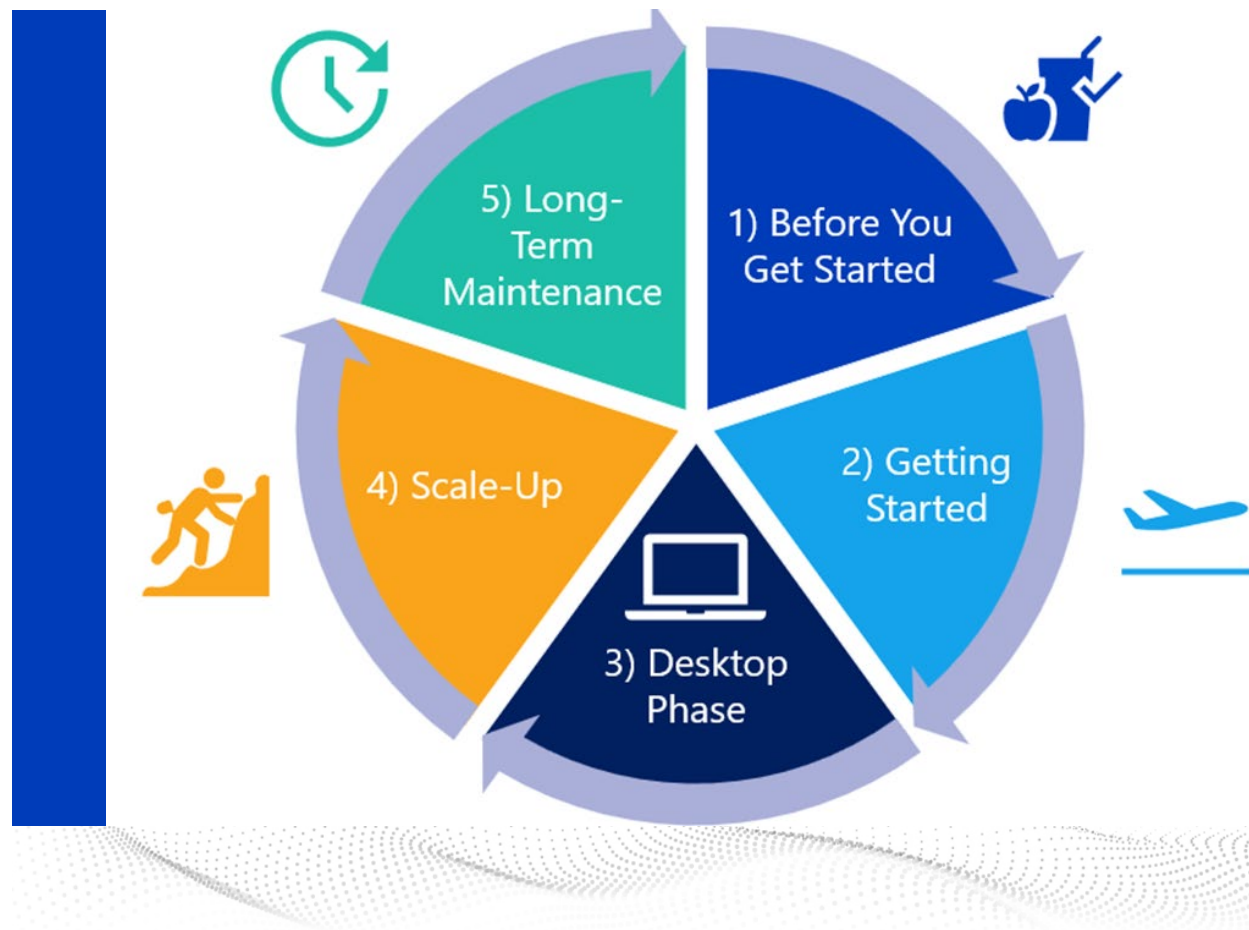




— BUREAU OF —
RECLAMATION

U.S. BUREAU OF RECLAMATION

Data-Driven Fault Detection and Process Control for Potable Reuse with Reverse Osmosis and Membrane Bioreactors



GUIDEBOOK

AI & Machine Learning Guidebook for Potable Reuse

May 2026 / FINAL



— BUREAU OF —
RECLAMATION

U.S. BUREAU OF RECLAMATION

Data-Driven Fault Detection and Process Control for Potable Reuse with Reverse Osmosis and Membrane Bioreactors

GUIDEBOOK

AI & Machine Learning Guidebook for Potable Reuse

May 2026 / FINAL

Prepared by:

Yoko Koyama, PE

Kyle Thompson, PhD, PE

Charles Ramey

Elise Moore, PE

Norm Anderson, PE

Acknowledgements

Principal Investigators

Andrew Salveson, PE
Carollo Engineers, Inc.

Yasuhiro Matsui, PhD, PE.Jp
Yokogawa

Kevin Hardy, JD
National Water Research Institute

Mandy Hering, PhD
Baylor University

Project Team

Kyle Thompson, PhD, PE
Carollo Engineers, Inc.

Amos Branch, PhD, PE
Carollo Engineers, Inc.

Yoko Koyama, PE
Carollo Engineers, Inc.

Carli Brucker, PhD
Carollo Engineers, Inc.

Charlie He, PE
Carollo Engineers, Inc.

Farshid Shoushtarian, PhD
Yokogawa

Rakheon Kim, PhD
Baylor University

Suzanne Sharkey
National Water Research Institute

Tianna Manzon
National Water Research Institute

NWRI Utility Advisory Board

Amber Boone
South Orange County Wastewater Authority (SOCWA)

Jesus Gonzalez
Los Angeles Department of Water and Power (LADWP)

Joe McDermott
Las Virgenes Municipal Water District (LVMWD)

Hubertus Cox
Los Angeles Sanitation & Environment (LASAN)

Hossein Ashktorab, PhD
Valley Water

NWRI Independent Expert Panel

Kathryn Newhart, PhD (Chair)
United States Military Academy

Amy Childress, PhD
University of Southern California

Zakir Hirani, PE
Stantec

Mehul Patel, PE
Orange County Water District

Kara Nelson, PhD
University of California, Berkeley

USBR Staff

Neal Gallagher, PE
US Bureau of Reclamation

Contents

EXECUTIVE SUMMARY	1
SECTION 1 BEFORE YOU GET STARTED	2
1.1 Why use Artificial Intelligence or Machine Learning	2
1.2 What are Artificial Intelligence and Machine Learning?	4
1.2.1 Artificial Intelligence vs Machine Learning	6
1.2.2 Statistics	8
1.2.3 Data Analytics	10
1.3 Types of ML	10
1.3.1 Supervised Learning	11
1.3.2 Unsupervised Learning	13
1.3.3 Reinforcement Learning	15
1.4 Terminologies	15
1.4.1 Inputs and Outputs	15
1.4.2 Training Set and Testing Set	15
1.5 Machine Learning Application Styles for Potable Reuse	15
1.6 Treatment Process Prioritization for Machine Learning	17
1.7 AI/ML Adoption Barriers	19
1.7.1 Overview	19
1.7.2 Data Management	19
1.8 Step-by-Step Machine Learning Overview	23
1.9 Partnership Strategy	26
1.10 Data Science Tools for Desktop Studies	27
1.11 The Cloud	28
1.11.1 What is the Cloud	28
1.11.2 Leveraging the Power of Cloud	28
1.11.3 Cybersecurity Concerns and Solutions	29
1.12 Edge Computing for Machine Learning	31
SECTION 2 DESKTOP PHASE MODELING AND EVALUATION	31
2.1 Metrics for Measuring ML Success	31
2.2 Data Splitting: Essential for Real-World Machine Learning Success	32
2.3 Model Cross-Validation	35
2.4 Data Leakage	36
SECTION 3 SCALE-UP, PILOTING, AND DEMONSTRATION	37
3.1 Scale-Up Pathways	37
3.2 Choosing the Right Scale-Up Pathway	37
3.3 Beyond Test Set Prediction: The Pilot Imperative	39

SECTION 4	THE LONG TERM	41
4.1	Overview	41
4.2	ML Solutions Deployment and Maintenance	42
4.2.1	The MLOps Paradigm	43
4.2.2	Building or Acquiring MLOps Capability	44
4.2.3	Technical Infrastructure for ML Deployment	45
4.3	Resource Requirements for Long-Term Success	45
4.3.1	Derisking AI/ML Implementation	46
4.3.2	Technical Infrastructure	47
4.3.3	Workforce and Organizational Readiness	49
4.3.4	Cost Considerations	49
4.4	What's Ahead	50
SECTION 5	REFERENCES	51

Tables

Table ES-1	Key Benefits of ML Implementation	1
Table 1-1	Comparison of Synonymous Terms for Input and Output in Statistics versus Machine Learning	15
Table 1-2	Machine Learning Application Styles for Potable Reuse	16
Table 1-3	Partnership strategy comparison	26
Table 1-4	Characteristics of the Programming Languages R and Python for Data Analysis and ML	27
Table 4-1	List of evaluated technologies for Machine Learning Operations platforms and tools, categorized by open-source and commercial examples (Kreuzberger et al., 2023)	45

Figures

Figure 1-1	Historical evolution of Artificial Intelligence	4
Figure 1-2	Conceptual Process Control Approach Applying Machine Learning and Artificial Intelligence	5
Figure 1-3	Relationship among Artificial Intelligence, Machine Learning, Deep Learning, Large Language Models, Natural Language Processing, Computer Vision, Data Science, Data Analytics, and Statistics.	6
Figure 1-4	The Difference between Mechanistic and Machine Learning Modeling	7
Figure 1-5	Types of Machine Learning and The Types of Mathematical Questions They Address	10
Figure 1-6	Data management framework for ML— Data management that is strong in each element of the framework will produce a more usable model	20
Figure 1-7	Improved data management de-silos data among utility departments	21
Figure 1-8	Data management hierarchy of needs pyramid	21
Figure 1-9	Machine learning solution development process	23

Figure 2-1	Illustration of models that are underfit, balanced fit, or overfit	33
Figure 2-2	K-Fold Cross-Validation Structure	34
Figure 4-1	Roles and their intersections contributing to the MLOps paradigm	43
Figure 4-2	Four interconnected categories of AI/ML adoption barriers identified in nuclear power; adapted from Hall and Agarwal (2024), applicable to water utilities facing similar regulatory and operational constraints	46

Abbreviations

AI	artificial intelligence
AMI	advanced metering infrastructure
AWS	Amazon Web Service
CMMS	computerized maintenance management system
CV	cross-validation
DL	deep learning
GAC	granular activated carbon
GIS	geographic information system
GCP	Google Cloud Platform
IoT	Internet of things
LIMS	laboratory information management system
LLM	large language models
ML	machine learning
MLOps	ML operations
OT	operational technology
PCA	principal component analysis
PFAS	per- and polyfluoroalkyl substances
R ²	R-squared
RMSE	Root Mean Squared Error
RO	reverse osmosis
ROI	return of investment
RTUs	remote terminal units
SCADA	supervisory control and data acquisition
TOC	total organic carbon
UF	Ultrafiltration
WRF	Water Research Foundation

EXECUTIVE SUMMARY

This guidebook serves as a resource for water utilities seeking to implement artificial intelligence (AI) and machine learning (ML) technologies in potable reuse operations. Herein, we define AI as machines with the ability to think, reason, and make decisions. Furthermore, we consider ML to be a sub-field as AI, defined a method within AI where algorithms learn from large data to identify patterns. Herein, we have done our best to differentiate between AI more broadly or ML specifically, or use “AI/ML” for sentences that could apply to both. While the focus of this Guidebook is on examples and applications for potable reuse, much of the same principles would apply just as well to conventional wastewater and drinking water treatment. Furthermore, we define the scope of potable reuse broadly to include wastewater flow and effluent quality monitoring, since these have direct implications for the overall economics and efficiency of water reuse projects. Integrating ML into potable reuse operation is a significant opportunity to enhance treatment efficiency and improve resource management while maintaining critical oversight by operations staff (Table ES-1).

Table ES-1 Key Benefits of ML Implementation

Characteristic of Machine Learning	Benefit
Data-Driven – Makes decisions based on data, not theory	Makes best use of the information available Leverages consistent correlations Can avoid certain human biases such as anchoring (i.e., over-reliance on first information received)
Adaptive – Can continue to learn over time. If used frequently or automatically, it can respond smoothly to sudden changes.	If retrained, the accuracy will increase over time Can improve resilience in emergencies as well as efficiency on easier days
Multivariate – Makes decisions based on many sources of information	Can consider more sources of information at once than a human Can balance out or ignore sensor errors based on other sources of information
Non-linear – Fits models to data without assuming straight lines	Works relatively well even in complex or non-linear systems like biological treatment

Upfront investment in data management that enables ML is significant. However, completed research demonstrates significant long-term savings that extend beyond the initial costs of implementation and integration. Meanwhile, there are costs to *not* investing in data management and ML, such as missing out on long-term savings from more energy-efficient operation and avoided equipment downtime. There are also benefits of using AI/ML that are harder to quantify, such as facilitating knowledge transfer from retiring workforce or instilling employees with the excitement of working at a leading-edge organization.

Guidebook Structure

The guidebook follows a systematic progression to mirror the key stages of the ML life cycle.

Chapter 1: Before You Get Started - Provides foundational knowledge about ML concepts, terminologies, and applications in advanced water treatment, along with data management fundamentals.

Chapter 2: Getting Started - Introduces a step-by-step ML implementation approach, data types, partnership strategies, and tools for desktop studies and cloud services that encompass a whole ML life cycle. In this chapter, we highlighted the four-step approach in starting an ML tool:

- **Problem Definition:** Identify specific operational challenges or opportunities for optimization that ML can address.
- **Data Collection:** Gather relevant data from diverse sources (supervisory control and data acquisition [SCADA], laboratory information management system [LIMS], and etc.).
- **Desktop Study:** Develop and test ML models.
- **Scale-Up:** Operationalize validated models.

Chapter 3: Evaluation - Covers model evaluation techniques during the desktop phase and piloting and demonstration strategies.

Chapter 4: The Long Term - Addresses solution deployment, maintenance requirements, and resource considerations for sustainable ML operations.

Key Takeaways

- Strong data management is crucial for ML implementation.
- Breaking down data silos between departments (e.g., operations, maintenance, laboratory) allows the integration of more data to the models in real-time or near-real-time. This improves the models especially if predicting something that does not have sensors available (e.g., soft sensors), or if useful input information does not have real-time sensors available (e.g., microbial plate counts).
- For proper modeling evaluation, data scientists should be encouraged not just to provide the most accurate prediction, but also to provide the fairest assessment of the confidence in that accuracy.
- The appropriate scale-up pathway (e.g., desktop, on-premise, cloud) depends on how frequently the model should be acted upon (e.g., continuous disinfection vs long-term membrane fouling).
- long-term success requires both technical infrastructure (e.g., data pipelines) and human expertise (e.g., data scientists). Operator involvement is critical, providing essential feedback for system usefulness, usability, safety, accuracy, and refinement.

SECTION 1 BEFORE YOU GET STARTED

1.1 Why use Artificial Intelligence or Machine Learning

The water sector faces rising operational demands while struggling to keep pace with organizational capacity. Competitive, fast-moving fields such as finance, pharma, and oil & gas have already widely adopted ML and AI. Manufacturing industries have seen 30-90 percent productivity gains (Ahmed et al., 2022; Plathottam et al., 2023; Sircar et al., 2021; Vora et al., 2023). AI/ML is now rapidly growing in the water sector as well (Vekaria & Sinha, 2024), with numerous promising applications including predicting de facto reuse (K. Thompson & Dickenson, 2021), predicting total organic carbon (TOC) removal by coagulation (Zhang et al., 2025) and predicting per- and polyfluoroalkyl substances (PFAS) removal from granular activated carbon (GAC) (Koyama et al. 2024).

Large language models (LLMs) like ChatGPT have brought unprecedented public awareness and excitement about AI, including among water professionals. LLMs are not yet directly applicable to control systems and alerts, the focus of this guidebook. Nonetheless, some of the principles and questions around LLMs also apply to ML for water process control. For example, (1) how to use it while respecting data privacy and security; and (2) how to QA/QC ML results in light of the fact that it can be highly accurate, yet also imperfect?

Water utilities encounter significant adoption barriers, leaving much of AI/ML's potential unrealized. Over half of water sector technology buyers report shortages in skilled staff, poor ML tools, data issues, and trust concerns (Sinha et al., 2025). Four main challenge areas for AI adoption in the energy sector were outlined in a previous study (Park, 2025): technical (data, cybersecurity, explainability), economic (cost, return of investment [ROI]), operational (integration, legacy systems), and workforce/social (skills gaps, ethics, public trust). These same types of barriers apply in the water sector. Critical infrastructure like energy and water must reliably reach strict regulatory requirements while contending with technical hurdles, high costs, and regulatory complexities (Chandrasekeran, 2026a; Hall & Agarwal, 2024; Park, 2025).

These barriers are worsened by an aging workforce (Chandrasekeran, 2026b), with one-third of water employees nearing retirement, risking loss of vital knowledge (Sinha et al., 2025). Utilities also spend 21 percent of their AI/ML life cycle just collecting and preparing data, limiting time for actual analytics (Sinha et al., 2025). Despite abundant operational data from SCADA, metering, and internet of things (IoT) sensors, most utilities lack the means to turn this data into actionable insights. Such capabilities are concentrated in wealthier, urban utilities with better resources (Sela et al., 2025), creating a divide where most utilities miss out on digital transformation benefits.

Today, water utilities face a critical decision regarding ML adoption. Costs for meaningful ML adoption are significant, including more sensors and their maintenance; data management; and model development and retraining. The benefits of ML for process control may appear as a marginal percentage, particular for utilities fortunate with relatively stable influent quality. Because ML is very site-specific, the exact benefits are hard to know until after a site-specific investigation. On the other hand, delaying the implementation of ML technologies could represent significant opportunity costs in terms of operational efficiency and resource optimization. Park (2025) demonstrates that while ROI uncertainty affects AI adoption, strategic approaches—including comprehensive scenario planning, phased pilot projects, and innovative financing through public-private partnerships—can effectively de-risk AI investment. These realized savings can then be strategically redirected to fund other essential facility upgrades, such as infrastructure repairs or capacity expansions, creating a compounding benefit for the utility's overall operations and financial health. This approach transforms ML investment from a cost center into a strategic enabler for broader facility improvements.

This mismatch between data availability and utilization signals a need for us to better raise our awareness on AI/ML to bridge the gap between where water utilities are and where they need to be. Cross-industry case studies broadly illustrate the barriers to and benefits from AI/ML adoption. In the rest of the guide, we aim to outline practical guidance for AI/ML adoption in water, grounded in potable reuse examples.

1.2 What are Artificial Intelligence and Machine Learning?

AI Has Been Evolving for Decades *But the Last Three Years Changed Everything*

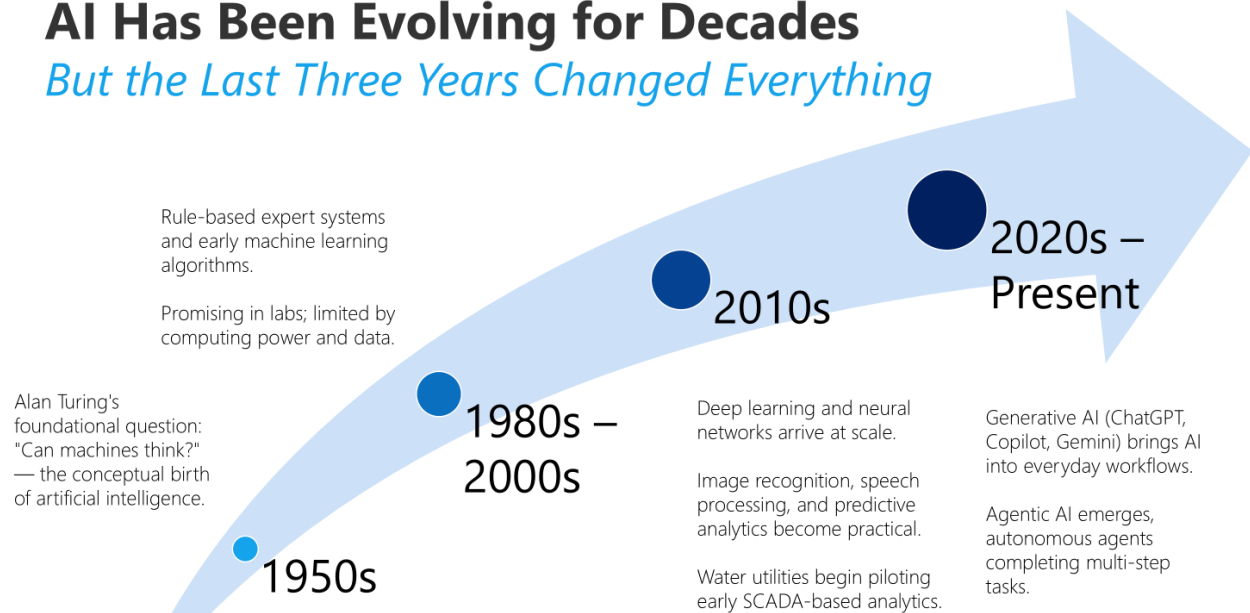


Figure 1-1 Historical evolution of Artificial Intelligence

To frame what follows, it helps to understand where AI and ML came from — because the technology we're hearing about today didn't appear overnight (Figure 1-1). The conceptual foundations go back to Alan Turing in the 1950s. What's changed dramatically in the last three years isn't the underlying theory — it's the scale, the accessibility, and the speed of deployment. The launch of generative AI tools like ChatGPT in late 2022 created a step-change in public awareness. Suddenly, AI wasn't just something that happened inside Google's data centers — it was in people's email drafts and meeting summaries. For utilities, the pressure to "do something with AI" intensified almost overnight. Our job today is to separate the durable signal from the short-term noise — and understand which parts of this evolution are genuinely relevant to water operations.

Figure 1-1 illustrates the relationship between AI and ML in an operational decision-making context. ML trains models using data to make predictions, such as forecasting contaminant concentration, equipment status, or future water flow. AI builds on this by simulating human intelligence to recommend or take action based on those predictions—for example, determining the required UV fluence or oxidant dose in response to a forecasted contaminant level. So far, in potable reuse and broader water applications, the focus has typically been on the ML step, which often suffices. The scope of this document focused on semi-autonomous ML implementation. However, there has been recent growing discussion of agentic AI for the water sector, which would be a paradigm shift towards fully autonomous control (Hosseini & Seilani, 2025; Lesnik, 2025).

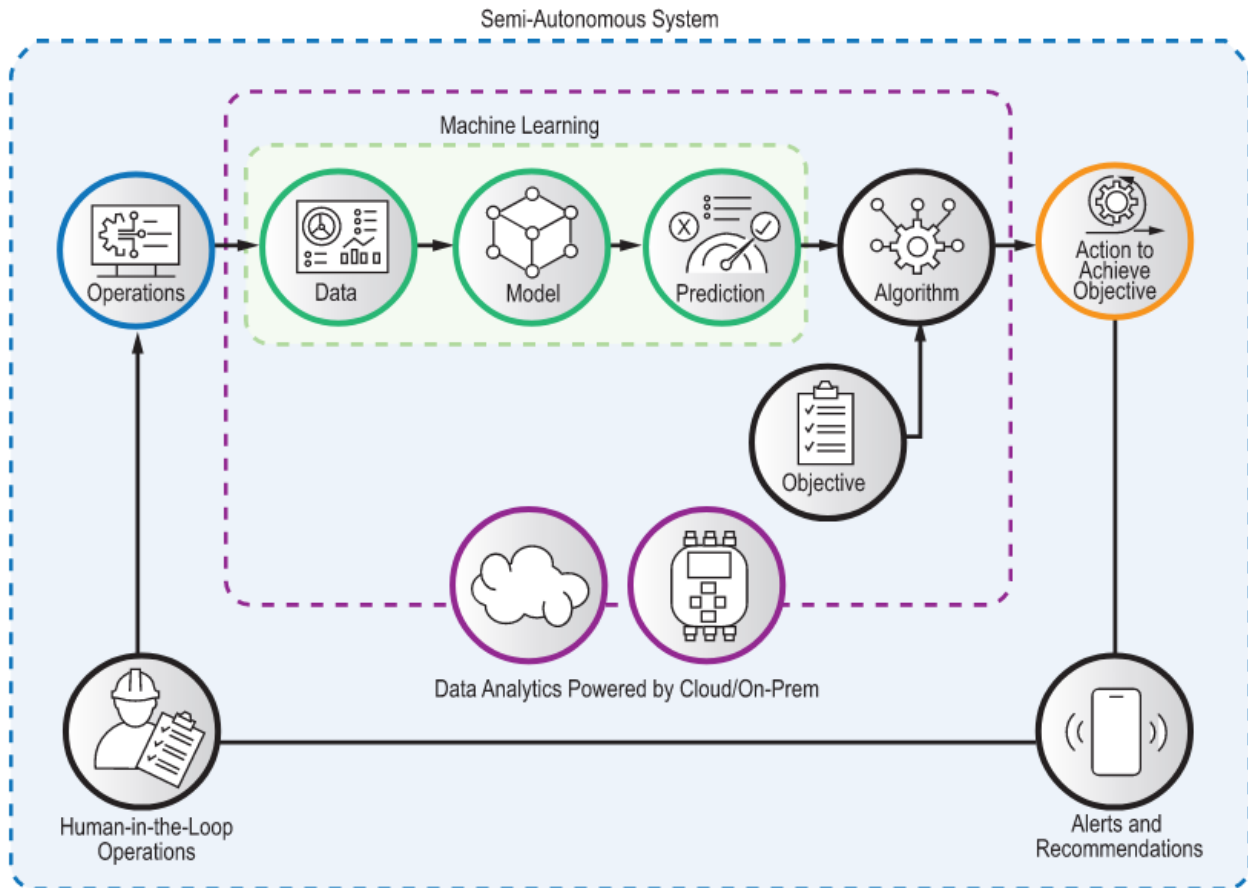


Figure 1-2 Conceptual Process Control Approach Applying Machine Learning and Artificial Intelligence

The relationships among AI, ML, and other related concepts, are illustrated in Venn diagram in Figure 1-2. Specifically, it positions AI as encompassing ML—which includes Deep Learning (DL)—and subsets such as Natural Language Processing, LLM, and Computer Vision. Data Science integrates Statistics and Data Analytics, with intersections highlighting synergies for applications like predictive modeling and site-specific optimization. For the next chapter and on, we focus ML due to more demonstrated applications for reuse monitoring and control, as opposed to Generative AI solutions.

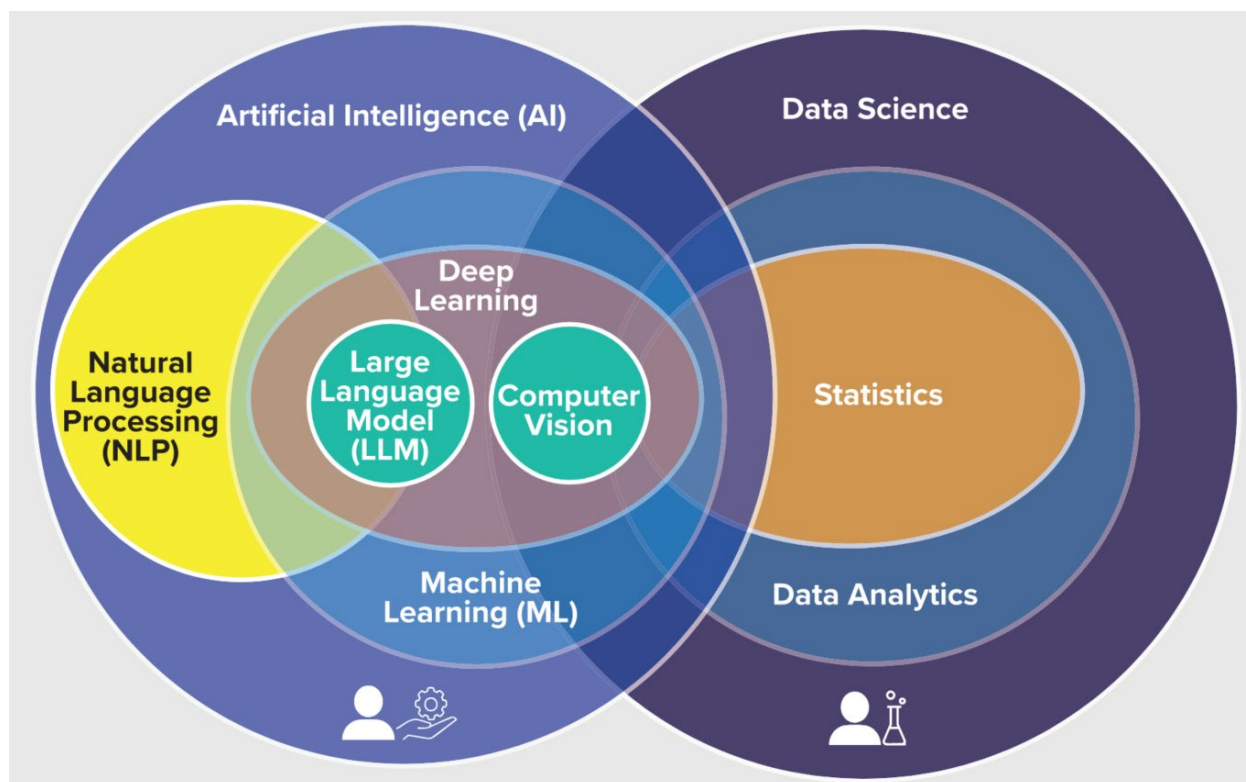


Figure 1-3 Relationship among Artificial Intelligence, Machine Learning, Deep Learning, Large Language Models, Natural Language Processing, Computer Vision, Data Science, Data Analytics, and Statistics.

1.2.1 Artificial Intelligence vs Machine Learning

AI is a broad category—the overarching field trying to create systems that can simulate human-like intelligence. It is like creating a digital brain that can reason, learn, and solve problems. ML is about learning from data, including DL algorithms that formulated foundation for computer vision, LLM, etc. AI is an all-encompassing umbrella that covers everything from lateral language processing, robotics, and decision-making systems.

To contrast with ML, conventional mathematical models (e.g., equations in environmental engineering textbooks like the Hazen-Williams equation or Ideal Gas Law) are called *mechanistic* models or *physics-based* models (Figure 1-3). Mechanistic modeling begins with a theory of how a system works (e.g., bacteria will eat food faster in the presence of more food), converts this theory into an algebraic equation (e.g., linear model), and then fits constants (e.g., slope) in this equation to data. Once this fitted equation is established, engineers plug situation-specific data into the equation for an answer.

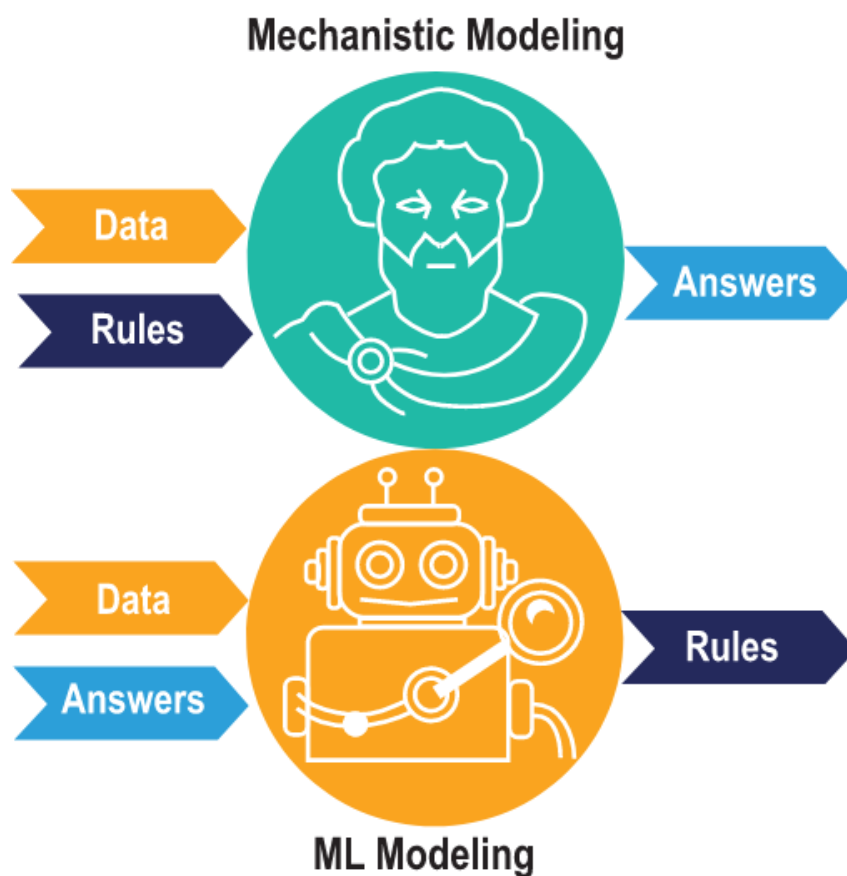
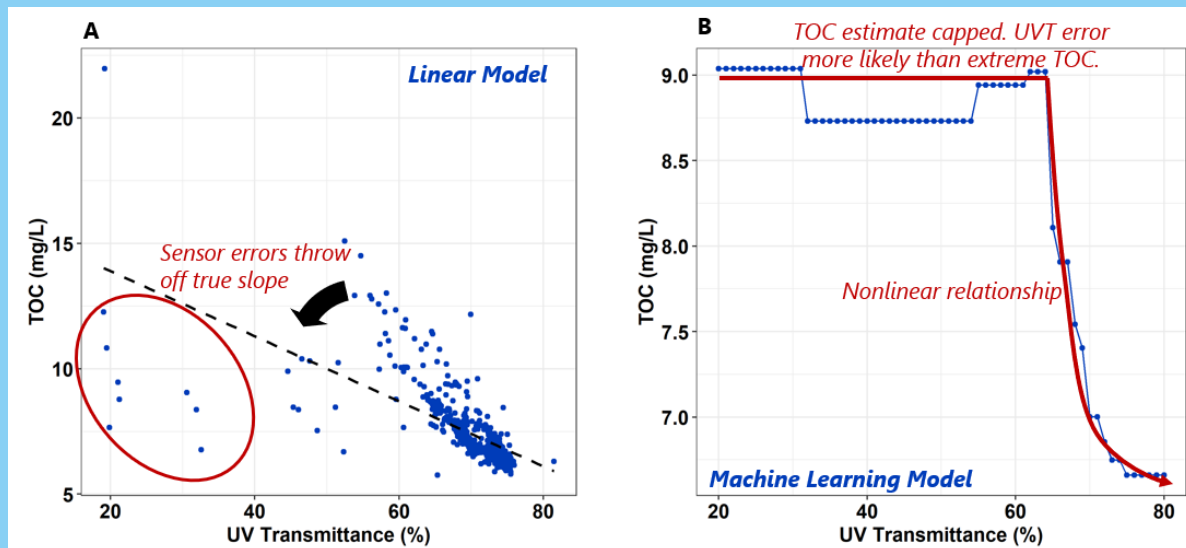


Figure 1-4 [The Difference between Mechanistic and Machine Learning Modeling](#)

ML modeling techniques differ in that ML models begin with the *data*, not a *theory* of how the system works. This is possible because many algorithms have been discovered that can extract insight from data emergently, or that can fit outputs to inputs an empirical, free-form way. This difference enables utilities to take into account site-specific patterns that would be overlooked by mechanistic models dictated by universal theories (Reuse Case Study #1;Salveson et al. 2025).

Reuse Case Study #1: Data-Driven, Nonlinear ML

As part of Water Research Foundation (WRF) Project #5129, Carollo developed a ML model for TOC between coagulation and ozone in a reuse system (Salveson et al., 2025). It is a known scientific theory that TOC correlates inversely with UV transmittance (UVT). However, in practice, UVT sensors often have low errors due to sensor fouling and sensor drift (Figure A). So, a linear model for TOC based on UVT would often overpredict TOC during these low UVT errors. In contrast, an ML model, by looking at additional variables, would not predict high TOC based on low UVT alone. It learns that if other variables are typical, a UVT error is more likely than a true TOC spike. Other variables considered by the ML algorithms included post-coagulation pH and ammonia.



Reference: Salveson, A., Kumar, P., Gerrity, D., Bott, C., Salazar-Benites, G., Bourgeois, K., Plumlee, M. H., Debroux, J., & Stanford, B. D. (2025). Demonstration of Innovation to Improve Pathogen Removal, Validation, and/or Monitoring in Carbon Based Advanced Treatment (CBAT) for Potable Reuse (No. 5129). The Water Research Foundation.

ML is the foundation that enables computers to learn from data without being explicitly programmed (i.e., coded with known equations or human-decided if-then rules only). Humans learn best when given a problem statement and encouraged to experiment or explore to find patterns and form our own answers, rather than given sets of rules to memorize. Analogously, ML works by feeding it data; it then discovers patterns to make predictions or decisions. ML includes many techniques and families of algorithms including random forests (Breiman, 2001), gradient boosting (Friedman, 2001), DL, and neural networks. Some data scientists may prefer to separate DL and traditional ML techniques that are simpler such as random forests and gradient boosting. These algorithms differ in the data structures they can handle and the dataset sizes they typically require.

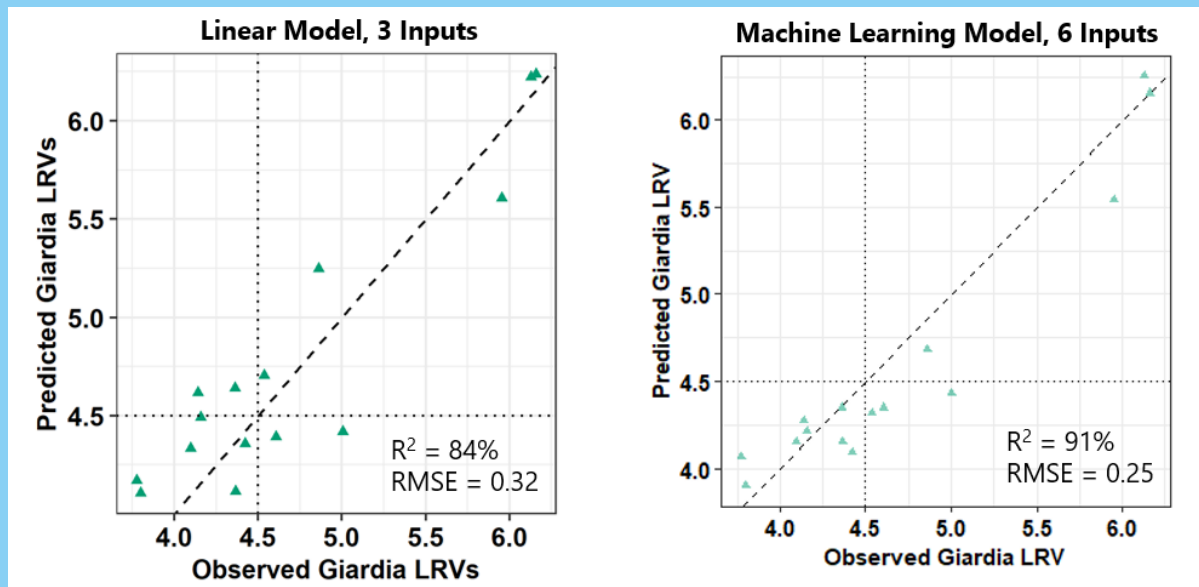
1.2.2 Statistics

Statistics, broadly, are the collection, analysis, description of data, and inference based on data. Many tools and purposes of ML and statistics overlap. However, ML and statistics differ in their focus, goals, and methodologies. The goals of statistics include understanding relationships and quantifying uncertainty and often times, there are underlying assumptions of data distributions to be made for statistical techniques. In contrast, ML can be more blackbox, and seeks to make the most accurate prediction

possible, even if the seemingly best model is only marginally more accurate than simpler models. As an example, traditional statistical modeling will only include inputs when there is a high probability that each input enhances the model. Meanwhile, ML will often include any input that improves the testing set accuracy or limit the number of inputs as a rule-of-thumb relative to the sample size. Despite these philosophical differences, statistics and ML are both key toolsets for modern data analytics (Section 1.2.4). To illustrate this point, we used a real-world example (Suresh et al., 2024) to demonstrate details (Reuse Case Study #2).

Reuse Case Study #2: Statistics vs ML

As part of the USBR project sponsoring this guidebook, Carollo developed a ML model for *Giardia* log reduction values (LRVs) in a membrane bioreactor (MBR) (Suresh et al., 2024). In conventional statistical multiple linear regression, inputs are selected based on *backward elimination*. That is, a linear model starts with all inputs, but the input with the highest p-value is dropped, and the model refit, until all remaining inputs have p-value < 0.05. Based on that criteria, *Giardia* LRV would have had three inputs: *E. coli*, *C. perfringens*, and maximum turbidity. However, based on a ML rule-of-thumb that the number of inputs should be less than the square root of the training set sample size, six inputs would be used—the same three plus 95th percentile turbidity, somatic coliphage, and median turbidity. The latter two did not significantly correlated with *Giardia* individually. Based on a fully separate test set, the ML model with six inputs had an R^2 of 0.91, compared to the linear model with three inputs having a lower R^2 of 0.84. However, in either case the root mean square error (RMSE) would be in rounding distance of 0.3 LRV. A traditionally trained statistician would question whether these arguably negligible accuracy gains are worthwhile or likely to hold up with a larger testing set.



References: Suresh, S., Thompson, K., Branch, A., & Salveson, A. (2024). Machine Learning to Predict Pathogen Removal in Potable Reuse MBRs Based on Microbial Surrogates. WEFTEC.

1.2.3 Data Analytics

Data analytics is the process or framework for starting with raw data and finding meaningful insights to make decisions. Data analytics is a broader umbrella term including both statistics and ML and steps like data preparation and visualization.

In AI, data analytics is crucial because it:

- Prepares and cleans data.
- Helps understand what's happening inside ML models.
- Provides criteria for measuring success.
- Bridges the gap between raw data and actionable insights.

Nonetheless, not all data analytics are AI. Often, advanced data analytics without AI/ML is sufficient or appropriate to solve the problem, particularly (1) when interpretability is required or (2) when working with single variables.

1.3 Types of ML

There are several types of ML methods and under each respective method, many unique algorithms exist for different types of problems (Figure 1-5).

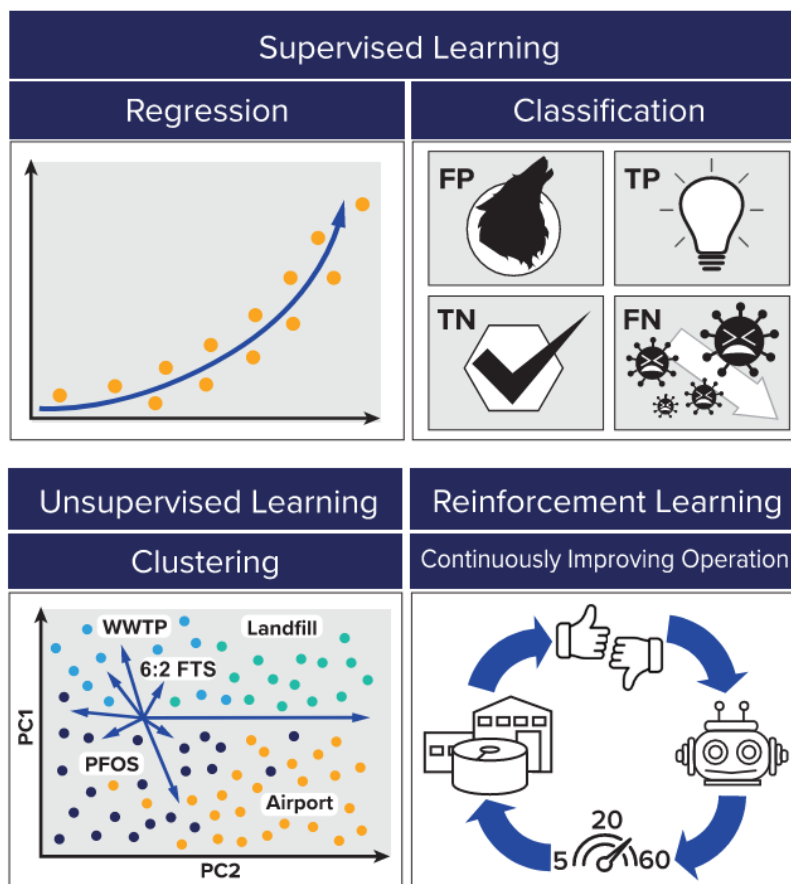


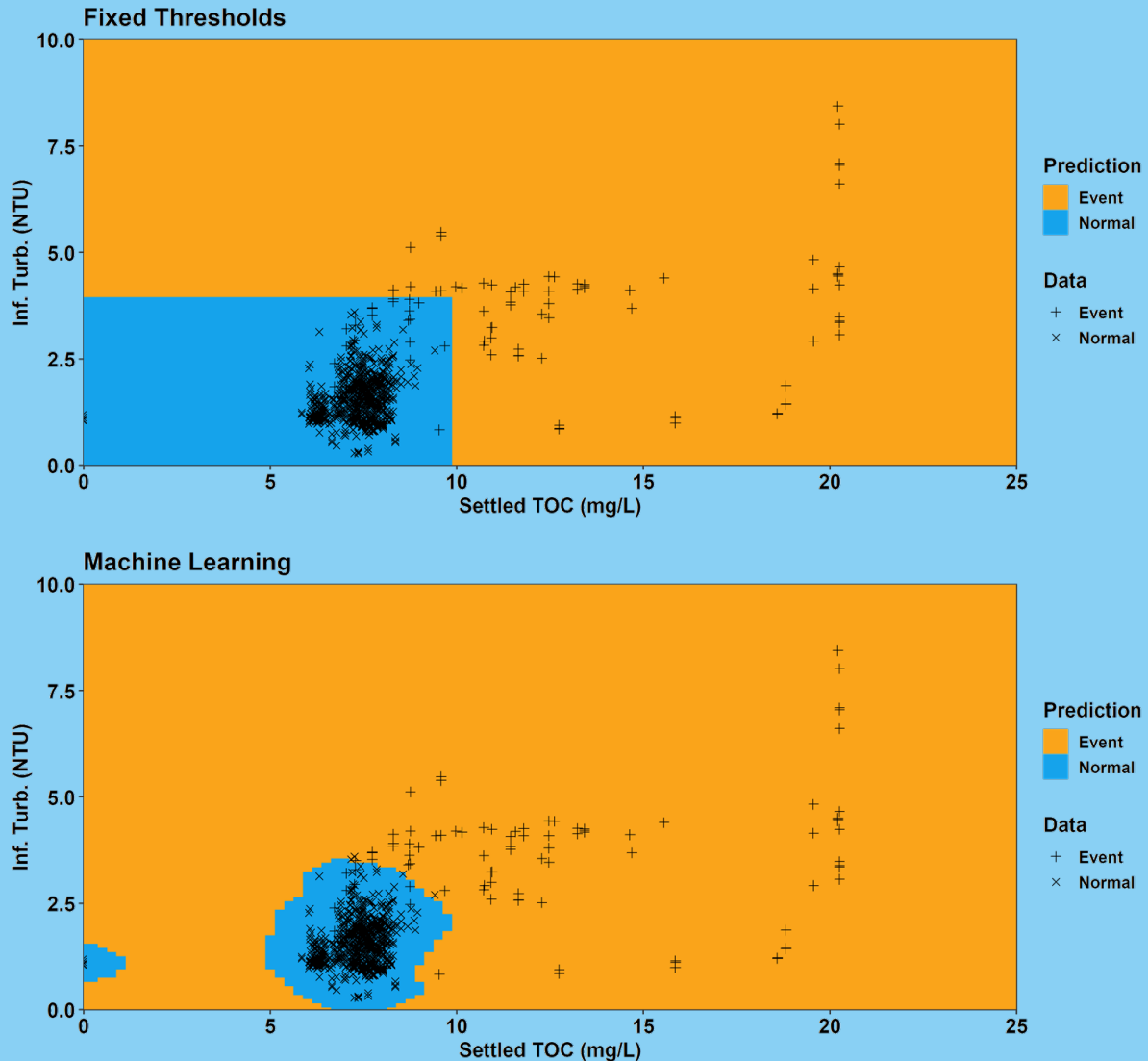
Figure 1-5 Types of Machine Learning and The Types of Mathematical Questions They Address

1.3.1 Supervised Learning

Supervised learning algorithms requires *labeled data*. Labeled data consists of input examples paired with their correct output values (the label). Labeled data are used to train supervised ML models by providing ground truth for the algorithm to learn from. Supervised learning includes regression, where the output is a number (e.g., what is the concentration of a target chemical?), and classification, where the output is a category (e.g., does this water quality look like Normal, Sensor Error, or True Event?). We have illustrated these details in a real-world example as Reuse Case Study #3 (Salveson et al., 2023; K. A. Thompson et al., 2022).

Reuse Case Study #3: Supervised Learning Classification

As part of WRF Project #5048, the project team applied supervised learning classification to detect industrial spike events in the feed water to a potable reuse system (Salveson et al., 2023; Thompson et al., 2022). The reuse system already had conventional alerts in place based on fixed thresholds, e.g. influent turbidity >4 NTU or settled TOC >10 mg/L. However, supervised learning classification was able to detect the industrial spike events sooner and with fewer false positives. The supervised learning classification could do this by wrapping the boundary between “Normal” and “Event” around the datapoints in a nonlinear, free-form way.



References: Salveson, A., Branch, A., Thompson, K., Weeks, B., Mansell, S., Nading, T., Thompson, K., Mohan, M., Dickenson, E., Thompson, B., Ackman, P., & Hsia, P. (2023). Integrating Real-Time Collection System Monitoring Approaches into Enhanced Source Control Programs for Potable Reuse.

<https://www.waterrf.org/research/projects/integrating-real-time-collection-system-monitoring-approaches-enhanced-source>; Thompson, K. A., Branch, A., Nading, T., Dziura, T., Salazar-Benites, G., Wilson, C., Bott, C.,

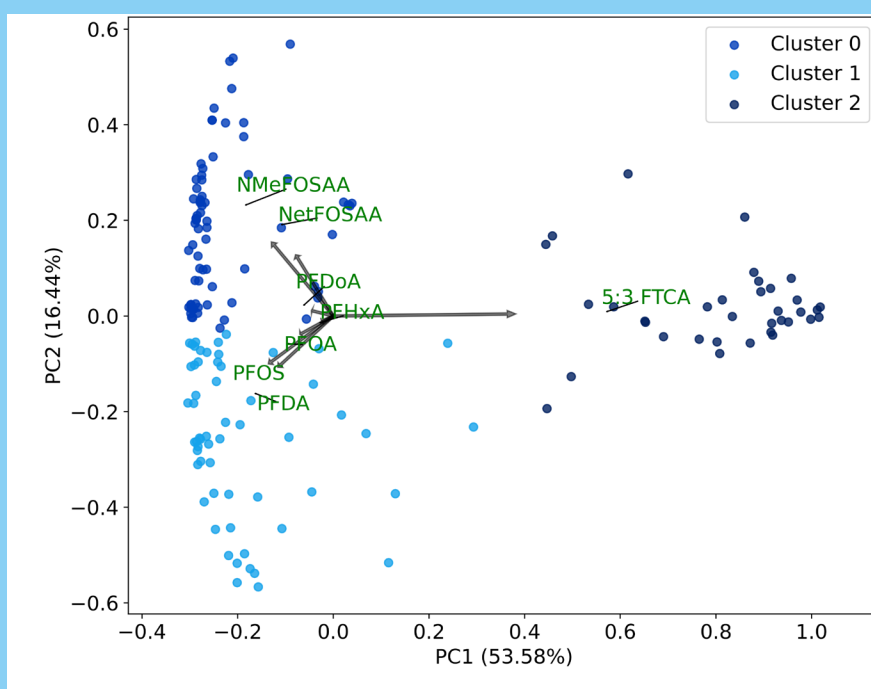
Salveson, A., & Dickenson, E. R. (2022). Detecting industrial discharges at an advanced water reuse facility using online instrumentation and supervised machine learning binary classification. *Frontiers in Water*, 4, 1014556.

1.3.2 Unsupervised Learning

Unlike supervised learning, unsupervised algorithms do not rely on known outputs or labeled training data but instead discover the inherent structure within datasets through mathematical relationships and statistical properties. Unsupervised learning algorithms operate with minimal human guidance, allowing the model to identify patterns independently from unlabeled data. As shown in Figure 1-4, this category of ML includes clustering techniques that group similar datapoints without predefined categories. We illustrated this concept in details in Reuse Case Study #4 (Lang et al., 2017; Ravindran et al., 2025; K. Thompson et al., 2025).

Reuse Case Study #4: Unsupervised Learning for PFAS Source Investigation

Considering their toxicity and challenging removal, PFAS have been ranked the most important chemicals for enhanced source control programs for potable reuse (Thompson et al., 2025). However, it can be challenging to discern whether there is a PFAS source and which type of PFAS source is most likely based only on PFAS data from WWTPs or reuse facilities. Carollo applied unsupervised learning clustering to influent, effluent, and biosolids data from the California PFAS database (Ravindran et al., 2025). Unsupervised learning was the appropriate choice since the presence or identity of PFAS point sources were not known for the majority of the WWTPs. The unsupervised learning revealed the clearest clusters when applied to the biosolids data. Cluster #0 had high NMeFOSAA and NEtFOSAA, Cluster #1 had high PFOS and PFDA, and Cluster #2 had high 5:3 FTCA. A previous study found 5:3 FTCA was the highest PFAS measured in landfill leachate (Lang et al., 2017). A geospatial analysis confirmed the WWTPs in Cluster #2 tended to be closer to more landfills. So, landfills could explain this PFAS data signature.



References: Thompson, K. & others. (2025). Review of Industrial Chemicals Impacting Potable Reuse: Hazards, Removal, and Recommendations for Enhanced Source Control Programs. *Total Environment Engineering*.
<https://doi.org/10.1016/j.teengi.2025.100027>; Ravindran, T., Koyama, Y., & Thompson, K. (2025). Analyzing California's PFAS Data: AI Insights in Source Tracking. *Clean Water*, 2025(2), 22–25.; Lang, J.R., Allred, B.M., Field, J.A., Levis, J.W. and Barlaz, M.A., 2017. National estimate of per- and polyfluoroalkyl substance (PFAS) release to US municipal landfill leachate. *Environmental science & technology*, 51(4), pp.2197-2205.

Unsupervised learning also includes dimensionality reduction methods (i.e., reducing the number of input variables) to simplify complex datasets while preserving essential information. A common unsupervised learning algorithm for dimensionality reduction is principal component analysis (PCA) (Thompson and Dickenson, 2021). Unsupervised learning techniques can also generate alerts for when water quality or equipment behavior as a whole is atypical (Grimm et al., 2024).

1.3.3 Reinforcement Learning

Reinforcement learning is a ML approach where an agent learns optimal behavior through trial-and-error. Reinforcement learning algorithms receive feedback in the form of rewards or penalties, such as thumbs up or down from an operator. The algorithm then adjusts its output or actions to maximize rewards over time. Reinforcement learning has not yet been applied to advanced treatment for potable reuse to the best of our knowledge. However, reinforcement learning has been applied to a simulated WWTP, resulting in 11 percent cost saving while controlling an overall water quality index (Croll et al., 2025).

1.4 Terminologies

1.4.1 Inputs and Outputs

In the world of data analysis, Statistics and ML often address similar concepts using different terminologies. Table 1-1 presents a comparison between these two fields, showing synonyms they use to refer to input and output data.

Table 1-1 Comparison of Synonymous Terms for Input and Output in Statistics versus Machine Learning

In Statistics	In ML	Definition
Independent Variables (x)	Features, inputs, predictors, attributes	Inputs to the model
Dependent Variable (y)	Target, outputs, label, response	That which is the goal to predict

1.4.2 Training Set and Testing Set

In ML model development, also called model training, data is first divided into two parts: the training set and the testing set. The training set is where the model learns. The training set is often further subdivided into smaller train-validate splits to preliminarily evaluate versions of the model on new data, in a process called cross-validation (CV). The original testing set remains untouched during this training and serves as the final, unbiased evaluation of the model's ability to handle new data. See Section 3.1 for more discussion on best practices for data splitting and CV in the context of potable reuse.

1.5 Machine Learning Application Styles for Potable Reuse

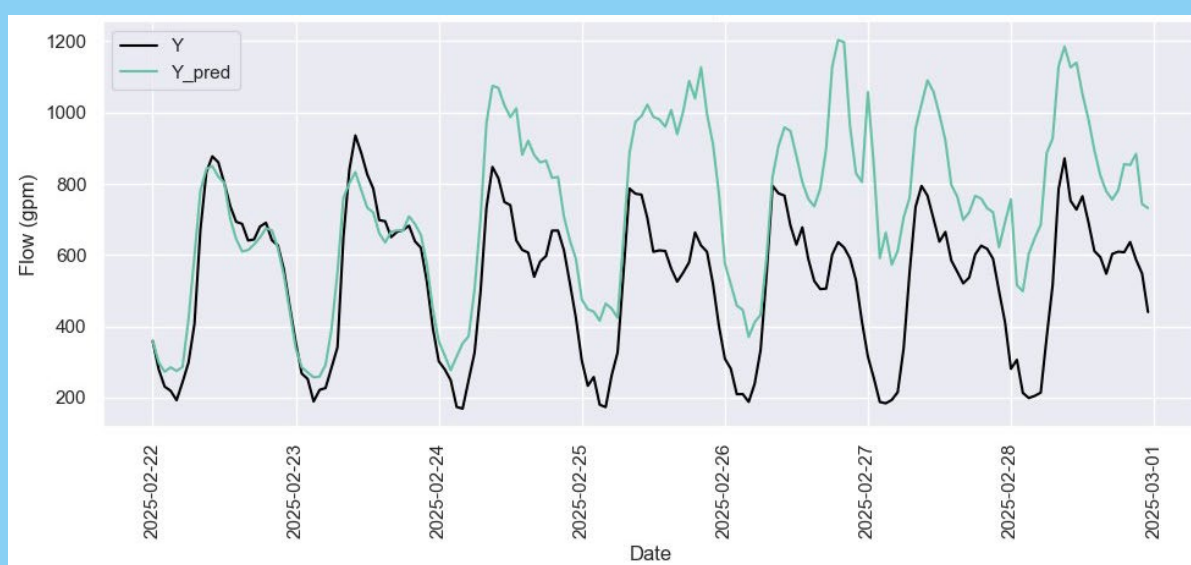
There are many ways to apply AI/ML for potable reuse operation. Table 1-2 non-exhaustively lists some of these strategies with examples.

Table 1-2 Machine Learning Application Styles for Potable Reuse

Application Style	Explanation and Examples
Multivariate Alerts	ML can consider data from multiple sensors or data sources to conclude whether the water quality as a whole is atypical or looks like a previous negative event (e.g., membrane integrity failure or industrial spill). See Reuse Case Study #3.
Soft Sensors	ML can predict the concentration of contaminants such as TOC, NDMA, or pathogens. This prediction can in turn trigger an alert or inform the treatment intensity (e.g., coagulant dose or UV fluence). See Reuse Case Study #1.
Forecasts	Predicting the future is a challenge for any model type, especially when the future situation depends in part on human actions (e.g., changes to upstream treatment). Nonetheless, ML models can forecast water quality, water quantity, or equipment status out to reasonable horizons in the future. This enables planning for operational choices that require lead time, such as filter backwash or membrane start-up. See Reuse Case Study #5.
Digital Twins	Strictly, a digital twin is defined as the real-time, two-way flow of data between a physical system and a digital model simulating it. Recommendations or automatic operational commands come from the digital model, and updated data from the physical system returns to the digital model, continuing the cycle. See Reuse Case Study #6.

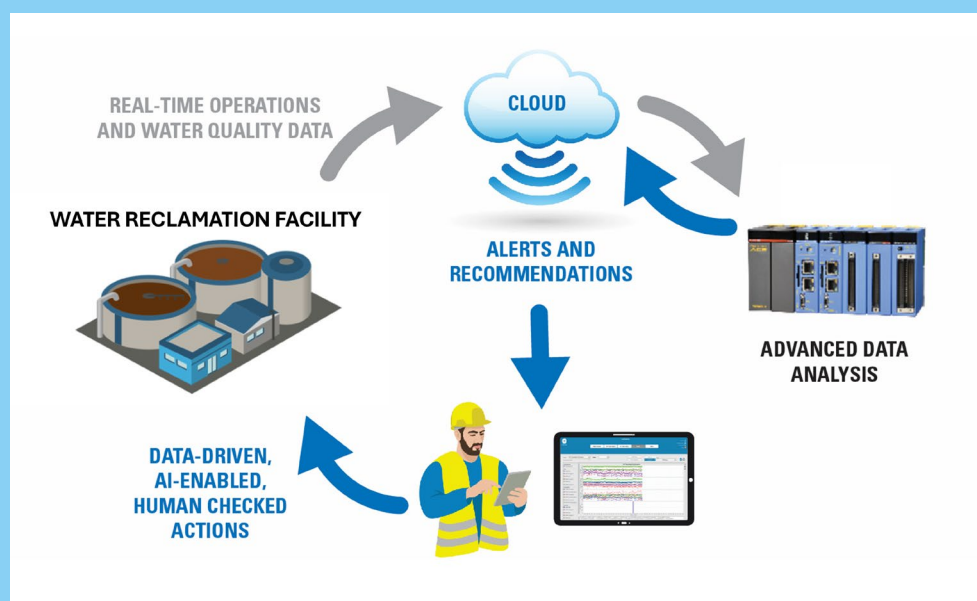
Reuse Case Study #5: ML Water Flow Forecast

For the USBR project sponsoring this guidebook, a participating utility has an MBR followed by three parallel RO trains, with little flow equalization post MBR. Excess flow from the MBRs would be discharged to the ocean outfall, losing the benefits of reuse. However, running all three RO trains when the flow was low would waste electricity and drop below the minimum rated flux for the RO equipment. Carollo applied ML to forecast the MBR water flow into the future, based on its historical patterns and local rain data. ML could predict the flow accurately out to 48 hours, beyond which a “butterfly effect” caused the prediction to drift. A proactive schedule was developed to turn on and off trains based upon this prediction, helping the utility meet an annual reclamation goal of 130 million gallons.



Reuse Case Study #6: Wastewater Digital Twin Upstream of Reuse

In a project sponsored by the Japanese Ministry of Economy, Trade and Industry, Carollo and Yokogawa Electric Corporation developed a full-scale digital twin for water reclamation facility upstream of a potable reuse demonstration facility. Data was sent automatically from the utility's data management provider to Yokogawa's cloud. Yokogawa then calculated the ideal dissolved oxygen setpoint based on empirical, data-driven modeling. The recommended dissolved oxygen setpoint then appeared hourly on a web dashboard for the operators. The utility trialed this system for three weeks at full scale. Aeration energy per kg of influent ammonia was 23% lower while using the digital twin compared to adjacent weeks.

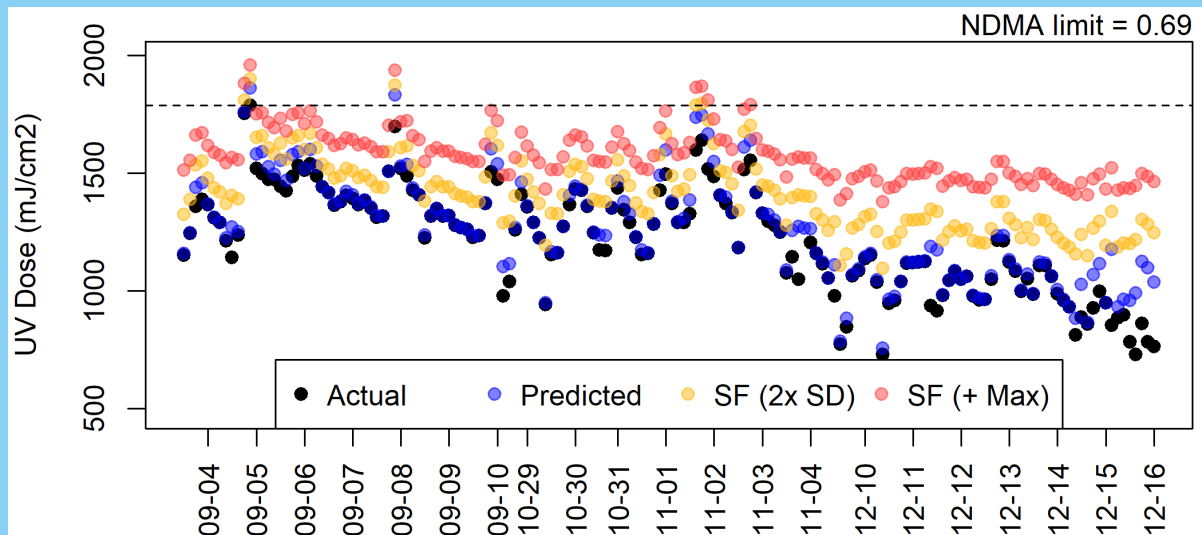


1.6 Treatment Process Prioritization for Machine Learning

In theory, savings from AI/ML-based optimization depend on two factors: (1) the difference between current practice and the true optimal operation and (2) ML model accuracy, which determines the ability of ML to meet that potential for improvement. Safety factors should be built-in in proportion to model uncertainty to ensure consistent compliance, but this reduces the actual savings. We illustrated this concept in details in Reuse Case Study #7 (Thompson et al., 2024).

Reuse Case Study #7: Safety Factors

In NAWI Project 5.17, Oregon State University and Carollo developed an ML soft sensor for NDMA between RO and UV/AOP (Thompson et al., 2024). The goal was that the UV dose in UV/AOP could be adapted in proportion to the predicted concentration of NDMA. The NDMA data used to train the model came from OCWD, but the team considered a hypothetical scenario where a utility would be treating to a lower NDMA standard (0.69 ng/L) for surface water discharge that is subject to the California Toxics Rule (CTR). As the benchmark (dashed horizontal line in the figure below), the project team assumed the UV dose would be set conservatively based on the highest actual NDMA in the dataset (black in the figure below). If the UV dose were set based on the ML-predicted NDMA concentration (blue in the figure below), the UV energy saving would be 26%. Setting the UV dose based on the predicted NDMA concentration plus a safety factor of two standard deviations (yellow in the figure below), would result in a 13% UV energy saving.



Reference: Newhart, K.B., Thompson, K.A., Branch, A., Salveson, A., Safarik, J. and Plumlee, M.H., 2025. NDMA soft-sensors for potable reuse: A model development study. *Water Research*, p.124451.

In practice, additional factors may affect ML implementation, such as process flexibility. Some processes, for example biofiltration, are mainly set in place during design and have few “levers” to adapt during operation. Others, like disinfection, may have regulatory constraints, especially for potable reuse. Even if research reveals safer, more efficient operations, rule changes might be necessary before implementation.

Absolute savings from a percentage improvement in the process of operation would also depend on the initial O&M cost. So, the priority for investigating and implementing AI/ML for a process would also depend on how it compares as a percentage to the overall operating costs of the utility. In potable reuse, O&M costs for unit processes can follow the order: RO > GAC > biofiltration > microfiltration > UV/AOP > coagulation ≈ ozone ≈ chlorination ≈ UV disinfection (Giammar et al., 2021). However, these relative costs depend on size and regional factors.

1.7 AI/ML Adoption Barriers

1.7.1 Overview

Water utilities face multifaceted ML adoption barriers including technical as well as organizational challenges—barriers examined in detail through Section 5.2's comprehensive derisking strategies. Data management sits at the intersection of these challenges: it's simultaneously a technical barrier requiring infrastructure investment, a business constraint demanding resource allocation, and a stakeholder readiness issue necessitating organizational capacity. Establishing robust data management practices therefore becomes the essential first step that enables progress across all other adoption dimensions.

1.7.2 Data Management

Data management is the essential first step for any ML project. It involves systematic collection, transfer, storage, organization, integration, analysis, visualization, and disposal of data from sources such as SCADA systems. SCADA architectures typically consist of multiple components: data collectors that gather real-time information, visualization for operator interfaces, and historians that store time-series data. Data acquisition methods vary—including time-sequenced polling at defined intervals, interface-based polling, or report-by-exception approaches that timestamp data only when values change. This infrastructure relies on digital tools and incorporates data governance, validation, and security to maintain integrity and reliability.

For ML-driven process optimization, sophisticated data management is crucial. These six elements (collection, transfer, storage, integration, analysis, and visualization) constitute a complete framework (Figure 1-6; Dent et al. 2023). Strong data management in each element enables effective AI/ML development and scale-up—even the most promising models often fail at operationalization due to poor data management (Kreuzberger et al., 2023a).

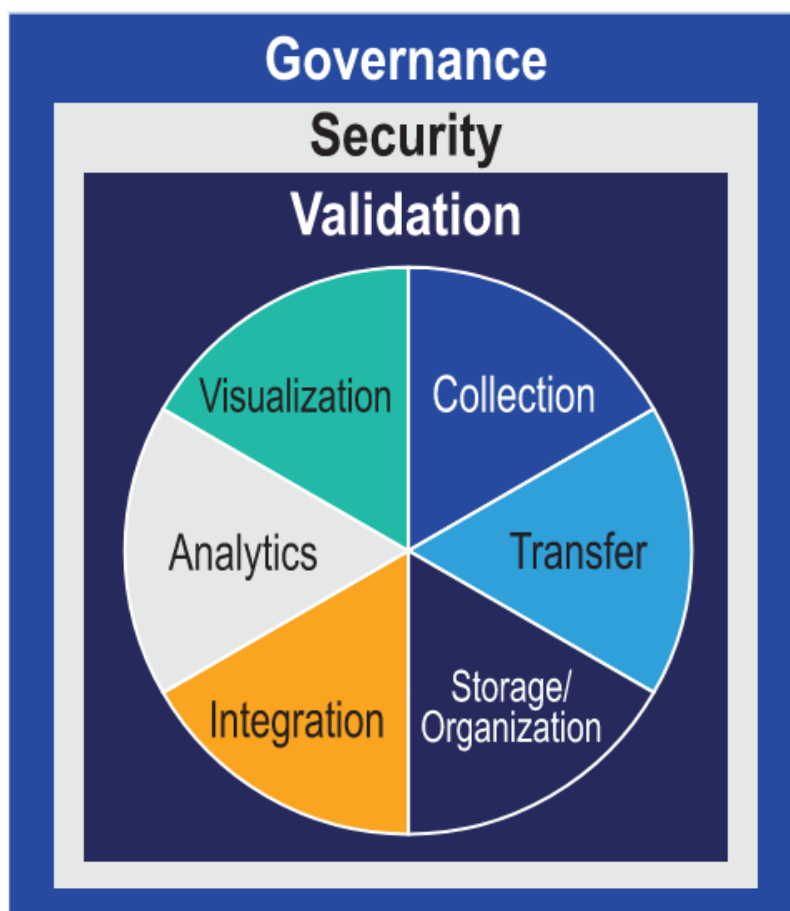


Figure 1-6 Data management framework for ML— Data management that is strong in each element of the framework will produce a more usable model

Water utilities historically managed data through siloed systems—LIMS, computerized maintenance management systems (CMMS), and SCADA—with limited interconnectivity (Figure 1-7). Many LIMS, CMMS, and advanced metering infrastructure (AMI) systems have transitioned to cloud-based platforms, positioning utility data increasingly in the cloud, while SCADA systems often remain the primary integration challenge. Modern utilities can now integrate diverse data sources including AMI, distributed water quality sensors, and geographic information systems (GIS). Newer SCADA systems are also evolving, offering APIs, ODBC connections, and direct connectors to bridge these data sources. This integration enables ML improvements and uncovers new insights; Figure 1-8 illustrates this progression from raw data to real-time decision support through a tiered AI/ML approach. In a water quality or process control context, “Data Source” in Figure 1-8 would refer to having sufficient, accurate, and well-maintained sensors. Achieving integration and moving up the pyramid requires modernizing data management systems to break down technological and organizational silos (Figure 1-8).



Figure 1-7 Improved data management de-silos data among utility departments

**Before You Can
Use AI/ML to
make decisions,
You Have to
Build
the Foundation
Beneath It**

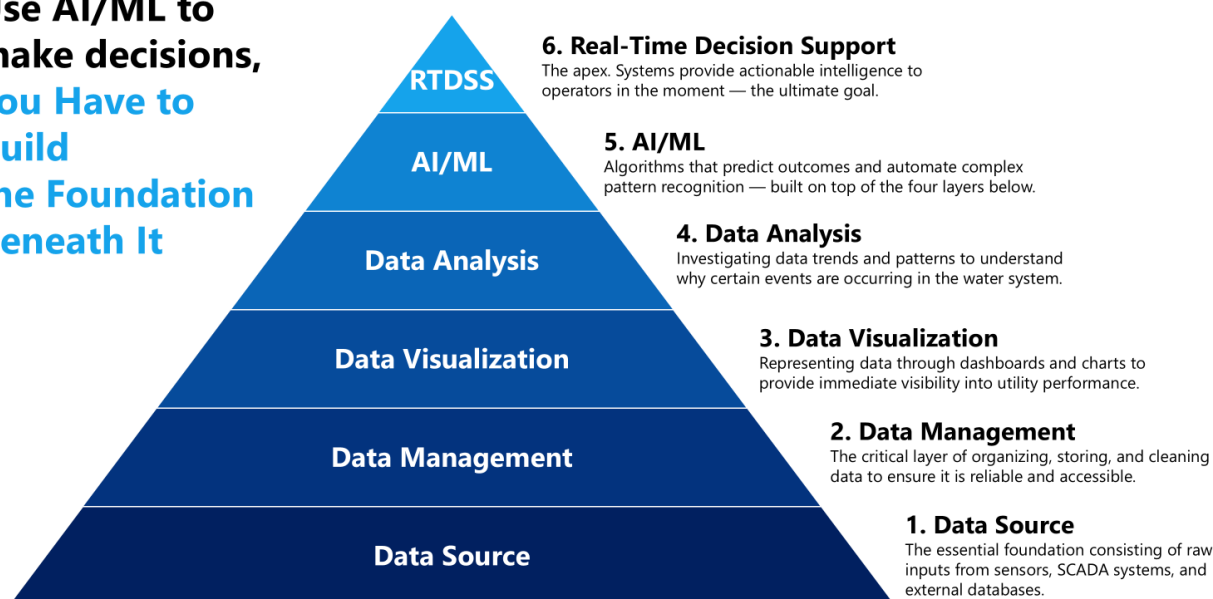


Figure 1-8 Data management hierarchy of needs pyramid

While data management is not this guidebook's primary focus, it is a critical need for AI/ML success. To this end, we recommend the following data management resources:

- **Foundational:** Earley, S., Henderson, D., & Sebastian-Coleman, L. (2017). *DAMA-DMBOK: Data Management Body of Knowledge* (2nd ed.). Technics Publications.
- **Water Utility-Specific:**
 - » Badruzzaman, M., Cherchi, C., Grenfell, R., Jacangelo, J. G., Gunaratnam, M., Bunn, S., & Gordon, M. (2020). *Managing Water and Wastewater Utility Data to Reduce Energy Consumption and Cost* (Research Report No. 4668). Water Research Foundation.

- » Martel, K., Hanson, A., Kirmeyer, G. J., Besner, M.-C., Carrière, A., Prévost, M., Lynggaard-Jensen, A., & Bazzurro, N. (2005). *Data Integration for Water Quality Management* (Research Report No. 91100). Awwa Research Foundation.
- **Intelligent Systems & Workforce:**
 - » Lehman, G., Chandrasekeran, P., Johnson, K., Loria, M., Cherchi, C., & Vitasovic, Z. (2022). *Definition, Framework, and Maturity Assessment for Intelligent Water Systems* (Technical Report No. 5039). Water Research Foundation.
 - » Oxenford, J., Hughes, D. M., Karasik, S., & Bukhari, Z. (2019). *Building Workforce Skills for Intelligent Water Operations* (Report No. 4663). Water Research Foundation.
 - » Thompson, K., & Macintosh, C. (2020). *Intelligent Water Networks Summit* (Research Report No. 4714). Water Research Foundation.
- **Regulatory & Cybersecurity:** U.S. Environmental Protection Agency. (2018). America's Water Infrastructure Act of 2018 (AWIA). Retrieved from <https://www.epa.gov/ground-water-and-drinking-water/americas-water-infrastructure-act-2018-awia>
- Cybersecurity and Infrastructure Security Agency (CISA). PII Authorized User Training. <https://www.cisa.gov/pcii-program>
- American Water Works Association. AWWA Cybersecurity Guidance Tool. <https://www.awwa.org/cybersecurity>

1.8 Step-by-Step Machine Learning Overview

Once adoption barriers are overcome, AI/ML implementation follows the five-step process shown in Figure 1-9.

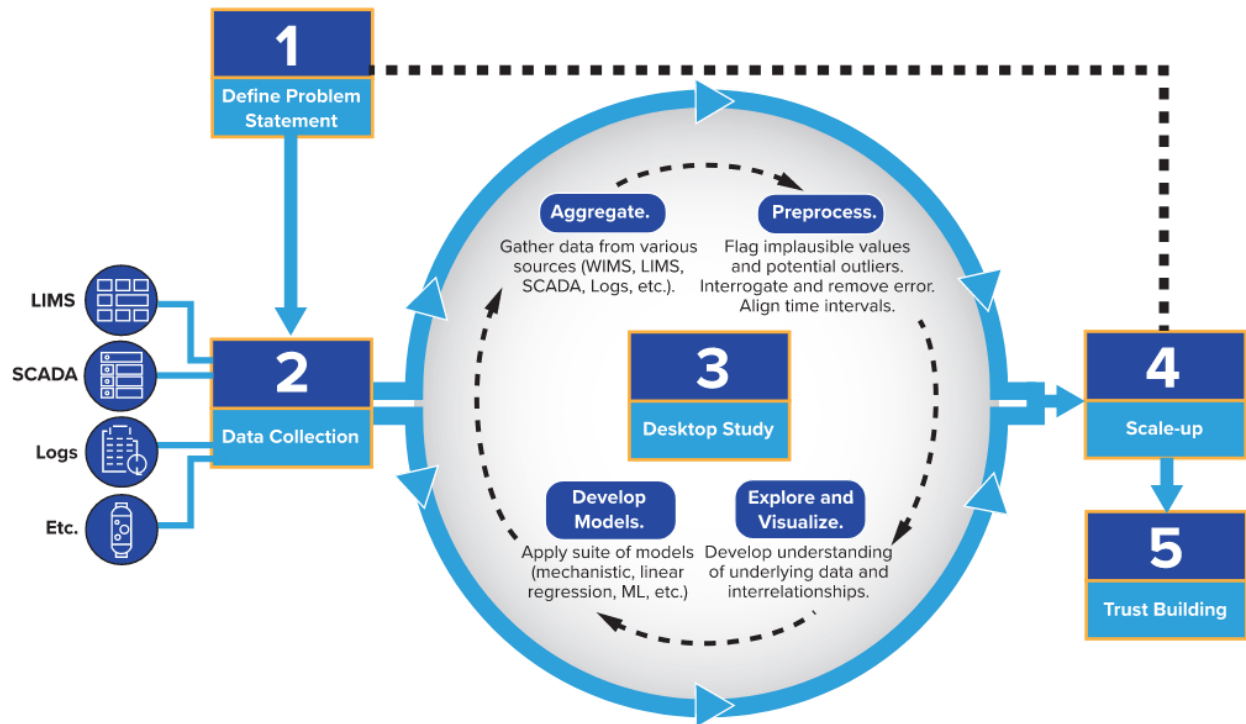


Figure 1-9 Machine learning solution development process

Phase 1: Problem Definition

To begin an ML or data analytics project, first identify the problem to solve that can benefit from applying ML modeling (i.e., what would be predicted?). Understand how operators or further algorithms would use that prediction. Consider whether simpler approaches could also solve the problem or should serve as benchmarks. The quality and availability of data vary greatly, introducing uncertainties that can exceed sensor accuracies (Chesterman et al., 2023), and existing historical data may be insufficient or inadequate, especially when input variables are constant. The training and testing sets should include a range of influent water quality and operational parameter variations to reveal relationships and calibrate the model over the plausible range (Section 1.4.2). Often, ML is more likely to be useful if staff develop an experimental plan with varied operational parameters.

Goal Setting & Scale-Up Vision:

Effective implementation requires clear definition of near- and long-term goals. Using the SMART goal framework:

- **Specific:** What process will ML be applied to? What water quality parameter will it predict.
- **Measurable:** How will we measure improvement? (e.g., energy, chemical use, fewer false alarms).
- **Actionable:** How will operators or algorithms use the ML prediction.

- **Realistic:** Is our target percent improvement possible.
- **Time-bound:** What's our deadline for desktop study and scale-up.

Goal setting happens concurrently with problem definition:

- **Identify** specific operational challenges that AI/ML can address, such as optimizing chemical dosing or predicting equipment maintenance needs.
- **Develop** metrics to quantify benefits of AI/ML based approach.
- **Prioritize** initiatives based on ROI and level of effort.
- **Articulate** with stakeholders why AI/ML is the appropriate solution versus traditional approaches. While clear initial goals are important for focused action, teams should also recognize that initial visions can evolve as the project progresses.

Optimization problems that do not touch regulatory requirements can be the lowest hanging fruit. Overall, we recommend defining end goals for scale-up that align with organizational capabilities and process timescales.

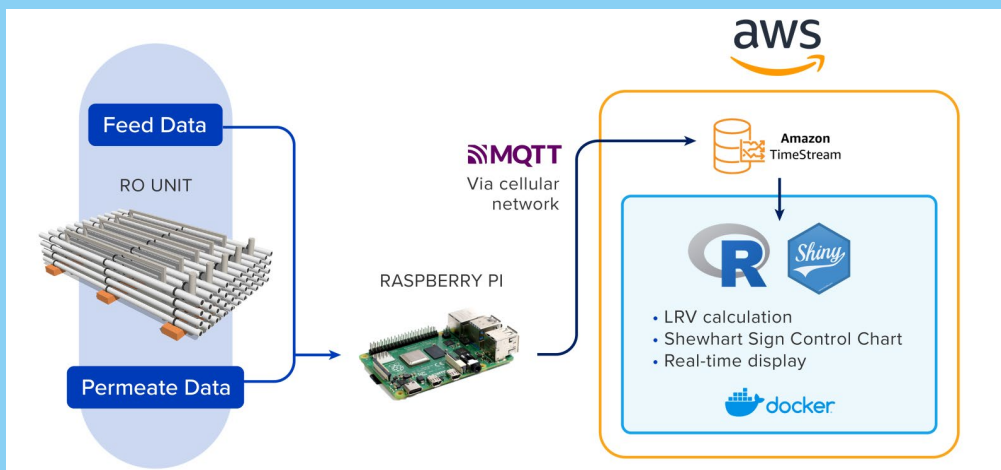
Phase 2-5: Data Collection, Desktop Study, Scale Up, and Trust Building

The Desktop Study is often an iterative process, where data scientists aggregate, preprocess, and analyze the data, and this step may reveal the need for additional data or require revisiting assumptions underlying data preprocessing. If AI/ML meets performance goals, the next step is Scale-Up (Reuse Case Study #8). A clear vision for scale-up enables early improvements to data management, perhaps kicking off while the Desktop Study is still iterating. Scale-up may involve hosting ML models online or on on-premise servers with human-machine interfaces such as graphical user interfaces, texts, emails, or SCADA screens. Lastly, we need to continuously validate ML outputs in response to operator or engineer feedback to build trust which leads to sustained AI/ML tools for the long term.

Reuse Case Study #8: The Step-by-Step Process from Problem to Scale-Up

To illustrate scale-up, the following is an example from NAWI Project 5.17 where an advanced statistical alarm was developed for RO LRV surrogate monitoring (Thompson et al, 2025). While this example is data analytics that is *not* technically ML model, the workflow is the same.

- 1) **Problem Statement** RO can remove pathogens by over 5 LRVs. However, this removal is challenging to verify in real-time. Surrogates like TOC or Peak C fluorescence are monitored before and after RO, and the LRV of pathogens is assumed higher than the LRV of these surrogates. However, these sensors have intermittent erroneous outliers, which could cause false alarms. A statistical method was needed that could detect true events promptly while avoiding false alarms.
- 2) **Data Collection** Historical data was provided from three utilities. This included both SCADA data and data collected directly from experimental sensors that OCWD was testing to see whether they could verify more LRVs.
- 3) **Desktop Study Aggregate** Highly frequent SCADA data (~1/second) was averaged to 15 minutes or hourly. **Preprocess** Outliers were left in, since testing the ability of the statistical tool to avoid alarms despite outliers was the purpose of the study. **Explore and Visualize** Data visualizations illustrated that the surrogate LRVs were not normally distributed, motivating the use of specialized statistical tools. **Develop Models** A statistical tool called a *Shewhart Sign Control Chart* was evaluated for its ability to avoid false alarms for singular outliers, but trigger when low LRVs continued for a few hours. The Shewhart Sign Control Chart met the goal of at least 50% false alarm reduction was still able to detect true events in under 12 hours.
- 4) **Scale-up** OCWD chose to scale this tool using cloud architecture. Designated TOC and fluorescence sensors were installed before and after a full-scale RO train. Data was transferred to the Amazon cloud via a raspberry pi and a cellular network. LRV calculations, the Shewhart Sign Control Chart analysis, and real-time visualizations were developed in an R Shiny application hosted on the cloud.



1.9 Partnership Strategy

Utilities often seek outside help when embarking on their AI/ML journey. The right partner depends on the utility's internal resources, step in the ML project life cycle, and desire for transparency or simplicity (Table 1-3).

Table 1-3 Partnership strategy comparison

Partnership Type	Implementation Step(s) ⁽¹⁾	Advantages	Limitations
Consultants	All steps	- Client-centric business model encourages site-specific models - Water sector expertise enables objective solution evaluation - Some firms use open-source approaches, facilitating collaboration with utility data scientists	- Limited AI/ML experience outside water sector - Hourly cost structures not optimized for long-term facility operation
Software Start-Ups	All steps	- Client-centric business model encourages site-specific models - Emerging AI/ML expertise tailored to water applications - Flexible deployment options	- Long-term viability not guaranteed - Lack of resources or diverse experts compared to large controls firms or consultants
Equipment Vendors	Scale-up step	- Integrated hardware-software solutions - Established service/support infrastructure	- Models often proprietary - Models may not be site-specific especially in the desktop phase
Academic Partners	Desktop step	- Access to cutting-edge research at low cost - Extensive literature and methodology resources	- May lack practical experience with real-world operational constraints - Project timelines aligned with academic calendars rather than utility needs
Non-Profit Organizations (e.g., AWWA, California Data Collaborative)	All steps	- Professional networking opportunities - Utility consortia and conferences for knowledge sharing - Relatively low-cost	- Services may vary by organization - May have limited personnel or resources for scale-up or long-term support
Utilities	Trust-building step	Utilities play critical role in building trust between end users and the AI/ML tool	Often have limitation on having full-time employee dedicated to AI/ML initiatives

1.10 Data Science Tools for Desktop Studies

Python and R have emerged as the primary programming languages for ML in the water sector. Other options include Julia, C++, and MATLAB. A comparison of R and Python is summarized in Table 1-4. Both R and Python have downloadable, add-on toolboxes of functions, data or code, called “packages” in R but “libraries” in python. The choice between them can boil down to ease of collaboration with internal teams or external partners, or the intended scale-up pathway.

Table 1-4 Characteristics of the Programming Languages R and Python for Data Analysis and ML

	R 	Python 
Primary Use	Statistical analysis and data visualization	General-purpose programming including data analysis
Learning Curve	Authors personally felt R was easier to learn, but a study says otherwise (Ozgur et al., 2017)	
Data Handling	Excellent for statistical computing and graphics with packages like tidyverse	Versatile data manipulation with data analysis libraries like Pandas
ML	Done with packages like caret	Done with libraries like Scikit-learn
Visualization	Good built-in plotting and ggplot2 package	Good visualization with libraries like Matplotlib and Seaborn
Speed	Generally slower for large-scale applications	Faster execution, especially for large datasets
Programmer Community	Fewer total packages but specialized in statistics, data analysis, and ML	Vast number of libraries for various purposes
Who uses this language	Widely used in academia, so suitable for R&D, cutting-edge statistics, and open-source collaboration	Larger overall community with more scale-up tools, support forums, and diverse function libraries

Learning Curve

R has a gentler learning curve due to its purpose-built focus on statistics and unified interface, RStudio. Python, on the other hand, has multiple interfaces (e.g., Spyder, Jupyter) that require re-training for some users.

Processing Power

Python generally runs faster, especially for larger datasets (> 10,000 rows).

Ease of Scale-Up

Python is regarded as a versatile language that is compatible with various programming needs beyond data science. On the other hand, R Shiny framework can be used for fast and easy web application development (Chang et al., 2024). Combining both languages (e.g., training a model in Python and deploying it in R) allows users to leverage the strengths of each language.

Conclusion

In today’s world with rapidly evolving GenAI landscape, Python, a more versatile language with both data analytics and GenAI capabilities would be a good long-term solution. That said, R is a perfect way to learn programming and data analytics for beginners trying to begin their digital journey.

1.11 The Cloud

1.11.1 What is the Cloud

Cloud computing delivers computing services over the internet, allowing organizations to access resources remotely instead of maintaining them on-site. Cloud providers like Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) Platform offer two essential functions: storage and computing power. They provide secure and scalable solutions for storing large volumes of operational data, as well as powerful processing capabilities for complex ML algorithms and big data analytics. Cloud services offer "just right" cost-effective computing costs by automatically tailoring resources to the task. They also provide built-in tools for data backup, security, and compliance, making them suitable for water utilities with strict regulatory standards and data protection protocols.

1.11.2 Leveraging the Power of Cloud

Cloud services help water utilities overcome data accessibility challenges across fragmented information systems without upfront hardware and engineering costs. While on-premises systems can connect SCADA, LIMS, CMMS, and IoT devices through proper network configuration, the primary challenge is maintaining custom connectors that require reconfiguration with each system upgrade and ensuring data accessibility across users and applications. Cloud platforms provide a centralized data environment with standardized ingestion tools and managed connectors. While these platforms still require significant configuration and governance, they also reduce maintenance overhead and streamline multi-user access. The elastic infrastructure efficiently manages large time-series data volumes while maintaining cost-effectiveness.

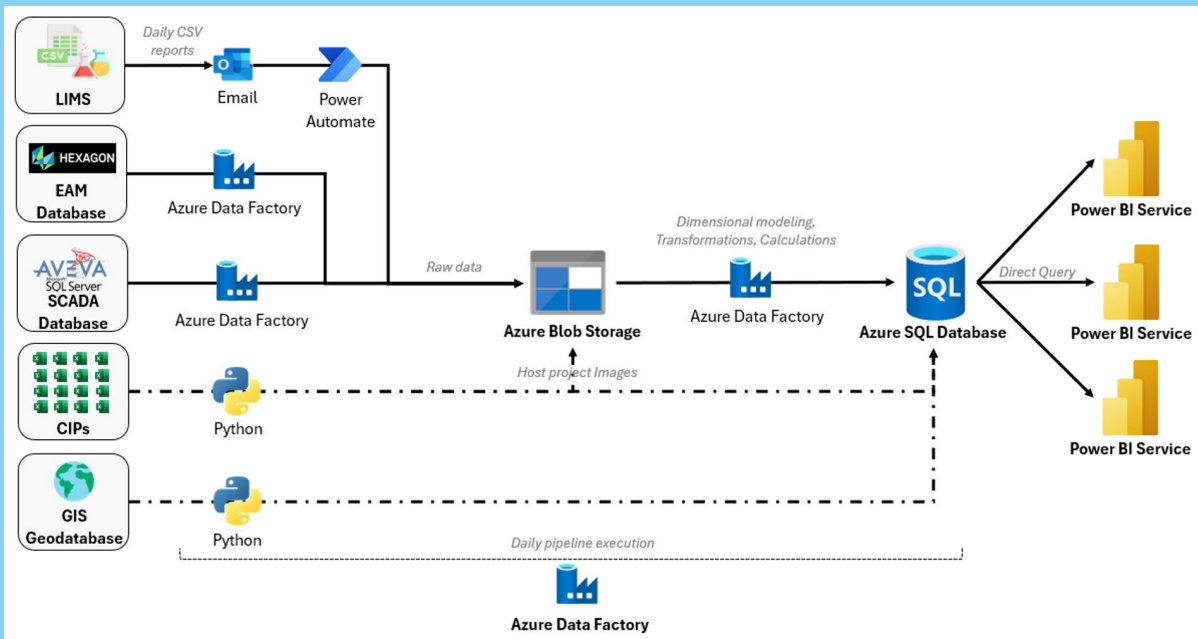
SaaS cloud systems shift costs from capital expenditure to O&M, which may integrate more easily into utility budgets. However, recurring subscription fees may fall outside traditional RFP processes. Many utilities address this by leveraging existing enterprise agreements (e.g., Microsoft licensing).

Beyond enabling real-time AI/ML operations, cloud dashboards deliver sophisticated visualizations that help operators identify patterns and anomalies faster. The Reuse Case Study 9 exemplifies the de-siloing exercise using the power of cloud-based infrastructure (Mishra et al., 2024).

Reuse Case Study #9: Leveraging the Cloud to Deliver Cross-Functional Analytics

To demonstrate the utility of the cloud for centralizing and maximizing the use of diverse data, the following presents a “Digital Water System” (DWS) – or network of cloud-connected visualization dashboards – developed by Carollo to support three key business functions for a water utility in Texas (Mishra, 2024).

- 1)**Asset Classification** Integrates data from GIS and EAM systems for asset risk assessment.
- 2)**Project Planning** Integrates multiple Excel-based capital investment project plans.
- 3)**Operations Management** Integrates SCADA and LIMS data to support operations decision-making.



While the data lives in a variety of locations with ownership spanning departments, the heart of the DWS is the Microsoft Azure Cloud infrastructure that enables the centralization of data from these systems. Azure Data Factory is used to automate data collection and transformation before ultimately landing the processed, production-level data in a secure Azure Blob cloud storage container.

Reference: Mishra, P. (2024, June). Centralized digital architecture to integrate utility data for analytics applications [Conference presentation]. AWWA Annual Conference & Exposition (ACE24), Anaheim, CA, United States.

1.11.3 Cybersecurity Concerns and Solutions

Cybersecurity is among the most commonly cited concerns about cloud-hosted AI/ML tools in the water sector, and understandably so — cloud adoption in water utilities lags significantly behind industries like finance, healthcare, and energy, and trust in cloud infrastructure is still developing across the sector. This hesitation deserves serious consideration, but so does the full picture of what cloud security entails.

When evaluating cloud service cybersecurity, utility managers should understand that the comparison is rarely between the cloud and a perfectly maintained, air-gapped system. For the most critical operational

technology (OT) systems — such as SCADA networks directly controlling treatment processes — physical air gaps remain a valid and highly secure option. However, most utility data management (IT systems, historical records, ML model inputs and outputs) does not operate in an air-gapped environment, and the practical comparison is between cloud-hosted solutions and typical on-premises IT infrastructure. In that context, major cloud providers such as AWS, Microsoft Azure, and GCP Platform offer enterprise-grade security capabilities that most utilities would find difficult to replicate independently. AWS, for example, manages all components from the host operating system and virtualization layer down to the physical security of its data center facilities — relieving a significant operational burden from customers (Amazon Web Services, 2024). Microsoft Azure maintains what it describes as the largest compliance portfolio in the industry, with independent audit reports verifying adherence to standards including ISO 27001, SOC 2, FedRAMP, and HITRUST (Microsoft Azure, 2024); AWS and Azure also hold government-specific certifications, including FedRAMP authorizations, (U.S. General Services Administration, 2024) for organizations requiring additional assurance.

Beyond security features, a cloud provider relationship shifts ongoing security maintenance and compliance responsibilities to an entity that specializes in exactly this work (Chenevey et al., 2021). Utilities with limited dedicated cybersecurity staff can benefit from this shared responsibility model. Rather than independently keeping pace with evolving AWIA, AWWA, and NIST cybersecurity requirements (American Water Works Association, 2024; National Institute of Standards and Technology, 2024b; U.S. Environmental Protection Agency, 2018), they can leverage the provider's continuous monitoring, automated patch management, and compliance tooling. This is not a replacement for a utility's own cybersecurity practices, but a meaningful complement to them.

Ultimately, the overall security posture of any cloud-based ML system will depend on the combination of the provider's controls and the utility's own risk management practices. Key utility responsibilities include enforcing appropriate access controls, providing staff cybersecurity training, maintaining strong separation between IT and OT networks, and establishing clear data governance and handling protocols. These layered defenses collectively determine system resilience.

Developing a cybersecurity plan is prudent for AI/ML scale-up, even if not using cloud services. Current industry standards list the following elements of a cybersecurity plan (International Organization for Standardization, 2022; National Institute of Standards and Technology, 2024a; Stouffer et al., 2023):

Risk Assessment: Identify and evaluate potential threats and vulnerabilities specific to the water system's data, applications, and supporting infrastructure.

1. **Access Control:** Implement multi-factor authentication, role-based access control, and least-privilege principles to restrict system access to only authorized users.
2. **Network Segmentation:** Separate critical databases and networks from less critical systems, using physical air gaps between IT and OT networks or implementing firewalls and demilitarized zones.
3. **Monitoring and Detection:** Increase network visibility to track user activity and data access patterns, enabling detection of unauthorized access or anomalous behavior.
4. **Incident Response and Crisis Management:** Establish, document, and periodically test procedures for responding to cyber incidents, including communication protocols and defined roles and responsibilities
5. **Employee Training:** Prioritize regular cybersecurity training for all staff to create a culture of cybersecurity awareness and vigilance.

1.12 Edge Computing for Machine Learning

Edge computing brings computation and data storage closer to data sources—directly at treatment plants, pumping stations, or monitoring sites rather than relying solely on centralized cloud servers. This enables real-time responses without constant cloud connectivity. Water utilities already implement edge processing extensively: programmable logic controllers (PLCs) and remote terminal units (RTUs) process almost all control logic, calculations, and alarms locally at the edge. Newer PLCs with Linux-based operating systems now allow Python scripts for embedded ML/AI algorithms alongside traditional control approaches, though utilities yet leverage this capability (Schneider Electric, 2024b, 2024a).

Cybersecurity considerations differ for edge versus cloud systems. Edge computing is often perceived to offer increased data security by limiting sensitive operational data transmitted across networks. However, distributed architectures can create larger attack surfaces. Edge deployments require protection at multiple points: device hardening, regular firmware updates, authentication protocols, and physical security for equipment that may be accessible to unauthorized actors. Both edge and cloud approaches require risk assessment, access control, network segmentation, monitoring, incident response planning, and employee training.

Hybrid architectures combining edge and cloud may provide the optimal approach:

- Edge devices could handle time-critical analytics (e.g., for disinfection control) and immediate operational decisions (Rockwell Automation, 2022, 2024).
- Cloud systems could provide computational power for especially complex ML models or cross-system optimization for recommendation-only setup (AVEVA, 2024).
- Data routing frameworks define what processes locally versus what transmits to the cloud, with platforms like Ignition Edge and AVEVA PI System providing the infrastructure for configurable hybrid data flows (AVEVA, 2024; Inductive Automation, 2024).

Successful implementations incorporate coordinated data governance defining how data moves between environments, federated learning approaches where edge devices contribute to model training without sending raw data centrally, and flexible deployment options as requirements evolve.

SECTION 2 DESKTOP PHASE MODELING AND EVALUATION

2.1 Metrics for Measuring ML Success

ML model performance is evaluated differently depending on whether the model predicts continuous numerical values (regression) or discrete categories (classification). Choosing the right metrics ensures your team can objectively assess whether a model meets operational requirements.

Regression Metrics: Measuring Numerical Accuracy

Regression models predict continuous numerical values like flow rates, chemical concentrations, or energy consumption. Common metrics include:

- **Root Mean Squared Error (RMSE):** Measures the average magnitude of prediction errors in the same units as your data (e.g., MGD, mg/L). Lower values indicate better performance. RMSE penalizes large errors more heavily than small ones.
- **Mean Absolute Error:** The average absolute difference between predicted and actual values. More interpretable than RMSE and less sensitive to outliers.
- **R-squared (R^2):** Represents the proportion of variance in the data that the model explains, ranging from 0 to 1. An R^2 of 0.85 means the model explains 85 percent of the variability in outcomes.
- **Mean Absolute Percentage Error:** Expresses errors as percentages, making it easier to compare performance across different scales of measurement. However, can give more weight to errors where the true value was low.

Classification Metrics: Evaluating Categorical Accuracy

Classification models categorize data into discrete groups, such as "normal operation" vs "alarm condition" or equipment failure risk levels. Key metrics include:

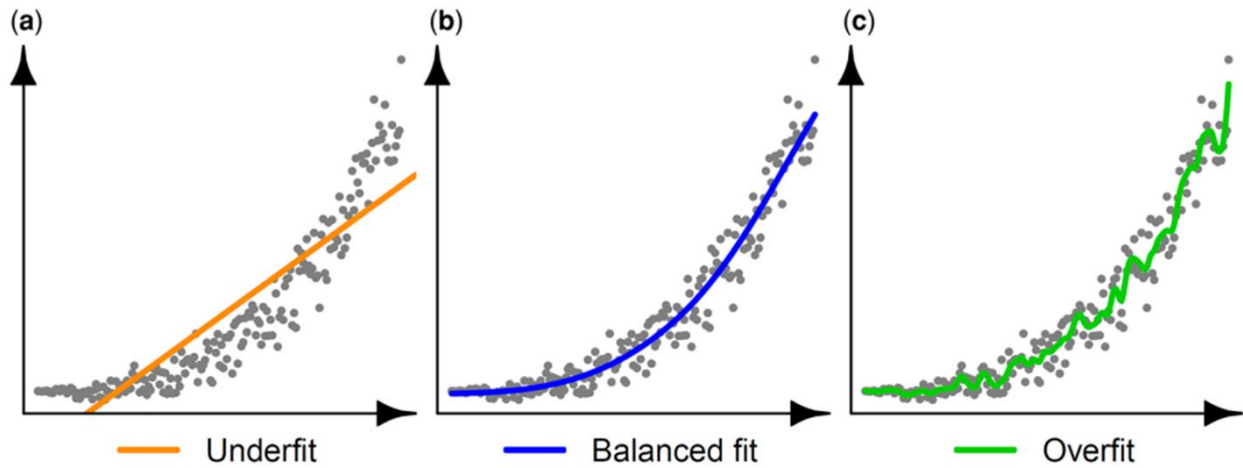
- **Accuracy:** The proportion of correct predictions out of all predictions. Simple to understand but can be misleading when classes are imbalanced (e.g., 99.99 percent normal operations, 0.01 percent failures).
- **Precision:** Of all instances predicted as positive (e.g., alarm conditions), what percentage were actually positive? Critical when false alarms are costly.
- **Recall (Sensitivity):** Of all actual positive instances, what percentage did the model correctly identify? Essential when missing true problems is dangerous.
- **F1 Score:** The harmonic mean of precision and recall, providing a balanced measure when you need to optimize both false positives and false negatives.
- **Confusion Matrix:** A table showing true positives, true negatives, false positives, and false negatives. Provides granular insight into where the model succeeds and fails.

The choice of metrics depends on operational priorities. For example, predicting influent flow requires minimizing RMSE to enable accurate chemical dosing decisions, while detecting membrane integrity breaches demands high recall to ensure no failures go undetected, even if it means accepting some false alarms.

2.2 Data Splitting: Essential for Real-World Machine Learning Success

When developing ML models, the goal is to ensure they perform well on new data they have not seen before, since that would be the case when using the models in practice. Because of the complexity and data-driven nature of ML models, they carry a risk of *overfitting*. When overfitting, models memorize random noise, errors, or spurious coincidences in the training data rather than learning the true generalizable relationships. Overfit models perform excellently on training data during development but

fail when deployed in real-world scenarios; that said, practitioners should also avoid underfit and aim for a balanced fit as illustrated in Figure 2-1.



Newhart, K.B., Hering, A.S. and Cath, T.Y., 2022. Data science tools to enable decarbonized water and wastewater treatment systems. *Pathways to Water Sector Decarbonization, Carbon Capture and Utilization*, p.275.

Figure 2-1 Illustration of models that are underfit, balanced fit, or overfit

To avoid overfitting and establish a reasonable model with balanced fit, we divide our complete dataset into two portions: a training dataset (blue in Figure 2-2) and a testing dataset (teal). The goal is to provide an unbiased evaluation of the final model, though in practice this requires careful implementation to avoid data leakage pitfalls such as random splits on time-series data or preprocessing across the full dataset (see Section 2.4 Data Leakage).

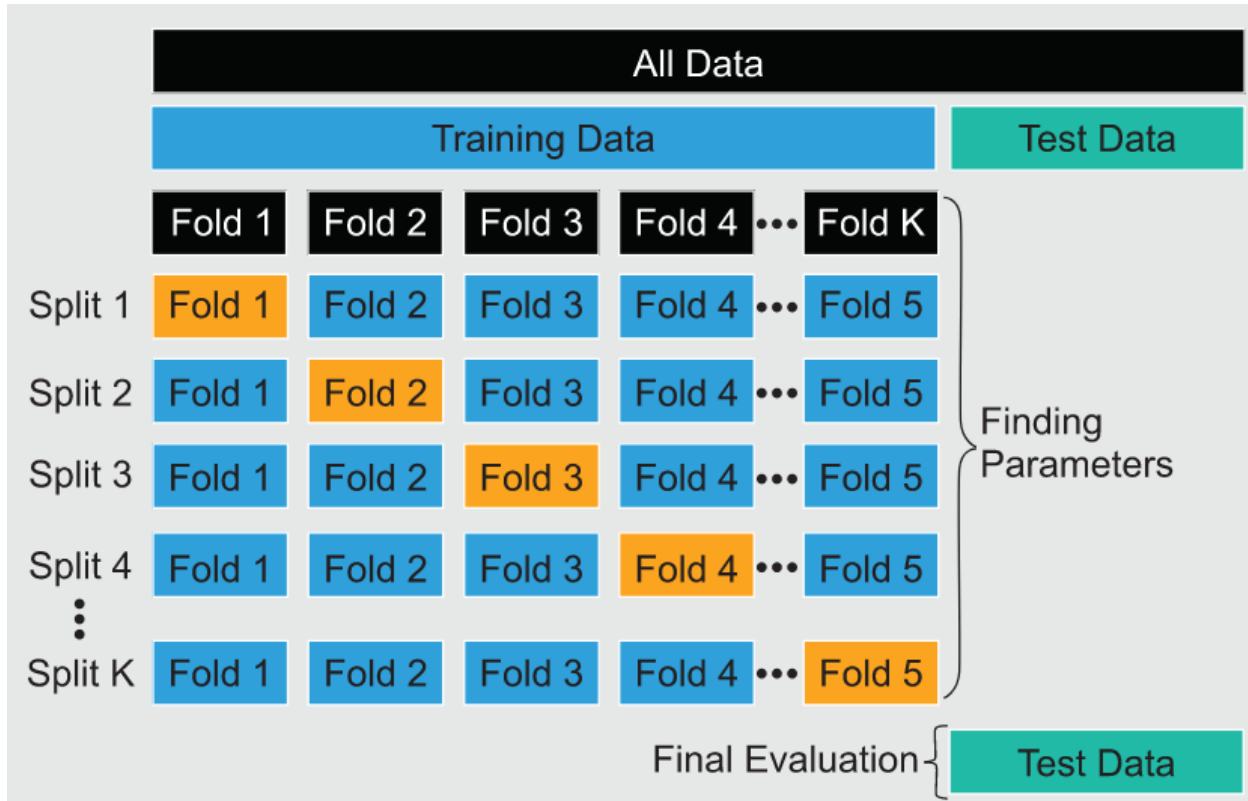


Figure 2-2 K-Fold Cross-Validation Structure

All data are divided into a fixed test set (teal) and a training set (blue), which is further split into K folds. In each of K splits, one-fold (orange) serves as the validation set while the remaining folds train the model — rotating until every fold has been validated. Final model performance is evaluated on the held-out test set only.

How we split the data depends on the nature of our data set. In many fields, randomly splitting the data is the default approach. There are also methods and functions for semi-randomly splitting the data while ensuring both the training and testing sets have representative range of key variables. However, for time series data (i.e., data with rows corresponding to chronological points in time) time-based splits are the best practice.

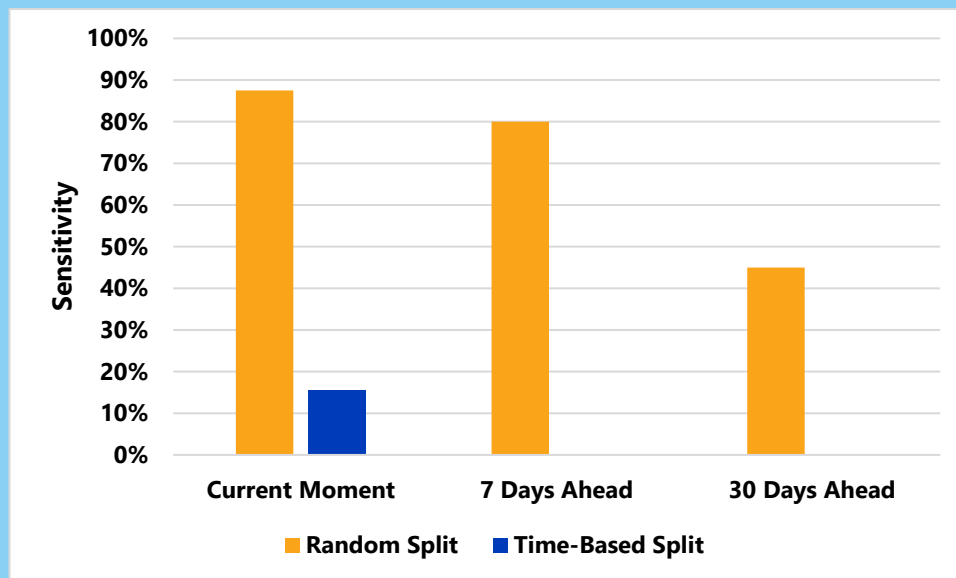
SCADA data is time series data. Time series data is usually *autocorrelated*. That is, water quality data will often correlate with its own recent past values, because it is recorded frequently in SCADA and does not usually change suddenly. Autocorrelated data is more at risk of coincidental, non-causal correlations with other autocorrelated data. For example, TOC and nitrate could appear correlated if they both had sustained peaks for separate reasons during summer of the same year. If randomly split, the training set and testing set would both include this fake correlation because they would both include data from the same summer. The ML model would appear successful in the desktop testing set but then fail in practice. However, if the testing set included data only from winter or the following summer, ML models relying on the fake correlation would have their true unreliability revealed. The project would be cancelled before wasting further resources, or more reliable models would be found and scaled up. An example of

unrealistic model performance expectations from apply random splits to time series data is given in Reuse Case Study #9.

Water quality can also drift or shift over time due to changes in the upstream wastewater treatment or factors like climate change. ML models would not be as well-calibrated for new data outside of the training set's range. Time-based data splits would give some indication of the accuracy loss from these drifts, if they cover a long enough time window. Random splits would hide this issue.

Reuse Case Study #9: Random vs Time-Based Data Splitting

As part of NAWI Project 5.17, Carollo strove to apply ML to forecast ammonia spikes over 3 mg/L in the feed water to a potable reuse demonstration. Ammonia spikes were defined as exceeding 3 mg/L. At first, it appeared ML could achieve this task. The ML was detecting over 80% of the ammonia spike datapoints, even seven days in advance. However, on closer look, these results were unrealistic due to a random data split. With a time-based data split, the ML did not predict any of the ammonia spikes seven days ahead. Most of the datapoints considered ammonia spikes were from a single, sustained event in the dataset. So, the ML with a random split was simply using coincidental water quality clues from that single time period, instead of finding real relationships with ammonia. After this realization, this ML tool was not chosen for scale-up.



2.3 Model Cross-Validation

While the initial training-test split helps prevent overfitting, we need additional validation during model development to tune parameters and evaluate performance. There are several approaches to validation, but this guide will illustrate one approach in-depth as an example.

K-fold CV involves dividing the training data randomly into K equal subsets or "folds." K can be different numbers, often 5 or 10, but Figure 2.1 shows an example with K=5 or "5-fold" CV. After these splits, we then perform K (in this case, 5) iterations of model training and validation:

- In each iteration or split, one fold serves as the validation set (orange in each split).

- The remaining 4 folds (blue) serve as the temporary training data.
- We train the model and evaluate its performance on the orange validation fold.
- We repeat this process until each fold has served as validation data once.

We then average the ML model performance on the orange validation sets as an estimate of the model performance on the teal final testing set or other new data in practice. We will measure ML model performance on the orange set based on the type of ML problem (see Section 2.1 Metrics for Measuring ML Success). This approach makes efficient use of limited data and reduces the variability in performance estimates.

For time series data, there are specialized CV approaches with similar workflow but that keep the data grouped chronologically instead of splitting into folds randomly. For example, in forward-chaining CV, model iterations are trained on a period of time and validated on the next period. Then, models are trained on the first two periods and trained on the third, and so on.

The model performance based on CV is often different but should be compared against the performance on the testing set. A small discrepancy is normal and informative. A significant drop in performance from CV to testing indicates overfitting. Conversely, comparable performance across validation and testing suggests a robust model that generalizes well. This comparison provides valuable insight into the model reliability and dictates if a model should be scaled up or additional refinement is needed before scaling up.

2.4 Data Leakage

Data leakage occurs during ML model development when information from the testing set influences the model creation process, resulting in unrealistically high performance metrics that will not translate to real-world applications. Warning signs include suspiciously high accuracy (>99 percent) on complex water phenomenon, significant performance gaps between CV and test set results, and inconsistent model behavior when deployed. At first glance, one might think data leakage can be avoided simply by doing a training/testing split, but there are subtle ways information from the testing set can influence an ML model. Preventing leakage requires time-based data splitting for time series data, isolated preprocessing pipelines, forecasting using only information available at prediction time, and regular reality checks with water experts. From a managerial perspective, the best ways to avoid data leakage are (1) incentivize developers based on how well the model works in practice, not in the desktop study phase or (2) reward developers based on how well the desktop study predicted the full-scale performance, not based on how the desktop model predicted the target variable.

Some common leakage issues are outlined below:

- **Target leakage:** Using variables unavailable at prediction time (e.g., downstream measurements to predict upstream processes despite significant differences in the water's residence time).
- **Time-series leakage:** Using future information to predict past events. E.g., interpolating backwards from LIMS data that would not have been collected yet, or using actual future rain data instead of the rain forecast at the earlier time.

- **Group leakage:** Related observations (e.g., from same lab batch) appearing in both training and test sets.
- **Preprocessing leakage:** Applying transformations (e.g., PCA or normalizing based on standard deviations) to the entire dataset before splitting.

SECTION 3 SCALE-UP, PILOTING, AND DEMONSTRATION

3.1 Scale-Up Pathways

The transition from ML model development and evaluation to real-world implementation can be implemented through various pathways:

- **On-premises deployment:** Computing resources located at your treatment facility (see Section 1.13).
- **Cloud deployment:** Models hosted on remote servers with data transmitted over secure connections (see Section 1.12).
- **Hybrid systems:** Combining local processing for time-sensitive operations with cloud capabilities for intensive computations.
- **Desktop models:** No scaling beyond the developer's computer, and the developer primarily runs the model upon request, in an isolated environment not connected to live data. The compute resource will simply be the computer that is used by the analyst.

Teams often associate the cloud with AI/ML tools and so may envision this form of deployment initially. However, desktop models or interactive webtools are sometimes sufficient, especially when the process being modeled is slow changing (e.g., irreversible RO fouling) or the input data is not real-time (e.g., microbial data from LIMS).

3.2 Choosing the Right Scale-Up Pathway

The specific storage and computing capabilities needed to run the ML model dictates what cloud-based or on-premise tools are necessary for scale-up.

A key factor in this decision is the response time requirements of the process being modeled (Rieger et al., 2025). In water treatment plants, real-time data availability is governed by SCADA scan rates—typically set to 1-second intervals—rather than by how fast the underlying process changes. For time-sensitive applications requiring near-immediate model predictions, on-premise deployment reduces transmission latency compared to cloud-based approaches. For such applications, there are also different levels of on-premise implementation: models can run from an external database outside the SCADA system, or—for the fastest response and fewest points of failure—directly within the PLC itself.

Typical response time needs and example scale-up pathways are listed below:

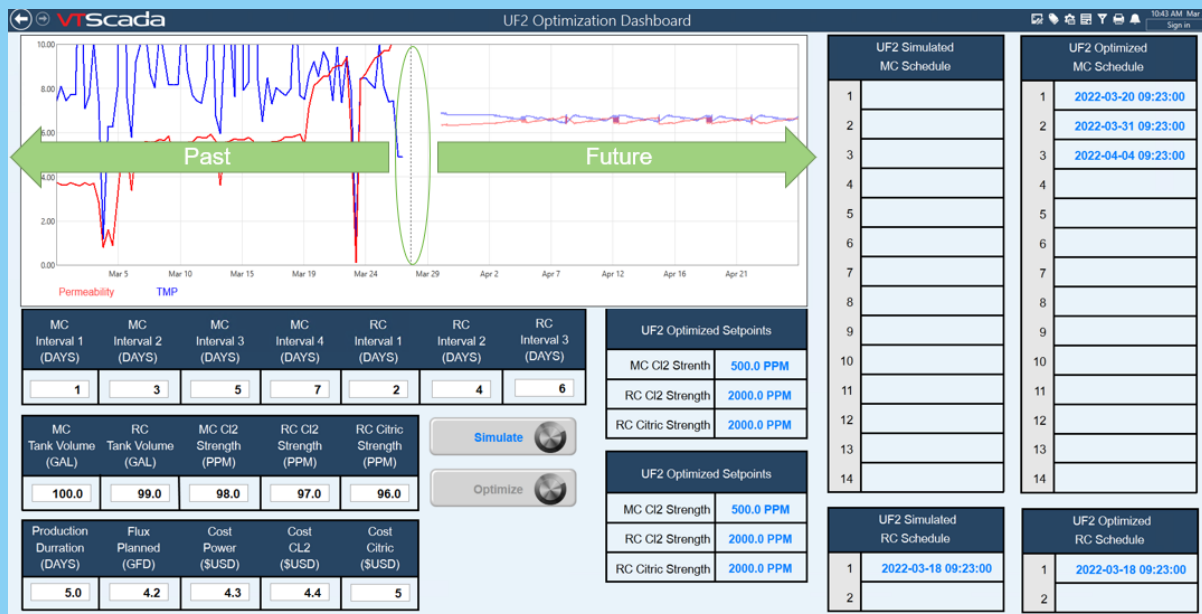
- **Fast-response processes** (e.g., disinfection with ozone, UV, chlorine):
 - » Require near-immediate model predictions.

- » Temporary lapses could have real consequences without hardwired defaults to conservative doses.
- » Often best suited to on-premise deployment, where transmission delays to and from the cloud are avoided. For fully automated control, PLC-level implementation provides the fastest response.
- **Medium-term processes** (e.g., Ultrafiltration [UF] fouling, biofiltration headloss, see Reuse Case Study #10):
 - » Predictions needed on daily or weekly cycles.
 - » Well-suited for cloud or hybrid approaches.
- **Long-term processes** (e.g., RO membrane fouling, GAC breakthrough):
 - » Prediction timelines measured in weeks or months.
 - » Less sensitive to occasional connectivity issues.
 - » Ideal for cloud-based solutions that facilitate SCADA-LIMS data integration, model retraining, and collaboration.
 - » Desktop models or non-real-time dashboards could suffice but require more manual inputs.

Data security requirements and utility cybersecurity policy can also determine the scale-up pathway (see Section 1.12.3). Advances in PLCs and edge computing devices are also raising the ceiling for the complexity of AI/ML solutions that could be implemented on-premise.

Reuse Case Study #10: Near-Real-Time Cloud-Based Scale-Up for a Medium-Speed Treatment Process

As part of an earlier USBR project (National Water Research Institute et al., 2023), Yokogawa and Carollo developed a cloud-based UF optimization dashboard for a potable reuse demonstration facility. Conventionally, UF membranes are cleaned at conservative, fixed intervals such as twice weekly. This cleaning frequency is likely more than needed and uses energy and chemicals like chlorine and citric acid. Yokogawa modeled the rate of UF fouling as a function of water quality and its permeability after the last clean. Water quality and operational data (e.g., transmembrane pressure) went to the cloud and from there to the VTScada dashboard. Operators could manually input financial data such as the cost of citric acid. With the click of an “Optimize” button, an AI optimization algorithm would then recommend a UF cleaning schedule, balancing the costs of cleaning vs the cost of pumping water through relatively fouled membranes. The optimization was with a click, not automatic, because of the computational intensity, requiring several minutes to run. However, this speed and set-up were sufficient considering UF only needed to be cleaned about weekly in typical conditions.



Reference: National Water Research Institute, Carollo Engineers, Inc., & Yokogawa Corporation of America. (2023). *Intelligent Human Machine Interface to Support Broader Adoption of Potable Water Reuse (Final Report Pitch to Pilot Final Report No. 1)*. U.S. Department of the Interior, Bureau of Reclamation. <https://www.usbr.gov/research/dwpr/>

3.3 Beyond Test Set Prediction: The Pilot Imperative

Desktop phase validation using the testing set is only the first step in evaluating an ML model. Successful implementation requires demonstration of ML model accuracy in actual operating conditions.

Why pilot demonstrations matter:

- Testing sets cover only a period of time. They cannot capture all real-world variability, particularly long-term drifts or unprecedented spikes.
- Operator feedback improves interface design and usability.
- Integration issues with existing systems may become apparent.

AI/ML piloting for reuse can take two forms:

- Demonstration on a potable reuse pilot or demonstration facility.
- Demonstration directly on a full-scale system but with public health guardrails in place

Demonstrating AI/ML on a potable reuse pilot may require a demonstration facility. In the reuse community, demonstration facilities are pilots large enough to use and validate full-scale components and equipment. As another benefit, demonstration facilities generally include more sophisticated controls and data management and more sensors than smaller scale pilots, enabling real-time AI/ML. Pilots or demonstration facilities can also vary operational settings widely with no risk to public health, generating a more robust training set.

However, ML is extremely site-specific. There have been incidents where ML models trained on a membrane or filter within a full-scale facility were not accurate for other membranes or filters within the same facility. The full-scale system may also have different sensors available, or more of the same sensors but replicated on parallel trains. So, any AI/ML operational tool developed and demonstrated at a demonstration facility would still require retraining and some degree of re-demonstration at the full-scale facility.

The following is a general pathway for conducting effective pilot to full-scale AI/ML demonstrations:

1. **Establish a public health protection plan.** Public safety is always the top priority in the water sector and potable reuse. As discussed above, full-scale AI/ML operational tools will always require some form of full-scale trial period to ensure their accuracy and reliability. ML alert systems do not directly impact the treatment intensity or water quality, so they do not carry any health risks as long as the prior alarm protocols remain in place in parallel. For any AI/ML tool that could potentially lower treatment intensity, safety measures could include:
 - a. Whenever the timescale allows, trial semi-autonomous operation (i.e., AI recommendations sent to operators) before full automation.
 - b. Leave all prior alert and alarm systems in place.
 - c. Trial the AI/ML control system on one parallel train first. This allows blending or discharge to sewer for the train controlled with AI/ML. This is also the best practice for benchmarking performance, since the efficiency and water quality of the AI/ML controlled train can be directly compared to the other train(s) treating the same quality of water at the same time with the prior control strategy.
 - d. Discharge the water treated with AI/ML control to sewer or recycle to headworks, especially for direct potable reuse.
 - e. Program constraints such that the AI/ML cannot set the treatment intensity lower than the old control system, or another safe minimum, until months of full-scale validation.
2. **Connect robust data pipelines and integration systems.** Reliable, well-structured data pipelines are the connective tissue of any operational ML system. Key requirements include:

- a. **Break down data silos.** ML models perform best when they can draw on data from across utility departments — operations, laboratory, and maintenance — in real time or near-real time. Integrating SCADA, LIMS, and other data sources into a unified pipeline expands the inputs available to models and reduces blind spots.
 - b. **Ensure seamless integration with existing control systems.** ML applications must interface cleanly with existing PLCs and SCADA systems, with clear protocols governing how ML outputs interact with control logic (Hall & Agarwal, 2024). For the majority of U.S. water utilities — which are small systems with limited IT resources — this integration should be designed to minimize operational disruption and not require extensive custom development (Sela et al., 2024).
 - c. **Build in flexibility and validation.** Data pipelines should be designed to adapt to changing conditions (e.g., seasonal variation, sensor replacement, process modifications) and include automated quality checks to flag anomalies or missing data early. Poor data quality is one of the most common causes of ML model degradation in operations.
 - d. **Implement version control.** Track changes to both data pipelines and ML models so that performance issues can be traced to specific updates. This is especially important during the validation period when the model is being tested against new field data.
3. **Deploy the ML model in a controlled, limited setting** such as a pilot or parallel train, as discussed above.
 4. **Validate model performance with new field data.** Compare ML predictions against new data as it rolls in real-time. Track and plot key performance metrics (accuracy, RMSE) automatically or daily. Check for model drift and investigate the cause of any substantial model errors. Assess not only whether the model was accurate but whether the desktop study correctly predicted full-scale accuracy, to refine future desktop methodology.
 5. **Quantify return on investment through multiple metrics.** Track energy, water loss, operator hours, or chemical use, as well as qualitative benefits, and compare to non-AI/ML trains. Document and visualize improvements to share with stakeholders.
 6. **Iterate and optimize based on pilot findings.** Refine or retrain model parameters based on real-world performance. Test the impact of enhanced sensor maintenance on AI/ML performance. Compare AI/ML models that performed similarly in the desktop phase.

Once an AI/ML tool has been demonstrated for months or a year without any adverse incidents, it could be transitioned from pilot to full-scale or from one parallel train to all trains. The efficiency and resulting water quality of the tool should remain closely tracked, and prior controls systems should be kept available as backups. Ongoing R&D could investigate whether additional sensors enhance the model.

SECTION 4 THE LONG TERM

4.1 Overview

Successfully implementing ML in water treatment extends far beyond pilot project success. The critical challenge lies in sustaining ML systems over years and decades—maintaining performance through seasonal variations, equipment upgrades, staff turnover, and evolving operational priorities. While

short-term success depends on technical factors like model accuracy, long-term success hinges on organizational capacity, workforce readiness, and strategic planning.

Water utilities face a fundamental decision: invest now in building ML capabilities, or risk falling behind as operational challenges intensify. This decision carries substantial uncertainty—upfront costs are immediate and tangible, while benefits may take years to materialize and can be difficult to quantify when they involve intangibles like enhanced reliability or avoided failures. The question is not simply "Does this ML application work?" but "Can we sustain it successfully over its operational lifetime?"

This chapter draws extensively on lessons from critical infrastructure sectors that have navigated ML adoption over the past decade. Energy utilities, nuclear power facilities, and advanced manufacturing have confronted remarkably similar barriers and developed evidence-based strategies for overcoming them.

These sectors share key characteristics with water utilities: critical infrastructure with zero tolerance for failures; risk-averse cultures shaped by regulatory oversight; long equipment lifecycles; multi-disciplinary operations; and aging workforces facing knowledge loss while adopting new technologies. The energy sector offers insights on derisking ML investments through scenario planning and innovative financing (Park, 2025). Nuclear power—operating under similarly strict oversight as water utilities—provides guidance on workforce transitions and organizational readiness (Hall & Agarwal, 2024). Manufacturing demonstrates balancing efficiency gains against operational risks.

Recent water-specific research reinforces these cross-sector lessons. Water utilities face barriers including steep learning curves, poor workflow integration, financial constraints (especially for smaller systems), and workforce readiness challenges (Sela et al., 2025; Vekaria & Sinha, 2024). The water sector can learn from other industries' successes and failures while adapting strategies to water's unique context.

This chapter addresses the full ML implementation lifecycle. Section 4.2 covers ML solutions deployment and maintenance, establishing the ML Operations (MLOps; Kreuzberger et al. 2023b) paradigm and technical platforms needed. Section 4.3 examines long-term resource requirements, including derisking strategies, technical infrastructure, workforce development, and cost considerations.

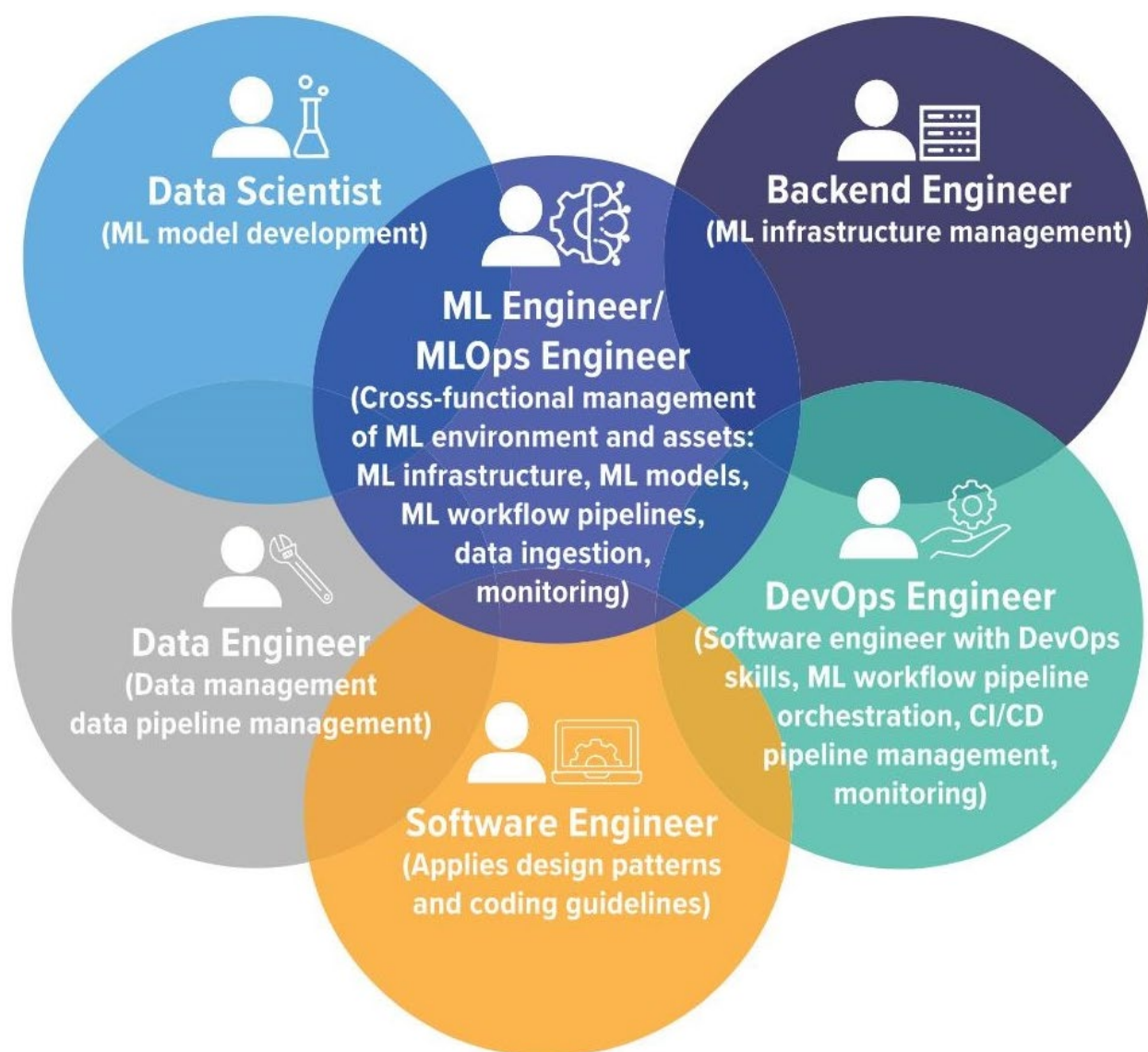
Throughout, we draw on cross-sector evidence (Sela et al., 2025; Sinha et al., 2025) while acknowledging that water utilities operate under distinct regulatory frameworks, predominantly public ownership structures, and more decentralized resource constraints than energy or manufacturing. The goal is not to blindly import approaches, but to provide utilities with a broader evidence base for making context-appropriate decisions about ML implementation.

4.2 ML Solutions Deployment and Maintenance

Successfully deploying ML solutions in water treatment requires understanding that model development is only the beginning. A highly accurate ML model provides operational value only to the extent it is consistently used. A model run manually and intermittently by a single analyst—however capable—cannot deliver the reliability, frequency, or organizational reach of a model integrated into decision-making workflows through user interfaces or automated systems. This transition from development to deployment represents a critical challenge that water utilities share with other sectors implementing ML at scale.

4.2.1 The MLOps Paradigm

The life-cycle of an AI/ML-based solution involves a team with diverse skillsets (Kreuzberger et al., 2023a), as illustrated in Figure 4-1. Kreuzberger et al. describe ML Operations (MLOps) as "a paradigm, including aspects like best practices, sets of concepts, as well as a development culture when it comes to the end-to-end conceptualization, implementation, monitoring, deployment, and scalability of ML products." MLOps leverages three contributing disciplines: ML, software engineering (especially DevOps), and data engineering, with the goal of bridging the gap between model development and operational deployment.



Adapted from Kreuzberger et al., 2023

Figure 4-1 Roles and their intersections contributing to the MLOps paradigm

This paradigm highlights that ML implementation is not purely a technical challenge—it requires coordination across multiple domains of expertise, from water treatment operations to data science to software engineering.

4.2.1.1 What Does This Mean for a Typical Water Utility?

A utility leader reviewing this diagram might reasonably ask: "What do I already have, and what's missing?" For most water utilities, the honest answer is that most of this diagram represents a genuine capability gap — not a hidden asset waiting to be repurposed.

The water sector faces a structural staffing reality that makes MLOps particularly challenging. More than half of all U.S. water and wastewater utilities have just one or two total employees, and 85 percent have three or fewer (American Public Works Association, 2020). Even at larger utilities, the typical IT staff member is a generalist simultaneously managing the network, maintaining a database, and troubleshooting dashboards — far from the specialized roles shown in Figure 4-1. In a survey of 680 water sector stakeholders, 54 percent of utilities cited staffing resources as their primary constraint to digital adoption, and 60 percent reported already outsourcing engineering and technical functions (Black & Veatch, 2025).

Across utility sizes, roles like Data Scientist, ML/MLOps Engineer, and Backend Engineer are uncommon to-date. Many utility engineers have spent careers working primarily in spreadsheets, with limited exposure to the programming environments and cloud infrastructure that MLOps requires (Prevos, 2024). The skills gap is not unique to water — even across industries, one in three IT leaders reports difficulty finding qualified ML staff (World Economic Forum, 2025)— but water utilities face this competition with smaller budgets and less brand recognition as major software firms.

This does not mean ML adoption is out of reach. Emerging no-code and low-code platforms are enabling utility operators and engineers to engage with data analytics without formal data science training, shifting some capability toward the frontline workforce (Water Research Foundation, 2025). For more sophisticated ML applications, however, most utilities will need to access expertise externally — through vendors, consulting firms, academic partners, or regional utility consortia — rather than relying on internal hiring alone. Section 4.2.2 discusses practical strategies for building this capability incrementally.

4.2.2 Building or Acquiring MLOps Capability

To efficiently develop and maintain ML solutions long-term, utilities face a strategic decision: retain external companies with comprehensive MLOps capabilities, or build internal teams. Many utilities have some elements of the required skillset—a growing number employ data scientists, and IT departments may include personnel with data engineering or software engineering backgrounds—but rarely possess the complete set of capabilities needed for sustained ML operations.

The most efficient approach to building internal MLOps capacity typically involves:

- **Assess existing capabilities:** Inventory current data science, IT, and operational expertise within the organization.
- **Add dedicated ML/MLOps expertise:** This central coordinating role bridges between all other functions.
- **Upskill existing personnel:** Train DevOps or backend engineers in ML-specific knowledge, and water treatment operators in data literacy.

Many utilities find success starting with small cross-functional teams that combine existing IT personnel with new ML specialists. This approach leverages institutional knowledge of IT infrastructure and treatment operations while adding specialized ML expertise.

Alternatively, utilities can retain consultants, software companies, or equipment/controls vendors through ongoing contracts to monitor and retrain ML models, maintain data pipelines, and provide technical support. This outsourced approach may be more cost-effective initially but requires careful attention to knowledge transfer and long-term sustainability.

4.2.3 Technical Infrastructure for ML Deployment

4.2.3.1 MLOps Platforms and Tools

Various MLOps tools have been developed to streamline the transition from desktop model development to full-scale deployment. These tools address model and data version management, workflow orchestration, and DevOps integration. Table 4-1 summarizes evaluated technologies for MLOps platforms, categorized as open-source or commercial options (Kreuzberger et al., 2023a).

Table 4-1 List of evaluated technologies for Machine Learning Operations platforms and tools, categorized by open-source and commercial examples (Kreuzberger et al., 2023)

Category	Technology Name
Open-source	TensorFlow Extended
Open-source	Airflow
Open-source	Kubeflow
Open-source	MLflow
Commercial	Databricks managed MLflow
Commercial	Amazon CodePipeline
Commercial	Amazon SageMaker
Commercial	Azure DevOps Pipelines
Commercial	Azure ML
Commercial	GCP + Vertex AI
Commercial	IBM Cloud Pak for Data (IBM Watson Studio)

For utilities with limited technical resources, understanding these tools at a high level and selecting appropriate commercial platforms for adoption is typically more practical than building custom solutions from open-source components.

4.3 Resource Requirements for Long-Term Success

Water utilities implementing ML solutions must address fundamental resource categories to ensure sustained operational success. Moving an ML project from desktop phase to full-scale operationalization requires not only diverse hardware and technical skillsets, but also strategic planning for long-term maintenance amid seasonal variations, equipment updates, and staff turnover. Success depends on whether facilities anticipate and plan for necessary resources throughout the entire ML lifecycle—from initial deployment through ongoing operations.

4.3.1 Derisking AI/ML Implementation

The substantial upfront investment and uncertain returns associated with ML adoption represent significant barriers for water utilities (Figure 4-2). Drawing from cross-sector experience in the energy industry, utilities can employ several strategic approaches to mitigate these risks and build confidence in ML investments.

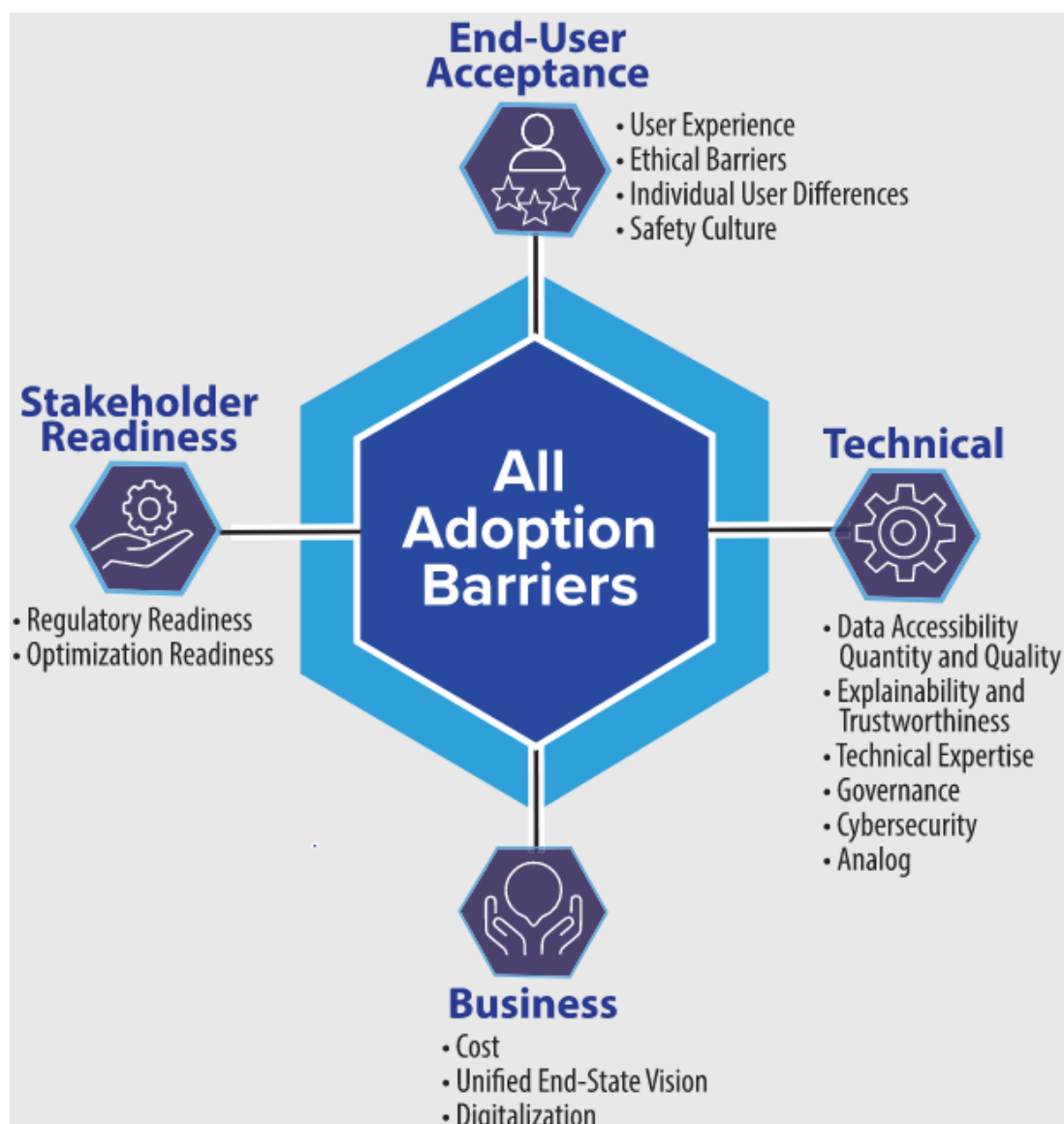


Figure 4-2 Four interconnected categories of AI/ML adoption barriers identified in nuclear power; adapted from Hall and Agarwal (2024), applicable to water utilities facing similar regulatory and operational constraints

In nuclear power — which shares water utilities' risk-averse culture and strict regulatory environment — no single barrier should be tackled in isolation; cost concerns, workforce readiness, regulatory uncertainty, and end-user trust are deeply interconnected and tend to reinforce each other (Hall & Agarwal, 2024).

Scenario Planning and Pilot Projects

Rather than committing to full-scale implementation, utilities should model different deployment scenarios to clarify potential ROI, then test ML applications through phased pilots where performance can be validated and adjustments made at lower cost (Park, 2025; Sela et al., 2025). Unlike treatment system pilots, these ML trials operate at full operational scale using existing data streams, without requiring upfront investment in complete software infrastructure.

Financing Strategies and Business Case Development

For high-capital-expenditure ML initiatives, securing alternative funding sources can reduce financial risk. Options include pursuing government grants targeted at water infrastructure modernization, forming public-private partnerships that share both investment costs and operational risks, and exploring financing arrangements that tie payments to demonstrated performance improvements (Park, 2025).

A critical challenge in building the business case is that ML benefits are often delayed — cost savings from fault detection or predictive maintenance may not materialize until months or years after deployment (Hall & Agarwal, 2024). Performance-based financing arrangements are particularly well-suited to this reality, as they align vendor incentives with demonstrated utility outcomes rather than requiring utilities to project uncertain future returns upfront. Overcoming this barrier also requires all stakeholders — from operations staff to senior management — to share a clear, unified vision of what ML implementation will look like across the organization (Hall & Agarwal, 2024).

Regulatory Sandboxes and Adaptive Governance

Where applicable, regulatory sandboxes — controlled environments for testing novel ML applications under oversight — allow utilities and regulators to collaboratively develop appropriate standards as technology evolves (Park, 2025). Nuclear power experience suggests that regulatory concern tends to diminish over time as evidence accumulates; early engagement is more productive than waiting for certainty (Hall & Agarwal, 2024).

Transitioning Workforce with Digital Upskilling

Because organizational resistance is as significant a barrier as technical or financial risk (Hall & Agarwal, 2024), utilities should develop reskilling pathways, create roles combining water treatment expertise with data literacy, and use public-private partnerships to build internal capacity over time (Park, 2025). Water utilities are already navigating this transition: treatment operators are shifting from manual adjustment to data interpretation and troubleshooting, while meter readers are being reassigned as data analysts and meter technicians (Oxenford et al., 2019). A particularly critical emerging role is the Data Manager — responsible for integrating data streams from multiple sources, translating outputs into operationally actionable information, and serving as the interface between operations staff and IT (Oxenford et al., 2019). At smaller utilities, this function may need to be outsourced rather than hired in-house. Reskilling pathways should also deliberately preserve manual operation skills: as automation increases, operators risk losing the hands-on proficiency needed when systems fail, making simulation-based training an important complement to digital upskilling (Oxenford et al., 2019).

4.3.2 Technical Infrastructure

Successful ML deployment requires robust technical infrastructure addressing three critical areas: data management, system integration, and operational reliability (Hall & Agarwal, 2024).

Data Accessibility and Quality

ML systems require substantial quantities of high-quality operational data. Water utilities face challenges similar to those in nuclear power, where real-time process data may not be readily available or exist in suitable formats (Hall & Agarwal, 2024). Within the water sector specifically, barriers include accessibility (specialized knowledge and coding requirements), usability (poor integration into existing workflows), and trainability (steep learning curves for advanced tools) (Sela et al., 2025; Vekaria & Sinha, 2024). Essential investments include enhanced instrumentation, flexible data pipelines that adapt to changing conditions, and robust validation protocols to identify quality issues early.

System Integration and Version Control

ML applications must integrate seamlessly with existing PLCs, SCADA systems, and OT while maintaining clear protocols for how ML recommendations interact with control logic (Hall & Agarwal, 2024). Equally important are version control systems that track model changes, allowing operators to trace performance issues to specific updates. For water utilities, this integration must account for the sector's predominantly small-scale operations—92.7 percent of U.S. water systems serve fewer than 10,000 people and often lack extensive IT resources (Sela et al., 2025).

Cybersecurity

Integration of ML systems into operational networks expands the attack surface at every lifecycle stage — from data collection through model deployment and ongoing updates (Hall & Agarwal, 2024). Water utilities should address cybersecurity in the formative planning stages of any ML project, not as an afterthought at deployment.

Each stage of building and running an ML system creates a different type of vulnerability (Hall & Agarwal, 2024). In plain terms:

- **Data collection/access** — when data is pulled from sensors, SCADA, or PLCs into a model, that connection is an entry point. Someone could intercept or manipulate the incoming data, causing the model to learn from corrupted inputs without anyone noticing.
- **Model training** — if training data is tampered with (called "data poisoning"), the model learns the wrong patterns. For a water treatment process control model, this could mean it confidently recommends the wrong chemical dose.
- **Model deployment** — when the model gets connected to operational systems to issue alerts or recommendations, that integration point can be exploited to push unauthorized commands toward control systems.
- **Model updates and retraining** — every time the model is retrained or updated with new data, this re-opens the data pipeline. If that process is not secured, it is another opportunity for interference.
- **Ongoing operation** — a running model that is continuously ingesting real-time sensor data is a persistent connection between your IT network and OT. That persistent link is exactly what attackers target in critical infrastructure.

4.3.3 Workforce and Organizational Readiness

Human factors surrounding ML adoption are as important as technical considerations (Hall & Agarwal, 2024). The water sector faces particular challenges including an aging workforce, high turnover rates, and institutional knowledge loss (Sela et al., 2025; Vekaria & Sinha, 2024).

Building Specialized Expertise

ML maintenance requires both data science capabilities and water treatment domain knowledge. Utilities need access to expertise for periodic model refinement, interface design support, and system architecture guidance. The multidisciplinary nature of successful ML deployment means no single domain expert can work in isolation (Hall & Agarwal, 2024; Park, 2025).

Operator Trust and Engagement

Building operator trust requires demonstrating that ML systems enhance rather than replace human judgment. Key strategies include establishing formal feedback channels, involving operators early in design processes, providing clear explanations of ML recommendations (explainable AI), and addressing ethical concerns transparently (Hall & Agarwal, 2024). Water-specific guidance emphasizes interpretable AI models that integrate seamlessly with existing workflows, enabling operators to question and validate AI decisions (Sela et al., 2024).

Organizational Culture

Successful ML adoption requires executive leadership framing ML as strategic investment, transparent communication about workforce transitions, and viewing implementation as ongoing learning rather than one-time deployment. The risk-averse safety culture in water utilities—similar to nuclear power—can create adoption resistance that must be overcome through careful validation and appropriate human oversight (Hall & Agarwal, 2024). Survey data from water utilities worldwide reveals a shared need for clearer understanding and guidance in AI applications, with particular challenges in strategy development and workforce readiness (Vekaria & Sinha, 2024).

Regulatory Coordination

For ML applications affecting compliance-critical systems, utilities should engage proactively with regulators by keeping them informed of pilot projects, participating in industry standards development, and documenting validation procedures (Hall & Agarwal, 2024).

4.3.4 Cost Considerations

Implementing a sustainable MLOps approach requires budget allocation across:

- Initial deployment costs (sensors, infrastructure setup, integration with existing systems).
- Ongoing operational expenses (sensor maintenance, cloud services, software licenses):
 - » Periodic retraining and model improvement cycles.
- Staff costs (hiring, training, retention of specialized personnel).

The investment in interdisciplinary MLOps, foundational data management, and robust sensors could be viewed as insurance against pitfalls such as promising models that are never scaled or scaled-up tools that drift or fall out of use over time.

4.4 What's Ahead

While this guidebook focuses on data analytics and predictive ML applications in water treatment, emerging AI paradigms warrant attention for future implementation. Cross-industry research identifies several critical gaps that water utilities should monitor: practical adoption of Generative AI including LLM Operations and associated data management requirements; agentic AI systems capable of autonomous decision-making and action; edge computing architectures that enable distributed ML processing closer to data sources (Plathottam et al., 2023); AI-native data management systems designed specifically for ML workflows; and water utility-specific strategies for mitigating adoption risks unique to GenAI and agentic AI deployments. For readers interested in exploring agentic AI applications in water systems, the International Water Association's white paper (Chandrasekeran, 2026a) on agentic AI provides an accessible introduction to use cases and value propositions specifically relevant to the water sector.

SECTION 5 REFERENCES

- Ahmed, S., Alshater, M. M., Ammari, A. E., & Hammami, H. (2022). *Artificial intelligence and machine learning in finance: A bibliometric review*. *Research in International Business and Finance*, 61, 101646. <https://doi.org/10.1016/j.ribaf.2022.101646>
- Amazon Web Services. (2024). *Shared Responsibility Model*. <https://aws.amazon.com/compliance/shared-responsibility-model/>
- American Public Works Association. (2020). *Water and Wastewater Utility Workforce Report*. American Public Works Association.
- American Water Works Association. (2024). *Cybersecurity & Guidance*. <https://www.awwa.org/resource/cybersecurity-guidance/>
- AVEVA. (2024). *AVEVA PI System*. <https://www.aveva.com/en/products/aveva-pi-system/>
- Black & Veatch. (2025). *Water Utilities Dive Deep into Data: Unlocking Resilience and ROI in a Digital Age*. Black & Veatch. <https://www.bv.com/en-US/perspectives/water-utilities-dive-deep-into-data-unlocking-resilience-and-roi-in-a-digital-age>
- Breiman, L. (2001). *Random Forests*. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chandrasekeran, P. (2026a). *GenAI and AgenticAI for Water Utilities*.
- Chandrasekeran, P. (2026b). *Preparing the Water Sector to Embrace Technology: Skillsets and Enterprise Management Approaches for the Digital Age* (No. 5178). <https://www.waterrf.org/research/projects/preparing-water-sector-embrace-technology-skillsets-and-enterprise-management>
- Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). *shiny: Web Application Framework for R [Computer software]*. <https://shiny.posit.co/>
- Chenevey, J., Ganesan, V., Richter, W., & Roth, E. (2021, July). *Security as Code: The Best (and Maybe Only) Path to Securing Cloud Applications and Systems*. <https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/security-as-code-the-best-and-maybe-only-path-to-securing-cloud-applications-and-systems>
- Chesterman, X., Verstraeten, T., Daems, P.-J., Nowé, A., & Helsen, J. (2023). *Overview of normal behavior modeling approaches for SCADA-based wind turbine condition monitoring demonstrated on data from operational wind farms*. *Wind Energy Science*, 8(6), 893–924. <https://doi.org/10.5194/wes-8-893-2023>
- Croll, H. C., Ikuma, K., Ong, S. K., & Sarkar, S. (2025). *Reinforcement learning optimization of a water resource recovery facility: Evaluating the impact of reward function design on agent training, control optimization, and treatment risk*. *Journal of Water Process Engineering*, 69, 106658.
- Dent, S., Meeks, C., Shindewolf, A., Victor, M., & Mishra, P. (2023). *Recognizing Value in an Open-Architecture Digital Water System: San Jacinto River Authority's Approach to Becoming a Smart Water Utility*. WEFTEC.

- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- Giammar, D. E., Greene, D. M., Mishra, A., Rao, N., Sperling, J. B., Talmadge, M., Miara, A., Sitterley, K. A., Wilson, A., Akar, S., & Kurup, P. (2021). Cost and energy metrics for municipal water reuse. *ACS ES&T Engineering*, 2(3), 489–507.
- Grimm, T. R., Branch, A., Thompson, K. A., Salveson, A., Zhao, J., Johnson, D., Hering, A. S., & Newhart, K. B. (2024). Long-Term Statistical Process Monitoring of an Ultrafiltration Water Treatment Process. *ACS ES&T Engineering*, 4(6), 1492–1506.
- Hall, A., & Agarwal, V. (2024). Barriers to adopting artificial intelligence and machine learning technologies in nuclear power. *Progress in Nuclear Energy*, 175, 105295. <https://doi.org/10.1016/j.pnucene.2024.105295>
- Hosseini, S., & Seilani, H. (2025). The role of agentic AI in shaping a smart future: A systematic review. *Array*, 26, 100399. <https://doi.org/10.1016/j.array.2025.100399>
- Inductive Automation. (2024). Ignition SCADA for the Water & Wastewater Industry. <https://page.inductiveautomation.com/industry/water-wastewater>
- International Organization for Standardization. (2022). Information security, cybersecurity and privacy protection – Information security management systems – Requirements (Standard ISO/IEC 27001:2022). International Organization for Standardization.
- Koyama, Y., Fasaee, M. A. K., Berglund, E. Z., & Knappe, D. R. U. (2024). Machine Learning Models to Predict Early Breakthrough of Recalcitrant Organic Micropollutants in Granular Activated Carbon Adsorbers. *Environmental Science & Technology*. <https://doi.org/10.1021/acs.est.4c01316>
- Kreuzberger, D., Kühl, N., & Hirschl, S. (2023a). Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE Access*, 11, 31866–31879. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2023.3262138>
- Kreuzberger, D., Kühl, N., & Hirschl, S. (2023b). Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE Access*, 11, 31866–31879. <https://doi.org/10.1109/ACCESS.2023.3262138>
- Lang, J. R., Allred, B. M., Field, J. A., Levis, J. W., & Barlaz, M. A. (2017). National estimate of per-and polyfluoroalkyl substance (PFAS) release to US municipal landfill leachate. *Environmental Science & Technology*, 51(4), 2197–2205.
- Lesnik, K. (2025, September). From Pilot to Production: Building Scalable Agentic AI Systems for Water Utilities. *WEFTEC 2025 Conference Proceedings*. https://weftec25.mapyourshow.com/8_0/sessions/session-details.cfm?scheduleid=313&
- Microsoft Azure. (2024). Azure Compliance Offerings. <https://azure.microsoft.com/en-us/explore/trusted-cloud/compliance>
- Mishra, P., Stinson, B., Newhart, K., & Nading, T. (2024, June). Centralized Digital Architecture to Integrate Utility Data for Analytics Applications. *Proceedings of the AWWA Annual Conference & Exposition (ACE24)*. <https://www.awwa.org/ace24>
- National Institute of Standards and Technology. (2024a). The NIST Cybersecurity Framework (CSF) 2.0 (NIST CSWP 29; p. NIST CSWP 29). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.CSWP.29>

National Institute of Standards and Technology. (2024b, April). NIST Begins Cybersecurity Project for Water and Wastewater Operations. <https://www.nist.gov/news-events/news/2024/04/nist-begins-cybersecurity-project-water-and-wastewater-operations>

Oxenford, J., Hughes, D. M., Karasik, S., & Bukhari, Z. (2019). *Building Workforce Skills for Intelligent Water Operations* (Report No. 4663). Water Research Foundation.

Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., & Myer-Tyson, E. "Bennie ". (2017). *MatLab vs. Python vs. R. Journal of Data Science*, 15(3), 355–371. [https://doi.org/10.6339/JDS.201707_15\(3\).0001](https://doi.org/10.6339/JDS.201707_15(3).0001)

Park, C. (2025). Addressing Challenges for the Effective Adoption of Artificial Intelligence in the Energy Sector. *Sustainability*, 17(13), 5764. <https://doi.org/10.3390/su17135764>

Plathottam, S. J., Rzonca, A., Lakhnori, R., & Illoeje, C. O. (2023). A review of artificial intelligence applications in manufacturing operations. *Journal of Advanced Manufacturing and Processing*, 5(3), e10159.

Prevos, P. (2024). *Principles of Strategic Data Science: Creating Value from Data, Big and Small*. Packt Publishing.

Ravindran, T., Koyama, Y., & Thompson, K. (2025). Analyzing California's PFAS Data: AI Insights in Source Tracking. *Clean Water*, 2025(2), 22–25.

Rieger, L., Papukchiev, U., Miletic, I., Gallien, D., Mason, T., & Johnson, B. (2025, May). *Designing and Implementing a Control Hierarchy for Full-Scale Hybrid Digital Twin Control: From Piloting to Safe and Highly Efficient Digital Twin Operation*.

Rockwell Automation. (2022). *Water Agency Tackles Sustainability Goals with AI: Eastern Municipal Water District Case Study*. <https://www.rockwellautomation.com/en-us/company/news/case-studies/water-agency-sustainability-goal-ai-ml.html>

Rockwell Automation. (2024). *FactoryTalk Analytics LogixAI*. <https://www.rockwellautomation.com/en-us/products/software/factorytalk/operationsuite/analytics-logixai.html>

Salveson, A., Kumar, P., Gerrity, D., Bott, C., Salazar-Benites, G., Bourgeois, K., Plumlee, M. H., Debroux, J., & Stanford, B. D. (2025). *Demonstration of Innovation to Improve Pathogen Removal, Validation, and/or Monitoring in Carbon Based Advanced Treatment (CBAT) for Potable Reuse* (No. 5129). The Water Research Foundation. <https://www.waterrf.org/research/projects/demonstration-innovation-improve-pathogen-removal-validation-andor-monitoring>

Salveson, A., Nading, T., Schimmoller, L., Assi, T., Desormeaux, E., Branch, A., Thompson, K., Weeks, B., & Mansell, S. (2023). *Integrating Real-Time Collection System Monitoring Approaches into Enhanced Source Control Programs for Potable Reuse* (No. 5048). The Water Research Foundation. <https://www.waterrf.org/research/projects/integrating-real-time-collection-system-monitoring-approaches-enhanced-source>

Schneider Electric. (2024a). *EcoStruxure Automation Expert*. <https://www.se.com/us/en/product-range/23643079-ecostruxure-automation-expert/>

Schneider Electric. (2024b). *SCADAPack 470i | 474i Remote Smart RTU and Controller with Integrated Edge Platform*. <https://www.se.com/us/en/product-range/356497944-ecostruxure-rtu-operations-expert/>

Sela, L., Sowby, R. B., Salomons, E., & Housh, M. (2025). Making waves: The potential of generative AI in water utility operations. *Water Research*, 272, 122935. <https://doi.org/10.1016/j.watres.2024.122935>

- Sinha, S., Vekaria, D., Vishwakarma, A., & Dadiala, R. (2025). *Artificial Intelligence (AI) Applications in the Water Sector* (No. 4797B). The Water Research Foundation. <https://www.waterrf.org/embedded-viewer/DRPT-4797B.pdf>
- Sircar, A., Yadav, K., Rayavarapu, K., Bist, N., & Oza, H. (2021). Application of machine learning and artificial intelligence in oil and gas industry. *Petroleum Research*, 6(4), 379–391. <https://doi.org/10.1016/j.ptlrs.2021.05.009>
- Stouffer, K., Pease, M., Tang, C., Zimmerman, T., Pillitteri, V., Lightman, S., Hahn, A., Saravia, S., Sherule, A., & Thompson, M. (2023). *Guide to Operational Technology (OT) Security* (NIST Special Publication Nos. 800-82 Revision 3). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-82r3>
- Suresh, S., Thompson, K., Branch, A., & Salveson, A. (2024). *Machine Learning to Predict Pathogen Removal in Potable Reuse MBRs Based on Microbial Surrogates*. WEFTEC.
- Thompson, K. A., Branch, A., Nading, T., Dziura, T., Salazar-Benites, G., Wilson, C., Bott, C., Salveson, A., & Dickenson, E. R. (2022). Detecting industrial discharges at an advanced water reuse facility using online instrumentation and supervised machine learning binary classification. *Frontiers in Water*, 4, 1014556.
- Thompson, K., & Dickenson, E. R. V. (2021). Using machine learning classification to detect simulated increases of de facto reuse and urban stormwater surges in surface water. *Water Research*, 204, 117556. <https://doi.org/10.1016/j.watres.2021.117556>
- Thompson, K., Newhart, K., Koyama, Y., Branch, A., & Salveson, A. (2024). *Machine Learning Soft Sensors for Energy Efficient Potable Reuse*. WEFTEC.
- Thompson, K., Steigerwald, J., Ray, H., Branch, A., Assi, T., Salveson, A., Nading, T., & Dickenson, E. (2025). *Review of Industrial Chemicals Impacting Potable Reuse: Hazards, Removal, and Recommendations for Enhanced Source Control Programs*. *Total Environment Engineering*, 4, 100027. <https://doi.org/10.1016/j.teengi.2025.100027>
- U.S. Environmental Protection Agency. (2018). *America's Water Infrastructure Act of 2018 (AWIA)*. <https://www.epa.gov/ground-water-and-drinking-water/americas-water-infrastructure-act-2018-awia>
- U.S. General Services Administration. (2024). *Federal Risk and Authorization Management Program (FedRAMP)*. <https://www.fedramp.gov/>
- Vekaria, D., & Sinha, S. (2024). aiWATERS: An artificial intelligence framework for the water sector. *AI in Civil Engineering*, 3(1), 6. <https://doi.org/10.1007/s43503-024-00025-7>
- Vora, L. K., Gholap, A. D., Jetha, K., Thakur, R. R. S., Solanki, H. K., & Chavda, V. P. (2023). Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design. *Pharmaceutics*, 15(7), Article 7. <https://doi.org/10.3390/pharmaceutics15071916>
- Water Research Foundation. (2025). *GenAI as a Catalyst for Water Sector Transformation: Summary of WRF Project 5321 Findings*. Water Research Foundation.
- World Economic Forum. (2025). *Future of Jobs Report: AI and Machine Learning Specialist Demand*.
- Zhang, J., Tylor, N., Zhao, W., Rui, R., & He, C. (2025). Enhancing Coagulation Prediction in Water Treatment Using a Similarity Score–Based Piecewise Machine Learning Model. *Journal of Environmental Engineering*, 151(6), 04025023.