The LIFT Intelligent Water Systems Challenge
Solution Submission
August 13, 2021

Utilizing Soft Sensor System for Process Control and Optimization

EXECUTIVE SUMMARY

Clean Water Services (CWS) operates four water resource recovery facilities that discharge to the Tualatin River watershed. During the wet weather season, the treatment facilities have toxicity-based effluent ammonia limits based on river flow and biomass inventory to meet the limits must be constantly managed to prepare for peak influent flow conditions. To facilitate effective management of nitrification at the facilities to achieve reliable permit compliance with optimized operation, a multidiscipline team with expertise in information technology (IT), operational technology (OT), engineering technology (ET), and data sciences (DS) was formed between Clean Water Services and Princeton University. The team identified and analyzed the critical needs on real-time water quantityprediction, developed and executed an intelligent water system solution that greatly enhanced our ability in operation control and optimization with tremendous benefits to process stability.

This work demonstrates that real-time monitoring and prediction can greatly improve utility operation efficiency, meet regulatory requirements with agility and flexibility, and prevent the construction of excess treatment capacity to maintain a nitrifying biomass year-round through brief wet-weather events where complete nitrification is not necessary. A complete artificial intelligence (AI) framework of soft sensor development based on a series of data mining and modeling steps and decision-making options was developed and applied in this project. Based on the framework, we compared and developed new machine learning decision support tools based on soft senor principles, which provided accurate predictions on wastewater facility influent flow one day in advance. Such information greatly helps CWS reduce the risk of secondary clarifier overload via timely adjustment of the flow directed to different trains. It also guides the deployment of wet weather operating modes before rain events, which leads to efficient operations and risk mitigation. The AI framework and IWS solution developed here can be readily applied to other utilities since many utilities face similar challenges to optimize the treatment process, especially during wet weather events to minimize conveyance overflow and overload to treatment processes.

THE TEAM

Ting Lu, Ph.D., P.E., Business Practice Leader – Digital Solutions, Clean Water Services
- **Area of expertise:** Digital twin design and software implementation
- **Role:** Team lead, project coordination, intelligent water system conceptual design

LuT@cleanwaterservices.org

Jeff Van Note, Digital Solutions Opportunities Manager, Clean Water Services
- **Area of expertise:** Programing, data integration, business intelligence
- **Role:** Data integration and dashboard development.

VanNoteJ@CleanWaterServices.org

Adrienne Menniti, Ph.D., P.E., Principal Engineer – Process, Technology Development and Research, Clean Water Services
- **Area of expertise:** Modeling, IoT sensors, process control
- **Role:** Model evaluation and validation, nitrification process improvement

MennitiA@CleanWaterServices.org

Peter Schauer, P.E., Principal Engineer – Process, Technology Development and Research, Clean Water Services
- **Area of expertise:** Nutrient removal and recovery, engineering design, process control, data analysis
- **Role:** Integration of model prediction for state point analysis, process control, and coordination with operations staff

SchauerP@CleanWaterServices.org

Junjie Zhu, Ph.D., Associate Research Scholar, Department of Civil and Environmental Engineering and Andlinger Center for Energy and the Environment, Princeton University
- **Area of expertise:** Statistical analysis/modeling, data mining, text mining, machine learning/deep learning
- **Role:** Data scientist, data driven model development and validation

junjiez@princeton.edu

Z. Jason Ren, Ph.D., P.E., Professor, Department of Civil and Environmental Engineering and Andlinger Center for Energy and the Environment, Princeton University
- **Area of expertise:** Energy and resource recovery, wastewater treatment and reuse
- **Role:** Data analysis and reporting

zjren@princeton.edu

This is a multidiscipline team with expertise in information technology (IT), operational technology (OT), and engineering technology (ET), and a data scientist's expertise to identify, analyze and communicate the information needed to develop and execute an intelligent water system solution. This team also represents the state-of-the-art knowledge from academic to practical application and implementation of an intelligent water system through rapid prototyping in the water sector.

THE PROBLEM STATEMENT

Clean Water Services (CWS) is a special service district that provides wastewater treatment, stormwater management, stream restoration and water resources management services to more than 620,000 residents and business in urban Washington County, Oregon.

CWS operates four resource recovery facilities that discharge to the Tualatin River watershed. During the wet weather season, the treatment facilities have toxicity-based effluent ammonia limits. The allowable ammonia limit, both daily max and monthly average, is tied to the Tualatin river flow and the toxicity limit gets more stringent as the river flow goes down (**Table 1**).

Table 1. Allowable ammonia limits

Tualatin River Flow at Farmington, cfs	RC Ammonia Limits, mg/L N (Nov-Apr)	
	Daily Max	Monthly Average
<500	11.5	4.8
500-1,000	23.2	11
>1,000	38.6	16.2

This effort is focused on the Rock Creek Water Resource Recovery Facility because it has the most stringent ammonia toxicity requirements. Rock Creek is an advanced treatment facility with multiple biological nutrient removal activated sludge configurations for phosphorus and nitrogen removal. Tertiary treatment is by chemical coagulation-flocculation-sedimentation and granular media filtration. The residuals treatment train includes primary sludge fermentation and thickening to support biological phosphorus removal, WASSTRIP™ - Ostara process for nutrient recovery, and anaerobic digestion with centrifuge dewatering. The facility treats an average annual flow of 36.2 MGD, however the coastal climate brings wet weather events with peaking factors of 3x or more, even with a separate sanitary sewer system. Nitrification is managed with two approaches at Rock Creek.

- The facility keeps one set of aeration trains in close to full nitrification throughout the winter to be prepared for dry weather and the corresponding more stringent ammonia limits. The flow directed to the fully nitrified trains is limited during wet weather events to avoid overloading the secondary clarifiers.
- The operating SRT of all the operating aeration trains is increased during dry periods of the wet weather season to increase the level of nitrification overall.

Having adequate biomass inventory to meet ammonia limits while minimizing the risk that the inventory could overload the clarifiers during a peak flow event must be constantly managed during the wet weather season. If the Rock Creek facility is managed to stay in complete nitrification the extra inventory needed creates a risk for secondary clarifier overload during wet weather events. Conversely, if a low biomass inventory is maintained and the river flow drop, signaling a more stringent ammonia limit, the facility could be at risk of not meeting the limit. On average, uncertainty in monthly average river flow and conservatism in process control has led to maintaining a higher biomass inventory than necessary to meet effluent limits. The additional nitrification also imposes operational costs including additional energy and chemical costs for alkalinity addition with lime.

In addition, the choice of when and how to use these wet weather inventory management tools has been dictated by flow based capacity of the secondary process without consideration of the biomass inventory or the settling characteristics. A common method to improve knowledge of clarifier capacity is by use of state-point analysis. Prior to this project operation staff had to manually enter all these data for analysis and process control into a separate spreadsheet. This is further complicated by the need to train new

operators that are unfamiliar with this analysis. During a wet weather event, they must be able to rely on streamlined data analysis. One major benefit of the operational dashboard developed through this challenge is to make state point analysis readily accessible to operations staff.

Having a tool that predicts the calendar monthly average ammonia limit as far in advance as possible would help inform choices on the level of nitrification to target. Additionally, predicting the influent flow accurately both one day in advance and three days in advance – especially during rain events – would help CWS reduce the risk of overloading secondary clarifiers by adjusting the flow directed to different trains and/or employing wet weather operating modes in advance of the rain events.

DESIRED OUTCOME AND BENEFITS

To facilitate cost-effective management of nitrification at the plant to achieve reliable permit compliance with optimized operation, three goals will be pursued in this proposed intelligent water system:

- Predict influent flows for the Rock Creek facility one day (and three days) in advance.
- Predict the next calendar month average Tualatin River flow.
- Develop a simple, easy-to-use dashboard to visualize the predictions of future influent flow and monthly average river flow and process control tools

CHARACTERIZATION OF THE EXISTING SYSTEM AND PROCESS CONTROL APPROACH

The daily maximum concentration limit is straightforward and depends on the daily average river flow. However, the monthly average ammonia concentration limit depends on the calendar month average river flow. The receiving stream flow can change dramatically in a single wet weather event and change the permit limit tier, and such flow variability has increased gradually due to climate change. Without confidence in a calendar month average river flow, Rock Creek operations staff does not know the plant effluent toxicity limit until the end of the month. Therefore, they must operate for the entire month without clear knowledge of the permit required effluent ammonia limit. Currently the process control personnel watch the trends in the river flow and the weather forecasts to anticipate the limit and adjust operation accordingly (**Figure 1(a)**). Spreadsheet tools are utilized to track the month to date average river flow and ammonia concentration and allow the analyst to manually run flow scenarios to predict the calendar month average river flow and calculate the ammonia concentration required to meet the different permit limit tiers. The final effluent ammonia is adjusted by changing the number of basins in full nitrification (**Figure 1(b)**). During dry periods when river flow is low, full nitrification is maintained throughout the secondary system. As the river flow increases, fewer basins are maintained in full nitrification.

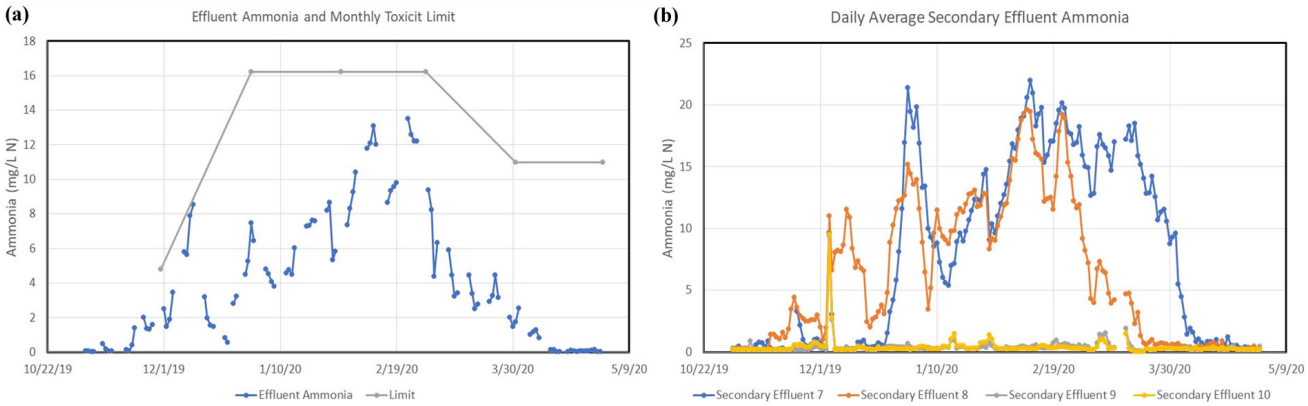


Figure 1. (a) Plant effluent ammonia and monthly average toxicity limit. (b) Daily average secondary effluent ammonia concentration as measured from an online

Managing Inventory During Rain Events

While the Rock Creek sewer system is not a combined system, it is impacted by infiltration and inflow. Therefore, rain events have a strong influence on both the influent flow and the river flow (**Figure 2**). Large rain events can increase the plant influent flow dramatically over just a single day, as can be seen by the rain event in early January 2020. When a rain event is expected, operations staff manage the inventory in the secondary system using two approaches:

1. **Primary effluent flow distribution:** The distribution of primary effluent can be actively adjusted to deliver different amounts between the east and west side secondary systems and between each online aeration basin on the east side.
2. **Wet weather step feed:** The east side aeration basins have wet weather step feed capabilities that direct primary effluent toward to the end of basin and protect the inventory from washout during a rain event. The amount of primary effluent step fed is also an operational variable.

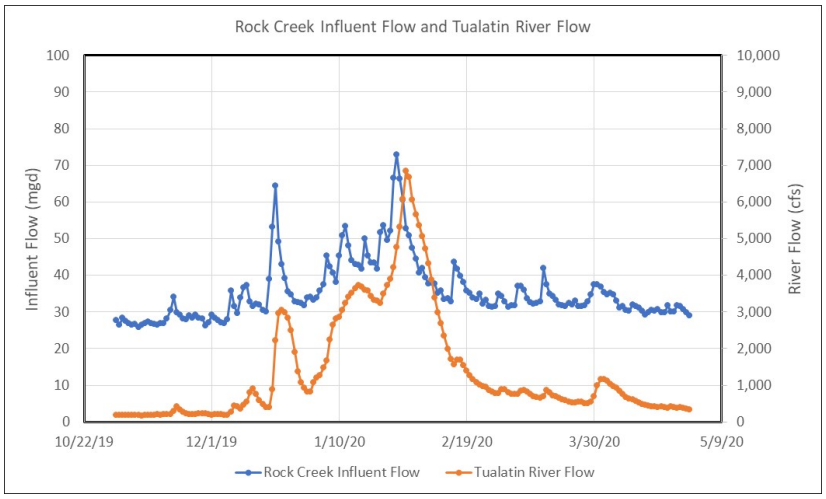


Figure 2. Rock Creek influent flow and Tualatin River flow through the 2019-2020 wet weather season.

These operational changes are highly prescriptive and laid out in a standard operating procedure for high flow events. Operations staff take action at influent flows of 60, 80, and 115 MGD to move step feed ratios toward contact stabilization. Return activated sludge rates are also increased to manage clarifier sludge blankets.

Historically, the choice of when and how to use these wet weather inventory management tools has been accomplished based on operator experience with historical clarifier performance. State point analysis is a highly useful tool for operational decision making. It is a visual model of secondary clarifier performance that accounts for all critical parameters driving clarifier operation: influent flow, mixed liquor concentration, return activated sludge (RAS) flow rate and sludge settleability. While state point analysis is available to operations staff, it is an offline spreadsheet tool that rarely has been used to guide decision-making.

The Intelligent Water System Plan (submitted in May, 2021)

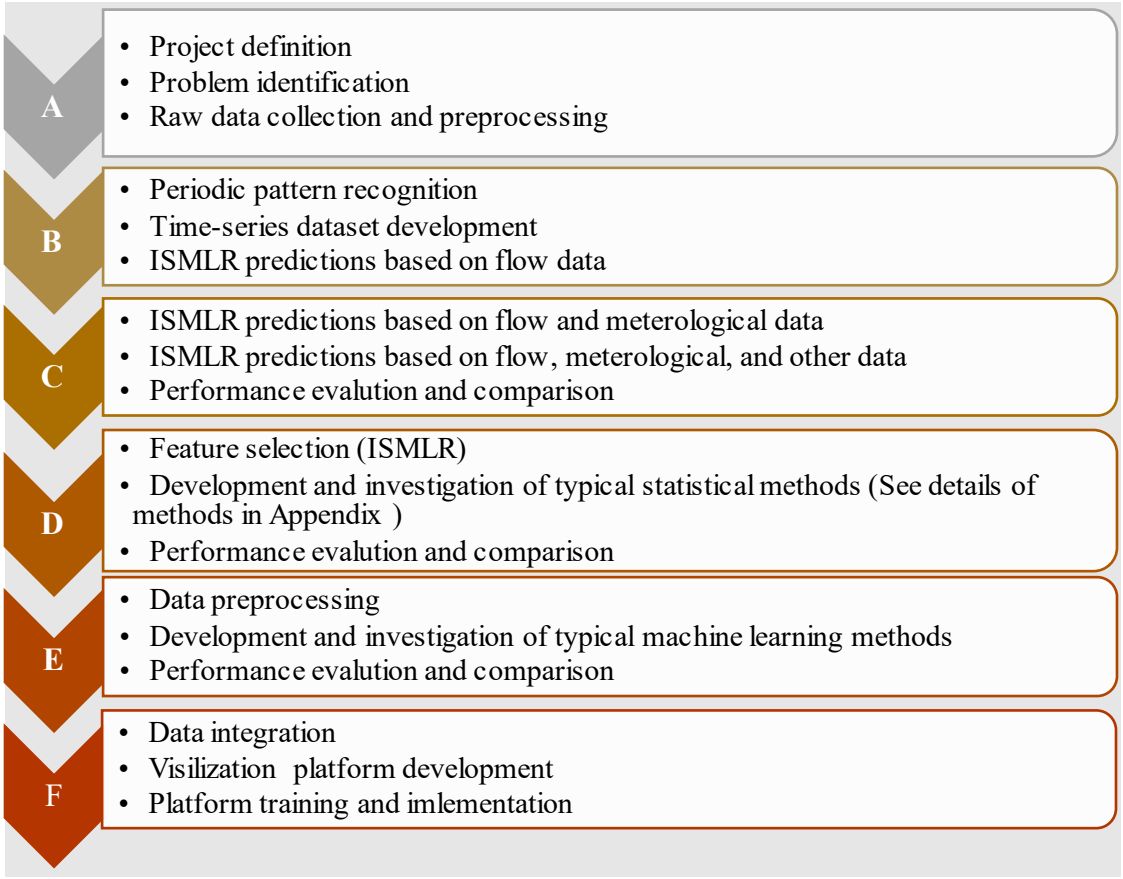
The work was planned to conduct in the following phases:

- **Soft sensor development:** Develop comprehensive offline soft sensors to predict deterministic values of future influent flow and river flow.
- **Visualization platform:** Develop a visualization platform to incorporate multiple data sources, model predictions and state point analysis to support process control and operational optimization

Specifically, the work was planned to conduct as the following workflow (see Scheme 1):

A. Define project, collect and preprocess data (April 11 – April 30, 2021).
B. Preliminary predictions (May 1 – May 15, 2021).
C. Data sources evaluation (May 16 – May 31, 2021).
D. Statistical methods investigations (June 1 – June 30, 2021).
E. Machine learning development and comparison (July 1 – July 31, 2021).
F. Visualization platform development, training and pilot implementation (July 1 – August 10, 2021)

Scheme 1. Work flowchart for Rock Creek soft-sensor development and implementation during this Challenge



SOLUTION (THE IWS SYSTEM)

1. Methodology

1.1. Framework of intelligent water system and the soft sensor development

The Clean Water Services Knowledge Development Model, shown below (**Figure 3**), was utilized to develop the IWS framework. This framework is consistent with the exiting industry framework, with additions of the collaboration process on the top. CWS staff and partner agencies and organizations are captured in the highest level of the Knowledge Development Model. Internal and external stakeholders from across the watershed work together to understand engineered and natural systems and measure the effectiveness of the organization's current practices and initiatives. Importantly, people involved in all these IWS layers make an impact for collaboration and impact on landscape.

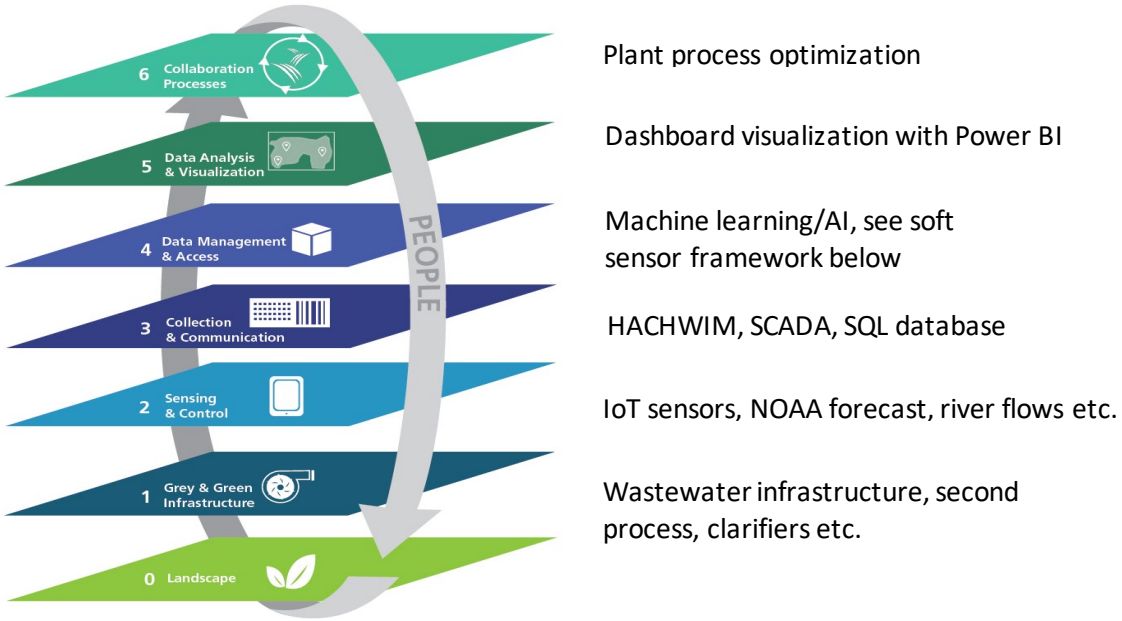


Figure 3. Overview of the Intelligent Water System framework.

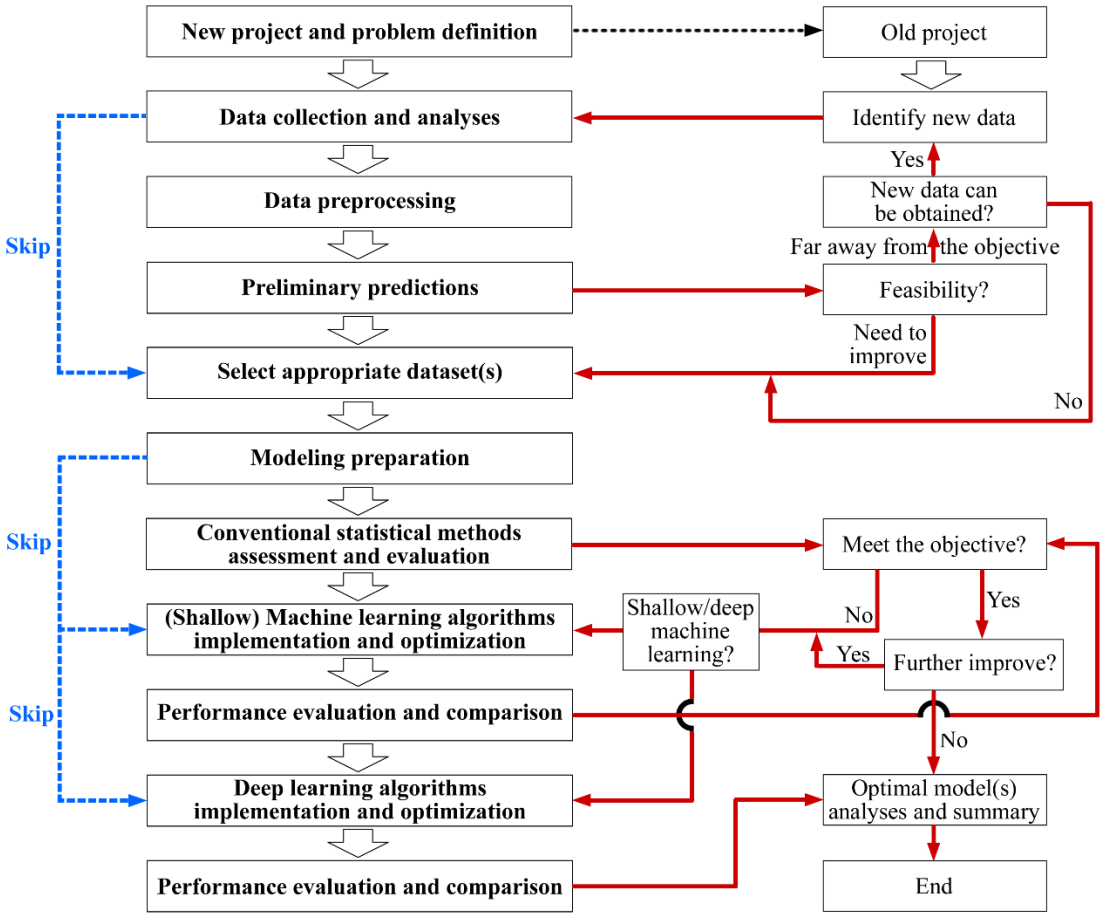


Figure 4. Recommended offline soft-sensor development framework for all water and wastewater utilities. We developed and applied the framework to develop soft sensors for both influent wastewater flow and river flow predictions. For a new project that starts from problem and scope definition, followed

by a series of data collection, analyses, preprocessing, and dataset selection. If a preliminary prediction does not demonstrate the feasibility, then additional data are required to increase information diversity. For any old projects, acquisition of up-to-date data is an important early step before further development. A complete development flow would be from assessment of conventional statistical methods, machine learning algorithms, and to deep learning approaches. However, one can skip any of the steps when the prediction results are satisfied or no need to further optimize depending on the local conditions.

Within this overall IWS framework, data analysis was performed by various machine learning methods. Current machine learning lacks a framework specifically for water and wastewater operation, it's more focused on innovation of individual methods. In this challenge project, we developed a new framework to fill this gap and also open the possibility for it to be used in a variety of applications, including time-series prediction, classification, and anomaly detection. We demonstrate how such a framework was successfully applied to the Rock Creek facility to accurately predict future influent wastewater flow and its receiving river flow. By implementing this intelligent water system, CWS can operate more efficiently through reduced chemical and energy consumption. More importantly, the system reduces the risk of clarifier failure and compliance issues by providing a tool to guide data driven decision making. This new framework on soft sensor development (**Figure 4**) has been able to predict 1) the next day's influent wastewater flow at Rock Creek and 2) the next month's average Tualatin River flow at Farmington. The framework delineates the major steps, including old/new project initialization, data collection and preprocessing, preliminary predictions, conventional statistical methods implementation, and shallow or/and deep machine learning implementation and optimization.

1.2. Data collection, analyses, preprocessing

To develop soft sensors for predicting the next day's influent wastewater flow, 11 years (2010-2020) of historical data, which describe typical influent conditions at the Rock Creek facility, river flowrate at Farmington, and local meteorological conditions from five nearby NOAA sites (Hillsboro, Oregon), were first collected (**Figure 5**). The data vary in temporal resolution (15-min, hourly, daily), missing data, and other miscellaneous issues, so the preliminary data analyses were taken to understand potential problems and to determine an initial list of variables to be investigated.

To develop soft sensors for prediction of the next month's average river flow, after an extensive data collection, we successfully identified possible representative data sources from two USGS river monitoring sites (Farmington and West Linn) and 40 NOAA meteorological sites based on eight ZIP areas for 81 years (10/1939 – 12/2020) (**Figure 5**). We included the 40 NOAA sites to cover rain, snow, and air temperature data in the area of the upper stream of the Tualatin River because no single site can provide complete data for any variable for such a long period. Average values of available data from the 40 sites were used as preprocessed daily data, and were later converted to monthly data for all meteorological variables.

A standard modeling preparation was applied to both cases of soft-sensor development: feature scaling, feature selection, cross-validation, hyperparameter optimization, and performance evaluation. Different statistical methods and (shallow) machine learning algorithms (**Table A1**) were evaluated and compared using the designing flowchart for both cases (**Figure 4**). A series of steps was taken to ensure QA/QC for the whole work. Details are attached in the appendix.

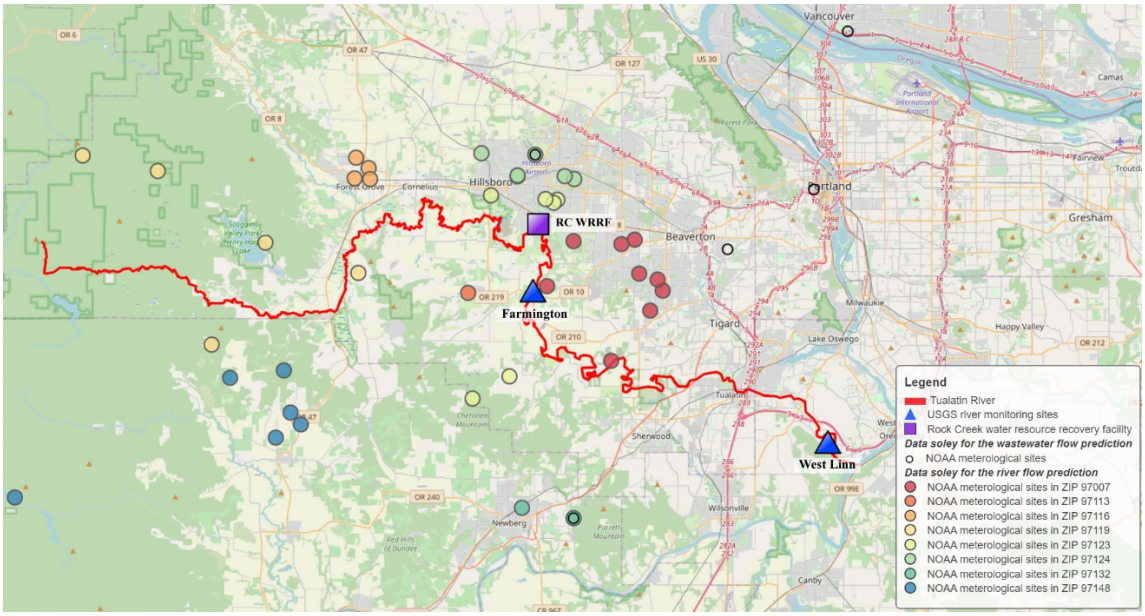


Figure 5. Map of the use cases: Rock Creek WRRF (RC, 2021), Tualatin River, USGS river monitoring sites (OWSC, 2021), and NOAA meteorological sites (NOAA, 2021). Data from Rock Creek WRRF, 1 USGS site (Farmington), and 5 NOAA sites (black void circles) were used to predict the next day's influent wastewater flow; data from two USGS sites and 40 NOAA sites (colored circles) were used to predict the month's river flow.

2. Results

2.1. Prediction of the next day's influent wastewater flow

Overall, machine learning algorithms are superior over conventional statistical methods. ETR, KRR, and SVR all demonstrated excellent prediction capability. **Figure 6** displays the example of using ETR to predict the flow with high accuracy based on the training and testing datasets; the well overlapping between the predicted values and real data demonstrates the functionality and feasibility of the soft sensor developed in this work. Details of the development and results can be found in the appendix.

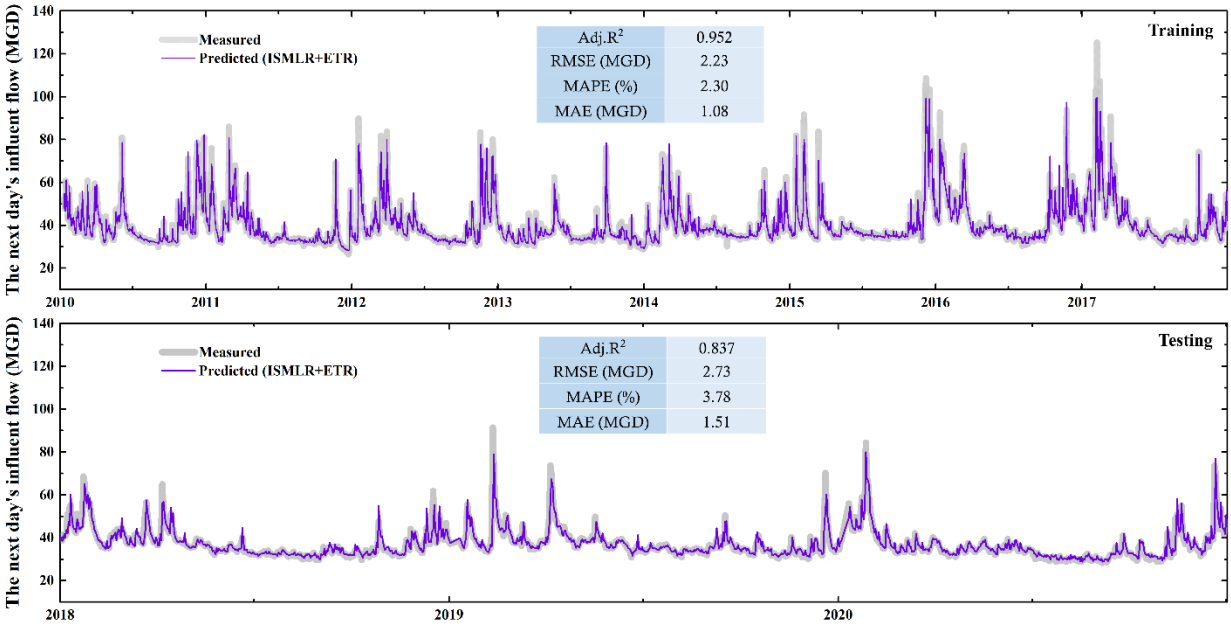


Figure 6. Comparison between measured and predicted (using ISMLR+ETR) flow accurately predicted the next day's influent flow at the Rock Creek facility based on training (2010-2017) and testing (2018-2020) datasets. The tables list the prediction performance.

2.2. Prediction of the next month's average river flow

While prediction of future river flow is more challenging because of bigger gaps (monthly) and less amount of available data (963 < 4007), the development of soft sensors followed the same procedure to pursue the best models. Overall, KNN, ETR, and KRR were the three methods that generated best candidates from the 999 tested models.

Figure 7(a) shows that the evolution of model prediction performance over the month based on the testing dataset. **Figure 7(b)** exhibits a good prediction performance (adj.$R^2 \approx 0.861$; RMSE ≈ 480 cfs; MAPE ≈ 19.1%; MAE ≈ 257 cfs) when 14 days passed in the month and most of the measured data are well predicted. A subsequent evaluation of classification accuracy based on the predicted data was taken over the month. Details of the development and results can be found in the appendix.

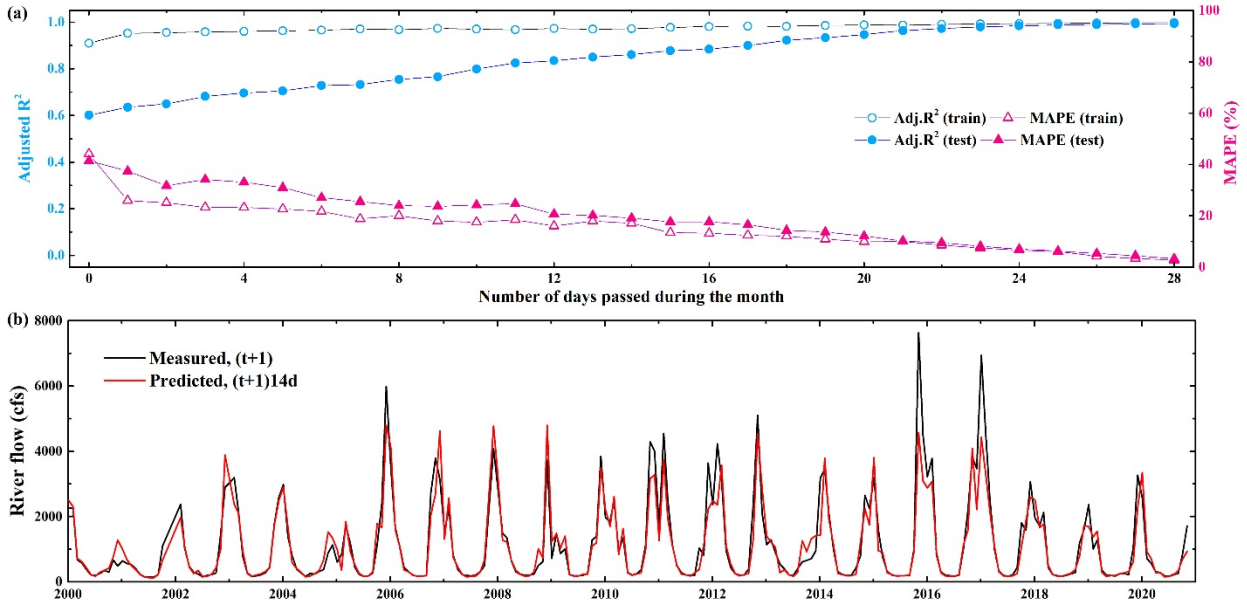


Figure 7. Summary of Tualatin River flow predictions. (a) Change of prediction performance (adjusted R^2 and MAPE) over the number of days passed during the next month. (b) Comparison between measured and predicted (14 days passed) the next month's river flow based on the testing dataset (2000-2020).

SOFT-SENSOR PREDICTION AND IWS SOLUTION IMPLEMENTATION

Once the model accuracy was accepted by the utility team, the dashboard was developed with Power BI as a decision support tool (**Figure 8**; **Figure 9**). In the figure below, both historical flow and model predictions were placed on the dashboard to inform Operations of the upcoming changes.

In addition, an existing state point analysis Excel file was used as the basis for a Power BI dashboard for Operations to perform what-if scenarios, and for more people to collaborate and decide operational changes. Information for the dashboard is retrieved from the District's Hach WIMS database.

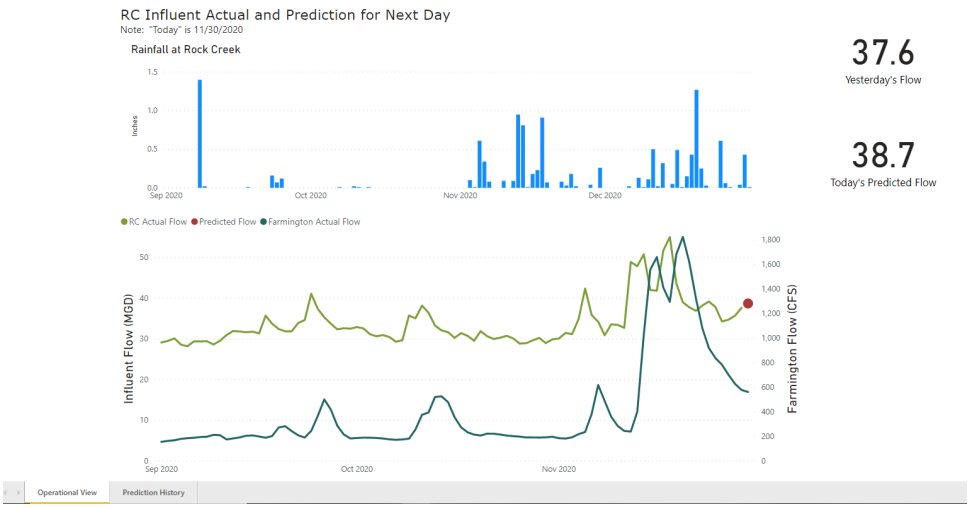


Figure 8. Power BI dashboard showing plant influent flow predictions.

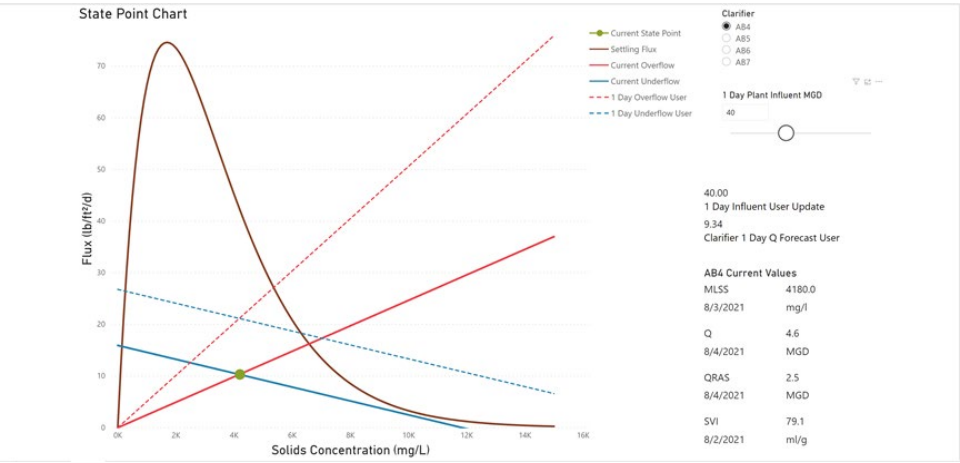


Figure 9. Power BI dashboard to show predictions and what if scenario based on soft senor program predicted flow.

The example presented in **Figure 9** shows the current state point as well as the expected state point with the predicted next day influent flow. The current state point for aeration basin 4 (connected to secondary clarifier 7) shows a critically loaded basin carrying a heavy inventory for full nitrification. With a high predicted flow due to a rain event, the state point clearly shows that the secondary clarifier will be overloaded if the operations staff uses the current setpoints for influent flow distribution, return activated sludge rate and without the use of a wet weather operating mode or increase of RAS rate to prevent thickening failure for the system. Based on the soft sensor data prediction, the operator reviewing this information therefore will know that they must adjust these critical operating parameters to ensure the clarifier does not fail during the rain event. Future iterations of the state point tool will provide suggested operational limits and allow the operator to adjust the set points for influent flow, MLSS and RAS rate to determine the best combination to avoid clarifier failure.

Figure 9 shows the direct benefits of state point adjustment for secondary clarifier operation based on the developed soft sensor data. The state-point analysis currently provides a single day look ahead to identify if influent flows could pose a risk for clarifier failure. The operations analyst must determine the potential corrective measure that would reduce risk for the following day. With the new dashboard and prediction data, the influent flow predictions can be made several days in advance, which would increase the number of corrective actions that can be taken by operations. In addition, the state point analysis tool will be

enhanced to provide additional troubleshooting details if predictions indicate clarifier failure. In a condition where the analysis shows a thickening failure, the tool will suggest the minimum RAS flow rate that would be required to prevent this from occurring. This value will be placed in context with the actual RAS pumping capacity. In the condition where the analysis indicates settling failure, the tool can display the maximum possible flow rate that the clarifier can be operated at or the maximum mixed liquor concentration that could be maintained while treating the entire flow rate.

Communication

During this project period, regular meetings have been held for discussion and training with operation staff on visualization platforms for process control and optimization. Once the model is validated and meets the operational requirement, it will be implemented in full scale for process control during wet weather events.

FUTURE IMPLEMENTATION AFTER IWS CHALLENGE

Future work will make efforts to further enhance model accuracy and feasibility by collecting forecast information of meteorological conditions and river flow (NWRFC, 2021), if any; and developing probabilistic models to provide additional information about confidence levels.

As this model is further refined, longer range predictions can be developed to provide a degree of certainty that the monthly average river flow will result in a given ammonia limit. This would allow for longer range planning and control of basin operation.

LONG TERM SUSTAINABILITY

The solution is built upon ML/AI algorithms developed by Princeton University, and data is integrated through Clean Water Services' Power BI visualization. The final solution will be automated for data transfer and update. To sustain the long-term solution, it's important to have data scientist, process engineer and operation expertise to interpret data and make informed decisions.

BENEFITS TO UTILITIES

The immediate application of this solution provides two benefits:

1. **More efficient basin operation and reduced risk of ammonia compliance issues.** The ability to project the river flow and subsequently the ammonia limit for the month will empower the operators with real-time data so they can operate the secondary system in the most efficient manner. With such information, utilities can adjust the number of basins in nitrification and the degree of nitrification achieved in the nitrifying basins to more confidently meet target effluent ammonia limits.
2. **Reduce risk of system failure**. Having a reliable prediction of the next day predicted influent flow and an online, easy-to-use state point tool gives the operator direct guidance on operational decision-making to avoid clarifier failure. This has never been demonstrated before in wastewater utilities.

By implementing this intelligent water system, the tool optimizes processes, and ensures efficient operations to provide certainty and reduces the risk of clarifier failure and compliance issues by providing a guide for data-driven decision-making.

LESSON LEARNED

Team members: It's important to have a multidisciplinary team to co-create an intelligent water solution that involves IT expertise, OT expertise, ET expertise, and data science. This ensures we have the practical problems to explore and leverage the latest machine learning techniques, and still use the foundational engineering knowledge such as State Point Analysis and IT expertise to build a dashboard to increase business efficiencies and consistency with multiple treatment facilities and operational staff.

Data Management: Better management of data issues such as various temporal resolutions and missing data issues is crucial for QA/QC. It is often required to balance the tradeoff between data quality and data quantity; we managed to resolve most cases by using tailored data analyses or trial and error. We also met a major challenge in the river flow prediction when the 11 years of data in the original plan did not meet the data need. We successfully overcame the challenge by following the framework to expand the data to 81 years and retrieve more data sources.

Learning journey: Machine learning techniques and artificial intelligence have been utilized in different sectors, however, it is still considered a black box solution and not fully utilized by utilities in the water sector. There are many questions around data quality and confidence in the ML model. Not only does this solution provide a decision support tool for operation optimization, the solution creation itself has been an education process for our team members to understand different types of the ML/AI model, their constrains and benefits, data quality requirement and use of Power BI dashboard. This is a good process and we recommend it for any utilities initiating new technology projects. Focus on the educational component and take advantage of the learning.

Relevance and application to other utilities: The IWS solution the team developed can be readily applied to other utilities since many utilities face the same challenges to optimize treatment processes, especially during wet weather events to minimize conveyance overflow and overload to treatment processes. The solution with predicting the influent flow, integrated with State Point Analysis ahead of time, will allow operators and process engineers to proactively adjust the flow and make necessary changes. The proposed framework can be applied to develop intelligent tools for time-series prediction, classification, or anomaly detection in all water and wastewater utilities that have regular monitoring and data acquisition systems. Therefore, we hope that our report will not only serve as a summary material to exhibit the results of our cases, but also can be used to showcase how the framework was used, what and how the data were collected and prepared, as well as how a robust soft sensor model was developed. Other water and wastewater utilities could use the information to develop similar tools by themselves.

REFERENCES

Angiulli, F., Basta, S., Pizzuti, C. (2006). Distance-based detection and prediction of outliers. *IEEE Transactions on Knowledge and Data Engineering*, 18(2), 145–160.

ISMLR. (2019). Iterated Stepwise Multiple Linear Regression (ISMLR) package version 1.0 was published on July 11, 2018. The latest version 1.1 was also published on March 01, 2019. https://junjiezhublog.wordpress.com/ismlr/

NOAA. (2021). National Oceanic and Atmospheric Administration, National Centers for Environmental Information. https://www.ncdc.noaa.gov/cdo-web/

NWRFC. (2021). Northwest River Forecast Center, River Information and Forecasts. TUALATIN AT FARMINGTON (FRMO3). https://www.nwrfc.noaa.gov/river/station/flowplot/flowplot.cgi?FRMO3

OWSC. (2021). Oregon Water Science Center, USGS. Tualatin River Basin Streamflow Page. https://or.water.usgs.gov/tualatin/streamflow/

RC. (2021). Historical Rock Creek influent flow data. The flow meters are recorded in our SCADA historian, so fairly fine increments of time can be downloaded if needed (e.g., 15-minute data or daily averages or something in between).

Zhu, J.-J., Anderson, P.R. (2016). Assessment of a soft sensor approach for determining influent conditions at the MWRDGC Calumet WRP. *Journal of Environmental Engineering*, 142(6), 04016023.

Zhu, J.-J., Kang, L., Anderson, P.R. (2018). Predicting influent biochemical oxygen demand: Balancing energy demand and risk management. *Water Research*, 128, 304-313.

Zhu, J.-J., Anderson, P.R. (2019). Performance evaluation of the ISMLR package for predicting the next day's influent wastewater flowrate at Kirie WRP. *Water Science and Technology*, 80 (4), 695-706.

APPENDIX

Additional IWS Plan details (submitted in May, 2021)

In Stage A (**Scheme 1**), eleven years of historical raw data from 2010 to 2020 will be ingested and processed. To develop a robust soft sensor, data preprocessing is an important step to exclude irrelevant and extreme values. Technical and statistical outliers and missing data will be detected and managed; conventional standard deviation and box plot based detection can overestimate the number of outliers, so a distance-based detection heuristic will be used to identify outliers.

In addition, 15-minute influent flow will be converted to both hourly and daily average flow data, which will be used as initial input variables. When there is a short gap (= 1) between adjacent values with respect to the time (15-minute, hourly or daily), linear interpolation will be applied to fill the gap; otherwise, the missing data will be remained before further treatment. A similar approach will also be applied to other variables.

The first step of Stage B will be to identify historical patterns of relevant variables based on periodogram analysis, so initial time-dependent regressors will be developed for different models in the later steps and stages. The pattern recognition and dataset development will be applied to all variables and data combinations. Iterated stepwise multiple linear regression (ISMLR, 2019) will be used as a base model to perform primary predictions using its MATLAB GUI software (Zhu and Anderson, 2019).

In Stage C, we will assess the effect of different data sources or variables on the prediction performance to better understand important variables and regressors. Three major combinations of data sources are: Base dataset (flow data), expanded dataset (flow + meteorological data), and full dataset (flow + meteorological + other useful data).

Stages D and E will mainly focus on developing and exploring various models, which are summarized in **Table A1**. For statistical methods, the dataset will be divided into training and testing; a validation part will be used in the machine learning and deep learning model development, so cross-validation will be applied to build more robust models. Grid search will be used as the hyperparameter optimization methods in machine learning model development. The overall objective of testing various methods is to evaluate and compare their respective prediction performances, helping to select appropriate and effective methods.

Stage F will focus on developing dashboards with time series as well as a representation of the state point analysis charts for process control. It will utilize SQL Server for data storage, Python for data integration, and Power BI for data visualization.

Integration processes will be created for these items:

- Import of the model's predictions to a database suitable for visualization.
- Import of rain, flow and river stage forecasts to a database suitable for visualization.
- A process to supply the model with recent SCADA measurements.

Data file

A supplementary data file that summarizes details of data sources, main prediction results and performances are attached as an Excel spreadsheet document.

Table A1. List of statistical and machine learning to be investigated in this work.

Stage	Type	Algorithm

D	Statistical methods	Multiple linear regression (MLR)
		Principal component regression (PCR)
		Partial least square regression (PLSR)
		Ridge regression
		Lasso regression
		Lasso least angle regression (LassoLars)
		Bayesian ridge regression (BRR)
E	Machine learning	Kernel ridge regression (KRR)
		Support vector regression (SVR)
		k neighbors regression (KNN)
		Gaussian process regression (GPR)
		Decision tree regression (DTR)
		Random forest regression (RFR)
		Extra trees regression (ETR)
		AdaBoost regression (ABR)
		Gradient boosting regression (GBR)
		HistGradient boosting regression (HGBR)

Additional abbreviation and acronym

Abbreviation	Full description
AI	Artificial intelligence
cfs	Cubic foot per second
cv	Cross validation
CWS	Clean Water Services
ET	Engineering technology
GUI	Graphical user interface
IoT	Internet of things
ISMLR	Iterated stepwise multiple linear regression
IT	Information technology
MAE	Mean absolute error
MAPE	Mean absolute percentage error
MGD	Million gallons per day
ML	Machine learning
MSE	Mean squared error
NOAA	National Oceanic and Atmospheric Administration

OT	Operational technology
PCA	Principal component analysis
PCs	Principal components
RAS	Return activated sludge
RMSE	Root mean squared error
SCADA	Supervisory control and data acquisition
SQL	Structured query language
SRT	Sludge retention time
USGS	United States Geological Survey
WIMS	Weather information management system

Prediction and classification performance metrics

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$Adjusted\ R^2 = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Details in data collection, analyses, and preprocessing

To develop soft sensors for prediction of the next day's influent wastewater flow, 11 years (2010-2020) of historical data, which describe typical influent conditions at the WRRF, river flowrate at Farmington, and local meteorological conditions from five nearby NOAA sites (Hillsboro, OR), were first collected (**Figure 5**). The data vary in temporal resolution (15-min, hourly, daily), missing data, and other miscellaneous issues, so the preliminary data analyses were taken to understand potential problems and to determine an initial list of variables to be investigated. A series of steps was taken to ensure QA/QC for the whole work. For example, wastewater flow data were recorded in an excellent condition (> 99.9% of days), while data of most typical wastewater compositions (e.g., CBOD, NH_3N, TSS) have a relatively large fraction of missing data and therefore they were not included in the subsequent dataset development. Similarly, the relatively complete daily records (> 99.2%) of rain, snow, and air temperature were successfully identified and retrieved from the five NOAA sites, site US1ORWS0007, and site USW00094261, respectively. Technical outliers were identified based on domain knowledge (e.g., flow > 0), and statistical outliers were detected based on a distance-based heuristic method (Angiulli et al., 2006; Zhu et al., 2015) to minimize the loss of data. A linear interpolation was used to fill the gaps only when the gaps were single days, so both data quality and data quantity were guaranteed. Wastewater flow data

were pretreated to three temporal resolutions, 15-min (raw), hourly, and daily, to capture more detailed flow dynamics. In addition, daily river flow data were also prepared to examine if they can provide useful, supplementary runoff information to the prediction model. The next step was to identify moving-window (the lookback period) of all the variables to develop time-dependent regressors based on the periodogram analysis (Zhu and Anderson, 2016; 2019). Different from typical wastewater variables that typically have seven-day periodic pattern, river flow and meteorological variables do not exhibit any patterns, so we set the window back to seven days ago for all variables (e.g., rain: rain(*t*), rain(*t-1*),…, rain(*t-7*), where *t* means the current day) except wastewater flow. 83 wastewater flow regressors were developed based on their respective temporal resolutions, including 11 daily regressors (*t*, *t-1*, …, *t-10*), 24 hourly regressors (*t-0h*, *t-1h*, …, *t-23h*, where *t* means the end of the current day), and 48 15-min regressors (*t-15min*, *t-30min*, …, *t-720min*). After data preprocessing, we adopted a step of preliminary predictions using a simple statistical method to determine appropriate dataset(s) before feature selection. Six datasets (build-up based on the following sequence: Wastewater flow, rain, river flow, snow, and/or air temperature) were assessed using iterated stepwise multiple linear regression (ISMLR), a simple and effective statistical method (Zhu and Anderson, 2016; 2019; Zhu et al., 2018). The assessment showed that air temperature was less important, so the final preprocessed dataset (n = 4007) was constituted of 147 initial regressors based on wastewater flow, rain, river flow, and snow data.

To develop soft sensors for prediction of the next month's average river flow, similar data except wastewater flow were first preprocessed (e.g., converting from days to months) to evaluate the model feasibility (**Figure 4**). However, a poor prediction performance (adj.R^2 < 0.24; MAPE > 117%) based on the data asked for a substantial need of new data to develop more accurate models. After an extensive data collection, we successfully identified possible, representative data sources from two USGS river monitoring sites (Farmington and West Linn) and 40 NOAA meteorological sites based on eight ZIP areas for a period of 81 years (10/1939 – 12/2020) (**Figure 5**). Missing data were frequently found in river data at Farmington; however, river data at West Linn were complete (≈ 100%) and data between West Linn and Farmington were highly correlated ($r \approx 0.98$). Therefore, most of the data gaps at Farmington were filled based on their correlation. Because none of a single site can provide complete data for any variable for such a long time, so 40 NOAA sites were used to cover rain, snow, and air temperature (TMAX and TMIN) data over the course of 81 years in the area of upper stream of the Tualatin River. Average values of available data form the 40 sites were used as preprocessed daily data, and were later converted to monthly data for all meteorological variables. In addition to the representative values (average values for river flow or temperature; sum for rain or snow), two statistical variables (daily maximum within the month and count of rain or snow events) were also prepared to enrich the input information. The periodogram analysis helped to determine a 12-month periodic pattern for river, rain, and temperature, so time-dependent regressors were developed from current month (*t*) to 12 month ago (*t-12*) for all variables. Following the recommended framework, preliminary predictions were conducted based on 16 different datasets, building up from river, rain, snow, temperature, and their statistical data. The evaluation showed that the best dataset (n = 963) was composed of 108 initial regressors based on river, rain, snow, air temperature, and statistical data of rain and snow.

Details in preparation for modeling and evaluation

All variables (regressors) were scaled to the range of [0, 1] based on their corresponding maximum and minimum values. Two feature selection methods, principal component analysis (PCA) and stepwise regression, were examined; the number of principal components (PCs) were determined based on both explained variance and mean squared error (MSE); the stepwise regression was conducted using ISMLR based on default *p*-values and regression types (Zhu and Anderson, 2019). Cross-validation with grid search was used to search the best models based on statistical methods and machine learning algorithms. In the case of future wastewater flow prediction, the 11 years of daily data were first split into training (2010-2017) and testing (2018-2020) datasets, and the training part was further used for cross-validation (cv = 3) for the statistical methods and machine learning algorithms. Machine learning algorithms have

more complex structure, so an additional finer model selection was taken subsequently with cv = 10 for three primarily selected machine learning methods. Similarly, in the case of future river flow prediction, the 81 years of monthly data were divided into training (1939-1999) and testing (2000-2020) datasets, and cross-validation with cv = 3 was applied in the training dataset for the statistical methods and a primary evaluation for machine learning algorithms. A finer model selection was taken subsequently with cv = 5 for three primarily selected machine learning methods. Prediction performance of all models was evaluated based on four typical metrics, adjusted R^2, mean absolute percentage error (MAPE), root mean squared error (RMSE), and mean absolute error (MAE). Because the goal of the future river flow prediction is to provide consulting information about the three flow classes (<500, 500-1000, and >1000), so classification was also performed using the predicted values based on classification accuracy.

Statistical methods and machine learning algorithms

Typical statistical methods do not have hyperparameters, but methods such as Ridge, Lasso, and Lasso Lars regressions have a regularization strength parameter, alpha. Therefore, grid search was also employed to search the best alpha values (0 – 1) for these methods. The ten machine learning algorithms can have different hyperparameters, resulting a large amount of models to be tested. For example, 10 numbers of neighbors (5, 10, 20, …, 100, 200), 2 weights (uniform and distance), and 3 metrics (Minkowski, Euclidean, and Manhattan) were tested in the KNN modeling, a total 60 models were tested. The GBR modeling included a total of 288 models based on 8 numbers of estimators (10, 20, 60, …, 300, 400), 6 max depth values (None, 5, 10, 20, 40, 60), and 6 learning rates (0.001, 0.005, 0.01, 0.05, 0.1, 0.5). Overall, 999 models were evaluated at the first stage of machine learning implementation and optimization (cv = 3). As describe previously that three best methods were selected from the first stage and a finer grid search was employed to search the final optimal models at the second stage of optimization, which helped to search more detailed high dimensional hyperparameter space. The case of river flow prediction involved an additional step of classification to evaluate the accuracy of predicted river flow classes. We then simulated model prediction and classification over the month being predicted (from day 0 to 28), to better understand how is the model performance evolving over the time.

Details of soft sensor development for the influent wastewater flow prediction

PCA of the initial dataset suggests that 22 is the minimum number of PCs that could achieve relatively low MSE (14.4 MGD^2) and high explained variance (> 96%). Among the 18 PCA-based statistical models, Ridge regression (alpha = 1) performed the best and was selected to re-evaluated based on the testing dataset, and it achieves an adj.R^2 of 0.777 and a MAPE of 4.06%. Different from PCA that tries to capture the most information from the data, ISMLR works in a way to retain important regressors and exclude irrelevant regressors from the model. In our case, ISMLR identified 15 important regressors and the top regressors were related to current day's rain and the most recent wastewater flow; snowfall and river flow also provided essential contributions. Ridge regression (alpha = 0.4) still works the best among the 18 ISMLR-based candidates, and exhibits even better performance based on testing dataset (adj.$R^2 \approx$ 0.824; RMSE ≈ 2.84 MGD; MAPE ≈ 4.03%; MAE ≈ 1.60 MGD). The prediction results based on statistical methods are promising, but it is worthwhile to pursue higher performance using machine learning algorithms. Therefore, the processed datasets based on PCA and ISMLR were subsequently used to evaluate the feasibility of the ten machine learning methods. PCA- and ISMLR-based hybrid models behave similarly, but ISMLR still obtains a better performance based on the training dataset using cross-validation grid search. Different machine learning algorithms rank the best depended on the evaluation metrics. For example, SVR dominates in the top rankings based on MAPE or MAE, whereas SVR, KRR, and ETR are frequently found in the top list based on adj.R^2 or RMSE. Among the ten methods, GPR is highly sensitive to the hyperparameters (the biggest variation), followed by GBR, HGBR, KRR, and SVR, whereas ETR, KNN, and RFR are much steadier and most of the model candidates achieve excellent results (**Figure A1(a)**). Other methods are between the two types and a good example is SVR that has a MAPE ranging from about 3.6% to 13.0% and an adj.R^2 ranging from about -0.146 to 0.852.

Overall, ETR, KRR, and SVR that generated best candidates were further investigated at the second stage.

The three selected methods were further explored by extending from 326 (the first stage) to 888 models at the second stage (**Figure A1 (b)**). The assessment shows that SVR achieves the highest overall performance (MAPE ≈ 3.6%), but could also obtain much worse results (e.g., 13.0%) without an appropriate hyperparameter optimization. The best SVR model was further evaluated based on testing dataset and exhibits an excellent performance (adj.R^2 ≈ 0.840; RMSE ≈ 2.71 MGD; MAPE ≈ 3.31%; MAE ≈ 1.34 MGD). Compared to SVR, ETR is able to achieve a comparable accuracy and is more likely to generate a favorable model (MAPE: 3.9 ~ 4.6%) in the cross-validation evaluation. Therefore, SVR is one of the best methods, whereas implementing ETR could reduce the complex of modeling process, which may be an advantage for onsite staffs who are less familiar with relevant knowledge.

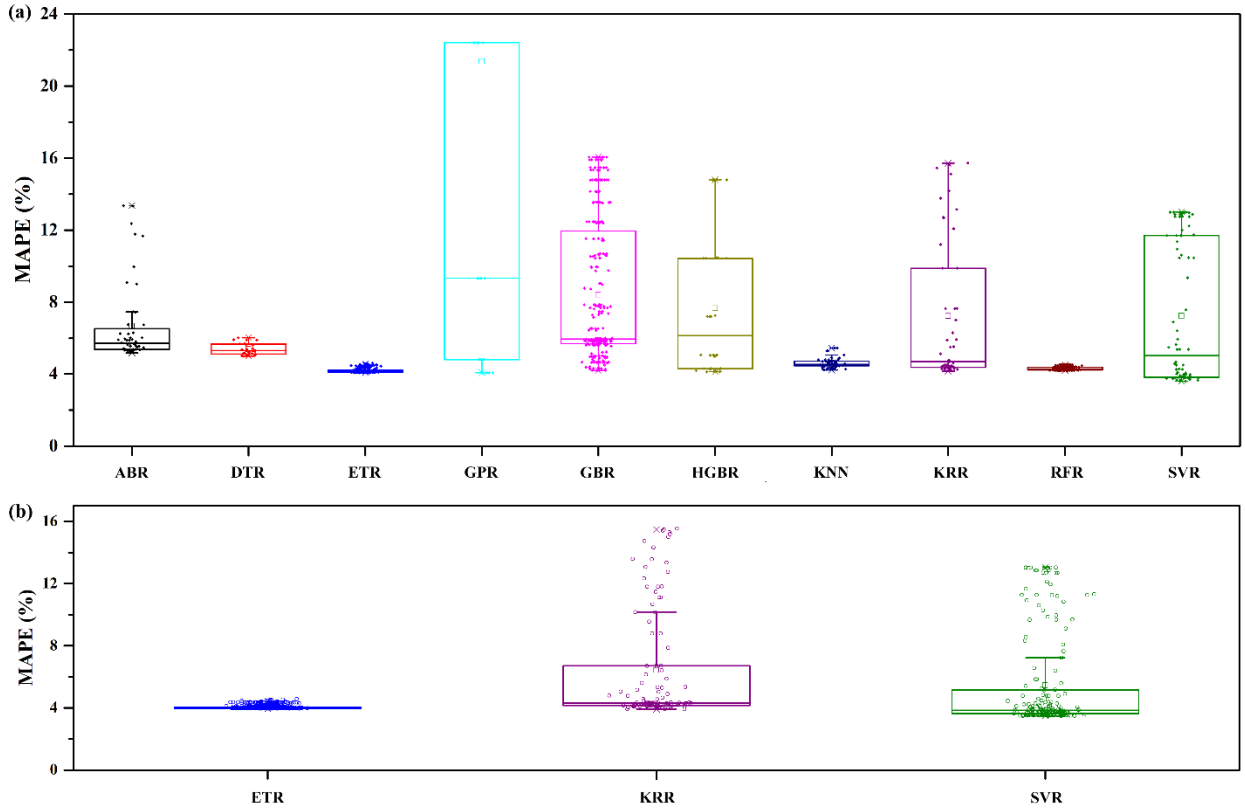


Figure A1. Box plots showing the prediction performance (e.g., MAPE) of (a) 999 model candidates (ten methods; cv = 3) at the first stage and (b) 888 models candidates (three best methods; cv = 10) using ISMLR, grid search, and cross-validation to determine the best machine learning models for prediction of the next day's influent wastewater flow at the Rock Creek WRRF.

Details of soft sensor development for the river flow prediction

PCA and ISMLR were able to identify 23 PCs (90% explained variance) and seven important regressors from the initial dataset, respectively. PCA-Bayesian Ridge regression and ISMLR-Lasso Lars regression (alpha = 1) are the best hybrid statistical models, respectively. The former model obtains an adj.R^2 of 0.545 and a MAPE of 80.6%, whereas the latter model could achieve an adj. R^2 of 0.602 and a MAPE of 60.8%. Similarly, ISMLR is superior over PCA in using machine learning algorithms, and the pattern of ISMLR-based machine learning models is similar to the case of wastewater flow prediction described in the section 2.1. For example, DTR, ETR, KNN, and RFR are less sensitive to the model structure, whereas GPR, GBR, HGBR, and KRR have much bigger variations. Overall, KNN, ETR, and KRR were

the three methods that generated best candidates from the 999 tested models at the first stage of machine learning optimization, so the second stage of cross-validation (cv = 5) grid search evaluated 776 models based on the three methods to refine their optimal model structures. The detailed assessment identified the best model is ISMLR-ETR with a fair performance based on the testing dataset (adj.$R^2 \approx 0.601$; RMSE ≈ 823 cfs; MAPE ≈ 41.5%; MAE ≈ 458 cfs). The monthly river flow prediction can be updating over the course of the month with the increasing amount of new daily data (river, rain, and snow), so onsite staffs can use the most recent, updated predicted river flow (flow*(t+1)nd*, where *t+1* means the next month and *nd* is the n^{th} day). As expected that adj.R^2 and MAPE values increases from 0.60 to 0.99 and decreases from 41.5% to 3.3% steadily over the month based on the testing dataset, respectively (**Figure 7(a)**). For example, the model could achieve a good prediction performance (adj.$R^2 \approx 0.861$; RMSE ≈ 480 cfs; MAPE ≈ 19.1%; MAE ≈ 257 cfs) when 14 days passed in the month and most of the measured data are well predicted (**Figure 7(b)**). The biggest errors are commonly found in peaks, but the underestimations of most peaks do not significantly affect the fact that they are classified into the high flow class (> 1000 cfs). For example, the biggest error (≈ 3061 cfs) is observed in November (predictions over December), 2015 when measured and predicted values are 7634 and 4573 cfs, respectively (**Figure 7(b)**); the predicted flow is still above 1000 cfs.

A subsequent evaluation of classification accuracy based on the predicted data was taken over the month (**Figure A2**). The overall accuracy increases from 0.8 to 1.0, and the classification accuracies in the low and high classes always maintain a high level. The result demonstrates that the soft sensor works well to provide the majority of accurate predictions, especially because the two classes account for more than 88% of historical records.

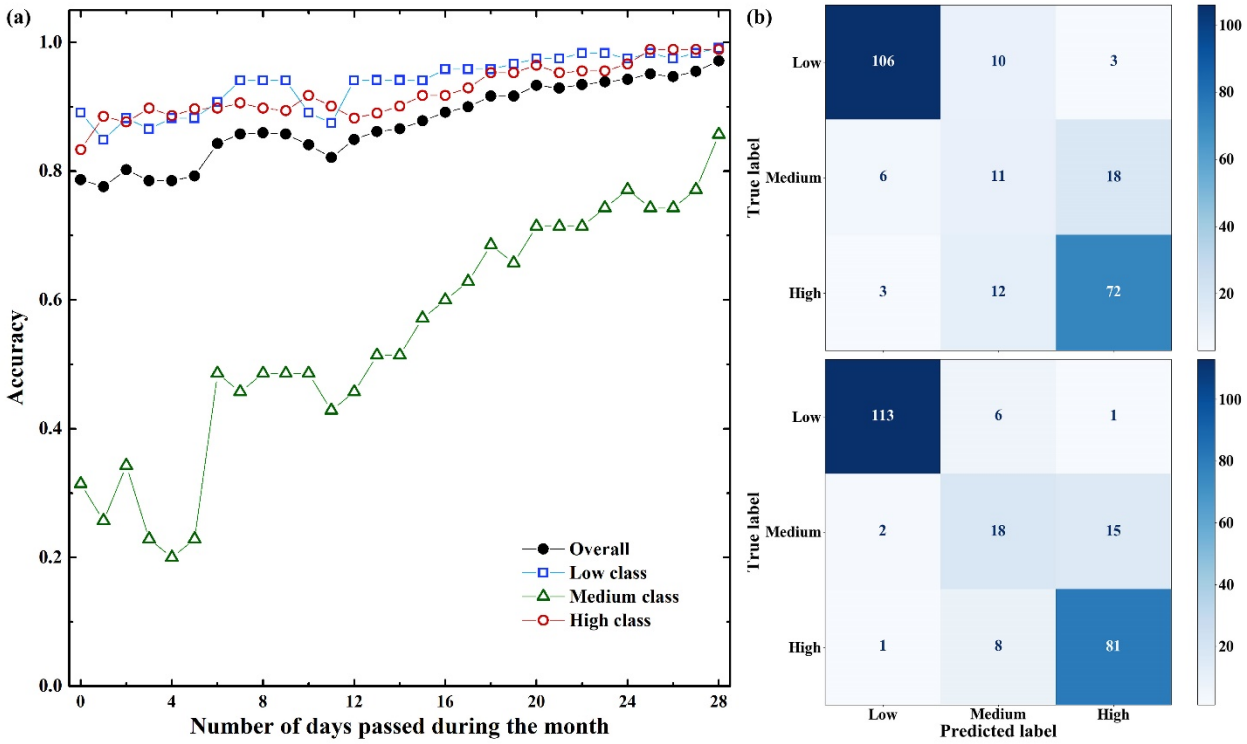


Figure A2. Classification results based on the predicted Tualatin River flow data. (a) Classification overall accuracy and accuracies for each of the three classes over the number of days passed during the next month. (b) Confusion matrix plots for classification results based on flow*(t+1)0d* (top) and flow*(t+1)14d* (bottom) models.