Cambridge Prisms: Water

www.cambridge.org/wat

A call for a fundamental shift from model-centric to data-centric approaches in hydroinformatics

Babak Zolghadr-Asli[1,2], Ahmad Ferdowsi[3,4] and Dragan Savić[5,2]

[1]The Sustainable Minerals Institute (SMI), The University of Queensland, Brisbane, Australia; [2]Centre for Water Systems, University of Exeter, Exeter, UK; [3]Department of Water Engineering and Hydraulic Structures, Faculty of Civil Engineering, Semnan University, Semnan, Iran; [4]University of Applied Science and Technology, Tehran, Iran and [5]KWR Water Research Institute, 3430 PE Nieuwegein, The Netherlands

Perspective

Cite this article: Zolghadr-Asli B, Ferdowsi A and Savić D (2024). A call for a fundamental shift from model-centric to data-centric approaches in hydroinformatics. *Cambridge Prisms: Water*, **2**, e7, 1–4
https://doi.org/10.1017/wat.2024.5

Received: 23 December 2023
Revised: 20 February 2024
Accepted: 05 March 2024

Keywords:
hydroinformatics; computational intelligence; artificial intelligence; data-centric approach

Corresponding author:
Babak Zolghadr-Asli;
Email: bz267@exeter.ac.uk

Abstract

Over the years, data-driven models have gained notable traction in water and environmental engineering. The adoption of these cutting-edge frameworks is still in progress in the grand scheme of things, yet for the most part, such attempts have been centered around the models themselves, and their internal computational architecture, that is, the model-centric approach. These endeavors can certainly pave the way for more tailor-fitted models capable of producing accurate results. However, such a perspective often neglects a fundamental assumption of these models, which is the importance of reliability, correctness, and accessibility of the data used in constructing them. This challenge arises from the prevalent model-centric paradigm of thinking in the field. An alternative approach, however, would prioritize placing data at the focal point, focusing on systematically enhancing current datasets and devising frameworks to improve data collection schemes. This suggests a paradigm shift toward more data-centric thinking in water and environmental engineering. Practically, this shift is not without challenges and necessitates smarter data collection rather than an excessive one. Equally important is the ethical and accurate collection of data, making it available to everyone while safeguarding the rights of individuals and other legal entities involved in the process.

Impact statement

In the realm of water and environmental engineering, the data-driven models have gained a lot of traction over the years. While the adoption of these advanced frameworks is an ongoing process, the predominant focus has traditionally centered on refining the models themselves and their internal computational architecture –a perspective encapsulated by the model-centric approach. While these are quite fundamental in reaching a more profound understanding about what these models are capable of, they often overlook a fundamental tenet: The reliability, correctness, and accessibility of the data underpinning these models. An alternative approach, advocating for a paradigm shift, prioritizes elevating data to the forefront. Emphasizing the systematic enhancement of existing datasets and the formulation of frameworks to optimize data collection schemes, this perspective advocates a move toward a more data-centric paradigm in water and environmental engineering. However, this transformative shift is not without its challenges, requiring a nuanced strategy for *smart data collection*. Equally critical is the ethical and accurate handling of data, ensuring universal availability while upholding the rights of individuals and other legal entities involved in the process. This article underscores the significance of embracing a data-centric perspective, anticipating its far-reaching impact on shaping the future trajectory of water and environmental engineering practices.

Introduction

Data-driven frameworks, including machine-learning (ML) models, have emerged as a prominent focus and a topical subject in various engineering disciplines, notably in the realm of water and environmental engineering (Solomatine and Ostfeld, 2008; Giustolisi and Savic, 2009; Araghinejad, 2013). Whether it involves a more efficient optimization algorithm (e.g., Jalili et al., 2023; Wu et al., 2023), employing meticulous data mining methods (e.g., Aslam et al., 2022; Beig Zali et al., 2023; Zolghadr-Asli et al., 2023), developing sophisticated ML models (e.g., Ray et al., 2023; Sun et al., 2023), or, more recently, utilizing large-language models such as ChatGPT (e.g., Foroumandi et al., 2023; Halloran et al., 2023), the core premise of this sub-discipline, often referred to as *hydroinformatics* within the domain of water and hydrology-related science, lies in the potential of computational intelligence (CI) and, possibly, artificial intelligence (AI) to reshape the future of this

field (Makropoulos and Savić, 2019; Loucks, 2023). In essence, hydroinformatics can be viewed as a management philosophy enabled by (CI/AI) technology, and its primary objective is to establish a systematic approach to representing and comprehending the intricate and multidimensional phenomena prevalent in water management. On that note, it is often believed that these technologies hold the promise of offering alternative perspectives on existing challenges, enabling more efficient problem-solving, and devising economically and environmentally sustainable solutions. Some prime examples of this include leakage detection (e.g., Rajasekaran and Kothandaraman, 2024), elucidating the underlying causes of abnormal hydro-climatological behaviors (e.g., Zolghadr-Asli et al., 2023), facilitating a better understanding of the impacts of extreme events such as floods (Adnan et al., 2023), and predicting droughts (Piri et al., 2023), among others. This subject remains topical, and rapidly evolving, with numerous researchers continually exploring novel approaches to leverage the potential of these frameworks within the context of water-related sciences.

When it comes to water-related challenges, a brief overview of the most current and trending topics in hydroinformatics reveals a significant focus on adopting and fine-tuning sophisticated models (e.g., Bozorg-Haddad et al., 2018; Yaseen et al., 2019) and/or comparing the performance of these models (e.g., Chen et al., 2020; Yaghoubzadeh-Bavandpour et al., 2022), that is, the *model-centric approach.* In theory, these model-centric efforts have yielded promising results (e.g., Sun and Scanlon, 2019; Aliashrafi et al., 2021; Ghobadi and Kang, 2023). Often, such approaches place significant emphasis on the 'model' component within the CI/AI-based frameworks, primarily concentrating on improving or comparing such models. While this focus is commendable in itself and offers valuable insights, it tends to overlook another pivotal element – the 'data.' This dichotomy gives rise to two distinct schools of thought regarding the perception and utilization of hydroinformatics. One approach is predominantly oriented toward the role and structure of models (i.e., models-centric), while an alternative perspective is mostly geared toward the data side of the equation (i.e., data-centric). This paper aimed to delve into the variations between these two schools of thought and argue for the long-term implications of an overreliance on model-centric approaches. Importantly, we explore how the alternative, or perhaps complementary, viewpoint of a data-centric approach can reshape the current paradigm of utilizing CI/AI-based frameworks in the context of water-related sciences.

Model-centric vs. data-centric paradigms

The widespread accessibility of computing power, particularly of cloud computing resources, has led to a substantial increase in the deployment of CI/AI-based models, garnering recognition for their efficacy across various domains. These models have demonstrated noteworthy advantages, featuring significantly reduced computation times and proving effective in addressing real-world challenges. Their applications span diverse fields, ranging from medicine (e.g., Rajpurkar et al., 2022) and economics (e.g., Qian et al., 2023) to water-related issues (e.g., Ray et al., 2023). Broadly speaking, one prevailing paradigm emphasizes the model-centric approach, placing a paramount focus on the model aspect of the equation. One of the foundational assumptions underpinning studies that are geared toward the model-centric paradigm is the reliability, correctness, and accessibility of the data used to construct data-driven models. While it can be argued that this assumption has been implicit in all models, including conceptual and physics-based ones, data-driven models take this reliance to a heightened level, where the model's configuration (i.e., structure and parametrization) and overall performance can significantly vary with different datasets (e.g., Beig Zali et al., 2023; Liu et al., 2024). This in-built adaptability of data-driven models is not inherently problematic in and of itself, but it raises a more profound question regarding the significance of data availability and data quality. Ultimately, it is essential to note that these models are only as reliable and effective as the data they are fed. Furthermore, their application beyond the confines of research papers depends heavily on the existence of reliable and factual datasets, which, more often than not, are lacking in most practical cases (Li et al., 2023).

The solution may seem straightforward – investing in collecting and preparing more reliable and comprehensive datasets, that is, a *data-centric approach* (DeepLearningAI, 2021; Liu et al., 2023). The primary distinction between these model-centric and data-centric paradigms lies not in the models themselves but in their perceived role. The model-centric approach seeks to leverage the computational structures of models to generate more accurate and applicable outcomes. In contrast, the data-centric paradigm emphasizes the crucial role of data in obtaining reliable results from such models.

In contrast to the model-centric paradigm, data-centric approaches emphasize the entire data value chain (e.g., data acquisition, analysis, curation, and storage) independently of its application. This allows for leveraging more information from existing datasets and promotes efficiency in expanding such datasets. Consequently, this paradigm prioritizes the data value chain, promoting the efficiency in the use and re-use of datasets. Here, the focus is not on modifying the model's internal architecture to produce general results but rather on systematically producing and altering datasets and data collection procedures to enhance the overall performance of the models, aiming for accurate and meaningful outcomes. The essence of this paradigm is to facilitate the establishment of a reliable and comprehensive dataset. It advocates for consistent and accurate data collection, coupled with a robust data quality-monitoring scheme tailored to the specific problem at hand. Table 1 summarizes the advantages and disadvantages of model-centric and data-centric paradigms.

Table 1. Comparison of data-centric and model-centric paradigms

	Data-centric	Model-centric
Advantages	Greater robustness in results can be attained in comparison to the model–centric approach. It is more straightforward to interpret the influential features or components of data on the results of the model.	The enhanced/proposed model exhibits applicability to alternative datasets. It is more straightforward to implement the models due to readily available code repositories.
Challenges	Data privacy Data scarcity No universal or ad hoc guidelines	Data privacy Model selection Model parameter tuning It is challenging to intuitively interpret the effects of model parameters on the results.

https://doi.org/10.1017/wat.2024.5 Published online by Cambridge University Press

The central premise of the data-centric paradigm within the context of water and environmental engineering seems easily obtainable. However, the practical implementation of this idea is far more challenging (Larsen et al., 2019; Pandeya et al., 2021). Both public and private water and environmental management organizations often face budgetary constraints that hinder their ability to create or acquire such datasets for their projects. This limitation stems from the fact that these endeavors do not immediately translate into revenue generation. The primary objective of prioritizing enhanced data is to establish more robust and dependable models. In the industry, unfortunately, it is often seen that investing in these datasets faces resistance, particularly in smaller organizations, owing to substantial cost and legal implications. In addition to these, larger organizations may also show hesitance due to potential public relations issues that could arise down the road. It is worth noting that real-world data tend to suffer from quality issues and undesirable flaws, such as missing values, erroneous readings, incorrect labels, and anomalies (Zha et al., 2023). Improvement of existing datasets and the adoption of data-centric approaches represent a paradigm shift from model design to data quality and reliability.

Another fundamental pillar of data-centric thinking is to move toward smarter data collection rather than an excessive one. Clearly, collecting data can be financially burdensome, and as demonstrated earlier, not without its challenges. Collecting excessive data without a clear idea of their use is arguably more harmful than having fewer data, as this approach drains financial resources that could have otherwise been directed toward better use. Overemphasis on collecting potentially irrelevant data can mislead the modeler and overwhelm the model. Other challenges with using data in data-driven models, for example, unjustified splitting of data into training, validation, and testing of models, indicate the need for educating modelers at the boundary of hydroinformatics, science, and engineering (Wagener et al., 2021). The reason for training individuals who are well-versed in both computer science and a targeted discipline, such as water and environmental engineering, as opposed to pure statisticians and applied mathematicians, is to provide the former group with a more in-depth understanding of the subtleties and nuances of the discipline. This insider knowledge enables them to adopt the most suitable computational model for a given problem. This emphasis on the data itself, characteristic of the data-centric paradigm, rewards investments in the underlying structure of the data over the architecture of the models.

As a final note on this topic, one should remember that while these two paradigms offer opposing viewpoints on leveraging CI/AI-based modeling, it is imperative to recognize their non-mutually exclusive nature, refraining from undermining one another. The fundamental premise is that an accurate, representative, and comprehensive dataset is indispensable for capturing the underlying structure of a phenomenon – a focal point of the data-centric paradigm. Nevertheless, the utility of such data is significantly enhanced when coupled with a reliable model, aligning with the objectives of the model-centric approach. In this context, it is important to emphasize that a sophisticated model does not obviate the need for a thorough and clean dataset. Similarly, focusing on high-quality data does not exempt the necessity of providing a reliable and robust model. In essence, the synergistic interplay between a capable model and a comprehensive dataset is vital to achieve reliable results. Therefore, the optimal perspective on these two paradigms is to appreciate their potential for complementarity, forming a synergistic framework where insights from one paradigm inform and enhance the other, thereby fostering the development of more robust strategies in the context of water and environmental engineering.

Concluding remarks

Due to the rapid development of AI/ML tools (e.g., Large-language models such as ChatGPT), the future of data-driven models, notably ML models, remains uncertain but is extremely exciting. Regardless of the outcomes, it is crucial to shift the perception among engineering professionals and scholars to emphasize the pivotal role of reliable datasets in the broader water industry. The paradigm shifts tend to spotlight the data rather than the models, highlighting the benefit of investing in improving our current datasets and systematically enhancing the data value chain, as opposed to trying to arbitrarily tamper with the model's architecture to achieve marginal improvements. This should not undermine the benefits of a more capable model; instead, it underscores the idea that a model is only as good and reliable as its input data. Meanwhile, it is equally vital to ensure that the data is collected intelligently, ethically, and accurately, is available to everyone, and safeguards the rights of individuals and other legal entities involved in the process. This is all also addressed by the objectives of the FAIR (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable) and SQUARE (**S**upporting, **QU**ality, **A**ction, and **RE**search) data principles (Cudennec et al., 2020). Achieving these objectives may necessitate new legislative initiatives and increased investments from the public sector to establish the necessary framework for responsible data collection. Considering the current and future landscape of this field, one can anticipate increased investment, not only from the academic sector but also from the water industry, in furthering data-centric approaches. Additionally, it is hopeful that both public and private companies will increasingly invest in smart data collection and monitoring protocols to ensure that data is not only reliable, but also repetitive, accurate, and readily available to relevant consumers.

Open peer review. To view the open peer review materials for this article, please visit http://doi.org/10.1017/wat.2024.5.

Data availability statement. All used data have been presented in the paper.

Author contribution. All authors have contributed equally to the conceptualization of the paper.

Financial support. **Dragan Savic** has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. (951424)).

Competing interest. The authors have no relevant financial or non-financial interests to disclose.

References

Adnan MSG, Siam ZS, Kabir I, Kabir Z, Ahmed MR, Hassan QK, Rahman RM and Dewan A (2023) A novel framework for addressing uncertainties in machine learning-based geospatial approaches for flood prediction. *Journal of Environmental Management* **326**, 116813.

Aliashrafi A, Zhang Y, Groenewegen H and Peleato NM (2021) A review of data-driven modelling in drinking water treatment. *Reviews in Environmental Science and Bio/Technology* **20**, 985–1009.

Araghinejad S (2013) *Data-Driven Modeling: Using MATLAB® in Water Resources and Environmental Engineering*, Vol. **67**. Dordrecht: Springer Science & Business Media.

Aslam B, **Maqsoom A**, **Cheema AH**, **Ullah F**, **Alharbi A and Imran M** (2022) Water quality management using hybrid machine learning and data mining algorithms: An indexing approach. *IEEE Access* **10**, 119692–119705.

Beig Zali R, **Latifi M**, **Javadi AA and Farmani R** (2023) Semisupervised clustering approach for pipe failure prediction with imbalanced data set. *Journal of Water Resources Planning and Management* **150**(2), 04023078.

Bozorg-Haddad O, **Latifi M**, **Bozorgi A**, **Rajabi MM**, **Naeeni ST and Loáiciga HA** (2018) Development and application of the anarchic society algorithm (ASO) to the optimal operation of water distribution networks. *Water Science and Technology: Water Supply* **18**(1), 318–332.

Chen K, **Chen H**, **Zhou C**, **Huang Y**, **Qi X**, **Shen R**, **Liu F**, **Zuo M**, **Zou X**, **Wang J**, **Zhang Y**, **Chen D**, **Chen X**, **Deng Y and Ren H** (2020) Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Research* **171**, 115454.

Cudennec C, **Lins H**, **Uhlenbrook S and Arheimer B** (2020) Editorial–towards FAIR and SQUARE hydrological data. *Hydrological Sciences Journal* **65**(5), 681–682.

DeepLearningAI and Landing AI (2021) Data-Centric AI Competition. Available at https://https-deeplearning-ai.github.io/data-centric-comp/ (last accessed 7 November 2023).

Foroumandi E, **Moradkhani H**, **Sanchez-Vila X**, **Singha K**, **Castelletti A and Destouni G** (2023) ChatGPT in hydrology and earth sciences: Opportunities, prospects, and concerns. *Water Resources Research* **59**(10), e2023WR036288.

Ghobadi F and Kang D (2023) Application of machine learning in water resources management: A systematic literature review. *Water* **15**(4), 620.

Giustolisi O and Savic DA (2009) Advances in data-driven analyses and modelling using EPR-MOGA. *Journal of Hydroinformatics* **11**(3–4), 225–236.

Halloran LJ, **Mhanna S and Brunner P** (2023) AI tools such as ChatGPT will disrupt hydrology, too. *Hydrological Processes* **37**(3), e14843.

Jalili AA, **Najarchi M**, **Shabanlou S and Jafarinia R** (2023) Multi-objective optimization of water resources in real time based on integration of NSGA-II and support vector machines. *Environmental Science and Pollution Research* **30**(6), 16464–16475.

Larsen MAD, **Petrovic S**, **Engström RE**, **Drews M**, **Liersch S**, **Karlsson KB and Howells M** (2019) Challenges of data availability: Analysing the water-energy nexus in electricity generation. *Energy Strategy Reviews* **26**, 100426.

Li C, **Sun SC**, **Wei Z**, **Tsourdos A and Guo W** (2023) Scarce data driven deep learning of drones via generalized data distribution space. *Neural Computing and Applications* **35**(20), 15095–15108.

Liu G, **Savic D and Fu G** (2023) Short-term water demand forecasting using data-centric machine learning approaches. *Journal of Hydroinformatics* **25** (3), 895–911.

Liu Z, **Zhou J**, **Yang X**, **Zhao Z and Lv Y** (2024) Research on water resource modeling based on machine learning technologies. *Water* **16**(3), 472.

Loucks DP (2023) Hydroinformatics: A review and future outlook. *Cambridge Prisms: Water* **1**, 1–26. https://doi.org/10.1017/wat.2023.10.pr3.

Makropoulos C and Savić DA (2019) Urban hydroinformatics: Past, present and future. *Water* **11**(10), 1959.

Pandeya B, **Buytaert W and Potter C** (2021) Designing citizen science for water and ecosystem services management in data-poor regions: Challenges and opportunities. *Current Research in Environmental Sustainability* **3**, 100059.

Piri J, **Abdolahipour M and Keshtegar B** (2023) Advanced machine learning model for prediction of drought indices using hybrid SVR-RSM. *Water Resources Management* **37**(2), 683–712.

Qian Y, **Liu J**, **Shi L**, **Forrest JYL and Yang Z** (2023) Can artificial intelligence improve green economic growth? Evidence from China. *Environmental Science and Pollution Research* **30**(6), 16418–16437.

Rajasekaran U and Kothandaraman M (2024) A survey and study of signal and data-driven approaches for pipeline leak detection and localization. *Journal of Pipeline Systems Engineering and Practice* **15**(2), 03124001.

Rajpurkar P, **Chen E**, **Banerjee O and Topol EJ** (2022) AI in health and medicine. *Nature Medicine* **28**(1), 31–38.

Ray SS, **Verma RK**, **Singh A**, **Ganesapillai M and Kwon YN** (2023) A holistic review on how artificial intelligence has redefined water treatment and seawater desalination processes. *Desalination* **546**, 116221.

Solomatine DP and Ostfeld A (2008) Data-driven modelling: Some past experiences and new approaches. *Journal of Hydroinformatics* **10**(1), 3–22.

Sun AY and Scanlon BR (2019) How can big data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. *Environmental Research Letters* **14**(7), 073001.

Sun L, **Zhu J**, **Tan J**, **Li X**, **Li R**, **Deng H**, **Zhang X**, **Liu B and Zhu X** (2023) Deep learning-assisted automated sewage pipe defect detection for urban water environment management. *Science of the Total Environment* **882**, 163562.

Wagener T, **Savic D**, **Butler D**, **Ahmadian R**, **Arnot T**, **Dawes J**, **Djordjevic S**, **Falconer R**, **Farmani R**, **Ford D and Hofman J** (2021) Hydroinformatics education–the water informatics in science and engineering (WISE) centre for doctoral training. *Hydrology and Earth System Sciences* **25**(5), 2721–2738.

Wu J, **Wang Z**, **Hu Y**, **Tao S and Dong J** (2023) Runoff forecasting using convolutional neural networks and optimized bi-directional long short-term memory. *Water Resources Management* **37**(2), 937–953.

Yaghoubzadeh-Bavandpour A, **Bozorg-Haddad O**, **Rajabi M**, **Zolghadr-Asli B and Chu X** (2022) Application of swarm intelligence and evolutionary computation algorithms for optimal reservoir operation. *Water Resources Management* **36**(7), 2275–2292.

Yaseen ZM, **Sulaiman SO**, **Deo RC and Chau KW** (2019) An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *Journal of Hydrology* **569**, 387–408.

Zha D, **Bhat ZP**, **Lai KH**, **Yang F**, **Jiang Z**, **Zhong S and Hu X** (2023) Data-centric artificial intelligence: A survey. Preprint, arXiv:2303.10158.

Zolghadr-Asli B, **Naghdyzadegan Jahromi M**, **Wan X**, **Enayati M**, **Naghdizadegan Jahromi M**, **Tahmasebi Nasab M**, **Tiefenbacher JP and Pourghasemi HR** (2023) Uncovering the depletion patterns of inland water bodies via remote sensing. *Data Mining, and Statistical Analysis. Water* **15**(8), 1508.