation and clustering represent two main methods of unsupervised learning; the former aims to create a new representation of the data which might be more comprehensive compared to the original representation while the latter divides data into specific groups of similar items to facilitate data visualisation and interpretation (Müller & Guido, 2016). Having no output features, unsupervised algorithms are often applied to explore and describe data; hence, it is difficult to assess their performance. Regarding supervised ML, linear models and ensemble methods are two classes of powerful predictive models that are widely applied in practice (Géron, 2017; Ho et al., 2018). Simply put, linear models use a linear function to predict output feature, while ensemble methods make use of the 'wisdom of the crowd' by aggregating the predictions of a group of predictors, that is, regressors or classifiers, to predict new incoming instances (Ho et al., 2018; Krawczyk et al., 2017). As new methods in river ML modelling, reinforcement learning, Naïve Bayes, associate rule and multiclass and multilabel algorithms have ranked the lowest in the spectrum of ML applications in river research. Note that while the Naive Bayes method requires a 'naive' assumption of independence between the input features (Liu et al., 2017), which is normally not the case with river research, the other algorithms are more popular in other disciplines, such as market basket analysis, game theory and multi-agent systems (Pedregosa et al., 2019).

3.4 | Trends in river research topics

Figure 6 shows the trends for 25 important topics in river research field. These topics can be classified under six major research hotspots, including environment (aquatic environment, eutrophication and biogeochemistry), ecology (biodiversity and movement), hydrology (groundwater, hydrology and hydropower and dam), sanitation (emerging contaminants, heavy metal and water quality/pollution), human health (drinking water and public health) and socioeconomics (economics and social development) (Ho & Goethals, 2020; Jiang et al., 2016). Across the study period, hydrology, hydropower and dam, public health and water quality/pollution have remained the most popular topics. The popularity of topics such as movement, drinking water and fisheries has decreased, while that of the other topics, such as heavy metal, gas fluxes, spatiotemporal trends, land-use change, eutrophication and climate change, has grown considerably. For example, starting at the bottom of the list before the 1980s, land use and climate change jumped up nine and seven ranks, respectively, to become the fastest rising topics in river research field during the 2010s and 2020. This surge demonstrates the increasing attention given to the substantial impact of changes in land use and climate on river ecosystem. Mantyka-Pringle et al. (2012) displayed the importance of understanding the synergistic effects of climate change and habitat loss on biodiversity to integrate climate change adaptation measures into decision-making processes. Similarly, Molina-Navarro et al. (2018) illustrated that the nutrient loads of rivers largely depend on land use management in interaction with climate change. What is striking is that microbial and antibiotic resistance have never been a major research topic to most river researchers as they have stayed on the lowest ranks ever since. It is also worth noting that our findings showed that economics and social development have been of moderate interest to river researchers, ranking 8 and 12 respectively. This moderate interest is in contrast to the findings of Vugteveen et al. (2014) that indicated few studies on social and engineering fields in river research.

The dendrogram of the clustering results based on research topic similarity is shown in Figure 7. Here the horizontal axis represents the disparity of topics: the lower the connection between two topics, the higher their similarity. It appears that, except for management, other research topics have high similarity with at least one other topic, indicating the multidisciplinary nature of river research (Vugteveen et al., 2014). Emerging contaminants, movement, fisheries, drinking water and antibiotic resistance share a high level of similarity in their research, which implies an interdisciplinary research field on the spreading of antibiotic resistance and emerging contaminants in rivers that can affect freshwater fish or end up in drinking water systems. To illustrate this point, Cassini et al. (2018) indicated that the remaining antibiotic resistance genes from freshwater bodies in drinking water systems are a serious threat to public health, leading to 33,000 casualties each year in Europe as a direct consequence of an infection due to bacteria resistant to antibiotics. A high level of similarity between spatiotemporal trends and aquatic environment indicates an interdisciplinary research field in spatiotemporal variations of river systems, which are caused by a continuous gradient of physical conditions from headwaters to mouth. This principle is illustrated in the well-known river continuum concept explaining how the gradient of physical variables along a river can affect the biological features and ecosystem structure of flowing water systems (Vannote et al., 1980). Mutual interactions between studies on land use and climate change are also indicated in the dendrogram. Specifically, land-use change can cause the elevation of greenhouse gas (GHG) emissions from the rivers Ho et al. (2022) while climate change has a considerable effect on the hydromorphology of a river, causing changes in surrounding landscapes (Akter et al., 2018). It is striking that a high proportion of studies on biodiversity in rivers have also considered economic aspects, indicating that considerable attention has been paid to the economic valuation of biodiversity and ecosystem services of rivers. This combination could be a result of the 'Economics of Ecosystems and Biodiversity', a global initiative that can help decision makers recognise the values of biodiversity and ecosystem services (Kumar, 2010). Please note that we would like to stress that the results of the clustering analysis should be used only as a suggestion of potential interdisciplinary research themes in ML applications in river research. Further research is needed to investigate the state-of-the-art of each of the interdisciplinary topics. The nature of interdisciplinarity of river research was also indicated via the number and percentage of shared publications among the research topics which can be found in the **Supplementary Material E**. In particular, around 70% of the river research studied more than one of the 25 proposed research topics, and of these public health, hydropower and dams, hydrology, and

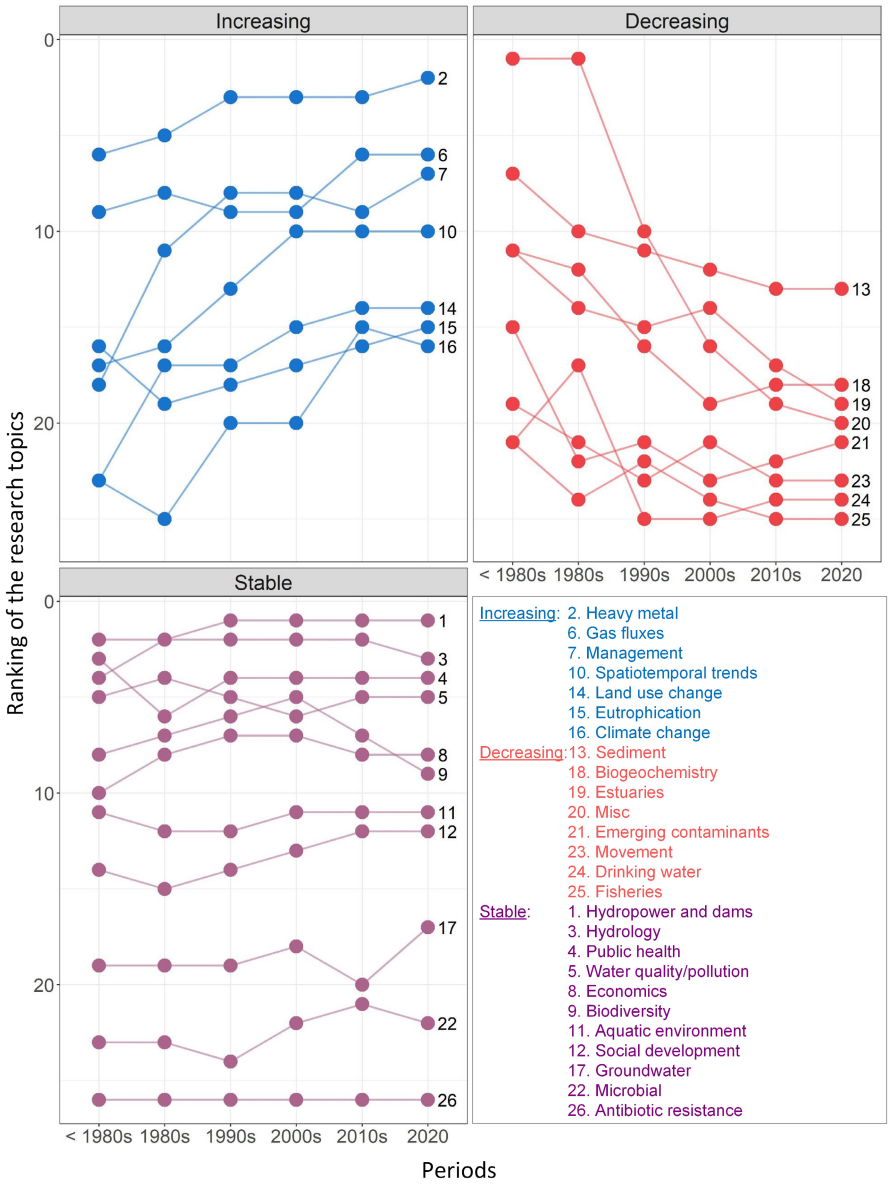

FIGURE 6 Changes in ranking of the major research topics over the study periods

heavy metal were the research topics that were linked to around 10% of the publications on other research topics.

4 | MACHINE LEARNING WORKFLOW IN RIVER RESEARCH: OPPORTUNITIES AND CHALLENGES

ML applications in river research should be considered as an end-to-end workflow consisting of various aspects of a data-intensive project. Figure 8 shows an ML pipeline or workflow which can be divided into four major blocks or steps: (1) data collection and preparation; (2) model development and optimisation; (3) model application; and (4) feedback loops. In each of the blocks, we illustrate the main components, including river systems, databases, ML algorithms, models and users, and the major processes needed to transform one component to the other, including data retrieval, data wrangling, feature engineering, model optimisation and tuning, model deployment and monitoring, and feedbacks. Note that the components and processes are not new in ML applications; however, to our knowledge, this is the first end-to-end ML workflow to be proposed in the context of river research. Building further on the framework of data science projects proposed by Wickham and Grolemund (2016), we highlighted the links between the systematic review and the workflow in the section of model development and optimisation. Beyond that, we elaborated in more detail below on important aspects within the ML workflow that river researchers should pay attention during the lifecycle of their ML applications.

4.1 | Data collection and preparation

4.1.1 | Data availability statement

Although gathering more data is often time-consuming and costly in river research, the availability of data is vital to the success of ML applications. Having sufficient data in all of the partitions of a given

dataset, including training, validation and test sets, is essential to ensure rigorous model training, optimisation and validation processes respectively. Note that training and test sets are used in cross validation as one of the common methods for model evaluation while a validation set is used in the hyperparameter optimisation step for model optimisation. Moreover, given the complex nature of river systems, ML applications in river research might have to deal with the curse of dimensionality which, together with overfitting, is the most challenging problem related to ML (Domingos, 2012). In particular, when considering additional factors, data dimensionality increases, leading to the tremendous expansion of data space, subsequently making the available data sparse. This sparsity in high-dimensional data hinders the ability of ML models to find patterns from the data that detect areas where samples form clusters with similar properties, consequently, causing statistically unreliable results. To avoid this, the required data must grow exponentially with the dimensionality (Trunk, 1979). As such, data have been a valuable yet limited resource for river researchers when conducting their studies. Recently, however, the state of affairs has changed drastically. While good quality data have lost none of their assets, there has been a rapid rise in the production of these data. Timely, reliable, accurate and comprehensive data can be collected at a relatively lower cost and they have thus become increasingly available (Kitchin, 2014). In general, 90% of the world's data volume has been generated in the last 5 years (Royal Society, 2017), leading to an increasing number of 'Big Data' and its research in various fields (Yaqoob et al., 2016). Note that Big Data refers to datasets with a large volume that cannot be stored, managed or analysed by common software tools (Vassakis et al., 2018). Moreover, following open data movement, open data platforms from several environmental agencies and organisations such as the European Environmental Agency, US EPA, and United Nations, have become available for researchers.

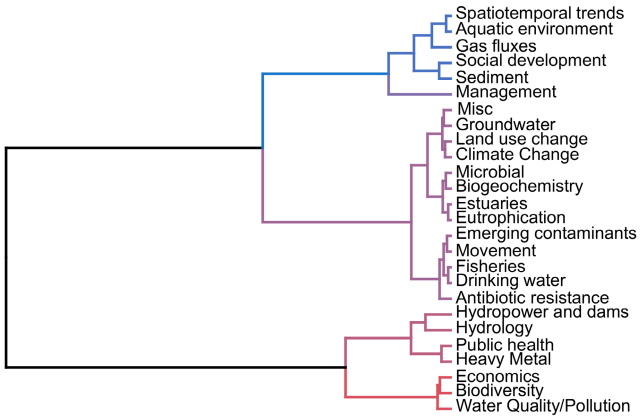

FIGURE 7 A cluster dendrogram of different topics in ML applications in river research

Together with the massive surge in observational field data, the number of benchmark datasets, which are often used for the comparison and testing of various ML algorithms that address specific questions in a specific field, has also been on the rise (Dueben

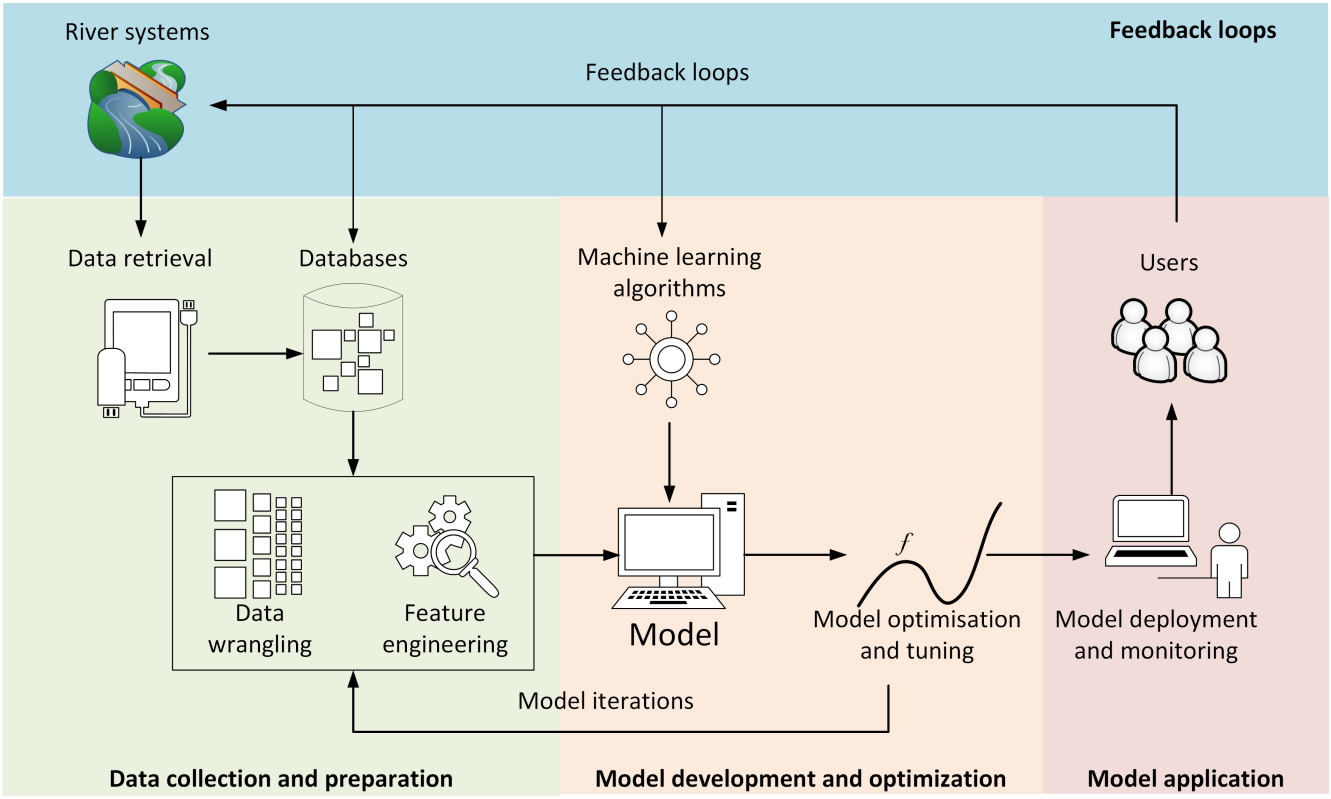

FIGURE 8 Workflow of machine learning applications in river research

et al., 2022). Despite this growth, only a handful of publications have investigated the topic of appropriate benchmarking datasets in general (Olson et al., 2017). In effect, high-quality benchmark datasets are valuable yet difficult, laborious and time-consuming to generate (Sarkar et al., 2020). In the context of river research, several benchmark datasets can be found, such as Caravan hydrology datasets (Kratzert et al., 2022), Global River Chemistry (GloRiCH) dataset (Hartmann et al., 2014), HYDROSHEDS digital database for hydro-ecological research (Dallaire et al., 2018), MethDB datasets of riverine methane concentrations and fluxes (Stanley et al., 2022). These datasets are often a result of meaningful cross-institutional collaborations which enables them grow over time in both complexity and size.

4.1.2 | Data collection

Alongside the rise of ML, researchers have witnessed the rapid development of technologies in data collection. Specifically, in contrast to monotonous and time-consuming processes of traditional monitoring systems, advanced monitoring tools are transforming the methods that river researchers use to collect data for their studies. For example, real-time and on-site data are being collected and monitored remotely using technology based on the Internet of Things (IoT), without human presence or intervention (Chowdury et al., 2019). Equipped with a low-power sensor device capable of wireless communication, this modern approach can provide high-quality data for water quality monitoring with reliability, scalability, speed and persistence, which have been widely applied in flood and irrigation (Adu-Manu et al., 2017). In addition to the IoT, remote sensing also allows researchers and scientists to rapidly and accurately obtain vast amounts of satellite data at high resolution. This tool has become very useful for fluvial hydrology and geomorphology to delineate and characterise river channels or to classify and monitor the changes in river landscape (Soulignac et al., 2018). Note that, unlike traditional remote sensing with multiple spectral sensors that have a limited number of broad spectral bands and were designed mainly to detect the concentration of the primary pigment in phytoplankton, hyperspectral remote sensing allows researchers to collect images across the full spectrum of visible and infrared light. Hence, hyperspectral remote sensing can offer more environmentally meaningful information, including assessments of aquatic biodiversity, habitats, water quality and environmental hazards (Dierssen et al., 2021). Come along with these benefits, the rapid increase in remote sensing technologies also poses new challenges regarding the right to privacy and personal data. For example, the observation of private spaces—including agricultural, residential and industrial areas near to river systems—with remote sensing technologies can interfere with rights to informational and location privacy (Maniadaki et al., 2021). As such, it is advisable for river researchers to ensure that they comply with the respective data protection and privacy requirements while using and sharing the data. Apart from the involvement of digital technologies, thousands of research projects, known as citizen science, are engaging millions of amateur scientists in collecting and analysing scientific data, generating a wealth of information (Bonney et al., 2014). In river research, the applications of citizen science can be found in international biodiversity monitoring (Chandler et al., 2017), water quality monitoring (Abbott et al., 2018) and hydrological monitoring (See, 2019), to name a few.

4.1.3 | Data quality

While having sufficient data is an obvious requirement for applying ML learners, ensuring the quality of data amenable to learning has typically not been discussed explicitly. One of the most important factors when deciding whether an ML application has been successful or not is the chosen features (Domingos, 2012). In fact, transforming and processing raw data into features with better representation of underlying problems is often an important and time-consuming step before feeding data to ML (Chicco, 2017). Specifically, this step of feature engineering includes cleaning and preprocessing input dataset, scale the data features into a normalised range, randomly shuffle dataset instances and imputation of missing values, to name a few. Note that ML algorithms are general purpose, while data are normally domain specific; hence, river researchers need to thoroughly understand their collected datasets and underlying problems of their studies. In this regard, Kosmala et al. (2016) recommended that ecological and environmental researchers should assess the quality of each citizen-science dataset individually according to the nature and purposes of research design and application. In addition to proper volunteer training and data management practices, expert validation and the application of relevant statistical techniques are important to guarantee high-quality data for citizen science. Bird et al. (2014) suggested that statistical methods, such as generalised linear models, mixed-effect models, hierarchical models and machine learning algorithms, can be used to mitigate random errors and systematic biases that can occur because of the potentially high variability among volunteers in terms of demographics, ability, effort and commitment.

4.2 | Model development and optimisation

4.2.1 | Algorithm selection

As shown in the systematic review, the number of machine learning algorithms from which researchers can choose to start a research project can seem overwhelming. Researchers may feel tempted to try sophisticated and up-to-date models to increase the quality and novelty of their studies, thereby increasing the chance of their research being published. Figures 2 and 3 show the sharp growth in the applications of neural networks and deep learning, support vector machines and ensemble modelling, which is different from the gradual increase in the application of well-established algorithms, such as manifold and linear models, which can seemingly be induced

by an increase in publications over time in general. This can be induced by a belief in the trade-off between model complexity and accuracy. In contrast, according to the widely acknowledged 'no free lunch' theorems, there is no guarantee that sophisticated models can perform better than simple linear models as no single technique will always do best (Wolpert & Macready, 1997). Besides, it is often more challenging to apply novel and complicated algorithms properly because of their complex and opaque internal structure (Domingos, 2015). As such, simple algorithms could be a more optimal option, and one that provides better generalisation insights, a lower chance of overfitting, and simple model training (Chicco, 2017; Domingos, 2012). This could be attributed to the fact that linear models and matrix factorisation remain the most commonly used ML algorithms in river research given their transparency and simplicity (Dobson & Barnett, 2011).

Noticeably, in the early days, most river research focused on one learner, while recently many efforts have been made to combine several learners, creating model ensembles. This change is indicated by the boost in the application of ensemble methods in river research during the last 5 years (Figure 2). The goal of the ensemble modelling is to balance the trade-off between bias, the tendency of an algorithm to learn in the same wrong way, and variance, the tendency of an algorithm to account for data variations (Zhang & Ma, 2012). Prominent among several ML techniques in ensemble modelling, *bagging*, *boosting* and *stacking* can often increase performance over a single model. However, the associated drawbacks of a loss of interpretability and increase in computational cost should be borne in mind (Witten & Frank, 2005).

4.2.2 | Model evaluation

One of the key concerns in model evaluation is the validation of model performance beyond the examples in the training set. In other words, to avoid the mistake of overestimating model performance by using the same test set, researchers need to evaluate model capacity in generalising their prediction to other test sets or new data (Greener et al., 2021). In effect, generalisation across studies is one of the fundamental yet difficult goals of ML in river research given that data are often featured by distinct study conditions with a large number of influencing factors which makes the fixed-size training set become a diminishing proportion of the data space (Domingos, 2012). On the other hand, within a case study, it is important to avoid contaminating learners by using test data during the training process (Domingos, 2012). Note that unlike supervised algorithms in which training and test sets are used in cross validation (CV) to test their predictive model (Browne, 2000), the evaluation of unsupervised algorithms consists of internal and external validations (Palacio-Niño & Berzal, 2019). In particular, while the goodness of the fitted model is evaluated in internal validation, the agreement between two patterns is compared in external validation: one pattern is a result of unsupervised algorithms, the other pattern is already defined in the a priori known dataset (Wang et al., 2009). Similar to the conflation between explanation and prediction purposes when choosing ML categories (Shmueli, 2010), researchers should avoid the conflation between the evaluation of supervised and unsupervised algorithms. This is particularly vital in the context of river research as it is demonstrated in both Figures 3 and 4 that researchers have paid attention evenly to both supervised and unsupervised ML categories.

It is important to note that in CV step, researchers have to confront the dilemma of choosing between exhaustive CV, including leave-p-out, leave-one-out, bootstrap and non-exhaustive CV, including holdout, *k*-fold and stratified *k*-fold. Researchers working on studies with small datasets might be tempted to apply exhaustive CV that tests all possible ways to split a given dataset. On the other hand, no benchmark can be found for non-exhaustive CV regarding the number of folds into which a given dataset should be partitioned and the number of times which researchers should repeat the CV process to avoid the effect of randomisation. Also noteworthy is an issue of data leakage in CV that leads to better model predictions than they actually are on the training or validation data. Data leakage happens when the information about the target gives models an impractical advantage to make better prediction in test sets or new data, such as test data are leaked into the training set or future data are leaked to the past (Zheng & Casari, 2018). For example, when a dataset contains features in different ranges, normalisation can be used to rescale them to the range 0–1 before feeding into an ANN model. This feature scaling requires minimum and maximum values of each feature in the whole dataset, whereby the information from the test set can affect the training set ultimately causing data leakage. This is also the case when we impute missing values using mean substitution of the feature in the full dataset or using regression to predict missing values based on other features in the full dataset. A simple solution for these subtle leakages is to implement these preprocessing steps separately in the training and test datasets. Other case of data leakage is when using features that hold information of the model outputs but will not naturally be available in new datasets or including feature that is a proxy of the other. For example, including electrical conductivity or total dissolved solid when predicting salinity of a water sample. To avoid this leakage, proper data preprocessing steps and model testing in rigorous benchmarking are necessary before a research can be considered for publication (Greener et al., 2021).

4.2.3 | Multi-objective optimisation

It is equally important that river researchers do not only focus on explanatory and predictor powers when evaluating a learner, since multiple performance criteria are also important, such as computational efficiency, generalisability, conceptual simplicity, robustness and required assumptions (Boulesteix, 2015). Traditionally, the model optimisation is implemented regarding a single objective; however, river modellers have increasingly encountered the

optimisation problem in model evaluation when multiple criteria must be simultaneously optimised. In the latter case, multi-objective or Pareto optimisation is normally used to evaluate the trade-off between predictive performance, simplicity and the number of selected features. Unlike single-objective algorithms, Pareto optimisation can deliver a set of optimal solutions rather than a single solution by exploring a considerably wider search space and tracking all possible solutions in the Pareto front (Freitas, 2004). In a comparison between Pareto optimisation and single-objective optimisation of species distribution models for river management by Gobeyn and Goethals (2019), the Pareto approach was two to four times more efficient in identifying a wide-range set of optimal models with only a 4% increase in runtime compared to the latter optimisation. Note that unlike multi-objective optimisation, multi-task optimisation aims to find the optimal solutions for multiple tasks in a single simulation (Xu et al., 2021). For example, instead of applying numerous models to predict single water quality variable, Zhang et al. (2019) applied a multi-task temporal convolution network to forecast various water quality constituents simultaneously, leading to a significantly reduced training time while retaining a promising predictive accuracy.

4.2.4 | Hyperparameter optimisation

After choosing a learner, selection and optimisation of its hyperparameters are the next important step because these parameters can strongly affect model complexity, efficiency and results (Probst et al., 2019). Notable examples of these hyperparameters are the learning rate for training a neural network, the number *k* of clusters in *k*-nearest neighbours and the size of a node in gradient boosting, to name a few. These hyperparameters are typically used for regularisation to reduce the overfitting effect or the balance between bias and variance in model predictions. One can reduce the depth of trees as an efficient regularisation parameter in gradient boosting or random forests since when the depth of trees increases, the model is likely to overfit the training data, leading to high variance and low bias. Note that another regularisation technique is to incorporate penalization in the optimisation routine or punish the model outcomes if overly complex models are preferred, such as smoothness, ridge and lasso penalty, and elastic net in linear models. Although this step remains largely omitted from river research, when applying this, researchers should provide sufficient information about what hyperparameters are selected, their ranges, and how sensitive model results are towards the tuning process of the hyperparameters (McCoy & Auret, 2019). In this case, apart from training set and test set, input dataset should include a validation set that is employed for hyperparameter tuning, while an independent test set should be withheld for checking model performance. This so-called 'lock-box' approach of data partition for CV has become a common and effective tool for data scientists in many fields to constrain model overfitting (Hosseini et al., 2016), hence, it is highly recommended to river researchers.

4.2.5 | Computational power, programming languages and platforms

Also noteworthy is the growth of computing power, programming languages and platforms whose main focus is on ML. Besides the availability of data, the rapid growth of computational power has been essential to the current breakthrough in ML applications. In effect, the IT industry has tended towards increasingly specialised platforms for ML applications because of fast growth of this field for commercial interests. A notable example of this specification is the development of graphics processing units (GPUs) which have become the current backbone of ML workflow (Hwang, 2018). Unlike traditional central processing units (CPUs) which feature a small number of cores used for handling a few tasks simultaneously, GPUs feature an architecture distributing tasks across a large number of cores that can be implemented in parallel. This parallel architecture is very compatible with ML applications, such as neural networks whose training and validation depend largely on the execution of numerous matrix multiplication calculations (Hwang, 2018). More recently, FPGAs (Field Programmable Gate Array) and ASICs (Application Specific Integrated Circuit) have also considerably increased the popularity of ML applications thanks to their ability to support massive parallel computation with a much lower power consumption (Itagi et al., 2021). To facilitate the use of the embedded hardware for processing ML applications, the largest companies in the IT industry, such as Microsoft Azure, Google Cloud, Amazon AWS, IBM Watson and DataRobot, have created several deployment platforms. The selection of platform and hardware system for processing AI typically depends on the trade-off between costs and model performance and scale.

Regarding programming languages and software, an increasing number of studies have preferred to use high-level, high-performance, open-source programming languages, such as R, Python, Torch, JavaScript, SQL and Julia, to develop their ML applications because of their reproducibility, flexibility and transparency (dos Santos et al., 2019). On the other hand, river researchers who are not interested in programming or algorithm development can use free software such as Waikato Environment for Knowledge Analysis (Weka) as a non-programing approach because of its friendly and self-explanatory user interface (dos Santos et al., 2019). Paid software tools such as MATLAB®, SPSS and STATA are still popular among river researchers due to their powerful toolbox (dos Santos et al., 2019).

4.3 | Model application

4.3.1 | Decision-making tools

Currently used for high-stakes decision-making in many fields, ML applications make a significant contribution to human society. However, some of these contributions are causing ethical concerns in healthcare, criminal justice and other domains (Varshney

& Alemzadeh, 2017). For example, significant racial bias was recently showed in a widely used algorithm in the US healthcare system: Black patients were assigned a lower score for receiving healthcare compared to equally at-risk White patients Obermeyer et al. (2019). This bias highlights the lack of transparency and accountability in the results of these black-box models (Varshney & Alemzadeh, 2017). Note that, in recent decades, researchers have witnessed the rise of uninterpretable ML algorithms, such as deep learning and neural networks, although it is easy to understand and interpret the mechanism and results of the well-established ML algorithms, such as linear models, decision trees and rule-based models (Barredo Arrieta et al., 2020). In river research, deep learning and neural networks now feature in 15%–21% of all publications during the last two decades. This rise has been based on the widespread belief in the trade-off between model accuracy and interpretability that is not always the case according to 'no free lunch' theorems. In contrast, while the difference in performance between ML algorithms is often small, the ability to interpret their results and processes becomes much more crucial for decision makers (Rudin, 2019).

On the other hand, there have also been considerable efforts to interpret 'black-box' ML models by means of explainable artificial intelligence (XAI) and integrated physics-ML models. First, aiming to produce details to clarify its functioning, XAI can be classified into two types: (1) transparent ML models, such as linear models, decision trees and nearest neighbours, and (2) post-hoc explainability techniques or model-agnostic methods, such as Partial Dependence Plots (PDPs), Individual Conditional Expectation (ICE) plots, Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) (Ribeiro et al., 2016). For further information on the concepts, taxonomies and opportunities of XAI, readers are referred to Barredo Arrieta et al. (2020). Second, the integration of physics-based modelling with ML aims to mitigate the drawbacks of both physics-based models and ML, some of which are the inaccuracy prediction of the physics-based models due to incomplete knowledge of complex systems, and the reduction of computational costs (Quaghebeur et al., 2022). In addition to these advantages, the increase in generalizability and interpretability of integrated physics-ML models is particularly valuable as the features are desirable but typically missing in ML models (Willard et al., 2020). For further information on the objectives, methods and architecture of integrated physics-ML models, readers are referred to Willard et al. (2020).

From another perspective to avoid the danger of making decisions without having a detailed explanation of the models, the EU introduced the revolutionary General Data Protection Regulation (GDPR) that governs 'right to an explanation', in particular the fact that 'The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her' (Article 22 of GDPR). This policy on the right of citizens to obtain an explanation for algorithmic-based decisions indicates the importance of human interpretability in algorithm designs (Goodman & Flaxman, 2016). While racial bias might not be of main concern in the ML applications in river research, involved stakeholders, such as farmers and fishermen, that are affected by the results of ML applications still have 'right to an explanation'. Hence, it is vital to integrate expert insights to evaluate the results of black-box ML algorithms before using them in decision support-making processes.

4.3.2 | Stakeholder participation

Equally important is the participation of involved stakeholders as the end-users of the ML algorithms to ensure the transfer of their methods into practice. From this point of view, the transparency and simplicity of models, such as decision trees and fuzzy models with simple if-then rules, are valuable assets. These models have proven their success in many fields, such as medicine (Ahmadi et al., 2018), electronics (Singh et al., 2013), (waste)water treatment (Porro et al., 2018) and climate change (Ho et al., 2021). On the other hand, having limited practical experience, researchers advance new statistical techniques that have appeared difficult for practitioners to learn due to their lack of participation in the development of these techniques (Corominas et al., 2017). To bridge this gap, it is advised that researchers should develop more intuitive data analysis methods while practitioners, including students, technical officers, consultants and academic, should either be more involved in the model development or they should contract the services of well-trained experts.

4.4 | Feedback loops

4.4.1 | Continuous ML applications

Feedback loops play an important role in boosting the performance of ML applications as they ensure that the results of ML applications will not become stagnant when the real-world data are continuously updated. In fact, acting as an approximation of what happens in the real world, training datasets of river ecosystems continuously evolves because of the nature of the systems as well as the unprecedented impacts of climate change; hence, with feedback loops, river researchers can reinforce and keep improving the performance of ML models over time. Fei and Lu (2018) suggested that the predictive performance of neural network models would be considerably higher if they have at least one feedback loop as compared to those with none. Feedback loops can be applied to both data collection and preparation as well as model development and optimisation, to optimise the methods applied in these steps in new cases. In other words, the methods that were optimal in the previous dataset can be inadequate when applied in the updated dataset, which can be the case in many rivers where real-time and on-site data are increasingly being retrieved because of the rapid development of IoT technologies.

4.4.2 | Avoid biases and limitations

It is especially worth noting that choices have to be made at every step of the proposed workflow, such as training/validation datasets, ML algorithms and model performance metrics, which can create biases and limitations in model results. River modellers hence must always be critical about their choices and data and explicitly discuss the biases and limitations in their publications; the users, policymakers, and the public can thus scrutinise the final deliverables of ML applications, whereby researchers keep providing feedback to model development process. For example, to overcome the shortcomings of ML applications on systematic underrepresentation or overrepresentation, proper experimental design, data collection scheme and sampling methodology should be followed (Ho et al., 2019). Specifically, sampling design and power analysis can be used to design a sampling programme that generates the most effective and precise estimates of the input features in the real world.

4.4.3 | System improvement

Another feedback loop is that the results of ML applications can be used to better manage and operate river basins. A notable example of this feedback loop is the use of ML models in flood prevention which has contributed to risk reduction, policy development, minimisation of the loss of human life and reduction of property damage associated with these destructive natural disasters (Mosavi et al., 2018). In fact, despite the early stage of research in this field, hydrologists and climate scientists have applied ML to more than 6000 scientific studies on flood forecast in which data including rainfall, water level and precipitation level, were increasingly collected by remote sensing technologies such as satellites, multisensory systems and radars (Mosavi et al., 2018). The effective, real-time continuous flow of data and feedback have enabled rapid and necessary actions for climate resilience in modern cities, such as constructing levees or reforming building codes. This is also the case with ML applications forecasting other natural disasters, such as droughts, storms, forest fires, to name a few (Goswami et al., 2018).

5 | CONCLUSIONS

Overall, there has been great excitement about machine learning (ML) applications in river research during the last two decades as the number of publications of most ML algorithms has rapidly increased. Trend analysis showed that unsupervised learning and supervised learning have dominated the field of river research while neural networks and deep learning have gained more attention in this field, featuring in 15%–21% of the total publications over the last two decades. The early dominance of unsupervised learning indicates the focus of river researchers on understanding river systems, while within the last decade, attention has turned towards predictive models for estimating and forecasting river states. A limited number of publications applying reinforcement learning, semi-supervised learning and self-supervised learning reveals an opportunity to exploit their application in river research in the future. Matrix factorisation and linear models have been the most popular ML algorithms with more than 1300 and 800 publications respectively appearing in 2020. Conversely, river researchers have had few applications in multiclass and multilabel algorithm, associate rule, Naïve Bayes and Gaussian processes.

Across the study periods, hydrology, hydropower and dam, public health and water quality/pollution have remained the most common topics. The popularity of topics such as movement, drinking water and fisheries has decreased, while that of other topics such as heavy metal, gas fluxes, spatiotemporal trends, land-use change, eutrophication and climate change has grown considerably. The interdisciplinary nature of river research was also indicated via the number and percentage of shared publications among the research topics as around 70% of the river research publications studied more than one of the 25 proposed research topics. Here, public health, hydropower and dams, hydrology and heavy metal were the research topics that were linked to around 10% of publications of other research topics.

While researchers might be tempted to choose sophisticated and advanced models, well-established ML such as decision trees and linear models, still retain their assets including generalisability, conceptual simplicity, robustness and transparency. On the other hand, substantial efforts have been made to interpret 'black-box' models by using explainable artificial intelligence (XAI) and integrated physics-ML models. Furthermore, data warehousing is becoming a cornerstone of ML's success in river management, while end-to-end project management with multiple-cycles is essential for the improvement of the integration of ML in decision-making in river and ecosystem management.

We proposed and recommended an innovative end-to-end workflow of ML applications in river research that includes four major steps: (1) data collection and preparation; (2) model evaluation and selection; (3) model application; and (4) feedback loops. Within this workflow, river modellers have to balance numerous trade-offs between: model complexity and accuracy; model interpretability and transparency; model bias and variance; data privacy and accessibility; model scale and required resources, such as computational cost, development time and data quantity. Any choices between these trade-offs can lead to different model outcomes that can affect model application. By considering modelling goals, understanding collected data and maintaining feedback loops, river researchers can continuously improve their model performance and eventually reach their research objectives.

AUTHOR CONTRIBUTIONS

Both authors conceived the ideas. Long Ho designed methodology, collected and analysed the data, and led the writing of the manuscript. Both authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

We thank Ngoc Quang Luong from the IDLab, University of Antwerp and IMEC for his comments on the development of our work. We are grateful for the insightful and constructive comments of Simon J Dixon from the University of Birmingham and other two reviewers that helped to improve the quality of our works. Long Ho is a post-doctoral fellow supported by the Research Foundation of Flanders (FWO) (project number 1253921N).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13992.

DATA AVAILABILITY STATEMENT

All data used in this manuscript are present in the Shiny application which is available online at https://env-research.shinyapps.io/ML_river/. Data are also published in the Dryad Digital Repository (https://doi.org/10.5061/dryad.z34tmpght).

ORCID

Long Ho https://orcid.org/0000-0002-2999-1691
Peter Goethals https://orcid.org/0000-0003-1168-6776

REFERENCES

Abbott, B. W., Moatar, F., Gauthier, O., Fovet, O., Antoine, V., & Ragueneau, O. (2018). Trends and seasonality of river nutrients in agricultural catchments: 18 years of weekly citizen science in France. *Science of the Total Environment*, *624*, 845–858.

Adu-Manu, K. S., Tapparello, C., Heinzelman, W., Katsriku, F. A., & Abdulai, J. D. (2017). Water quality monitoring using wireless sensor networks: Current trends and future research directions. *ACM Transactions on Sensor Networks*, *13*, 1–41.

Ahmadi, H., Gholamzadeh, M., Shahmoradi, L., Nilashi, M., & Rashvand, P. (2018). Diseases diagnosis using fuzzy logic methods: A systematic and meta-analysis review. *Computer Methods and Programs in Biomedicine*, *161*, 145–172.

Akter, T., Quevauviller, P., Eisenreich, S. J., & Vaes, G. (2018). Impacts of climate and land use changes on flood risk management for the Schijn River, Belgium. *Environmental Science & Policy*, *89*, 163–175.

Albalawi, F., Chahid, A., Guo, X., Albaradei, S., Magana-Mora, A., Jankovic, B. R., Uludag, M., Van Neste, C., Essack, M., Laleg-Kirati, T. M., & Bajic, V. B. (2019). Hybrid model for efficient prediction of poly(a) signals in human genomic DNA. *Methods*, *166*, 31–39.

Ayodele, T. O. (2010). Types of machine learning algorithms. *New Advances in Machine Learning*, *3*, 19–48.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.

Beaulieu, J. J., Tank, J. L., Hamilton, S. K., Wollheim, W. M., Hall, R. O., Mulholland, P. J., Peterson, B. J., Ashkenas, L. R., Cooper, L. W., Dahm, C. N., Dodds, W. K., Grimm, N. B., Johnson, S. L., McDowell, W. H., Poole, G. C., Valett, H. M., Arango, C. P., Bernot, M. J., Burgin, A. J., … Thomas, S. M. (2011). Nitrous oxide emission from denitrification in stream and river networks. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 214–219.

Bird, T. J., Bates, A. E., Lefcheck, J. S., Hill, N. A., Thomson, R. J., Edgar, G. J., Stuart-Smith, R. D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J. F., Pecl, G. T., Barrett, N., & Frusher, S. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, *173*, 144–154.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., & Parrish, J. K. (2014). Next steps for citizen science. *Science*, *343*(6178), 1436–1437. https://doi.org/10.1126/science.1251554

Boulesteix, A. L. (2015). Ten simple rules for reducing overoptimistic reporting in methodological computational research. *PLoS Computational Biology*, *11*, e1004191.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*, 108–132.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics*, *22*, 155–205. https://doi.org/10.1007/BF02019280

Cassini, A., Högberg, L. D., Plachouras, D., Quattrocchi, A., Hoxha, A., Simonsen, G. S., Colomb-Cotinat, M., Kretzschmar, M. E., Devleesschauwer, B., & Cecchini, M. (2018). Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European economic area in 2015: A population-level modelling analysis. *The Lancet Infectious Diseases*, *19*, 56–66.

Chandler, M., See, L., Copas, K., Bonde, A. M. Z., Lopez, B. C., Danielsen, F., Legind, J. K., Masinde, S., Miller-Rushing, A. J., Newman, G., Rosemartin, A., & Turak, E. (2017). Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, *213*, 280–294.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2015). *Package 'shiny'*. http://citeseerx.ist.psu.edu/viewdoc/download

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *Biodata Mining*, *10*, 35.

Chowdury, M. S. U., Bin Emran, T., Ghosh, S., Pathak, A., Alam, M. M., Absar, N., Andersson, K., & Hossain, M. S. (2019). IoT based real-time river water quality monitoring system. *16th International Conference on Mobile Systems and Pervasive Computing (Mobispc 2019), the 14th International Conference on Future Networks and Communications (Fnc-2019), the 9th International Conference on Sustainable Energy Information Technology*, *155*, 161–168.

Cooper, S. D., Lake, P. S., Sabater, S., Melack, J. M., & Sabo, J. L. (2013). The effects of land use changes on streams and rivers in mediterranean climates. *Hydrobiologia*, *719*, 383–425.

Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., & Poch, M. (2017). Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. *Environmental Modelling & Software*, *106*, 89–103.

Dallaire, C. O., Lehner, B., Sayre, R., & Thieme, M. (2018). A multidisciplinary framework to derive global river reach classifications at high spatial resolution. *Environmental Research Letters*, *14*, 024003.

Dierssen, H. M., Ackleson, S. G., Joyce, K. E., Hestir, E. L., Castagna, A., Lavender, S., & McManus, M. A. (2021). Living up to the hype of hyperspectral aquatic remote sensing: Science, resources and outlook. *Frontiers in Environmental Science*, *9*.

Dobbelaere, M. R., Plehiers, P. P., Van de Vijver, R., Stevens, C. V., & Van Geem, K. M. (2021). Machine learning in chemical engineering: Strengths. *Weaknesses*, *Opportunities*, *and Threats Engineering*, *7*, 1201–1211.

Dobson, A. J., & Barnett, A. (2011). *An introduction to generalized linear models*. CRC Press.

Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. *Proceedings of the IEEE International Conference on Computer Vision*, 1422–1430.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*, 78–87.

Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Penguin Books Limited.

dos Santos, B. S., Steiner, M. T. A., Fenerich, A. T., & Lima, R. H. P. (2019). Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. *Computers & Industrial Engineering*, *138*, 138.

Drake, T. W., Raymond, P. A., & Spencer, R. G. M. (2018). Terrestrial carbon inputs to inland waters: A current synthesis of estimates and uncertainty. *Limnology and Oceanography Letters*, *3*, 132–142.

Dueben, P. D., Schultz, M. G., Chantry, M., Gagne, D. J., Hall, D. M., & McGovern, A. (2022). Challenges and benchmark datasets for machine learning in the atmospheric sciences: Definition, status, and outlook. *Artificial Intelligence for the Earth Systems*, *1*, e210002.

Fei, J. T., & Lu, C. (2018). Adaptive sliding mode control of dynamic systems using double loop recurrent neural network structure. *IEEE Transactions on Neural Networks and Learning Systems*, *29*, 1275–1286.

Foley, J., Yuan, Z., Keller, J., Senante, E., Chandran, K., Willis, J., Shah, A., van Loosdrecht, M., & van Voorthuizen, E. (2015). N_2O and CH_4 emission from wastewater collection and treatment systems: State of the science report and technical report. *Water intelligence online* (Vol. *14*). IWA Publishing.

Freitas, A. A. (2004). A critical review of multi-objective optimization in data mining: A position paper. *SIGKDD Explorations Newsletter*, *6*, 77–86.

Géron, A. (2017). *Hands-on machine learning with Scikit-learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Gibert, K., Horsburgh, J. S., Athanasiadis, I. N., & Holmes, G. (2018). Environmental data science. *Environmental Modelling & Software*, *106*, 4–12.

Gobeyn, S., & Goethals, P. L. M. (2019). Multi-objective optimisation of species distribution models for river management. *Water Research*, *163*, 114863.

Goethals, P. L. M., Dedecker, A. P., Gabriels, W., Lek, S., & De Pauw, N. (2007). Applications of artificial neural networks predicting macroinvertebrates in freshwaters. *Aquatic Ecology*, *41*, 491–508.

Goodman, B., & Flaxman, S. (2016). *EU regulations on algorithmic decision-making and a 'right to explanation'*. ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY. http://arxiv.org/abs/1606.08813v1

Goswami, S., Chakraborty, S., Ghosh, S., Chakrabarti, A., & Chakraborty, B. (2018). A review on application of data mining techniques to combat natural disasters. *Ain Shams Engineering Journal*, *9*, 365–378.

Greener, J. G., Kandathil, S. M., Moffat, L., & Jones, D. T. (2021). A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, *23*, 40–55.

Guneralp, B., Guneralp, I., & Liu, Y. (2015). Changing global patterns of urban exposure to flood and drought hazards. *Global Environmental Change-Human and Policy Dimensions*, *31*, 217–225.

Hao, J. G., & Ho, T. K. (2019). Machine learning made easy: A review of Scikit-learn package in python programming language. *Journal of Educational and Behavioral Statistics*, *44*, 348–361.

Hartmann, J., Lauerwald, R., & Moosdorf, N. (2014). A brief overview of the GLObal RIver CHemistry database, GLORICH. *Geochemistry of the Earth's Surface Ges-10*, *10*, 23–27.

Hauer, C., Siviglia, A., & Zolezzi, G. (2017). Hydropeaking in regulated rivers—From process understanding to design of mitigation measures. *Science of the Total Environment*, *579*, 22–26.

Ho, L., & Goethals, P. (2020). Research hotspots and current challenges of lakes and reservoirs: A bibliometric analysis. *Scientometrics*, *124*, 603–631.

Ho, L., Jerves-Cobo, R., Barthel, M., Six, J., Bode, S., Boeckx, P., & Goethals, P. (2022). Greenhouse gas dynamics in an urbanized river system: Influence of water quality and land use. *Environmental Science and Pollution Research*, *29*, 37277–37290.

Ho, L., Jerves-Cobo, R., Eurie Forio, M. A., Mouton, A., Nopens, I., & Goethals, P. (2021). Integrated mechanistic and data-driven modeling for risk assessment of greenhouse gas production in an urbanized river system. *Journal of Environmental Management*, *294*, 112999.

Ho, L., Pompeu, C., Van Echelpoel, W., Thas, O., & Goethals, P. (2018). Model-based analysis of increased loads on the performance of activated sludge and waste stabilization ponds. *Water*, *10*, 1410.

Ho, L., Thas, O., Van Echelpoel, W., & Goethals, P. (2019). A practical protocol for the experimental Design of Comparative Studies on water treatment. *Water*, *11*, 162.

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., & Wyble, B. (2016). I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience and Biobehavioral Reviews*, *119*, 456–467.

Hotchkiss, E. R., Hall, R. O., Jr., Sponseller, R. A., Butman, D., Klaminder, J., Laudon, H., Rosvall, M., & Karlsson, J. (2015). Sources of and processes controlling CO_2 emissions change with the size of streams and rivers. *Nature Geoscience*, *8*, 696–699.

Hwang, T. (2018). *Computational power and the social impact of artificial intelligence*. Available at SSRN 3147971.

Itagi, A., Krishvadana, S., Bharath, K. P., & Kumar, M. R. (2021). FPGA architecture to enhance hardware acceleration for machine learning applications. *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1716–1722). IEEE. https://doi.org/10.1109/ICCMC51019.2021.9418015

Jiang, H. C., Qiang, M. S., & Lin, P. (2016). A topic modeling based bibliometric exploration of hydropower research. *Renewable & Sustainable Energy Reviews*, *57*, 226–237.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. SAGE Publications.

Kosmala, M., Wiggins, A., Swanson, A., & Simmons, B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, *14*, 551–560.

Kostopoulos, G., Karlos, S., Kotsiantis, S., & Ragos, O. (2018). Semi-supervised regression: A recent review. *Journal of Intelligent & Fuzzy Systems*, *35*, 1483–1500.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, *26*, 159–190.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., & Nevo, S. (2022). Caravan—A global community dataset for large-sample hydrology.

Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Wozniak, M. (2017). Ensemble learning for data stream analysis: A survey. *Information Fusion*, *37*, 132–156.

Kroeze, C., Dumont, E., & Seitzinger, S. P. (2005). New estimates of global emissions of N_2O from rivers and estuaries. *Environmental Sciences*, *2*, 159–165.

Kumar, P. (2010). *The economics of ecosystems and biodiversity: Ecological and economic foundations*. Earthscan.

Kummu, M., de Moel, H., Ward, P. J., & Varis, O. (2011). How close do we live to water? A global analysis of population distance to freshwater bodies. *PLoS ONE*, *6*.

Lee, K. F. (2018). *AI superpowers: China, Silicon Valley, and the New World order*. HMH Books.

Lehner, B., Liermann, C.R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., & Magome, J. (2011). *Global reservoir and dam (grand) database*. Technical Documentation, Version, 1.

Lehner, B., Liermann, C. R., Revenga, C., Vorosmarty, C., Fekete, B., Crouzet, P., Doll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J. C., Rodel, R., Sindorf, N., & Wisser, D. (2011). High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. *Frontiers in Ecology and the Environment*, *9*, 494–502.

Linden, A., & Fenn, J. (2003). *Understanding Gartner's hype cycles*. Strategic analysis report no R-20-1971. Gartner, Inc, 88.

Liu, R., Chen, Y., Wu, J. P., Gao, L., Barrett, D., Xu, T. B., Li, X. J., Li, L. Y., Huang, C., & Yu, J. (2017). Integrating entropy-based naive Bayes and GIS for spatial evaluation of flood Hazard. *Risk Analysis*, *37*, 756–773.

Maniadaki, M., Papathanasopoulos, A., Mitrou, L., & Maria, E. A. (2021). Reconciling remote sensing technologies with personal data and privacy protection in the European Union: Recent developments in Greek legislation and application perspectives in environmental law. *Laws*, *10*, 33.

Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica*, *13*, 245–259.

Mantyka-Pringle, C. S., Martin, T. G., & Rhodes, J. R. (2012). Interactions between climate and habitat loss effects on biodiversity: A systematic review and meta-analysis. *Global Change Biology*, *18*, 1239–1252.

McCallen, E., Knott, J., Nunez-Mir, G., Taylor, B., Jo, I., & Fei, S. L. (2019). Trends in ecology: Shifts in ecological research themes over the past four decades. *Frontiers in Ecology and the Environment*, *17*, 109–116.

McCoy, J. T., & Auret, L. (2019). Machine learning applications in minerals processing: A review. *Minerals Engineering*, *132*, 95–109.

McLeod, A., & McLeod, M. A. (2015). *Package 'Kendall'*.

Meijer, L. J. J., van Emmerik, T., van der Ent, R., Schmidt, C., & Lebreton, L. (2021). More than 1000 rivers account for 80% of global riverine plastic emissions into the ocean. *Science Advances*, *7*, eaaz5803.

Molina-Navarro, E., Andersen, H. E., Nielsen, A., Thodsen, H., & Trolle, D. (2018). Quantifying the combined effects of land use and climate changes on stream flow and nutrient loads: A modelling approach in the Odense Fjord catchment (Denmark). *Science of the Total Environment*, *621*, 253–264.

Mosavi, A., Ozturk, P., & Chau, K. W. (2018). Flood prediction using machine learning models: Literature review. *Water*, *10*, 1536.

Mouselimis, L., Sanderson, C., Curtin, R., Agrawal, S., Frey, B., & Dueck, D. (2019). *ClusterR: Gaussian mixture models, K-means, mini-batch-Kmeans, K-medoids and affinity propagation clustering*. R package version, 1.

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with python: A guide for data scientists*. O'Reilly Media, Incorporated.

Neff, M. W., & Corley, E. A. (2009). 35 years and 160,000 articles: A bibliometric exploration of the evolution of ecology. *Scientometrics*, *80*, 657–682.

Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., Lopez Garcia, A., Heredia, I., Malík, P., & Hluchý, L. (2019). Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey. *Artificial Intelligence Review*, *52*, 77–124.

Nunez-Mir, G. C., Iannone, B. V., Pijanowski, B. C., Kong, N. N., & Fei, S. L. (2016). Automated content analysis: Addressing the big literature challenge in ecology and evolution. *Methods in Ecology and Evolution*, *7*, 1262–1272.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*, 447–453.

Olson, R. S., La Cava, W., Orzechowski, P., Urbanowicz, R. J., & Moore, J. H. (2017). PMLB: A large benchmark suite for machine learning evaluation and comparison. *Biodata Mining*, *10*, 36.

Palacio-Niño, J.-O., & Berzal, F. (2019). Evaluation metrics for unsupervised learning algorithms. *arXiv preprint arXiv:1905.05667*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2019). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Porro, J., De Mulder, C., Amerlinck, Y., Torfs, E., Balemans, S., Weijers, S., Nopens, I., Rodriguez-Roda, I., & Comas, J. (2018). *Integrated artificial intelligence and mathematical modelling for online supervision and control of water resource recovery facilities. 6th IWA/WEF water resource recovery modelling seminar*.

Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*, e1301.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.

Qian, F., He, M. C., Song, Y. H., Tysklind, M., & Wu, J. Y. (2015). A bibliometric analysis of global research progress on pharmaceutical wastewater treatment during 1994–2013. *Environmental Earth Sciences*, *73*, 4995–5005.

Quaghebeur, W., Torfs, E., De Baets, B., & Nopens, I. (2022). Hybrid differential equations: Integrating mechanistic and data-driven techniques for modelling of water systems. *Water Research*, *213*, 118166.

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. isbn:3-900051-07-0.

R Core Team, & Worldwide Contributors. (2002). *The R stats package*. R Foundation for Statistical Computing. http://www.R-project.org

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Model-agnostic interpretability of machine learning*. arXiv preprint arXiv:1606.05386.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*, 503–520.

Royal Society. (2017). *Machine learning: The power and promise of computers that learn by example: An introduction*. Royal Society.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215.

Sarkar, A., Yang, Y., & Vihinen, M. (2020). Variation benchmark datasets: Update, criteria, quality and applications. *Database—The Journal of Biological Databases and Curation*.

Saunois, M., Stavert, A. R., Poulter, B., Bousquet, P., Canadell, J. G., Jackson, R. B., Raymond, P. A., Dlugokencky, E. J., Houweling, S., Patra, P. K., Ciais, P., Arora, V. K., Bastviken, D., Bergamaschi, P., Blake, D. R., Brailsford, G., Bruhwiler, L., Carlson, K. M., Carrol, M., … Zhuang, Q. (2020). The global methane budget 2000–2017. *Earth System Science Data*, *12*, 1561–1623.

Schimel, D. S. (1995). Terrestrial ecosystems and the carbon cycle. *Global Change Biology*, *1*, 77–91.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Scikit-learn. (2019). *Scikit-learn user guide*.

Scopus Elsevier. (2016). *Scopus content coverage guide*. Elsevier BV.

See, L. (2019). A review of citizen science and crowdsourcing in applications of pluvial flooding. *Frontiers in Earth Science*, *7*, 44.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310.

Singh, H., Gupta, M. M., Meitzler, T., Hou, Z.-G., Garg, K. K., Solo, A. M. G., & Zadeh, L. A. (2013). Real-life applications of fuzzy logic. *Advances in Fuzzy Systems*, *2013*, 581879.

Soulignac, F., Danis, P. A., Bouffard, D., Chanudet, V., Dambrine, E., Guenand, Y., Harmel, T., Ibelings, B. W., Trevisan, D., Uittenbogaard, R., & Anneville, O. (2018). Using 3D modeling and remote sensing capabilities for a better understanding of spatio-temporal heterogeneities of phytoplankton abundance in large lakes. *Journal of Great Lakes Research*, *44*, 756–764.

Stanley, E., Loken, L., Crawford, J., Casson, N., Oliver, S., Gries, C., & Christel, S. (2022). *A global database of methane concentrations and*

atmospheric fluxes for streams and rivers ver 5375843. Environmental Data Initiative.

Stanley, E. H., Casson, N. J., Christel, S. T., Crawford, J. T., Loken, L. C., & Oliver, S. K. (2016). The ecology of methane in streams and rivers: Patterns, controls, and global significance. *Ecological Monographs*, *86*, 146–171.

Trunk, G. V. (1979). A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*, 306–307.

United Nations, D.o.E. (2015). *The millennium development goals report 2015*. United Nations Publications.

van Vliet, M. T. H., Franssen, W. H. P., Yearsley, J. R., Ludwig, F., Haddeland, I., Lettenmaier, D. P., & Kabat, P. (2013). Global river discharge and water temperature under climate change. *Global Environmental Change-Human and Policy Dimensions*, *23*, 450–464.

Vannote, R. L., Minshall, G. W., Cummins, K. W., Sedell, J. R., & Cushing, C. E. (1980). The river continuum concept. *Canadian Journal of Fisheries and Aquatic Sciences*, *37*, 130–137.

Varshney, K. R., & Alemzadeh, H. (2017). On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big Data*, *5*, 246–255.

Vassakis, K., Petrakis, E., & Kopanakis, I. (2018). Big data analytics: Applications, prospects and challenges. In G. Skourletopoulos, G. Mastorakis, C. X. Mavromoustakis, C. Dobre, & E. Pallis (Eds.), *Mobile big data: A roadmap from models to technologies* (pp. 3–20). Springer International Publishing.

Vugteveen, P., Lenders, R., & Van den Besselaar, P. (2014). The dynamics of interdisciplinary research fields: The case of river research. *Scientometrics*, *100*, 73–96.

Wang, K., Wang, B., & Peng, L. (2009). CVAP: Validation for cluster analyses. *Data Science Journal*, *8*, 88–93.

Wang, L. M., Yuan, S. M., Ling, L., & Li, H. J. (2004). Improving the performance of decision tree: A hybrid approach. *Conceptual Modeling—Er 2004*, *Proceedings*, *3288*, 327–335.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236–244.

Wickham, H., & Grolemund, G. (2016). *R for data science: Import, tidy, transform, visualize, and model data*. O'Reilly Media.

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2020). Integrating physics-based modeling with machine learning: A survey. *arXiv preprint arXiv:2003.04919*, *1*, 1–34.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wojciechowski, J., Hopkins, A. M., & Upton, R. N. (2015). Interactive Pharmacometric applications using R and the shiny package. *CPT: Pharmacometrics & Systems Pharmacology*, *4*, 146–159.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*, 67–82.

Xu, Q. Z., Wang, N., Wang, L., Li, W., & Sun, Q. (2021). Multi-task optimization and multi-task evolutionary computation in the past five years: A brief review. *Mathematics*, *9*, 864.

Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management*, *36*, 1231–1247.

Zhang, C., & Ma, Y. (2012). *Ensemble machine learning: Methods and applications*. Springer.

Zhang, Y. F., Thorburn, P. J., & Fitch, P. (2019). Multi-task temporal convolutional network for predicting water quality sensor data. *Neural information processing (Iconip 2019)*, *Pt Iv*, *1142*, 122–130.

Zhao, L., Dai, T. J., Qiao, Z., Sun, P. Z., Hao, J. Y., & Yang, Y. K. (2020). Application of artificial intelligence to wastewater treatment: A bibliometric analysis and systematic review of technology, economy, management, and wastewater reuse. *Process Safety and Environmental Protection*, *133*, 169–182.

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: Principles and techniques for data scientists*. O'Reilly Media.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ho, L., & Goethals, P. (2022). Machine learning applications in river research: Trends, opportunities and challenges. *Methods in Ecology and Evolution*, *13*, 2603–2621. https://doi.org/10.1111/2041-210X.13992