Water Science & Technology

© 2022 The Authors Water Science & Technology Vol 85 No 9, 2503 doi: 10.2166/wst.2022.115

Hybrid modelling of water resource recovery facilities: status and opportunities

Mariane Yvonne Schneider[a,*], Ward Quaghebeur[b,c,d], Sina Borzooei[b,c], Andreas Froemelt[e], Feiyi Li[f], Ramesh Saagi[g], Matthew J. Wade[h], Jun-Jie Zhu[i] and Elena Torfs[b,c]

[a] Next Generation Artificial Intelligence Research Center & School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[b] Centre for Advanced Process Technology for Urban Resource recovery (CAPTURE), Frieda Saeysstraat 1, Gent 9000, Belgium
[c] BIOMATH, Department of Data Analysis and Mathematical Modelling, Ghent University, Coupure Links 653, Ghent 9000, Belgium
[d] KERMIT, Department of Data Analysis and Mathematical Modelling, Ghent University, Coupure Links 653, Ghent 9000, Belgium
[e] Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf 8600, Switzerland
[f] modelEAU, CentrEau, Département de génie civil et de génie des eaux, Pavillon Adrien-Pouliot, Université Laval, Quebec City, Canada
[g] Division of Industrial Electrical Engineering and Automation (IEA), Department of Biomedical Engineering, Lund University, P.O. Box 118, Lund SE-22100, Sweden
[h] School of Engineering, Newcastle University, Newcastle-upon-Tyne NE1 7RU, UK
[i] Department of Civil and Environmental Engineering and Andlinger Center for Energy and the Environment, Princeton University, Princeton, NJ 08544, USA
*Corresponding author. E-mail: myschneider@isi.imi.i.u-tokyo.ac.jp

MYS, 0000-0003-3397-2773; WQ, 0000-0002-6162-2124; SB, 0000-0002-0694-3064; AF, 0000-0001-9388-7816; FL, 0000-0003-4278-730X; RS, 0000-0003-4373-2562; MJW, 0000-0001-9824-7121; J-JZ, 0000-0002-7546-2870; ET, 0000-0002-5629-6950

ABSTRACT

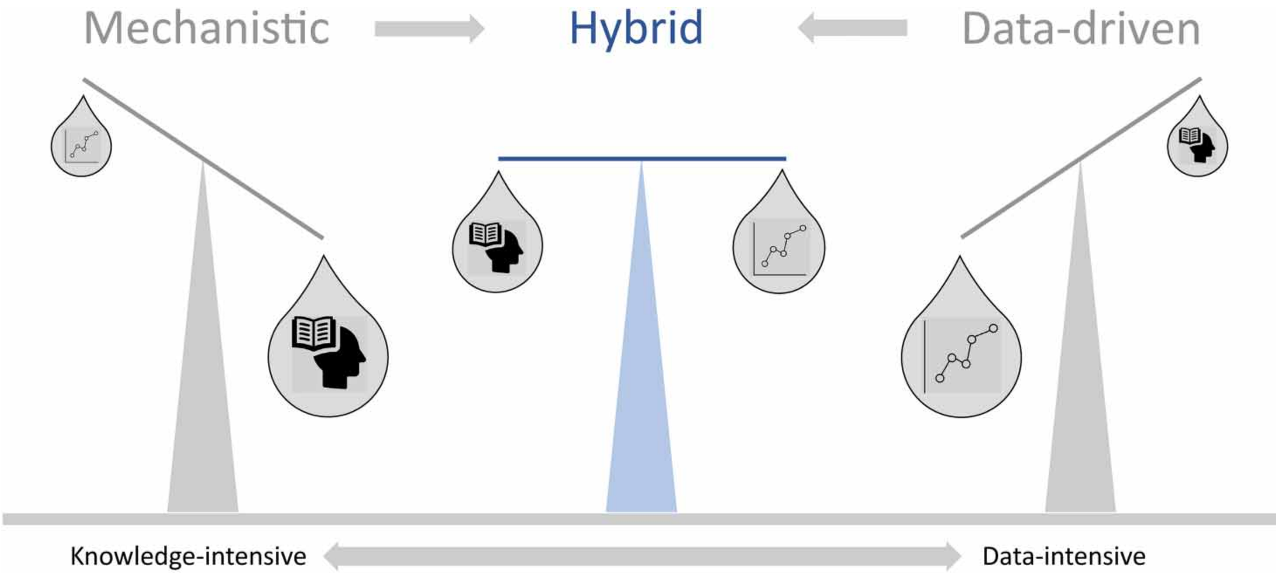
Mathematical modelling is an indispensable tool to support water resource recovery facility (WRRF) operators and engineers with the ambition of creating a truly circular economy and assuring a sustainable future. Despite the successful application of mechanistic models in the water sector, they show some important limitations and do not fully profit from the increasing digitalisation of systems and processes. Recent advances in data-driven methods have provided options for harnessing the power of Industry 4.0, but they are often limited by the lack of interpretability and extrapolation capabilities. Hybrid modelling (HM) combines these two modelling paradigms and aims to leverage both the rapidly increasing volumes of data collected, as well as the continued pursuit of greater process understanding. Despite the potential of HM in a sector that is undergoing a significant digital and cultural transformation, the application of hybrid models remains vague. This article presents an overview of HM methodologies applied to WRRFs and aims to stimulate the wider adoption and development of HM. We also highlight challenges and research needs for HM design and architecture, good modelling practice, data assurance, and software compatibility. HM is a paradigm for WRRF modelling to transition towards a more resource-efficient, resilient, and sustainable future.

Key words: data-driven model, hybrid model, mechanistic model, process control, urban water management, wastewater

HIGHLIGHTS

- HM combines mechanistic and data-driven models.
- WRRFs are too sensitive to rely only on data-driven techniques; HM is an alternative.
- Hybrid models have a high predictive power at low computational cost, proving them useful for online optimisation and control.
- Major challenges for more widespread implementation of HM are discussed.
- HM supports the transition towards a digital and resource-efficient water sector.

GRAPHICAL ABSTRACT



ABBREVIATIONS

ADM	anaerobic digestion model
AI	artificial intelligence
ANN	artificial neural networks
Anammox	anaerobic ammonium oxidation
AOB	ammonia-oxidising bacteria
API	application programming interface
ARIMA	autoregressive integrated moving average
ASM	activated sludge model
BSM	benchmark simulation model
CFD	computational fluid dynamics
CSTR	continuously stirred tank reactor
DAE	differential-algebraic equation
GLUE	generalised likelihood uncertainty estimation
GMP	good modelling practice
HM	hybrid modelling
kNN	k-nearest neighbours
LSTM	long short-term memory
ML	machine learning
MLP	multilayer perceptrons
MPC	model predictive control
NLP	natural language processing
NOB	nitrite-oxidising bacteria
ODE	ordinary differential equation
PCA	principal component analysis
PDE	partial differential equation
PLS	partial least square
RBF	radial basis functions
SVM	support vector machine
WRRF	water resources recovery facility
WRRmod	water resource recovery modelling
WTP	water treatment plant

1. INTRODUCTION AND MOTIVATION

Urban water management has a critical function for maintaining human and environmental health by ensuring potable water quality and treated wastewater. This goal can be reached by operating within defined safety margins or through adequate monitoring, control, and interventions, or a combination of the two. Conservative design and operation, employing high safety margins, typically lead to oversized plants and energy-intensive processes, resulting in increased costs and resource consumption. Such approaches can improve reliability or resilience for future changes. However, urban water management increasingly and critically faces systemic challenges that cannot be ignored, such as resource shortages (e.g., clean water, energy, nutrients), population growth, and climate change. Therefore, new approaches that support activities to address these challenges and improve the sustainability and resilience of urban water management systems (Butler *et al.* 2017) are critical research needs. Here, we focus on the considerable utility of mathematical modelling, specifically the potential of hybrid models.

Water resources recovery facilities (WRRFs) are an important part of urban water management and have complex, partially observable processes that require monitoring and control to ensure stable, safe, and appropriate operation (Olsson & Newell 1999). Modelling biological, chemical, and physical processes is useful to acquire process understanding, simulate and test control strategies, predict future behaviour under changing conditions (Gernaey *et al.* 2004), or model-predictive control (Lund *et al.* 2018). In recent decades, mechanistic approaches to model wastewater treatment processes have been favoured by engineers due to the first-principles approach well established across many industries (Henze *et al.* 2000; Newhart *et al.* 2019). These mechanistic models are based on process knowledge. These models usually simplify the complex processes they represent, such as aeration and mixing in activated sludge plants or aggregation of particulates. A concrete example is the diversity of nitrifying bacteria known to function in these systems (e.g., Daims 2014), which are typically summarised in models as nitrifying bacteria or differentiated into a few species, i.e., ammonia-oxidising bacteria (AOB) and nitrite-oxidising bacteria (NOB). Furthermore, and especially with greater model complexity, parameterising, calibrating, and validating mechanistic models usually require laborious experiments that may not always be possible due to economic, time, or measurability constraints (Vanrolleghem *et al.* 2005). While activated sludge models (ASMs) (Henze *et al.* 2000) are widely accepted in both research and practice, modelling several underlying processes (e.g., nitrous oxide production, phosphorus removal processes) remains incomplete (Regmi *et al.* 2019). Similarly, a standard and widely accepted modelling framework for several novel treatment processes (e.g., anaerobic ammonium oxidation (Anammox) processes (Baeten *et al.* 2019), membrane treatment (Mannina *et al.* 2011)) is not yet available. Such modelling bottlenecks can result in hindered innovation and rapid system implementation.

While data-driven methods have always been part of a modeller's toolbox, they are currently receiving greater interest, motivated by cheaper and smarter sensor technologies that enable ubiquitous data collection as well as online and real-time monitoring (Vanrolleghem & Lee 2003; Corominas *et al.* 2018). This has resulted in an increasing amount of high-resolution data, which together with available computational storage, cheap, edge computation, modern database technology, data processing (e.g., Hadoop (Apache Software Foundation n.d.)), and transmission capacity maximise the extraction of useful information available to plant operators (Wade 2004; Samuelsson *et al.* 2021). The need for novel and improved methods in combination with greater opportunities for data-driven approaches has induced a significant body of literature discussing the use of such methods in the water processing industry (Wade *et al.* 2021). These methods are particularly successful on problems involving large and high-quality data sets (Dobbelaere *et al.* 2021). However, data quality might be a barrier, especially in the wastewater treatment sector, where sensors are exposed to harsh conditions (e.g., Cecconi *et al.* 2020; Samuelsson *et al.* 2021). An additional issue of the data-driven approach is a lack of mechanistic-based interpretability of the result and limitations in extrapolation power (Newhart *et al.* 2019).

In this article, one solution to bring forth the advantages of both mechanistic and data-driven models and minimise their deficits is hybrid modelling (HM). HM is a combination of mechanistic and data-driven modelling. The mechanistic models ensure the preservation of the process knowledge and extrapolation capabilities of the model. The data-driven models offer improved predictive capabilities that can alleviate the weaknesses in mechanistic models, for example, by learning unknown, hidden relationships.

The water sector's transition towards a circular economy has created a shift in modelling objectives. For example, more case-specific quality targets for different water reuse purposes might be defined, which require tailor-made monitoring and modelling (Reynaert *et al.* 2021). Simultaneously, the digitalisation of the water sector, characterised by developments in sensor and IT technology, is creating vast amounts of data and, as such, the need for increased computational capacity and storage. This provides both more input data for modelling but also allows models to be evaluated at finer timescales, e.g., real-time or near real-time, which is particularly beneficial for improving process automation and control. These combined benefits create an environment for hybrid models to be used to their full potential by simultaneously harnessing and integrating the available data, computational power, and enhanced information on the studied processes. As such, the application of HM in the domain of WRRFs has the potential to foster automation (Rodriguez-Roda *et al.* 2002), increase efficiency, and increase the predictive power of models (von Stosch *et al.* 2014). Furthermore, we argue that HM helps keep up with new technologies and processes by being, for example, a preceding step in developing a mechanistic model prior to acquiring sufficient process understanding. However, uptake of hybrid models in the WRRF modelling community is hampered by the absence of expertise, trust in the novel (i.e. data-driven) approaches, and documented methods for the development and robust implementation of HM.

To better coordinate advances of HM in the WRRF sector, a seminar on HM was organized at the specialist WRR modelling (WRRmod) 2021 conference in Arosa, Switzerland. As an output from this, we provide a comprehensive overview of HM philosophy and architecture. We give examples of successful implementations in the WRRF sector and discuss challenges, opportunities, and development needs for hybrid models to develop and be integrated by the industry as a powerful framework for WRRF applications.

In this article, we use the terminology mechanistic, data-driven, and hybrid modelling to refer to the different modelling paradigms. This deviates from previous work, e.g., von Stosch *et al.* (2014), who refer to mechanistic models as parametric since their parameters have a physical meaning. This is not the case in data-driven models, hence their designation as non-parametric. Hybrid models, falling somewhere on the spectrum between the two approaches, are referred to as semi-parametric. However, this usage can be confusing. In statistics, parametric is defined as *having a finite number of parameters.* Thus, both mechanistic models and certain types of data-driven models (e.g., artificial neural networks (ANNs)) fall within this definition. In contrast, other types of data-driven models (e.g., k-nearest neighbours (kNNs) or support vector machine (SVM)) are considered non-parametric. To limit confusion, we do not use the parametric terminology in this manuscript but refer to von Stosch *et al.* (2014) as an appropriate review of HM on the broader process industry.

2. BACKGROUND ON HM

Following Schubert *et al.* (1994), we define two types of knowledge: mechanistic knowledge and knowledge learned from any patterns encapsulated within data. To enable a deep discussion of the intricacies of hybrid models, it is necessary first to discuss two modelling paradigms based on these types of knowledge: mechanistic and data-driven models.

Mechanistic models are models based on foundational knowledge. Mechanistic knowledge provides a representative description of a system or process derived from a fundamental understanding of its underlying mechanisms based on physical or chemical laws or causal relations. This type of knowledge is then encapsulated in a set of mathematical functions, whose behaviour can be described by one or more mathematical equations. We suggest that mechanistic models comprise elements of first-principles (i.e. the foundational understanding of a process or system) and phenomenological (i.e. describing empirical relationships of process/system components) models. First-principles models are completely based on natural laws without relying on data, whereas phenomenological or empirical models are structured according to domain knowledge, but data determine their parameters. Examples of the latter are Monod kinetics in ASM (Henze *et al.* 2000) or settling velocity functions in settler models (Takács *et al.* 1991).

Simple algebraic equations can often adequately describe the system. However, for more complex systems, the underlying dynamics are often explicitly described using ordinary differential equations (ODEs), differential-algebraic equations (DAEs) or partial differential equations (PDEs). In addition, there is a wide range of alternatives for equation-based models, including agent-based models, population balance models, and equation-free modelling (DeAngelis & Yurek 2015). However, these are not commonly applied as models for water and resource recovery applications, so we focus the discussion on models based on algebraic and differential equations.

However, process knowledge is often limited and incomplete, where the model does not consider latent or unobservable mechanisms. This typically results in only a partial description of the process or system, which may have value but often is not robust or resilient to significant changes.

Data-driven models derive knowledge from patterns encapsulated by measured data and metadata. In contrast to mechanistic models, no domain knowledge is assumed. It is important to distinguish between data-driven models and empirically derived phenomenological models. The former does not assume any structure in the data; instead, it learns the structure from the data itself. At the same time, the latter defines a structure from current domain knowledge, determining the values of its parameters from data. This inclusion of domain knowledge classifies them as mechanistic.

Examples of frequently used data-driven methods are Gaussian processes, polynomials, multivariate adaptive regression spline, partial least squares (PLS), principal component analysis (PCA), SVM, decision tree-based methods, and ANNs. In the context of machine learning (ML), data-driven methods and models can be categorised based on their learning algorithms, i.e. supervised, unsupervised, and reinforcement learning. For methods based on supervised learning, including classification and regression algorithms, the mapping function is learned from input and output data and their associations (Razavi 2021). For unsupervised learning problems, such as clustering and association, a model learns from unlabelled input data and discovers common aspects in unknown patterns (Abrahart *et al.* 2008; Zhu *et al.* 2018; Borzooei *et al.* 2020). Reinforcement learning (Sutton & Barto 2018), where the model or agent learns using trial and error or through a reward-punishment process, has found favour more recently as a more knowledge-based and adaptive method for plant control (Hernández-del-Olmo *et al.* 2012; Pang *et al.* 2019; Lazăr *et al.* 2020). We recommend, for example, Mitchell (1997) for a more comprehensive discussion of data-driven methods.

As no domain knowledge is included, the success of the data-driven approach depends heavily on the quality and quantity of the data and relies on an appropriate choice of method for extraction. Moreover, predictions are not designed to obtain a causal relation. Therefore, the interpretability of data-driven models is a critical bottleneck.

HM emerged to combine the use of the two aforementioned modelling paradigms (Figure 1). The first research to combine these paradigms can be traced back to the early 1990s (compare section 3, text mining). Early hybrid models consisted of a combination of either ANNs or radial basis functions (RBFs) with mathematical expressions describing mass and energy balances, respectively data-driven and mechanistic components (Johansen & Foss 1992; Kramer *et al.* 1992; Psichogios & Ungar 1992; Su *et al.* 1993). As hybrid models were simultaneously developed in different research fields, the terminology used differed; The de facto name hybrid model was initially used in the field of chemical engineering. Other terms for HM include but are not limited to: (i) Hybrid semi-parametric modelling (e.g., von Stosch *et al.* 2014), mainly to emphasise the use of both parametric (i.e. mechanistic) and non-parametric (i.e. data-driven) models; (ii) Knowledge-based HM, highlighting the fact

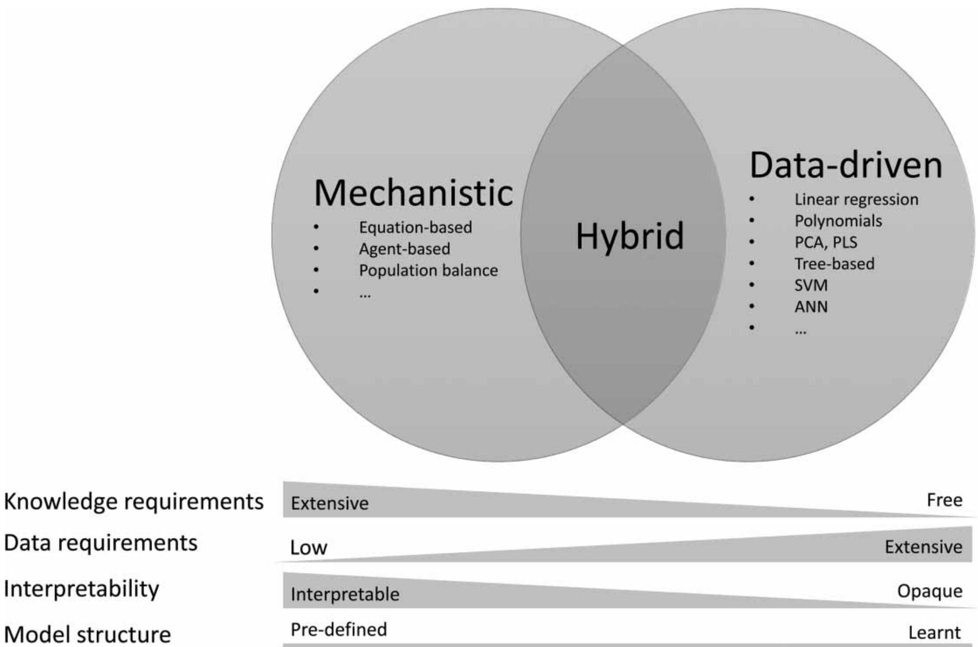


Figure 1 | Venn diagram of hybrid modelling, schematically showing the requirements of expert knowledge and data, the interpretability, and the model structure.

that all knowledge available (data, information, and domain expert knowledge) is incorporated into the model; (iii) Integrated neural network is used for an approach similar to HM (Su *et al.* 1993). (iv) Grey-box modelling, where the terminology arose, reflects a mixture of the black- and white-box concepts (Tulleken 1993; Van Can *et al.* 1996). Black comes from the black-box model to indicate that the exact relation of the mapping between input and output is often difficult to explain and does not necessarily have physical interpretability. In contrast, a white-box model corresponds to a mechanistic model, where the relationships themselves are defined by modelling experts and therefore assumed to be known. The black-box model is often associated with the data-driven approach; however, this use is controversial, both in the word's negative connotations and the confusing terminology, e.g., with linear regression or decision trees, the mapping is transparent, apropos our suggested terminology. Therefore, we refrain from using this terminology in this manuscript.

The original intention of hybrid models was to impute information absent from mechanistic models using available (i.e. measurable) data (Psichogios & Ungar 1992). The domain knowledge used to construct a mechanistic model is often incomplete and subject to several simplifying assumptions. The inverse motivation, i.e. to integrate mechanistic components into data-driven models, is for example, present in climate research (Karpatne *et al.* 2017a). Here, techniques such as theory-guided data science and physics-guided or physics-informed neural networks are built on high data volumes and subjected to certain physical constraints (Karpatne *et al.* 2017b; Raissi *et al*. 2017, 2019; Luo & Bao 2018).

Central to a hybrid model is using a data-driven component that does not assume a structure. Solely using a phenomenological or empirical component within a mechanistic model does not make it a hybrid model, as this component still assumes a structure derived from domain knowledge (e.g., the usage of an empirical Monod function in the ASM model (Henze *et al.* 2000)). Different architectures have been investigated for combining mechanistic and data-driven components with hybrid models. We distinguish between a strict and a loose definition for HM. The strict definition requires that both sub-models must be present in the final hybrid model. Two architectures, serial and parallel (von Stosch *et al.* 2014), fit this description. When loosening this definition by requiring both paradigms to be used during model development, other paradigms such as surrogate modelling (Razavi *et al.* 2012), also known as emulation, fit under the HM umbrella. The latter is specifically useful when the function of existing models is constrained by the computational ability or observability of the outputs (Oyebamiji *et al.* 2019). Our recommendations apply to the loose definition as they are addressed towards any expert using both approaches. Figure 2 schematically displays the three model architectures.

The first architecture adhering to the strict definition, a **serial** hybrid model, is arguably the most used. In this architecture, the output of one model is used as input to the other. Usually, the output of the data-driven model is used as input of the mechanistic model (e.g., Côté *et al.* 1995). The mechanistic component incorporates the known dynamics, usually chemical and physical phenomena, such as mass and energy balances. The data-driven component outputs dynamics that are not explicitly modelled by the mechanistic component, e.g., ill-defined or poorly understood reaction kinetics. An early example is a work done by Psichogios & Ungar (1992), where a serial hybrid model showed higher accuracy than either a mechanistic or data-driven model alone. Less often used is the inverse approach, where the output of a mechanistic model is embedded into or used by a data-driven model. This approach aims to augment the features of a data-driven model with domain knowledge, directly feeding the data-driven model with relevant data transformations. A pertinent example of this is the representation of inter-connected processes at different spatial or temporal scales, where a multiscale approach utilises both mechanistic representation for dynamic simulation and data-driven methods for higher=scale inference of process behaviour (Li *et al.* 2019; Hannaford *et al.* 2021). Tsen *et al.* (1996) show that this architecture outperforms other architectures and the mechanistic model itself. Nevertheless, this approach has not yet found much application in practice.

The **parallel** architecture is compelling when a mechanistic model is limited in describing the dynamics, e.g., due to incomplete domain knowledge or oversimplification of the processes. This contrasts with the serial hybrid model, which excels in situations where a specific subprocess is not defined. In a first variant, a cooperative parallel hybrid model, the data-driven model is trained to learn the mismatch between the mechanistic model and historical data, i.e. the residual error. Then both results are fused, usually by addition. A recent improvement in this approach is to learn the dynamics' residuals rather than the state's residuals (Quaghebeur *et al.* 2022). In a second variant, a competitive parallel hybrid model, the mechanistic and data-driven models are trained to make the same prediction. In other words, this technique can be considered an ensemble approach, in which different components seek to solve the same problem independently to improve the ensemble model performance (Hu *et al.* 2011; Abba *et al.* 2020). These predictions are then weighted and combined in a final output (Dors *et al.* 1995; Peres *et al.* 2001; Galvanauskas *et al.* 2004; Ghosh *et al.* 2019).

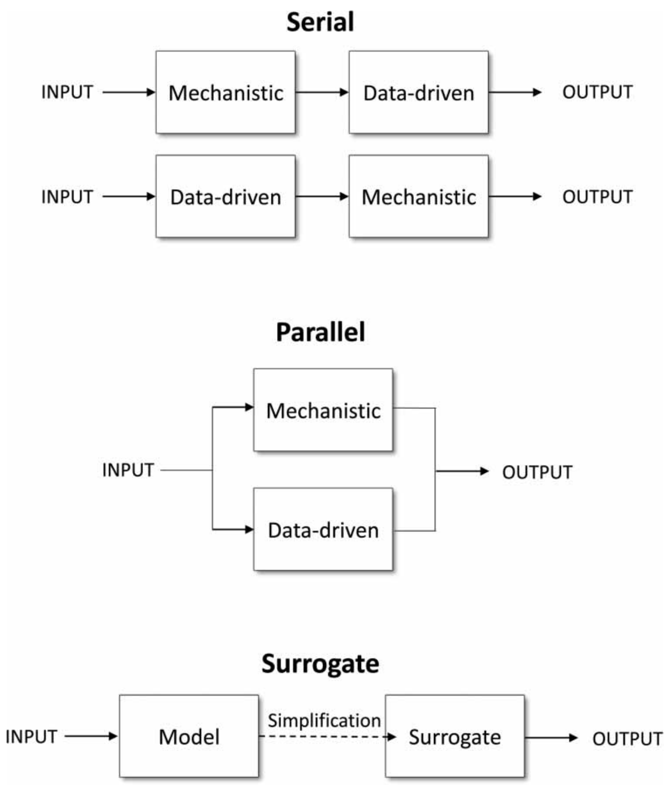


Figure 2 | Schematics of the three model architectures used in HM.

When we apply the loose definition of hybrid models, it typically implies the use of **surrogate models**, also known as metamodels (Jin *et al.* 2001), substitute models (Romijn *et al.* 2008), emulators (Conti & Hagan 2010; Mahmoodian *et al.* 2018), or response surface models (Wang & Georgakis 2019). These models are trained on the output of a mechanistic model to create a computationally less demanding model for rapid or large-scale simulations, such as required by, for example, real-time model predictive control, uncertainty analysis, or optimisation problems.

After defining the model architecture, whether serial, parallel, or surrogate, the parameters of the model need to be estimated from data, a process also known as parameter identification. Mechanistic modellers refer to this process as calibration, while data-driven modellers usually prefer the term training. The identification procedure usually aims to minimise the error between data and model output, i.e. by optimising the model fit. In both serial and parallel hybrid models, the parameter values of both components can be identified either separately or in conjunction.

In the separate approach, the parameters of one of the models are identified before identifying the other model. Usually, the mechanistic model is identified first, with the data-driven model identified afterwards. If needed, this procedure can be iterated several times. However, convergence is not guaranteed nor stable. Azarpour *et al.* (2017) suggest good modelling practices for hybrid model identification following the separate approach.

In the conjunction approach, the parameter values of both models are identified simultaneously. For example, the hybrid model is trained by backpropagating errors through the data-driven and mechanistic models (Psichogios & Ungar 1992). This involves tracking the gradients of both components with either nonlinear optimisation algorithms, such as sequential quadratic programming. Recently, methods such as automatic differentiation or adjoint sensitivity calculation have been adapted to the context of differential equations (Chen *et al.* 2018), making them applicable in the context of HM (Quaghebeur *et al.* 2021).

3. HM IN THE WATER INDUSTRY

A deep, comprehensive text mining analysis was conducted to inventory HM research publications related to the field of wastewater treatment. First, we used general terms (e.g., ‘wastewater’, ‘sewage’, ‘activated sludge’, and ‘anaerobic digestion’) to retrieve more than 320,000 wastewater related publication records (such as title, abstract, author keywords, etc.; updated on January 1, 2022) from the Web of Science (WOS 2021). We adopted a series of natural language processing (NLP) pretreatment steps to inventory the data for year, title, abstract, author keywords, and correspondence (Zhu *et al.* 2021). The NLP-based data preprocessing included N-grams generation (N=1–5, or tokenisation to one, two, three, four, and five adjacent words), lower-case/punctuation/stop-words treatment, and stemming (convert a word or term to its root form; Manning *et al.* (2008)). Meanwhile, we prepared four lists of terminologies (L) that were used as surrogates to represent (L1) wastewater treatment (>90 terms; e.g., wastewater, sewage, activated sludge, etc.), (L2) data-driven modelling only (>180 terms; e.g., regression, ML, ANN, etc.), (L3) mechanistic modelling only (>70 terms; e.g., mechanistic model, first-principles model, ASM, etc.), and (L4) hybrid modelling (e.g., gray-/grey-box, etc.). Two combinations ('L1 AND L2 AND L3' and 'L1 AND L4') of the key lists were then used to identify relevant papers from the initial 320,000 raw data. Specifically, the data of year and correspondence were used to identify variations of temporal (1990–2022) and spatial distribution. The spatial distribution analysis was performed based on the origin of the corresponding author(s) and countries or regions (Zhu *et al.* 2021). The dictionary of preprocessed textual corpora was used to identify hybrid model-based (the two combinations described above), data-driven model-based (e.g., supervised or/and unsupervised learning), and mechanistic model-based (e.g., variations of ASM and the anaerobic digestion model No. 1 (ADM)) publications based on their corresponding lists of terminology-surrogates. Furthermore, 11 representative terminologies (and their synonyms) were selected and used to showcase their research popularity in the HM publications. It is worth noting that model predictive control (MPC) studies were also identified from the raw wastewater database; not all MPC studies used a HM framework, but MPC can be a typical HM application in wastewater management.

In the 320,000 wastewater-related publications, more than 9,800 papers studied data-driven relevant methods and more than 3,900 papers focused on mechanistic models; their intersection, hybrid models, accounts for 534 publications. The number of relevant papers for HM increased steadily over the years from less than 10 in the early 1990s to more than 50 in 2021 (Figure 3(a)). Among the 534 papers, the majority of the HM research papers (506 or 94.8%) are identified based on the combination of data-driven and mechanistic terminologies. Specific HM terminologies, such as grey-box modelling, were used to retrieve 24 papers (4.5%) (Figure 3(c)). From the spatial perspective, the relevant papers were authored from 56 countries or regions; and the annual spatial popularity increased from less than five in the early 1990s to more than 20 countries or regions by 2021 (Figure 3(b)). Specifically, most of the papers (corresponding) were authored from Europe (224 or 41.9%), Asia (212 or 39.7%), and North America (56 or 10.5%) (Figure 3(d)). From the data-driven methods perspective, 324 papers studied supervised learning method(s), whereas only 32 papers explicitly used unsupervised learning method(s), and only six studies were related to reinforcement learning ((Figure 3(e)). The remaining papers did not specify the data-driven methods, including grey-box modelling and other miscellaneous terms, such as machine learning, big data, artificial intelligence, etc. A similar analysis applied to mechanistic models shows that 166 papers focused on the ASMs or ADM, 111 papers studied benchmark simulation models (BSMs) or related, and the remaining 257 papers did not specify any mechanistic models in their abstracts. A closer analysis of the 257 'non-explicit' papers shows that 57 papers studied Monod kinetics, and 12 described relevant process simulation software. As previously mentioned, MPC can be one HM application, and 276 studies have investigated MPC applied to wastewater management. In addition, the specific terminology inventory also shows that more studies used ML (250) than classical statistical methods (126) (e.g., multiple linear regression, autoregressive integrated moving average (ARIMA), PLS); among the ML methods, neural networks (146) are the dominant type (Figure 3(g)).

The text mining analyses reveal several valuable observations. First, the term 'grey-box modelling' (28) was not frequently used, while the majority of HM studies were identified based on a combination of using both data-driven and mechanistic models (Figure 3(g)). Second, a rapidly increasing number of papers based on both temporal and spatial variations started about 10–20 years ago, matching the period when the data-driven methods gained increasing popularity due to the recent advances in ML. Third, supervised learning is the dominant data-driven application, probably because data obtained from this approach can be more readily used for subsequent process management, while unsupervised learning is typically used to cluster different scenarios, so its applications are more often limited to off-site scenario analyses (Zhu *et al.* 2015; Borzooei *et al.* 2019). Though the number of papers is low, reinforcement learning or active learning may potentially improve process control when labelled data is not largely available (Russo *et al.* 2020). Furthermore, it seems that there are two ways to implement and test HM; (i) adaptation of ASM/ADM to hybridise with data-driven methods when data is not a limiting factor; (ii) testing process optimisation based on benchmarking models, such as BSM, with simulated data.

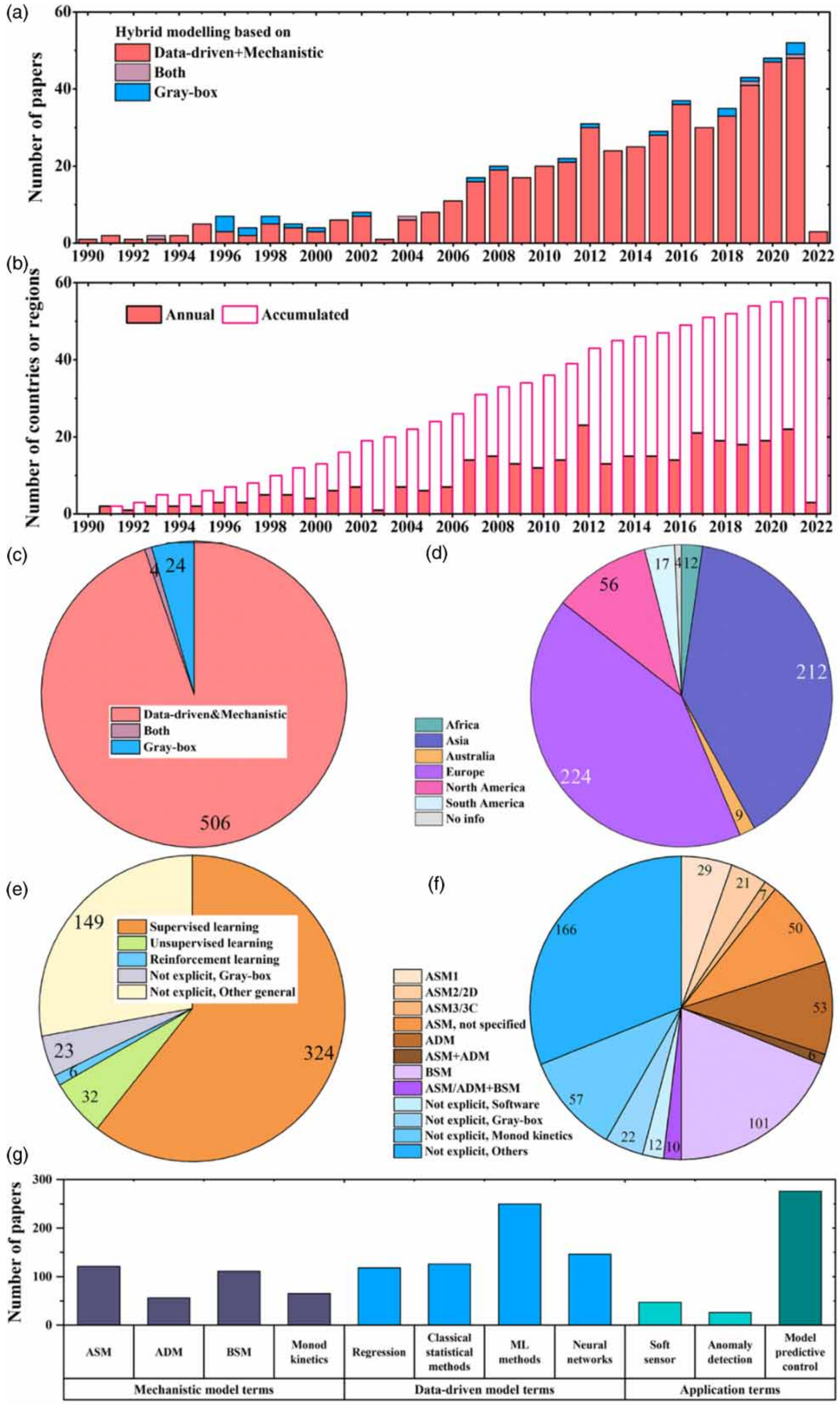


Figure 3 | Text mining analysis of the 534 relevant papers focused on HM research in wastewater management (data updated on January 1, 2022). Temporal variations in (a) the number of papers and (b) the number of countries or regions (annual and accumulated) associated with the papers published from 1990 to 2021. The numbers of HM papers that (c) were identified based on data-driven and mechanistic methods, gray-box, and both; (d) were authored from different areas based on the corresponding authors; (e) used supervised learning (neural networks are included here), unsupervised learning, reinforcement learning, not explicit but used gray-box, and other not explicit; and (f) studied ASM/ADM, BSM, and various not explicit (simulation software, gray-box, Monod kinetics, and others); (g) number of papers associated with 11 selected terminologies in the relevant papers.

3.1. HM in wastewater treatment

ASM mainly dominates WRRF modelling for the primary biological process, along with other unit processes such as sedimentation. ASM1, ASM2d, and ASM3 (Henze *et al.* 2000) have provided a standard framework for modelling based on which several other variants are developed. The use of hybrid models in wastewater treatment is also extensively documented in the literature (e.g. Boger 1992; Côté *et al.* 1995), long before the utilisation of ML methods and the current digitalisation trends related to the so-called, 4th Industrial Revolution (Corominas *et al.* 2018). Due to the increasing number of studies focusing on HM of WRRFs, the primary aspects covered in the existing studies are broadly mentioned below, emphasising the role of HM compared to using only mechanistic or data-driven models.

- **Influent generation:** The availability of dynamic influent data is a major bottleneck for applying ASMs to evaluate the design and operational scenarios for WRRFs. The WRRF influent generation concerns many complex processes, i.e. from at-source to the outlet of the sewer system. Hence, the development of influent generators focusing on the wastewater dynamics at the entrance of WRRFs, which can also be coupled with ASMs, continues to be a topic of interest. Influent generator models that use phenomenological or empirical approaches (De Keyser *et al.* 2010; Gernaey *et al.* 2011), as well as ML methods (Ahnert *et al.* 2016; Zhu *et al.* 2018; Ansari *et al.* 2020) based on historical data, are widely available. In addition, limited hybrid models are also developed that combine conceptual and stochastic models for dynamic influent generation (Talebizadeh *et al.* 2016). Such models aim to improve the model predictability for the key input state variables when compared to solely using a mechanistic or data-driven approach. Another generally employed approach is to introduce randomness or stochasticity to the predicted outputs using phenomenological methods to replicate real data (De Keyser *et al.* 2010; Flores-Alsina *et al.* 2014; Martin & Vanrolleghem 2014). In this case, the aim is to generate realistic influent data that is especially important for dynamic process modelling and control design. Influent generation is not only limited to predicting the incoming flow rate and pollutant loads to WRRF but is also valuable for generating dynamic inputs from other sources like domestic households (Penn *et al.* 2017; Sitzenfrei *et al.* 2017; Wärff *et al.* 2020) and combined sewer overflows (Keupers & Willems 2015).
- **Process modelling:** Hybrid models can be used to fill the gaps in prediction performance and process knowledge arising mainly from the use of mechanistic models (Anderson *et al.* 2000). Several serial approaches are successfully demonstrated, where ANNs and their variants are commonly used (e.g. Côté *et al.* 1995; Cong *et al.* 2020; Cheng *et al.* 2021). Studies using other approaches like Gaussian processes (Hvala & Kocijan 2020) and those using an ensemble of several methods (Lee *et al.* 2005) are also performed. Other studies use parallel HM, where the underlying dynamics and the relevant process are modelled based on the ASM family, and a data-driven model is used to improve the prediction capability (Thompson & Kramer 1994; Lee *et al.* 2005; Villez *et al.* 2020). HM is further applied to improve parameter identification in process modelling using ASM (Mašić *et al.* 2017; Villez *et al.* 2019). Zhu *et al.* (2015) and Borzooei *et al.* (2019) developed a HM approach to improve weather resilience and energy efficiency of a biological nutrient removal activated sludge system at a large-scale WRRF. An unsupervised clustering method was used on an extensive historical climate dataset to identify a series of weather-based influent scenarios, which were further used as an input of the process simulation model. The authors reported that the implementation of weather-based aeration strategies improved the energy efficiency of a WRRF.
- **Process monitoring and control:** HM can support greater plant automation via online, real-time control by taking advantage of the low computational cost of data-driven models. A HM framework can be realised by applying adaptive plant models through the use of digital twins, for example, for the training of surrogate models to be used in model-based control algorithms. Also, hybrid models can present an important added value for other parts of the control loop development. For instance, it is known that data-driven models (such as multilayer perceptrons (MLP)), long short-term memory (LSTM), ARIMA) have excellent performance in time-series prediction (Fernández *et al.* 2009; Zhu *et al.* 2018) and fault/anomaly detection (Wade *et al.* 2005; Haimi *et al.* 2016; Cheng *et al.* 2019). As a result, the forecasting capability of soft sensors can enhance process monitoring when working in parallel with traditional measurements or even reduce the necessity of some direct measurements (Haimi *et al.* 2013). A soft sensor is a model that predicts a variable of interest using the input from physical sensor(s). Soft sensors can predict either real-time/future influent conditions (Wang *et al.* 2019; Zhu & Anderson 2019) or the effluent quality/system performance (Shi & Xu 2018; Niu *et al.* 2020), which can be then used as input information to the feedback controller (utilised with the ASM/BSM models).

Overall, with the increasing model complexity of mechanistic models and the growing acceptance of data-driven approaches, research and application of hybrid modelling approaches to real-world case studies are expected to increase in the future (Sin & Al 2021).

3.2. HM in water treatment

As the wastewater treatment sector is transitioning to a more circular economy, decentralized treatment and water reuse concepts are gaining more attention. It thus becomes more and more challenging to separate wastewater treatment and water treatment into two completely unconnected domains. Hence, to provide a comprehensive overview of the applications and potential of HM in current and future (waste)water treatment processes, we briefly touch upon HM in drinking and process water treatment modelling studies.

The treatment of surface and groundwater resources to maintain a certain level of potable water quality imposed by regulators is based on a wide range of unit operations in water treatment plants (WTPs). Considering the importance of the continuous and efficient operation of WTPs in public health, several phenomenological models have been developed to simulate water treatment processes (Bhattacharjee & Tollner 2016). These models provide mathematical interpretations of various mechanisms, including the flow of water through reactor, equilibrium between water and gas, or solid phases in, for example, flotation processes (Yang *et al.* 2021), decay in the water and/or solid-phase in for example disinfection processes (Crapulli *et al.* 2014), chemical reactions (e.g., coagulation processes), and solution-diffusion processes in for example membrane processes (Gaublomme *et al.* 2020). Mechanistic models include programs developed for individual treatment processes as well as a few simulation packages providing the plant-wide simulation of WTPs such as Stimela (Van der Helm & Rietveld 2002), OTTER (Head *et al.* 2002), Metrex (Mälzer & Nahrstedt 2002), and TAPWAT (Versteegh *et al.* 2001). One of the main limitations for more widespread use of mechanistic models is that some of the input and/or calibration parameters of several unit process models, such as floc porosity, colloid charge, micropollutant concentration, and pathogen abundance, are unknown, case-specific, and hard to measure with sufficient precision (Juntunen *et al.* 2013; Gomes *et al.* 2015).

On the other hand, several studies applied data-driven approaches to monitor, model, and predict source water quality (e.g., Sahoo *et al.* 2005; Stedmon *et al.* 2011; Perelman *et al.* 2012; Mei *et al.* 2014; Debnath *et al.* 2015; Keskin *et al.* 2015; Deng & Wang 2017; Tesoriero *et al.* 2017; Mohammed *et al.* 2018; Delpla *et al.* 2019; Guo *et al.* 2019; Jin *et al.* 2019; Panidhapu *et al.* 2020; Li & Peleato 2021) and to model and control water treatment operations (e.g., Gagnon *et al.* 1997; Maier *et al.* 2004; Chen & Hou 2006; Wu & Lo 2008; Griffiths & Andrews 2011; Heddam *et al.* 2012; Juntunen *et al.* 2013; Gomes *et al.* 2015). Parallel to the increasing interest in using data-driven models, concerns have been raised about the liability and accountability of these models implemented in safety-critical systems such as water treatment, in which the slightest malfunction and failures may damage human health and/or the environment (Aliashrafi *et al.* 2021).

Recent studies have investigated the combination of data-driven models or hybridisation in response to explainability, auditability, and interpretability requirements (Doshi-Velez & Kim 2017; Abba *et al.* 2020) of data-driven approaches and first principles for water treatment processes (Karniadakis *et al.* 2021). An integrated framework was proposed in Wang *et al.* (2016), combining process-based and data-driven models to improve the forecasting of salinity and dissolved oxygen parameters for a real case study. The hybrid platform consisted of two methods based on chaos theory coupled with water quality models. It was reported that hybridization could improve the forecast ability of the model and reduce the computation time. Jia *et al.* (2021) investigated the application of a physics-guided recurrent neural network in the context of modelling the dynamics of the temperature of water resources. It was reported that hybridization of the law of energy conservation with LSTM could improve the forecasting accuracy of the model. By applying the hybrid linear-state-space dynamic Langmuir model, Nordstrand & Dutta (2021) achieved a more robust and flexible control of capacitive-deionisation desalination processes. The authors reported that combining a general modelling framework with physical insights could offer flexibility in calibration data and better predictability over a wide range of operating models.

It should be noted that digital adaptation has been far quicker in the water treatment sector than in wastewater, mainly due to more straightforward functionality and higher reliability of water quantity measuring devices used in WTPs than wastewater quality sensors in WRRFs (Therrien *et al.* 2020). However, to the authors' knowledge, only a few studies have been found implementing a HM approach in WTPs. This can be since there is still a lack of a standardised mechanistic modelling framework similar to ASM models in water treatment processes. In contrast, the water treatment sector has employed a wide variety of mechanistic and data-driven models to address the modelling needs. This might influence the development of hybrid models in these two sectors.

4. CHALLENGES AND DEVELOPMENT NEEDS

By combining the fundamental and generalisable process knowledge built into mechanistic models and the predictive power of data-driven approaches, HM offers an opportunity to overcome inherent weaknesses in both and significantly improve the overall model performance. However, the application of HM as a powerful tool for application in WRRF modelling studies is hampered by several critical challenges and associated development needs, which are discussed in this section.

4.1. Good modelling practice frameworks for HM

The WRRF community has a strongly developed modelling community (evidenced by the organisation of specialist WRRmod seminars and the IWA Specialist Group on Modelling and Integrated Assessment, MIA), which has stimulated the development of robust modelling frameworks and protocols within the domain. Standardised model frameworks, such as ASM, have been used for decades, and different protocols for Good Modelling Practice (GMP), e.g., for calibration and validation (Rieger *et al.* 2012) or uncertainty analysis (Belia *et al.* 2021), have been suggested to encourage a systematic workflow for model development and application. Integrating new modelling paradigms into existing WRRF modelling frameworks raises several questions on extending the existing protocols and practices to continue compliance with GMP guidelines. Azarpour *et al.* (2017) suggest GMP be used for hybrid model identification, but the application of these guidelines should be assessed within the context of existing frameworks for WRRF modelling. Questions to be answered in future research include: Are current protocols for WRRFs transferable to HM, or does the addition of a data-driven component require separate considerations? Should the data-driven and mechanistic components be considered together for each sub-process (e.g., hydraulics, aeration, sludge balance), or should the calibration first focus on the mechanistic models before including any data-driven components?

4.2. How to deal with uncertainty?

Simulation studies are sensitive to different sources of uncertainty, i.e. input data quantity and quality, inadequate model structure, inaccurately estimated model parameters and the inherent uncertainty of the system (Belia *et al.* 2013). Some uncertainty estimation methods are available, such as Monte Carlo and Generalised Likelihood Uncertainty Estimation (GLUE). The sources of uncertainty in HM originate, naturally, from both the mechanistic and data-driven components. However, many previous HM studies lack the quantification of the inherent uncertainty. This makes it difficult to assess whether the model prediction is supported by the training data or due to an uncertain extrapolation. In HM, the accuracy and uncertainty of the results depend on the data quantity and quality used for training/calibration (Refsgaard *et al.* 2007). This means that when process data changes, then the uncertainty of the results will change. The uncertainty quantification is an example of why a consistent collection of meta-data is essential, for example, to consider the wear-and-tear of sensor technologies (Ohmura *et al.* 2019). Dynamically quantifying uncertainty in the context of HM is an essential challenge to assess the trust in hybrid models and, thus, to fully establish and benefit from these powerful tools.

4.3. Model selection – balancing complexity

Determining optimal model complexity for a specific objective is equally relevant for both mechanistic and data-driven models. An overly complex model makes parameter identification impractical or even infeasible for mechanistic models and increases computational cost, rendering the model less user-friendly. Depending on the system and the intended purpose, the modeller must often strike a careful balance between model complexity and usability. An example of this is the wide range of existing models to describe mixing behaviour in a reactor. They range from a simple continuously stirred tank reactor (CSTR) that assumes instant mixing and tanks-in-series and compartmental models that simplify mixing behaviour to advanced computational fluid dynamics (CFD) models, which explicitly model all underlying mixing dynamics. A CFD model would be impractical for most real-time applications, while a CSTR might not yield sufficiently detailed results for an in-depth study.

Similarly, for data-driven models, an appropriate level of complexity needs to be carefully considered. In data-driven terminology, this is known as the bias-variance trade-off. If the model is not complex enough, it cannot capture the relationships, resulting in the underfitting of the model with a high bias. As the complexity of a model is increased, more complex relationships can be considered. For example, a simple linear regression model can only fit a straight line, while a polynomial can fit to nonlinear dynamics. It is not only possible to increase the complexity by switching between techniques; most methods include certain hyperparameters controlling the complexity, e.g., the number of layers and neurons in a neural network; however, as complexity increases, the probability of fitting to noise increases (Fayyad *et al.* 1996). Consequently, the model learns a relationship that it cannot generalise to new data.

The application of HM brings the additional challenge of balancing complexity between the mechanistic and data-driven components. For example, in the serial HM approach, there is a risk that one model compensates for a suboptimal structure of the other model. Since mechanistic and data-driven modelling paradigms are rooted in different engineering practices, only a few experts are trained to a sufficient level in both, which might lead to sub-optimally balanced models, biased to the methods most comfortable with the practitioner. A sub-optimally balanced model is one in which the prediction power is lower than its potential prediction capacity. For example, the data-driven model will try to compensate for the lack of identifiability in the mechanistic model, thus hampering its ability to accurately capture the process dynamics for the subprocess for which it is being developed.

4.4. Data quality and metadata

A fundamental prerequisite for the data-driven part of the HM application is ensuring a high-quality data pipeline to avoid a 'garbage in – garbage out' scenario (Therrien *et al.* 2020; Dobbelaere *et al.* 2021). We put special emphasis on the data quality because, e.g., if there is an operational change from routine, manual sampling at a laboratory to the use of an online sensor, then a wholesale change in protocol for screening, processing and analysing the data is likely necessary. However, Schneider *et al.* (2019) and Thürlimann *et al.* (2019) demonstrate that inaccurate data can be beneficial, for example, in soft-sensor development, where choosing an approach robust to the disturbances will require knowledge of uncertainties and errors in the data. An additional challenge is descriptive data, commonly referred to as meta-data. Especially for large amounts of data, meta-data is invaluable and necessary to ensure the data and metadata are in a consistent, machine-interpretable form, preferably with an automated collection (Aguado *et al.* 2021). Automatic collection forces a clear definition and helps to avoid gaps in the metadata collection.

4.5. Architecture challenge

As discussed in section 2, HM can be constructed according to different architectures and choosing the right one is an important step in the construction of HM frameworks. Choosing which architecture to use is mainly problem-dependent, conditional on the data available and the knowledge missing. Given that theory lags behind the practice in ML, trial and error are a large part of a model's development process. According to von Stosch *et al.* (2014), a serial architecture is preferred in cases where the confidence in the overall mechanistic model structure is relatively high, but specific subprocesses are insufficiently known or understood. One could argue that this is precisely the case for WRRF modelling. The ASM provide a backbone model structure, but modellers often struggle to accurately capture the variability in specific subprocesses such as microbial composition (Ni & Yu 2010), mixing behaviour (Arnaldos *et al.* 2015), or oxygen transfer over a wide range of operating conditions (Amaral *et al.* 2019). A nice example is a recent effort to develop ASM-based model extensions to capture the dynamics of short cut N-removal processes (Sharif Shourjeh *et al.* 2021) or efforts to increase the realism of mixing behaviour deriving compartmental models from CFD models (Le Moullec *et al.* 2011). HM could provide an interesting solution here as they present benefits derived from knowledge incorporated in existing mechanistic wastewater modelling frameworks, such as ASM/ADM, while capturing unobserved dynamics flexibly. An important challenge for such applications of serial hybrid models is that direct and frequent measurements of the relevant variables for the unknown subprocesses (such as microbial composition, dynamic oxygen transfer efficiency, specific growth rates) are often not possible. Thus, the data-driven component of the hybrid model in such a case can only be trained indirectly, or the hybrid model should be calibrated as a whole on available data through the use of, e.g., neural differential equations. Benchmarking the potential of HM for several commonly known ill-defined subprocesses is an interesting and pragmatic avenue to advance the uptake of HM in the water and resource domain.

4.6. The reputation that data-driven methods impede process understanding

The data-driven approach is often associated with a lack of interpretability, such as for SVM (Angelov *et al.* 2021), but it is not always the case. Data-driven modelling can support the understanding and interpretation of data. Examples are dimension reduction methods such as self-organising maps (Dürrenmatt & Gujer 2011) or PCA (Lee & Vanrolleghem 2003) in combination with clustering methods (Villez *et al.* 2008). Furthermore, explainable AI (Samek *et al.* 2017; Adadi & Berrada 2018) or decision trees (Quinlan 1990) are data-driven methods with higher interpretability (Angelov *et al.* 2021). Hybrid models suffer much less from the lack of interpretability. Since the data-driven part is inherently coupled to a model structure with physical meaning, this allows us to evaluate the specific dynamics of the data-driven part as a measure of model structure error in the mechanistic model. Such hybrid models have the potential to provide insights into ill-defined processes. Therefore, applying interpretable ML tools such as Shapley values (Shapley 2016; Lundberg & Lee 2017) to derive mechanistic insights from HM deserves attention in future research.

4.7. Common API and open source software

For hybrid models to be used to their full potential as standard practice for WRRF modelling, it is imperative to have access to a compatible platform or application programming interface (API). However, the data-driven approaches are often hard coded using open-source software, e.g., Python, R, MATLAB, Julia, while mechanistic models, such as ASM and the ADM No. 1, have been packaged in off-the-shelf commercial simulators. Some WRRF simulation software providers (Sumo, WEST, Simba#, GPS-X, dDockDT) embrace open data philosophy and build APIs for open-source scripting languages. Improvement of the compatibility and flexibility of these simulation platforms is still in progress. Furthermore, maintaining the pre-release research code is an issue, especially when linking several different software. As many developers leave academia, a similar effort might be required several times.

5. VISION ON HYBRID MODELS AS AN ENABLING FACTOR IN THE TRANSITION TOWARDS A MORE RESOURCE-EFFICIENT WATER SECTOR

There has been a clear trend of increasing utilisation of data-driven models in the water sector over the last two decades (Eggimann *et al.* 2017). Many data-driven approaches have the potential to break through and become routinely used tools in the industry or have been ranked as an important topic for research and development by experts in the WRRF field (Blumensaat *et al.* 2019). Some data-driven models have replaced, or will soon replace, more conventional knowledge-based models (Hadjimichael *et al.* 2016). Much research exists on the application of ML in the urban water sector. However, mechanistic models will remain necessary as long as high-quality process data is laborious, difficult, and costly to collect and significant uncertainty remains unaccounted for in treatment processes. On the other hand, data-driven approaches are viewed with scepticism due to the lack of process explainability (Newhart *et al.* 2019). However, we see a huge potential for hybrid model development and application due to the combination of full or partial mechanistic explainability, predictive power, and computational efficiency. As such, HM can lead to significant process improvements, especially in the context of multi-objective decision-making, which is already a significant challenge for WRRF operators. Specifically, we see a large potential for the application of HM in the following areas.

5.1. Soft-sensor development for decentralised water treatment and reuse

WRRF have an important societal function to make urban water management more resource-efficient, sustainable, reliable, and safe (Larsen 2011; Sedlak 2014; van Loosdrecht & Brdjanovic 2014). Decentralised WRRF play an important role to reach this goal as they stimulate a more circular water economy through the production of water fit for purpose (Larsen *et al.* 2016). However, a true paradigm shift is only possible with appropriate monitoring, control, and optimization (Eggimann *et al.* 2017). Unfortunately, the amount of undetected failures at decentralised WRRFs is significant (Moelants *et al.* 2008), which results in reduced performance. To monitor such plants, soft sensors (which provide an estimation of variables that are difficult to measure automatically due to sensor availability and cost limitations), using data from low-cost, unmaintained, and likely faulty sensors (80% accuracy in a real-world implementation) can be developed (Schneider *et al.* 2020). There is, however, still significant scope for improvement of these types of soft sensors. We think that HM can improve the acuity and reliability of soft sensors and can be used more generally at unsupervised WRRF, for example. Additionally, a parallel HM approach can be considered to increase the robustness of the monitoring, especially for plants where an undetected failure has a critical process impact. Moreover, a more widespread use, or optimal placement, of sensors (Villez *et al.* 2020) will allow the collection of more valuable datasets that can be utilised to populate the data-driven component of hybrid models with higher resolution and better quality data.

5.2. Harnessing the power of advanced, computationally expensive modelling frameworks

Additionally, HM can facilitate the transition of advanced model frameworks to the WRRF modelling practice. Examples are models describing particle heterogeneity, such as population balance models (Nopens *et al.* 2015). This model framework has the theoretical potential to describe many of the dynamics related to particle heterogeneity in WRRF processes, such as floc and granule formation, struvite crystal growth, or precipitation (Nopens *et al.* 2015; Elduayen 2020) but is dependent on functions that describe sub-processes such as growth, aggregation, and break-up for which current mechanistic knowledge is insufficient. HM can be an enabling factor for the application of these models by combining population balance models (to gain insight into processes driven by variables with important distributed properties, e.g., microbial diversity, floc size and structure, precipitation) with data-driven models to describe the sub-processes (e.g., particle interactions) (Torfs *et al.* 2012).

5.3. Advanced control and optimisation of urban water systems

Recent trends in the digitalisation of the water sector have led to the emergence of digital twins as powerful decision-making tools. A digital twin is a virtual representation of a physical entity (i.e. a model) that has a live, automated data connection to its physical counterpart (Fuller *et al.* 2020). As such, a digital twin allows for process simulations in (near) real-time, opening up the potential for real-time operational decision-making, advanced process control, and online process optimisation at a scale beyond existing automation to date. The use of digital twins brings along an important shift in model objectives. For real-time scenario analysis or control, operational digital twins require a very high predictive power, typically on a short time-scale (hours to days in the future) (Wright & Davidson 2020). Assuming sufficient high-quality data is available, this time frame is exactly the ecosystem in which data-driven models can perform very well. However, despite the consensus on their great potential and substantial research efforts, purely data-driven models are not yet widely used for full-scale WRRF operations. This is due to several factors such as the non-stationary behaviour of these systems, sizeable daily flow fluctuations (Lowe *et al.* 2010), variations in sludge residence time or dilution effects. Hybrid models have the potential to boost the predictive power in real-time applications by leveraging the power of data-driven techniques without losing the process knowledge incorporated in the mechanistic model. An example is the use of reinforcement learning or active learning (Hernández-del-Olmo *et al.* 2018; Filipe *et al.* 2019; Alves Goulart & Dutra Pereira 2020), which will bring benefits to better deployment of real-time control. This can accelerate the transition of WRRF models to digital twin applications and, as such, stimulate real-time and online optimisation of WRRF concerning energy consumption, effluent quality (tailored for discharge or specific reuse purposes), or greenhouse gas emissions.

5.4. Beyond the combination of two paradigms

Most current applications of HM are either based on the development of surrogate models or the traditional serial or parallel architecture. In the near future, such approaches have a great potential to improve the management of WRRF significantly. However, in these approaches, the contributions of the data-driven and mechanistic components are still relatively separated. Taking a step forward, we propose truly integrated approaches rather than combining two distinct models, whether serially or in parallel. This means an iterative approach in which one model profits and learns from the other internally and intrinsically. For example, it can be used to explore a wide range of potential mechanistic model pathways. An unknown process could be evaluated first using a data-driven methodology. With a human-in-the-loop approach, information can be checked for consistency and acuity prior to accepting the knowledge gained from the model. This information would then flow into the design for the next experiment to gain more detailed or structured mechanistic knowledge. Hence, in addition to improving the prediction power, greater process understanding can be obtained at a relatively lower cost than by mechanistic modelling alone.

ACKNOWLEDGEMENTS

We thank Kris Villez for critically reviewing the manuscript and all the HM workshop participants at the WRRmod 2021 conference in Arosa, Switzerland, for input and discussion.

FUNDING SOURCES

MYS received funding from the Japanese Society for Promotion of Science (JSPS) Grant P20763. WQ received funding from the Research Foundation – Flanders (FWO) through Grant 3S79219.

AUTHOR CONTRIBUTIONS

All authors contributed equally to the development and finalization of this manuscript.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

REFERENCES

Abba, S. I., Pham, Q. B., Saini, G., Linh, N. T. T., Ahmed, A. N., Mohajane, M., Khaledian, M., Abdulkadir, R. A. & Bach, Q.-V. 2020 Implementation of data intelligence models coupled with ensemble machine learning for prediction of water quality index. *Environ. Sci. Pollut. Res.* **27**, 41524–41539.

Abrahart, R. J., See, L. M. & Solomatine, D. P. 2008 *Practical Hydroinformatics: Computational Intelligence and Technological Developments in Water Applications*. Springer Science & Business Media, Berlin, Heidelberg, Germany.

Adadi, A. & Berrada, M. 2018 Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* **6**, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Aguado, D., Blumensaat, F., Baeza, J. A., Villez, K., Ruano, M. V., Samuelsson, O. & Plana, Q. 2021 *Digital Water, The Value of Meta-Data for Water Resource Recovery Facilities*. IWA.

Ahnert, M., Marx, C., Krebs, P. & Kuehn, V. 2016 A black-box model for generation of site-specific WWTP influent quality data based on plant routine data. *Water Science and Technology* **74** (12), 2978–86. https://doi.org/10.2166/wst.2016.463.

Aliashrafi, A., Zhang, Y., Groenewegen, H. & Peleato, N. M. 2021 A review of data-driven modelling in drinking water treatment. *Rev. Environ. Sci. Biotechnol.* **20**, 985–1009.

Alves Goulart, D. & Dutra Pereira, R. 2020 Autonomous pH control by reinforcement learning for electroplating industry wastewater. *Comput. Chem. Eng.* **140**, 106909. https://doi.org/10.1016/j.compchemeng.2020.106909.

Amaral, A., Gillot, S., Garrido-Baserba, M., Filali, A., Karpinska, A. M., Plósz, B. G., De Groot, C., Bellandi, G., Nopens, I., Takács, I., Lizarralde, I., Jimenez, J. A., Fiat, J., Rieger, L., Arnell, M., Andersen, M., Jeppsson, U., Rehman, U., Fayolle, Y., Amerlinck, Y. & Rosso, D. 2019 Modelling gas–liquid mass transfer in wastewater treatment: when current knowledge needs to encounter engineering practice and vice versa. *Water Sci. Technol.* **80**, 607–619. https://doi.org/10.2166/wst.2019.253.

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I. & Atkinson, P. M. 2021 Explainable artificial intelligence: an analytical review. *WIREs Data Min. Knowl. Discovery* **11**, e1424. https://doi.org/10.1002/widm.1424.

Ansari, M., Othman, F. & El-Shafie, A. 2020 Optimized fuzzy inference system to enhance prediction accuracy for influent characteristics of a sewage treatment plant. *Science of The Total Environment* **722**, 137878. https://doi.org/10.1016/j.scitotenv.2020.137878.

Anderson, J. S., McAvoy, T. J. & Hao, O. J. 2000 Use of hybrid models in wastewater systems. *Industrial & Engineering Chemistry Research* **39** (6), 1694–1704. https://doi.org/10.1021/ie990557r.

Apache Software Foundation n.d. *Hadoop (Version 3.3.1)*. Available from: https://hadoop.apache.org. (accessed 1 July 2022).

Arnaldos, M., Amerlinck, Y., Rehman, U., Maere, T., Van Hoey, S., Naessens, W. & Nopens, I. 2015 From the affinity constant to the half-saturation index: understanding conventional modeling concepts in novel wastewater treatment processes. *Water Res.* **70**, 458–470. https://doi.org/10.1016/j.watres.2014.11.046.

Azarpour, A., Borhani, T. N. G., Wan Alwi, S. R., Manan, Z. A. & Abdul Mutalib, M. I. 2017 A generic hybrid model development for process analysis of industrial fixed-bed catalytic reactors. *Chem. Eng. Res. Des.* **117**, 149–167. https://doi.org/10.1016/j.cherd.2016.10.024.

Baeten, J. E., Batstone, D. J., Schraa, O. J., van Loosdrecht, M. C. M. & Volcke, E. I. P. 2019 Modelling anaerobic, aerobic and partial nitritation-anammox granular sludge reactors – a review. *Water Res.* **149**, 322–341. https://doi.org/10.1016/j.watres.2018.11.026.

Belia, E., Johnson, B., Benedetti, L., Bott, C. B., Martin, C., Murthy, S., Neumann, M. B., Rieger, L., Weijers, S. & Vanrolleghem, P. A. 2013 *Uncertainty Evaluations in Model Based WRRF Design for High Level Nutrient Removal*. WERF NUTR1R06q, 54.

Belia, E., Neumann, M. B., Benedetti, L., Johnson, B., Murthy, S., Weijers, S. & Vanrolleghem, P. A. 2021 *Uncertainty in Wastewater Treatment Design and Operation*. IWA Publishing.

Belia, Evangelia, Lorenzo Benedetti, Bruce Johnson, Sudhir Murthy, Marc Neumann, Peter Vanrolleghem, and Stefan Weijers, eds. Uncertainty in Wastewater Treatment Design and Operation: Addressing Current Practices and Future Directions. Scientific and Technical Report 21. London, UK: IWA Publishing, 2021. https://doi.org/10.2166/9781780401034.

Bhattacharjee, N. V. & Tollner, E. W. 2016 Improving management of windrow composting systems by modeling runoff water quality dynamics using recurrent neural network. *Ecol. Modell.* **339**, 68–76.

Blumensaat, F., Leitão, J. P., Ort, C., Rieckermann, J., Scheidegger, A., Vanrolleghem, P. A. & Villez, K., 2019 How urban storm- and wastewater management prepares for emerging opportunities and threats: digital transformation, ubiquitous sensing, new data sources, and beyond – a horizon scan. *Environ. Sci. Technol.* acs.est.8b06481. https://doi.org/10.1021/acs.est.8b06481

Boger, Z. 1992 Application of neural networks to water and wastewater treatment plant operation. *ISA Trans.* **31**, 25–33. https://doi.org/10.1016/0019-0578(92)90007-6.

Borzooei, S., Miranda, G. H., Teegavarapu, R., Scibilia, G., Meucci, L. & Zanetti, M. C. 2019 Assessment of weather-based influent scenarios for a WWTP: application of a pattern recognition technique. *J. Environ. Manage.* **242**, 450–456.

Borzooei, S., Miranda, G. H., Abolfathi, S., Scibilia, G., Meucci, L. & Zanetti, M. C. 2020 Application of unsupervised learning and process simulation for energy optimization of a WWTP under various weather conditions. *Water Sci. Technol.* **81**, 1541–1551.

Butler, D., Ward, S., Sweetapple, C., Astaraie-Imani, M., Diao, K., Farmani, R. & Fu, G. 2017 Reliable, resilient and sustainable water management: the Safe & SuRe approach. *Global Challenges* **1** (1), 63–77. https://doi.org/10.1002/gch2.1010.

Cecconi, F., Reifsnyder, S., Sobhani, R., Cisquella-Serra, A., Madou, M. & Rosso, D. 2020 Functional behaviour and microscopic analysis of ammonium sensors subject to fouling in activated sludge processes. *Environ. Sci. Water Res. Technol.* **6**, 2723–2733. https://doi.org/10.1039/D0EW00359 J.

Chen, C.-L. & Hou, P.-L. 2006 Fuzzy model identification and control system design for coagulation chemical dosing of potable water. *Water Sci. Technol. Water Supply* **6**, 97–104.

Chen, R. T., Rubanova, Y., Bettencourt, J. & Duvenaud, D. 2018 *Neural Ordinary Differential Equations*. ArXiv Prepr. ArXiv180607366.

Cheng, G., Ramirez-Amaro, K., Beetz, M. & Kuniyoshi, Y. 2019 Purposive learning: robot reasoning about the meanings of human activities. *Sci. Rob.* **4**. https://doi.org/10.1126/scirobotics.aav1530.

Cheng, X., Guo, Z., Shen, Y., Yu, K. & Gao, X. 2021 Knowledge and data-driven hybrid system for modeling fuzzy wastewater treatment process. *Neural Comput. Appl.* https://doi.org/10.1007/s00521-021-06499-1.

Cong, Q., Xi, X., Su, C., Deng, S. & Zhao, Y. 2020 Hybrid integrated model of water quality in wastewater treatment process via RBF neural network. In: *Robotics and Rehabilitation Intelligence, Communications in Computer and Information Science* (Qian, J., Liu, H., Cao, J. & Zhou, D., eds). Springer, Singapore, pp. 333–341. https://doi.org/10.1007/978-981-33-4932-2_24.

Conti, S. & Hagan, A. O. 2010 Bayesian emulation of complex multi-output and dynamic computer models. **140**, 640–651. https://doi.org/10.1016/j.jspi.2009.08.006.

Corominas, L., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U. & Poch, M. 2018 Transforming data into knowledge for improved wastewater treatment operation: a critical review of techniques. *Environ. Modell. Software* **106**, 89–103. https://doi.org/10.1016/j.envsoft.2017.11.023.

Côté, M., Grandjean, B. P. A., Lessard, P. & Thibault, J. 1995 Dynamic modelling of the activated sludge process: improving prediction using neural networks. *Water Res.* **29**, 995–1004. https://doi.org/10.1016/0043-1354(95)93250-W.

Crapulli, F., Santoro, D., Sasges, M. R. & Ray, A. K. 2014 Mechanistic modeling of vacuum UV advanced oxidation process in an annular photoreactor. *Water Res.* **64**, 209–225. https://doi.org/10.1016/j.watres.2014.06.048.

Daims, H. 2014 The family nitrospiraceae. In: *The Prokaryotes* (Rosenberg, E., DeLong, E. F., Lory, S., Stackebrandt, E. & Thompson, F., eds). Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 733–749. https://doi.org/10.1007/978-3-642-38954-2_126.

DeAngelis, D. L. & Yurek, S. 2015 Equation-free modeling unravels the behavior of complex ecological systems. *Proc. Natl. Acad. Sci.* **112**, 3856–3857.

Debnath, A., Majumder, M. & Pal, M. 2015 A cognitive approach in selection of source for water treatment plant based on climatic impact. *Water Resour. Manage.* **29**, 1907–1919.

De Keyser, W., Gevaert, V., Verdonck, F., De Baets, B. & Benedetti, L. 2010 An emission time series generator for pollutant release modelling in urban areas. *Environ. Model. Software* **25**, 554–561. https://doi.org/10.1016/j.envsoft.2009.09.009.

Delpla, I., Florea, M. & Rodriguez, M. J. 2019 Drinking water source monitoring using early warning systems based on data mining techniques. *Water Resour. Manage.* **33**, 129–140.

Deng, W. & Wang, G. 2017 A novel water quality data analysis framework based on time-series data mining. *J. Environ. Manage.* **196**, 365–375.

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. https://doi.org/10.1016/j.eng.2021.03.019.

Dors, M., Simutis, R. & Lübbert, A. 1995 *Hybrid Process Modeling for Advanced Process State Estimation, Prediction, and Control Exemplified in a Production-Scale Mammalian Cell Culture*. ACS Publications.

Doshi-Velez, F. & Kim, B. 2017 Towards a rigorous science of interpretable machine learning. ArXiv:1702.08608 [Cs, Stat]. https://doi.org/10.48550/arXiv.1702.08608.

Dürrenmatt, D. J. & Gujer, W. 2011 Data-driven modeling approaches to support wastewater treatment plant operation. *Environ. Model. Software* **30**, 47–56. https://doi.org/10.1016/j.envsoft.2011.11.007.

Eggimann, S., Mutzner, L., Wani, O., Schneider, M. Y., Spuhler, D., Moy de Vitry, M., Beutler, P. & Maurer, M. 2017 The potential of knowing more: a review of data-driven urban water management. *Environ. Sci. Technol.* **51**, 2538–2553. https://doi.org/10.1021/acs.est.6b04267.

Elduayen, E. B. 2020 *New Mass-Based Population Balance Model Including Shear Rate Effects: Application to Struvite Recovery*. University of Navarra, Donostia.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. 1996 From data mining to knowledge discovery in databases. *AI Mag.* **17**, 37. https://doi.org/10.1609/aimag.v17i3.1230.

Fernández, C., Vega, J. A., Fonturbel, T. & Jiménez, E. 2009 Streamflow drought time series forecasting: a case study in a small watershed in North West Spain. *Stochastic Environ. Res. Risk Assess.* **23**, 1063–1070.

Filipe, J., Bessa, R. J., Reis, M., Alves, R. & Póvoa, P. 2019 Data-driven predictive energy optimization in a wastewater pumping station. *Appl. Energy* **252**, 113423. https://doi.org/10.1016/j.apenergy.2019.113423.

Flores-Alsina, X., Saagi, R., Lindblom, E., Thirsing, C., Thornberg, D., Gernaey, K. V. & Jeppsson, U. 2014 Calibration and validation of a phenomenological influent pollutant disturbance scenario generator using full-scale data. *Water Res.* **51**, 172–185. https://doi.org/10.1016/j.watres.2013.10.022.

Fuller, A., Fan, Z., Day, C. & Barlow, C. 2020 Digital twin: enabling technologies, challenges and open research. *IEEE Access* **8**, 108952–108971. https://doi.org/10.1109/ACCESS.2020.2998358.

Gagnon, C., Grandjean, B. P. & Thibault, J. 1997 Modelling of coagulant dosage in a water treatment plant. *Artif. Intell. Eng.* **11**, 401–404.

Galvanauskas, V., Simutis, R. & Lübbert, A. 2004 Hybrid process models for process optimisation, monitoring and control. *Bioprocess Biosyst. Eng.* **26**, 393–400.

Gaublomme, D., Strubbe, L., Vanoppen, M., Torfs, E., Mortier, S., Cornelissen, E., De Gusseme, B., Verliefde, A. & Nopens, I. 2020 A generic reverse osmosis model for full-scale operation. *Desalination* **490**, 114509. https://doi.org/10.1016/j.desal.2020.114509.

Gernaey, K. V., Flores-Alsina, X., Rosen, C., Benedetti, L. & Jeppsson, U. 2011 Dynamic influent pollutant disturbance scenario generation using a phenomenological modelling approach. *Environ. Modell. Software* **26**, 1255–1267. https://doi.org/10.1016/j.envsoft.2011.06.001.

Gernaey, K. V., van Loosdrecht, M. C. M., Henze, M., Lind, M. & Jørgensen, S. B. 2004 Activated sludge wastewater treatment plant modelling and simulation: state of the art. *Environmental Sciences and Artificial Intelligence* **19** (9), 763–83. https://doi.org/10.1016/j.envsoft.2003.03.005.

Ghosh, D., Hermonat, E., Mhaskar, P., Snowling, S. & Goel, R. 2019 Hybrid modeling approach integrating first-principles models with subspace identification. *Ind. Eng. Chem. Res.* **58**, 13533–13543.

Gomes, L. S., Souza, F. A. A., Pontes, R. S. T., Fernandes Neto, T. R. & Araújo, R. A. M. 2015 Coagulant dosage determination in a water treatment plant using dynamic neural network models. *Int. J. Comput. Intell. Appl.* **14**, 1550013.

Griffiths, K. A. & Andrews, R. C. 2011 The application of artificial neural networks for the optimization of coagulant dosage. *Water Sci. Technol. Water Supply* **11**, 605–611.

Guo, D., Lintern, A., Webb, J. A., Ryu, D., Liu, S., Bende-Michl, U., Leahy, P., Wilson, P. & Western, A. W. 2019 Key factors affecting temporal variability in stream water quality. *Water Resour. Res.* **55**, 112–129.

Hadjimichael, A., Comas, J. & Corominas, L. 2016 Do machine learning methods used in data mining enhance the potential of decision support systems? A review for the urban water sector. *AI Commun.* **29**, 747–756. https://doi.org/10.3233/AIC-160714.

Haimi, H., Mulas, M., Corona, F. & Vahala, R. 2013 Data-derived soft-sensors for biological wastewater treatment plants: an overview. *Environ. Modell. Software* **47**, 88–107. https://doi.org/10.1016/j.envsoft.2013.05.009.

Haimi, H., Mulas, M., Corona, F., Marsili-Libelli, S., Lindell, P., Heinonen, M. & Vahala, R. 2016 Adaptive data-derived anomaly detection in the activated sludge process of a large-scale wastewater treatment plant. *Eng. Appl. Artif. Intell.* **52**, 65–80.

Hannaford, N. E., Heaps, S. E., Nye, T. M. W., Curtis, T. P., Allen, B., Golightly, A. & Wilkinson, D. J. 2021 *A Sparse Bayesian Hierarchical Vector Autoregressive Model for Microbial Dynamics in a Wastewater Treatment Plant*. ArXiv210700502 Q-Bio Stat.

Head, R., Shepherd, D., Butt, G. & Buck, G. 2002 OTTER mathematical process simulation of potable water treatment. *Water Sci. Technol. Water Supply* **2**, 95–101.

Heddam, S., Bermad, A. & Dechemi, N. 2012 ANFIS-based modelling for coagulant dosage in drinking water treatment plant: a case study. *Environ. Monit. Assess.* **184**, 1953–1971.

Henze, M., Gujer, W., Mino, T. & van Loosdrecht, M. C. M. 2000 *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3*. IWA publishing, Londen, UK.

Hernández-del-Olmo, F., Llanes, F. H. & Gaudioso, E. 2012 An emergent approach for the control of wastewater treatment plants by means of reinforcement learning techniques. *Expert Syst. Appl.* **39**, 2355–2360. https://doi.org/10.1016/j.eswa.2011.08.062.

Hernández-del-Olmo, F., Gaudioso, E., Dormido, R. & Duro, N. 2018 Tackling the start-up of a reinforcement learning agent for the control of wastewater treatment plants. *Knowledge Based Syst.* **144**, 9–15. https://doi.org/10.1016/j.knosys.2017.12.019.

Hu, G., Mao, Z., He, D. & Yang, F. 2011 Hybrid modeling for the prediction of leaching rate in leaching process based on negative correlation learning bagging ensemble algorithm. *Comput. Chem. Eng.* **35**, 2611–2617. https://doi.org/10.1016/j.compchemeng.2011.02.012.

Hvala, N. & Kocijan, J. 2020 Design of a hybrid mechanistic/Gaussian process model to predict full-scale wastewater treatment plant effluent. *Comput. Chem. Eng.* **140**, 106934. https://doi.org/10.1016/j.compchemeng.2020.106934.

Jia, X., Willard, J., Karpatne, A., Read, J. S., Zwart, J. A., Steinbach, M. & Kumar, V. 2021 Physics-guided machine learning for scientific discovery: an application in simulating lake temperature profiles. *ACMIMS Trans. Data Sci.* **2**, 1–26.

Jin, R., Chen, W. & Simpson, T. W. 2001 Comparative studies of metamodelling techniques under multiple modelling criteria. *Struct. Multidiscip. Optim.* **23**, 1–13.

Jin, T., Cai, S., Jiang, D. & Liu, J. 2019 A data-driven model for real-time water quality prediction and early warning by an integration method. *Environ. Sci. Pollut. Res.* **26**, 30374–30385.

Johansen, T. A. & Foss, B. A. 1992 Representing and learning unmodeled dynamics with neural network memories. In: *1992 American Control Conference*. IEEE, pp. 3037–3043, Chicago, IL, USA.

Johansen, Tor A. & Bjarne, A. 1992 Foss. "Representing and Learning Unmodeled Dynamics with Neural Network Memories.". In 1992 American Control Conference, 3037–43. Chicago, IL, USA, 1992. https://doi.org/10.23919/ACC.1992.4792705.

Juntunen, P., Liukkonen, M., Lehtola, M. & Hiltunen, Y. 2013 Cluster analysis by self-organizing maps: an application to the modelling of water quality in a treatment process. *Appl. Soft Comput.* **13**, 3191–3196.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S. & Yang, L. 2021 Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N. & Kumar, V. 2017a Theory-guided data science: a new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.* **29**, 2318–2331.

Karpatne, A., Watkins, W., Read, J. & Kumar, V. 2017b *Physics-guided Neural Networks (pgnn): An Application in Lake Temperature Modeling*. ArXiv Prepr. ArXiv171011431.

Keskin, T. E., Düğenci, M. & Kaçaroğlu, F. 2015 Prediction of water pollution sources using artificial neural networks in the study areas of Sivas, Karabük and Bartın (Turkey). *Environ. Earth Sci.* **73**, 5333–5347.

Keupers, I. & Willems, P. 2015 CSO water quality generator based on calibration to WWTP influent data. In: *Proceedings of the 10th International Conference on Urban Drainage Modelling*. Québec, pp. 97–104.

Kramer, M. A., Thompson, M. L. & Bhagat, P. M. 1992 Embedding theoretical models in neural networks. In: *1992 American Control Conference*. IEEE, pp. 475–479, Chicago, IL, USA.

Kramer, M. A., Thompson, M. L. & Bhagat, P. M. 1992 "Embedding Theoretical Models in Neural Networks." In: *1992 American Control Conference*, 475–79. Chicago, IL, USA, 1992. https://doi.org/10.23919/ACC.1992.4792111.

Larsen, T. A. 2011 Redesigning wastewater infrastructure to improve resource efficiency. *Water Sci. Technol.* **63**, 2535–2541. https://doi.org/10.2166/wst.2011.502.

Larsen, T. A., Hoffmann, S., Lüthi, C., Truffer, B. & Maurer, M. 2016 Emerging solutions to the water challenges of an urbanizing world. *Science* **352**, 928–933. https://doi.org/10.1126/science.aad8641.

Lazăr, D. C., Avram, M. F., Faur, A. C., Goldiş, A., Romoşan, I., Tăban, S. & Cornianu, M. 2020 The impact of artificial intelligence in the endoscopic assessment of premalignant and malignant esophageal lesions: present and future. *Medicina (Mex.)* **56**, 364.

Lee, D. S. & Vanrolleghem, P. A. 2003 Monitoring of a sequencing batch reactor using adaptive multiblock principal component analysis. *Biotechnol. Bioeng.* **82**, 489–497. https://doi.org/10.1002/bit.10589.

Lee, D. S., Vanrolleghem, P. A. & Park, J. M. 2005 Parallel hybrid modeling methods for a full-scale cokes wastewater treatment plant. *J. Biotechnol.* **115**, 317–328. https://doi.org/10.1016/j.jbiotec.2004.09.001.

Le Moullec, Y., Potier, O., Gentric, C. & Leclerc, J. P. 2011 Activated sludge pilot plant: comparison between experimental and predicted concentration profiles using three different modelling approaches. *Water Res.* **45**, 3085–3097. https://doi.org/10.1016/j.watres.2011.03.019.

Li, Z. & Peleato, N. M. 2021 Comparison of dimensionality reduction techniques for cross-source transfer of fluorescence contaminant detection models. *Chemosphere* **276**, 130064.

Li, B., Taniguchi, D., Gedara, J. P., Gogulancea, V., Gonzalez-Cabaleiro, R., Chen, J., McGough, A. S., Ofiteru, I. D., Curtis, T. P. & Zuliani, P. 2019 NUFEB: A massively parallel simulator for individual-based modelling of microbial communities. *PLOS Comput. Biol.* **15**, e1007125. https://doi.org/10.1371/journal.pcbi.1007125.

Lowe, K. S., Tucholke, M. B., Tomaras, J. M. B., Conn, K., Hoppe, C., Drewes, J. E., McCray, J. E. & Munakata-Marr, J. 2010 *Influent Constituent Characteristics of the Modern Waste Stream from Single Sources*. IWA Publishing. https://doi.org/10.2166/9781780403519.

Lund, N. S. V., Falk, A. K. V., Borup, M., Madsen, H. & Steen Mikkelsen, P. 2018 Model predictive control of urban drainage systems: a review and perspective towards smart real-time water management. *Crit. Rev. Environ. Sci. Technol.* **48**, 279–339. https://doi.org/10.1080/10643389.2018.1455484.

Lundberg, S. M. & Lee, S.-I. 2017 A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. pp. 4768–4777.

Luo, L. & Bao, S. 2018 Knowledge-data-integrated sparse modeling for batch process monitoring. *Chem. Eng. Sci.* **189**, 221–232.

Mahmoodian, M., Carbajal, J. P., Bellos, V., Leopold, U., Schutz, G. & Clemens, F. 2018 A hybrid surrogate modelling strategy for simplification of detailed urban drainage simulators. *Water Resour. Manage.* **32**, 5241–5256. https://doi.org/10.1007/s11269-018-2157-4.

Maier, H. R., Morgan, N. & Chow, C. W. 2004 Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters. *Environ. Modell. Software* **19**, 485–494.

Mälzer, H. J. & Nahrstedt, A. 2002 *Modellierung mehrstufiger Trinkwasseraufbereitungsanlagen mittels eines expertensystem-basierten Simulationsmodells (Metrex) am Beispiel von oberflächenwasser*.

Mannina, G., Di Bella, G. & Viviani, G. 2011 An integrated model for biological and physical process simulation in membrane bioreactors (MBRs). *J. Membr. Sci.* **376**, 56–69. https://doi.org/10.1016/j.memsci.2011.04.003.

Manning, C., Raghavan, P. & Schütze, H. 2008 The term vocabulary and postings lists. In: Mogotsi, I. C. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (eds), *Introduction to Information Retrieval*. Cambridge University Press, Cambridge.

Martin, C. & Vanrolleghem, P. A. 2014 Analysing, completing, and generating influent data for WWTP modelling: a critical review. *Environ. Modell. Software* **60**, 188–201. https://doi.org/10.1016/j.envsoft.2014.05.008.

Mašić, A., Srinivasan, S., Billeter, J., Bonvin, D. & Villez, K. 2017 Shape constrained splines as transparent black-box models for bioprocess modeling. *Comput. Chem. Eng.* **99**, 96–105. https://doi.org/10.1016/j.compchemeng.2016.12.017.

Mei, K., Liao, L., Zhu, Y., Lu, P., Wang, Z., Dahlgren, R. A. & Zhang, M. 2014 Evaluation of spatial-temporal variations and trends in surface water quality across a rural-suburban-urban interface. *Environ. Sci. Pollut. Res.* **21**, 8036–8051.

Mitchell, T. 1997 *Machine Learning*. New York, NY, USA: McGraw-Hill.

Moelants, N., Janssen, G., Smets, I. & Van Impe, J. 2008 Field performance assessment of onsite individual wastewater treatment systems. *Water Sci. Technol.* **58**, 1–6. https://doi.org/10.2166/wst.2008.325.

Mohammed, H., Hameed, I. A. & Seidu, R. 2018 Comparative predictive modelling of the occurrence of faecal indicator bacteria in a drinking water source in Norway. *Sci. Total Environ.* **628**, 1178–1190.

Mogotsi, I. C. "Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze: Introduction to Information Retrieval." Information Retrieval 13, no. 2 (April 1, 2010): 192–95. https://doi.org/10.1007/s10791-009-9115-y.

Newhart, K. B., Holloway, R. W., Hering, A. S. & Cath, T. Y. 2019 Data-driven performance analyses of wastewater treatment plants: a review. *Water Res.* **157**, 498–513. https://doi.org/10.1016/j.watres.2019.03.030.

Nordstrand, J. & Dutta, J. 2021 Flexible modeling and control of capacitive-deionization processes through a linear-state-space dynamic Langmuir model. *Npj Clean Water* **4** (1), 1–7. https://doi.org/10.1038/s41545-020-00094-y.

Ni, B.-J. & Yu, H.-Q. 2010 Mathematical modeling of aerobic granular sludge: a review. *Biotechnol. Adv.* **28**, 895–909. https://doi.org/10.1016/j.biotechadv.2010.08.004.

Niu, G., Yi, X., Chen, C., Li, X., Han, D., Yan, B., Huang, M. & Ying, G. 2020 A novel effluent quality predicting model based on genetic-deep belief network algorithm for cleaner production in a full-scale paper-making wastewater treatment. *J. Cleaner Prod.* **265**, 121787. https://doi.org/10.1016/j.jclepro.2020.121787.

Nopens, I., Torfs, E., Ducoste, J., Vanrolleghem, P. A. & Gernaey, K. V. 2015 Population balance models: a useful complementary modelling framework for future WWTP modelling. *Water Sci. Technol.* **71**, 159–167. https://doi.org/10.2166/wst.2014.500.

Ohmura, K., Thürlimann, C. M., Kipf, M., Carbajal, J. P. & Villez, K. 2019 Characterizing long-term wear and tear of ion-selective pH sensors. *Water Sci. Technol.* **80**, 541–550. https://doi.org/10.2166/wst.2019.301.

Olsson, G. & Newell, B. 1999 *Wastewater Treatment Systems*. London, UK: IWA Publishing.

Oyebamiji, O. K., Wilkinson, D. J., Li, B., Jayathilake, P. G., Zuliani, P. & Curtis, T. P. 2019 Bayesian emulation and calibration of an individual-based model of microbial communities. *J. Comput. Sci.* **30**, 194–208. https://doi.org/10.1016/j.jocs.2018.12.007.

Pang, J., Yang, S., He, L., Chen, Y. & Ren, N. 2019 Intelligent control/operational strategies in WWTPs through an integrated Q-learning algorithm with ASM2d-Guided reward. *Water* **11**, 927. https://doi.org/10.3390/w11050927.

Panidhapu, A., Li, Z., Aliashrafi, A. & Peleato, N. M. 2020 Integration of weather conditions for predicting microbial water quality using Bayesian Belief Networks. *Water Res.* **170**, 115349.

Penn, R., Schütze, M., Gorfine, M. & Friedler, E. 2017 Simulation method for stochastic generation of domestic wastewater discharges and the effect of greywater reuse on gross solid transport. *Urban Water J.* **14**, 846–852. https://doi.org/10.1080/1573062X.2017.1279188.

Perelman, L., Arad, J., Housh, M. & Ostfeld, A. 2012 Event detection in water distribution systems from multivariate water quality time series. *Environ. Sci. Technol.* **46**, 8212–8219.

Peres, J., Oliveira, R. & De Azevedo, S. F. 2001 Knowledge based modular networks for process modelling and control. *Comput. Chem. Eng.* **25**, 783–791.

Psichogios, D. C. & Ungar, L. H. 1992 A hybrid neural network-first principles approach to process modeling. *AIChE J.* **38**, 1499–1511.

Quaghebeur, W., Nopens, I. & De Baets, B. 2021 Incorporating unmodeled dynamics into first-principles models through machine learning. *IEEE Access* **9**, 22014–22022. https://doi.org/10.1109/ACCESS.2021.3055353.

Quinlan, J. R. 1990 Decision trees and decision-making. *IEEE Trans. Syst. Man Cybern.* **20**, 339–346. https://doi.org/10.1109/21.52545.

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. https://doi.org/10.1016/j.watres.2022.118166

Raissi, M., Perdikaris, P. & Karniadakis, G. E. 2017 Machine learning of linear differential equations using Gaussian processes. *J. Comput. Phys.* **348**, 683–693.

Raissi, M., Perdikaris, P. & Karniadakis, G. E. 2019 Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707.

Razavi, S. 2021 Deep learning, explained: fundamentals, explainability, and bridgeability to process-based modelling. *Environ. Modell. Software* **144**, 105159.

Razavi, S., Tolson, B. A. & Burn, D. H. 2012 Review of surrogate modeling in water resources. *Water Resour. Res.* **48**. https://doi.org/10.1029/2011WR011527.

Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L. & Vanrolleghem, P. A. 2007 Uncertainty in the environmental modelling process – a framework and guidance. *Environ. Modell. Software* **22**, 1543–1556. https://doi.org/10.1016/j.envsoft.2007.02.004.

Regmi, P., Stewart, H., Amerlinck, Y., Arnell, M., García, P. J., Johnson, B., Maere, T., Miletić, I., Miller, M., Rieger, L., Samstag, R., Santoro, D., Schraa, O., Snowling, S., Takács, I., Torfs, E., van Loosdrecht, M. C. M., Vanrolleghem, P. A., Villez, K., Volcke, E. I. P., Weijers, S., Grau, P., Jimenez, J. & Rosso, D. 2019 The future of WRRF modelling – outlook and challenges. *Water Sci. Technol.* **79**, 3–14. https://doi.org/10.2166/wst.2018.498.

Reynaert, E., Hess, A. & Morgenroth, E. 2021 Making waves: why water reuse frameworks need to co-evolve with emerging small-scale technologies. *Water Res. X* **11**, 100094. https://doi.org/10.1016/j.wroa.2021.100094.

Rieger, L., Gillot, S., Langergraber, G., Ohtsuki, T., Shaw, A., Takacs, I. & Winkler, S. 2012 *Guidelines for Using Activated Sludge Models*. London, UK: IWA publishing.

Rodriguez-Roda, I. R., Sànchez-Marrè, M., Comas, J., Baeza, J., Colprim, J., Lafuente, J., Cortes, U. & Poch, M. 2002 A hybrid supervisory system to support WWTP operation: implementation and validation. *Water Sci. Technol.* **45**, 289–297. https://doi.org/10.2166/wst.2002.0608.

Romijn, R., Özkan, L., Weiland, S., Ludlage, J. & Marquardt, W. 2008 A grey-box modeling approach for the reduction of nonlinear systems. *J. Process Control* **18**, 906–914. https://doi.org/10.1016/j.jprocont.2008.06.007.

Russo, S., Lürig, M., Hao, W., Matthews, B. & Villez, K. 2020 Active learning for anomaly detection in environmental data. *Environ. Modell. Software* **134**, 104869. https://doi.org/10.1016/j.envsoft.2020.104869.

Sahoo, G. B., Ray, C. & Wade, H. F. 2005 Pesticide prediction in ground water in North Carolina domestic wells using artificial neural networks. *Ecol. Modell.* **183**, 29–46.

Samek, W., Wiegand, T. & Müller, K.-R. 2017 *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*. ArXiv170808296 Cs Stat.

Samuelsson, O., Olsson, G., Lindblom, E., Björk, A. & Carlsson, B. 2021 Sensor bias impact on efficient aeration control during diurnal load variations. *Water Sci. Technol.* **83**, 1335–1346. https://doi.org/10.2166/wst.2021.031.

Schneider, M. Y., Carbajal, J. P., Furrer, V., Sterkele, B., Maurer, M. & Villez, K. 2019 Beyond signal quality: the value of unmaintained pH, dissolved oxygen, and oxidation-reduction potential sensors for remote performance monitoring of on-site sequencing batch reactors. *Water Res.* **161**, 639–651. https://doi.org/10.1016/j.watres.2019.06.007.

Schneider, M. Y., Furrer, V., Sprenger, E., Carbajal, J. P., Villez, K. & Maurer, M. 2020 Benchmarking soft sensors for remote monitoring of on-site wastewater treatment plants. *Environ. Sci. Technol.* **54**, 10840–10849. https://doi.org/10.1021/acs.est.9b07760.

Schubert, J., Simutis, R., Dors, M., Havlík, I. & Lübbert, A. 1994 Hybrid modelling of yeast production processes–combination of a priori knowledge on different levels of sophistication. *Chem. Eng. Technol. Ind. Chem.-Plant Equip.-Process Eng.-Biotechnol.* **17**, 10–20.

Sedlak, D. 2014 *Water 4.0: The Past, Present, and Future of the World's Most Vital Resource*. Yale University Press, New Haven, CT, USA.

Shapley, L. S. 2016 *17. A Value for n-person Games*. 307–18, Princeton University Press, 1953. Princeton. https://doi.org/10.1515/9781400881970-018.

Shi, S. & Xu, G. 2018 Novel performance prediction model of a biofilm system treating domestic wastewater based on stacked denoising auto-encoders deep learning network. *Chem. Eng. J.* **347**, 280–290. https://doi.org/10.1016/j.cej.2018.04.087.

Sharif Shourjeh, M., Kowal, P., Lu, X., Xie, L. & Drewnowski, J. 2021Development of strategies for AOB and NOB competition supported by mathematical modeling in terms of successful deammonification implementation for energy-efficient WWTPs. *Processes* **9** (3), 562. https://doi.org/10.3390/pr9030562

Sin, G. & Al, R. 2021 Activated sludge models at the crossroad of artificial intelligence – a perspective on advancing process modeling. *Npj Clean Water* **4**, 1–7. https://doi.org/10.1038/s41545-021-00106-5.

Sitzenfrei, R., Hillebrand, S. & Rauch, W. 2017 Investigating the interactions of decentralized and centralized wastewater heat recovery systems. *Water Sci. Technol.* **75**, 1243–1250. https://doi.org/10.2166/wst.2016.598.

Stedmon, C. A., Seredyńska-Sobecka, B., Boe-Hansen, R., Le Tallec, N., Waul, C. K. & Arvin, E. 2011 A potential approach for monitoring drinking water quality from groundwater systems using organic matter fluorescence as an early warning for contamination events. *Water Res.* **45**, 6030–6038.

Su, H.-T., Bhat, N., Minderman, P. & McAvoy, T. 1993 Integrating neural networks with first principles models for dynamic modeling. In: *Dynamics and Control of Chemical Reactors, Distillation Columns and Batch Processes*. Elsevier, pp. 327–332. IFAC Symposia Series. Oxford: Pergamon, 1993. https://doi.org/10.1016/B978-0-08-041711-0.50054-4.

Su, Hong-Te, N. Bhat, P. A. Minderman, and T. J. McAvoy. "Integrating Neural Networks with First Principle Modles for Dynamic Modeling." In Dynamics and Control of Chemical Reactors, Distillation Columns and Batch Processes, edited by J. G. Balchen, 327–32. IFAC Symposia Series. Oxford: Pergamon, 1993. https://doi.org/10.1016/B978-0-08-041711-0.50054-4

Sutton, R. S. & Barto, A. G. 2018 *Reinforcement Learning: An Introduction*. Cambridge: MIT press.

Takács, I., Patry, G. G. & Nolasco, D. 1991 A dynamic model of the clarification-thickening process. *Water Res.* **25**, 1263–1271. https://doi.org/10.1016/0043-1354(91)90066-Y.

Talebizadeh, M., Belia, E. & Vanrolleghem, P. A. 2016 Influent generator for probabilistic modeling of nutrient removal wastewater treatment plants. *Environ. Modell. Software* **77**, 32–49. https://doi.org/10.1016/j.envsoft.2015.11.005.

Tesoriero, A. J., Gronberg, J. A., Juckem, P. F., Miller, M. P. & Austin, B. P. 2017 Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification. *Water Resour. Res.* **53**, 7316–7331.

Therrien, J.-D., Nicolaï, N. & Vanrolleghem, P. A. 2020 A critical review of the data pipeline: how wastewater system operation flows from data to intelligence. *Water Sci. Technol.* **82**, 2613–2634. https://doi.org/10.2166/wst.2020.393.

Thompson, M. L. & Kramer, M. A. 1994 Modeling chemical processes using prior knowledge and neural networks. *AIChE J.* **40**, 1328–1340. https://doi.org/10.1002/aic.690400806.

Thürlimann, C. M., Udert, K. M., Morgenroth, E. & Villez, K. 2019 Stabilizing control of a urine nitrification process in the presence of sensor drift. *Water Res.* **165**, 114958. https://doi.org/10.1016/j.watres.2019.114958.

Torfs, E., Dutta, A. & Nopens, I. 2012 Investigating kernel structures for Ca-induced activated sludge aggregation using an inverse problem methodology. In: *4th Int. Conf. Popul. Balance Model*. Vol. 70, pp. 176–187. https://doi.org/10.1016/j.ces.2011.06.069.

Tsen, A. Y.-D., Jang, S. S., Wong, D. S. H. & Joseph, B. 1996 Predictive control of quality in batch polymerization using hybrid ANN models. *AIChE J.* **42**, 455–465.

Tulleken, H. J. 1993 Grey-box modelling and identification using physical knowledge and Bayesian techniques. *Automatica* **29**, 285–308.

Van Can, H. J., Hellinga, C., Luyben, K. C. A., Heijnen, J. J. & Te Braake, H. A. 1996 Strategy for dynamic process modeling based on neural networks in macroscopic balances. *AIChE J.* **42**, 3403–3418.

Van der Helm, A. W. C. & Rietveld, L. C. 2002 Modelling of drinking water treatment processes within the Stimela environment. *Water Sci. Technol. Water Supply* **2**, 87–93.

van Loosdrecht, M. C. M. & Brdjanovic, D. 2014 Anticipating the next century of wastewater treatment. *Science* **344**, 1452–1453. https://doi.org/10.1126/science.1255183.

Vanrolleghem, P. A. & Lee, D. S. 2003 On-line monitoring equipment for wastewater treatment processes: state of the art. *Water Sci. Technol.* **47**, 1–34. https://doi.org/10.2166/wst.2003.0074.

Vanrolleghem, P. A., Benedetti, L. & Meirlaen, J. 2005 Modelling and real-time control of the integrated urban wastewater system. *Vulnerability Water Qual. Intensiv. Dev. Urban Watersheds* **20**, 427–442. https://doi.org/10.1016/j.envsoft.2004.02.004.

Versteegh, J. F. M., Van Gaalen, F. W., Rietveld, L. C., Evers, E. G., Aldenberg, T. A. & Cleij, P. 2001 *TAPWAT: Definition Structure and Applications for Modelling Drinking Water Treatment*.

Villez, K., Ruiz, M., Sin, G., Colomer, J., Rosén, C. & Vanrolleghem, P. A. 2008 Combining multiway principal component analysis (MPCA) and clustering for efficient data mining of historical data sets of SBR processes. *Water Sci. Technol.* **57**, 1659. https://doi.org/10.2166/wst.2008.143.

Villez, K., Billeter, J. & Bonvin, D. 2019 Incremental parameter estimation under Rank-Deficient measurement conditions. *Processes*. https://doi.org/10.3390/pr7020075.

Villez, K., Vanrolleghem, P. A. & Corominas, L. 2020 A general-purpose method for Pareto optimal placement of flow rate and concentration sensors in networked systems – with application to wastewater treatment plants. *Comput. Chem. Eng.* **139**. https://doi.org/10.1016/j.compchemeng.2020.106880.

von Stosch, M., Oliveira, R., Peres, J. & Feyo de Azevedo, S. 2014 Hybrid semi-parametric modeling in process systems engineering: past, present and future. *Comput. Chem. Eng.* **60**, 86–101. https://doi.org/10.1016/j.compchemeng.2013.08.008.

Wade, M. 2004 *Process Monitoring and Knowledge Extraction in Wastewater Treatment Plants. PhD Thesis*.

Wade, M. J., Sánchez, A. & Katebi, M. R. 2005 On real-time control and process monitoring of wastewater treatment plants: real-time process monitoring. *Trans. Inst. Meas. Control* **27**, 173–193. London, UK. https://doi.org/10.1191/0142331205tm140oa.

Wade, M. J., Keedwell, E. C., Steyer, J.-P. & Ruano Garcia, M. V. 2021 *Making Water Smart, In Focus – Special Book Series*. IWA Publishing.

Wang, Z. & Georgakis, C. 2019 A dynamic response surface model for polymer grade transitions in industrial plants. *Ind. Eng. Chem. Res.* **58**, 11187–11198.

Wang, X., Zhang, J. & Babovic, V. 2016 Improving real-time forecasting of water quality indicators with combination of process-based models and data assimilation technique. *Ecol. Indic.* **66**, 428–439.

Wang, X., Kvaal, K. & Ratnaweera, H. 2019 Explicit and interpretable nonlinear soft sensor models for influent surveillance at a full-scale wastewater treatment plant. *J. Process Control* **77**, 1–6. https://doi.org/10.1016/j.jprocont.2019.03.005.

Wärff, C., Arnell, M., Sehlén, R. & Jeppsson, U. 2020 Modelling heat recovery potential from household wastewater. *Water Sci. Technol.* **81**, 1597–1605. https://doi.org/10.2166/wst.2020.103.

Wright, L. & Davidson, S. 2020 How to tell the difference between a model and a digital twin. *Advanced Modeling and Simulation in Engineering Sciences* **7** (1), 13. https://doi.org/10.1186/s40323-020-00147-4.

WOS 2021 *Web of Science Core Collection Help*.

Wu, G.-D. & Lo, S.-L. 2008 Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. *Eng. Appl. Artif. Intell.* **21**, 1189–1195.

Yang, M., del Pozo, D. F., Torfs, E., Rehman, U., Yu, D. & Nopens, I. 2021 Numerical simulation on the effects of bubble size and internal structure on flow behavior in a DAF tank: a comparative study of CFD and CFD-PBM approach. *Chem. Eng. J. Adv.* **7**, 100131. https://doi.org/10.1016/j.ceja.2021.100131.

Zhu, J.-J. & Anderson, P. R. 2019 Performance evaluation of the ISMLR package for predicting the next day's influent wastewater flowrate at Kirie WRP. *Water Sci. Technol.* **80**, 695–706. https://doi.org/10.2166/wst.2019.309.

Zhu, J.-J., Segovia, J. & Anderson, P. R. 2015 Defining influent scenarios: application of cluster analysis to a water reclamation plant. *J. Environ. Eng.* **141**, 04015005. https://doi.org/10.1061/(ASCE)EE.1943-7870.0000934.

Zhu, J.-J., Kang, L. & Anderson, P. R. 2018 Predicting influent biochemical oxygen demand: balancing energy demand and risk management. *Water Res.* **128**, 304–313. https://doi.org/10.1016/j.watres.2017.10.053.

Zhu, J.-J., Dressel, W., Pacion, K. & Ren, Z. J. 2021 *ES&t* in the 21st century: a data-driven analysis of research topics, interconnections, and trends in the past 20 years. *Environ. Sci. Technol.* **55**, 3453–3464. https://doi.org/10.1021/acs.est.0c07551.

First received 10 January 2022; accepted in revised form 24 March 2022. Available online 6 April 2022