processes

MDPI

Article

Machine Learning-Based Prediction of Coagulant Dosing in Drinking Water Treatment Plants Using Polynomial Regression with Lasso Regularization

Jusuk An [1,2], Joonhong Park [2,*], Seungjae Yeon [1], Changseog Oh [1], Bokjin Lee [1], Woosik Jung [2], Jeongmin Yun [2] and Hyun Je Oh [1,*]

[1] Department of Environmental Research, Korea Institute of Civil Engineering and Building Technology (KICT), Goyang 10223, Republic of Korea; jusuk@kict.re.kr (J.A.); yeon@kict.re.kr (S.Y.); csoh@kict.re.kr (C.O.); bokjinlee@kict.re.kr (B.L.)
[2] Department of Civil and Environmental Engineering, Yonsei University, Seoul 03722, Republic of Korea; woosikjung1929@gmail.com (W.J.); jmy0204@yonsei.ac.kr (J.Y.)
* Correspondence: parkj@yonsei.ac.kr (J.P.); hjoh@kict.re.kr (H.J.O.)

Abstract

Coagulation is a critical unit process in drinking water treatment plants (DWTPs), where accurate dosing of coagulants such as polyaluminum chloride (PAC) and polyaluminum hydroxide chloride silicate (PACS) directly determines turbidity removal and operational stability. However, nonlinear interactions among water-quality variables complicate dosage prediction, and jar tests or operator heuristics cannot support real-time control. This study presents a scientifically interpretable and operationally transferable framework based on polynomial multiple linear regression (PMLR) with Lasso regularization, which was specifically developed for full-scale DWTP environments. While conventional PMLR rapidly overfits beyond polynomial degrees of 4–5, the Lasso-regularized model maintained stable generalization even at a degree of 10 by automatically pruning redundant terms and suppressing multicollinearity, thereby minimizing the need for manual hyperparameter tuning. Using 8303 hourly operational records from a full-scale DWTP in Korea, the Lasso-PMLR achieved R^2 = 0.951, RMSE = 0.120, and MAPE = 7.02%, outperforming traditional linear regression (R^2 = 0.896; MAPE = 8.64%). This proportional stability across increasing polynomial degrees, demonstrated directly using long-term real-world data, is particularly valuable for practical deployment because it ensures robustness without complex model-selection procedures. The transparent coefficient structure enables operators—who typically rely on jar tests—to understand and adjust dosing behavior, offering a field-ready and interpretable alternative to black-box models and supporting more efficient coagulant use, reduced sludge production, and sustainable automation in DWTP operation.

Keywords: coagulant dosing; drinking water treatment plants; interpretable machine learning; polynomial multiple linear regression (PMLR); lasso regularization; real-time control; sustainable operation

check for updates

Academic Editor: Iqbal M. Mujtaba

Received: 5 November 2025
Revised: 20 November 2025
Accepted: 25 November 2025
Published: 26 November 2025

Citation: An, J.; Park, J.; Yeon, S.; Oh, C.; Lee, B.; Jung, W.; Yun, J.; Oh, H.J. Machine Learning-Based Prediction of Coagulant Dosing in Drinking Water Treatment Plants Using Polynomial Regression with Lasso Regularization. *Processes* **2025**, *13*, 3829. https://doi.org/10.3390/pr13123829

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Coagulation in drinking water treatment plants (DWTPs) is a key process for removing colloids, natural organic matter, and turbidity-causing particles through the addition of chemical coagulants. Its effectiveness directly influences downstream treatment performance, as improper dosing can result in residual turbidity, excessive chemical consumption,

increased sludge production, and higher operational costs [1]. In many regions, seasonal fluctuations and climate-driven variability further complicate this challenge by inducing rapid changes in raw-water quality. Such conditions make accurate and stable coagulant dosing essential for maintaining both technical reliability and economic sustainability.

The operational data used in this study were obtained from a full-scale DWTP with a treatment capacity of approximately 200,000 m^3/day, which treats surface water through a conventional process train consisting of coagulation, flocculation, sedimentation, sand filtration, and final disinfection. Online sensors measuring turbidity, pH, electrical conductivity (EC), temperature, and color are installed immediately upstream of the coagulation stage, providing continuous monitoring of raw-water conditions that directly influence chemical dosing decisions. This operational context is crucial for understanding the dataset and assessing the practical applicability of the proposed model; further details on data acquisition and quality control are provided in Section 2.1.

Although automation in DWTPs has advanced considerably, determining the optimal coagulant dosage in real time remains difficult because of the highly nonlinear interactions among multiple water-quality parameters and operational variables [2]. Such complexity limits the reliability of empirical dosing formulas or operator heuristics, which remain widely used in practice. In this study, the applied coagulant was polyaluminum hydroxide chloride silicate (PACS), a widely adopted agent known for strong turbidity-removal performance and stable operation under variable influent conditions. However, PACS dosage is still commonly determined using jar-test lookup tables or operator experience, both of which struggle to respond to rapid influent fluctuations and are therefore unsuitable for real-time control [3].

To address these limitations, recent studies have explored advanced statistical and machine-learning (ML) approaches for predicting coagulant demand [4–6]. Given the nonlinear relationships between influent characteristics and dosage, fuzzy inference systems (FIS) [5–9] and artificial neural networks (ANN) [10–14] have achieved strong predictive performance under diverse operating conditions [15–17]. More recently, ensemble-based and explainable-AI methods such as random forest (RF), Adaboost, gradient boosting machine (GBM), and XGBoost have been introduced, often in combination with SHAP-based post hoc interpretation frameworks to enhance model transparency [18–21]. Despite their advantages, these models require substantial parameter tuning and computational resources and still operate largely as "black boxes," limiting operator trust and reducing their applicability in full-scale DWTP operation [17]. Deep-learning models trained on large-scale operational datasets have also been explored for coagulant dose optimization; however, these approaches remain computationally intensive and lack the transparency required for routine use in full-scale DWTP operations [22].

A recent systematic review also summarized the rapid expansion of AI-based optimization methods for coagulation processes, highlighting persistent challenges related to model complexity, interpretability, and real-world deployability in DWTPs [23].

These challenges highlight the need for an approach that simultaneously balances predictive accuracy, scientific and operational interpretability, and simplicity suitable for real-time deployment. Motivated by this gap, the present study develops a scientifically interpretable yet practically deployable framework based on polynomial multiple linear regression (PMLR) with regularization. Polynomial expansion captures nonlinear interactions among key water-quality parameters, while Ridge and Lasso penalties constrain model complexity to enhance generalization [19–21]. This formulation quantifies each variable's contribution to coagulant demand, revealing mechanistic dependencies among turbidity, pH, and EC and providing coefficient-level transparency rarely achieved in ML-based control frameworks. The model's simplicity enables seamless integration with

existing PLC–SCADA systems and allows operators to validate or adjust dosing decisions using familiar process-control concepts.

Compared with black-box models such as ANN or XGBoost—which require complex hyperparameter tuning and post hoc interpretation—the proposed PMLR + Lasso framework offers explicit coefficient-based interpretability while maintaining comparable predictive accuracy [19]. Previous research has primarily emphasized performance optimization using complex ML architectures, often at the expense of interpretability and operational usability. However, no study has systematically examined how an interpretable regression model behaves under high-degree polynomial expansion when applied to long-term full-scale DWTP data. This distinction is critical, because DWTPs require models that remain stable as raw-water conditions fluctuate seasonally and during turbidity shocks, without demanding frequent retuning or specialized ML expertise. In this study, we show that Lasso-regularized PMLR uniquely maintains generalization even as the polynomial degree increases—a behavior not observed with unregularized PMLR or Ridge—thereby reducing the tuning burden and enabling robust real-time deployment. This proportional stability under increasing polynomial degrees constitutes a key practical advantage for DWTP automation, where model maintainability is as important as predictive accuracy.

To the best of our knowledge, no previous study has systematically evaluated an interpretable and SCADA-ready regression framework for coagulant dosing using one year of full-scale DWTP data, particularly under extreme turbidity events and seasonal variability. Accordingly, the objectives of this study are threefold: (1) to develop a scientifically interpretable yet operationally simple regression-based framework for PACS coagulant dosing; (2) to evaluate its predictive accuracy and stability using one year of full-scale operational data; and (3) to demonstrate its applicability as a transparent and field-ready decision-support tool for DWTP operators.

2. Methodology

2.1. Data Collection

Operational data were obtained from a full-scale drinking water treatment plant (DWTP) in A City, Republic of Korea, with a treatment capacity of approximately 200,000 m^3/day. The plant treats surface water using a conventional process train consisting of coagulation, flocculation, sedimentation, sand filtration, and disinfection. Online instruments measuring turbidity, pH, electrical conductivity (EC), temperature, and color are installed immediately upstream of the coagulation stage and continuously record raw-water conditions that guide chemical dosing decisions.

All measurements were acquired through the plant's PLC (Programmable Logic Controller)–HMI (Human–Machine Interface) system in 2020. Instruments included a turbidity analyzer (1720E, Hach Company, Loveland, CO, USA), pH and electrical conductivity (EC) sensors (TP 44,900 series, Endress+Hauser, Reinach, Switzerland), and temperature and color sensors integrated into the SC200 controller platform (Hach Company, Loveland, CO, USA). These sensors comply with national water-quality monitoring standards defined by the Korea Water and Wastewater Works Association (KWWA) and Korean Industrial Standards (KS), and were calibrated as part of the plant's routine quality-management program to ensure accurate and reliable operation.

Hourly measurements were automatically logged in the PLC–HMI database and screened to remove missing or erroneous values arising from sensor maintenance, cleaning, or communication faults. Quality-control rules were applied consistently across all parameters: records containing any missing fields were removed, and physically implausible values—such as turbidity < 0 NTU, pH < 4 or >10, EC < 0 μS/cm, or temperature < 0 °C or >40 °C—were excluded based on established operational ranges. Short-duration spikes

caused by maintenance or flushing events were identified by cross-checking with plant logs and discarded. After these steps, a total of 8303 complete hourly observations remained for model development.

The monthly distribution of influent water quality—including turbidity, pH, EC, temperature, and color—is summarized in Table 1, illustrating the seasonal variability captured in the dataset and the range of raw-water conditions under which the model was evaluated.

Table 1. Monthly influent water quality at the DWTP in A city, Republic of Korea.

Sampling Point	**Raw Water**											
Month	**JAN**	**FEB**	**MAR**	**APR**	**MAY**	**JUN**	**JUL**	**AUG**	**SEP**	**OCT**	**NOV**	**DEC**
Temperature (°C)	6.4 [a] (5.7–7.6) [b]	7.0 (5.8–8.2)	9.6 (7.8–11.9)	13.2 (11.9–15.0)	17.8 (15.1–20.7)	23.2 (20.4–25.0)	23.5 (22.1–24.7)	22.5 (21.3–24.1)	21.7 (20.7–24.1)	18.1 (15.8–20.7)	13.7 (11.4–16.6)	7.4 (4.4–12.0)
Turbidity (NTU)	5.8 (1.7–13.4)	3.8 (1.5–11.2)	3.3 (1.2–12.1)	4.3 (2.1–8.5)	3.9 (1.6–9.0)	4.4 (1.6–11.8)	6.6 (1.1–83.9)	32.2 (2.3–100.0)	14.6 (2.5–39.3)	12.0 (0.1–100.0)	5.6 (1.2–69.2)	2.2 (0.3–8.4)
pH (-)	7.5 (7.1–8.0)	7.9 (7.2–8.6)	7.6 (6.9–8.7)	7.8 (6.7–8.7)	7.0 (6.4–7.9)	6.5 (6.0–7.4)	6.4 (6.0–7.2)	6.1 (5.6–6.5)	6.2 (5.8–7.4)	6.6 (5.9–7.5)	6.6 (6.0–7.1)	7.2 (6.6–7.8)
E.C. (μS/cm)	145.7 (110–192)	146.6 (115–199)	140.6 (112–182)	147.9 (124–179)	145.5 (108–176)	157.0 (108–194)	164.8 (104–204)	112.5 (101–144)	111.2 (101–123)	133.2 (101–162)	141.6 (113–181)	193.6 (175–213)
Color (°)	1.2 (1.0–2.0)	1.0 (1.0–2.0)	1.1 (1.0–2.0)	1.0 (1.0–2.0)	1.0 (1.0–2.0)	1.0 (1.0–2.0)	3.7 (1.0–42.0)	31.8 (3.0–321)	5.4 (2.0–21.0)	2.6 (2.0–7.0)	1.7 (1.0–4.0)	1.0 (1.0–2.0)
Previous-day PACS dosing rate((mg/L)·h)	0.8 (0.7–1.4)	0.7 (0.5–1.2)	0.9 (0.5–1.5)	0.9 (0.7–1.3)	1.0 (0.6–1.7)	1.1 (0.8–1.6)	1.4 (0.9–2.7)	2.3 (1.1–4.1)	1.6 (1.0–2.2)	1.3 (0.9–1.9)	1.4 (1.1–2.0)	1.0 (0.9–1.6)
Current PACS dosing rate ((mg/L)·h)	0.8 (0.7–1.4)	0.7 (0.5–1.2)	0.9 (0.5–1.5)	0.9 (0.7–1.3)	1.0 (0.6–1.7)	1.1 (0.8–1.6)	1.5 (0.9–4.0)	2.2 (1.1–4.1)	1.6 (1.0–2.2)	1.3 (0.9–1.9)	1.4 (1.1–2.0)	1.0 (0.9–1.6)

[a] Average data, [b] Range of data.

The monitoring system provides raw-water measurements at minute-level resolution; however, practical coagulation control in full-scale DWTPs typically operates at 30–60 min update intervals. Minute-level turbidity or pH signals exhibit substantial short-term fluctuations due to hydraulic instability and sensor noise, whereas chemical dosing pumps respond on the scale of tens of minutes rather than seconds. Therefore, aligning all parameters to an hourly interval offers a stable and operationally meaningful basis for model training and evaluation. The proposed framework is not restricted to 60-min data; if utilities deploy higher-frequency measurements (e.g., 10–15 min), the model can generate predictions at any desired control interval.

2.2. Model Fitting

Six input variables—temperature, turbidity, pH, electrical conductivity (EC), color, and the previous-day PACS dosing rate—were used as predictors, while the current PACS dosing rate served as the target variable. The previous-day dosing term (t–1 d) was included to capture slow-response and carry-over effects commonly observed in full-scale coagulation systems. It was defined as the mean PACS dosage applied during the preceding 24 h and aligned to the subsequent day at hourly resolution.

A total of 8303 observations were divided chronologically into two subsets: 75% for model development and 25% for validation. All predictors were standardized (zero mean, unit variance) to prevent scale imbalance and ensure numerical stability during the regression optimization. After model training, the learned coefficients were inversely transformed to their original units to allow direct physical interpretation of each parameter's influence on coagulant demand.

Because the dataset consists of sequential operational records from a continuously operating DWTP, random k-fold cross-validation was not used, as it would break temporal structure and introduce information leakage. Instead, a chronological train–test split was adopted to preserve temporal dependence, and the final model used for coefficient interpretation was refitted on the full dataset after validation to obtain the most stable coefficient estimates.

To avoid leakage, data preprocessing steps—including *StandardScaler* fitting, *PolynomialFeatures* expansion, and the construction of all interaction terms—were performed after the *train_test_split*. All transformations were fitted exclusively on the training subset and subsequently applied to the validation subset.

Finally, polynomial expansion was applied to capture nonlinear interactions among predictors. The purpose of the degree exploration in this study was not to identify a single "optimal" polynomial degree, but to evaluate how regularization behaves under increasingly complex feature spaces. Therefore, polynomial degrees from 1 to 10 were examined to assess stability, generalization, and the scalability of regularization methods under high-dimensional expansion. A detailed explanation of the regularization strategies is provided in Section 2.7.

2.3. Implementation

All model development and analyses were performed in Python 3.8 using the Jupyter Notebook environment (Anaconda distribution). Data handling and preprocessing were conducted with the *pandas* and *NumPy* libraries. Input variables were standardized using *scikit-learn's StandardScaler*, and the dataset was divided into training and testing subsets using the *train_test_split* function in accordance with the chronological framework described in Section 2.2.

Polynomial expansion and model fitting were implemented using *scikit-learn's PolynomialFeatures, LinearRegression, Ridge,* and *Lasso* classes. Model evaluation employed standard statistical metrics—R^2, RMSE, and MAPE—as defined in Section 2.4, computed with scikit-learn functions and cross-verified using custom Python scripts to ensure numerical consistency.

All visualizations, including prediction performance and residual analyses, were generated using *matplotlib*. This open-source configuration enables full reproducibility of the modeling workflow, facilitates transparent examination of regression coefficients, and supports straightforward integration of the trained model into SCADA or decision-support environments.

2.4. Performance Metrics

Model performance was evaluated using four standard statistical metrics: the Pearson correlation coefficient (r), root mean squared error (RMSE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2). The correlation coefficient (r) quantifies the linear agreement between observed and predicted dosing rates, while RMSE and MAPE assess absolute and relative prediction errors, respectively. Together, these indices provide a comprehensive evaluation of model accuracy.

The Pearson correlation coefficient (r) quantifies the linear association between the input and output variables (Equation (1)):

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]} \quad (1)$$

where n is the number of data points, x represents input variable values, and y represents output variable values.

The root-mean-square error (RMSE) measures the deviation between predicted and observed values (Equation (2)):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(O_p(i) - O_a(i)\right)^2} \quad (2)$$

where $O_p(i)$ is the predicted dosing rate, $O_a(i)$ is the observed dosing rate, and n is the number of data points.

The mean absolute percentage error (MAPE) expresses relative prediction error as a percentage (Equation (3)):

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{O_p(i) - O_a(i)}{O_a(i)}\right| \quad (3)$$

The coefficient of determination (R^2) evaluates the proportion of variance in the observed data explained by the model (Equation (4)):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(O_p(i) - O_a(i)\right)^2}{\sum_{i=1}^{n}\left(O_a(i) - O_{mean}(i)\right)^2} \quad (4)$$

where O_{mean} is the mean of the observed dosing rate values.

2.5. Multiple Linear Regression

The general form of the MLR model is expressed in Equation (5):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (5)$$

where Y is the predicted dosing rate ((mg/L)·h), X_1–X_n are the input parameters, β are the regression coefficients, n is the number of predictors, and ε is the error term.

2.6. Polynomial Multiple Linear Regression

Because the coagulation process in DWTPs is inherently nonlinear, PMLR was employed by expanding predictors into higher-order and interaction terms. This allows the model to reflect physical and chemical interactions, such as the combined effects of turbidity and pH on dosing requirements.

Polynomial degrees ranging from 1 to 10 were examined not to identify a single "optimal" polynomial order, but to evaluate how model stability and generalization change as feature-space complexity increases. This design allows the study to systematically assess the scalability of PMLR under high-dimensional expansion and the role of regularization (Ridge and Lasso) in maintaining numerical stability—a critical requirement for deployable real-time control in DWTPs.

The general polynomial form is expressed in Equation (6):

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_n X^n + \varepsilon \quad (6)$$

where Y is the predicted dosing rate, X is an input variable, n is the polynomial degree, and ε is the error term.

When the polynomial degree is 1, the PMLR reduces to a standard MLR model with six predictors (Equation (7)):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 \quad (7)$$

When the degree is increased to 2, squared terms and pairwise interaction terms are added, resulting in 27 predictors (Equation (8)):

$$\begin{aligned} Y = \; & \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \\ & \beta_7 {X_1}^2 + \beta_8 X_1 X_2 + \beta_9 X_1 X_3 + \beta_{10} X_1 X_4 + \beta_{11} X_1 X_5 + \beta_{12} X_1 X_6 + \\ & \beta_{13} {X_2}^2 + \beta_{14} X_2 X_3 + \beta_{15} X_2 X_4 + \beta_{16} X_2 X_5 + \beta_{17} X_2 X_6 + \\ & \beta_{18} {X_3}^2 + \beta_{19} X_3 X_4 + \beta_{20} X_3 X_5 + \beta_{21} X_3 X_6 + \\ & \beta_{22} {X_4}^2 + \beta_{23} X_4 X_5 + \beta_{24} X_4 X_6 + \\ & \beta_{25} {X_5}^2 + \beta_{26} X_5 X_6 + \\ & \beta_{27} {X_6}^2 \end{aligned} \tag{8}$$

This explicit expansion illustrates how the number of predictors grows rapidly with polynomial degree. Beyond degree 2, the number of higher-order and interaction terms becomes very large; therefore, only the degree-2 expansion is shown here, while higher-degree expansions follow the same principle. Polynomial degrees from 1 to 10 were examined to investigate the scalability and stability of PMLR under increasingly complex feature spaces, rather than to determine a single optimal degree. Among these, degree 4 yielded the lowest validation RMSE for the unregularized PMLR model, and this selection criterion has been explicitly applied in the baseline evaluation.

2.7. Lasso and Ridge Regularization

Polynomial expansion significantly increases the number of predictors, which can lead to overfitting when the model complexity becomes comparable to or exceeds the available sample size. To address this issue, two regularization techniques—Ridge and Lasso—were applied.

- Ridge regression penalizes the squared magnitude of coefficients, effectively reducing variance and improving numerical stability in the presence of multicollinearity. However, Ridge does not eliminate redundant variables, as all coefficients remain nonzero.
- Lasso regression, in contrast, imposes an absolute (L_1) penalty that can shrink less influential coefficients exactly to zero, thereby simplifying the model and enabling embedded variable selection. This sparsity provides interpretability by identifying only the most physically meaningful contributors to coagulant demand and reducing the dimensionality of the expanded polynomial space.

Both methods were systematically tested to identify the balance between predictive.

Although the six raw input variables exhibited limited multicollinearity, the high-degree polynomial expansion introduced substantial collinearity among interaction terms. Under this high-dimensional setting, regularization is essential to prevent overfitting and to maintain numerical stability.

Both methods were systematically tested to evaluate the balance between predictive accuracy and model simplicity. In this study, Ridge regression slightly improved model stability but retained weakly influential predictors. In contrast, the Lasso-regularized PMLR model demonstrated superior generalization and provided the most consistent predictive performance across polynomial degrees, achieving the best trade-off between accuracy and parsimony. This behavior aligns with the study's objective to develop a transparent, operator-oriented framework, as the reduced set of active predictors can be interpreted directly in terms of measurable physicochemical parameters.

Model performance was evaluated using R^2, RMSE, and MAPE, confirming that Lasso regularization produced the most robust and physically interpretable coefficients. For

the Lasso model, the regularization parameter was set to α = 1.0, which yielded stable performance without degrading predictive accuracy.

3. Results and Discussion

3.1. Multiple Linear Regression

Multiple linear regression (MLR) was first developed using six input variables—temperature, turbidity, pH, electrical conductivity (EC), color, and the previous-day PACS dosing rate—to predict the current PACS dosage. The Pearson correlation coefficients for these variables are summarized in Table 2. Turbidity and the previous-day dosing rate exhibited relatively strong positive correlations with the output variable, whereas pH and EC showed negative correlations, consistent with their known physicochemical influence on coagulation effectiveness.

Table 2. Pearson correlation coefficients among input and output variables.

Parameters	Temperature (°C)	Turbidity (NTU)	pH (-)	E.C. (μS/cm)	Color (°)	Precious-Day PACS Dosing Rate ((mg/L)·h)	Current PACS Dosing Rate (mg/L)·h
Temperature (°C)	1.000	0.296	−0.739	−0.309	0.201	0.510	0.516
Turbidity (NTU)		1.000	−0.297	−0.387	0.349	0.590	0.573
pH (-)			1.000	0.287	−0.268	−0.608	−0.601
E.C. (μS/cm)				1.000	−0.237	−0.418	−0.416
Color (°)					1.000	0.668	0.660
Previous-day PACS dosing rate ((mg/L)·h)						1.000	0.944
Current PACS dosing rate (mg/L)·h)							1.000

The predictive performance of the MLR model is shown in Figure 1, with R^2 values of 0.902 (training), 0.880 (testing), and 0.896 (overall). Corresponding MAPE values were 8.57%, 8.85%, and 8.64%, and RMSE values were 0.167, 0.193, and 0.174, respectively. These metrics indicate that MLR achieved moderately high predictive accuracy. However, residuals were not uniformly distributed across operating conditions. Errors increased markedly during high-turbidity events (>30 NTU) and low-temperature conditions (<10 °C), reflecting the inability of linear models to capture the nonlinear interactions that dominate coagulation behavior under extreme influent variability. From an operational standpoint, this suggests that relying solely on MLR could lead to under- or overestimation of coagulant demand during challenging water-quality episodes, with potential risks to treated water quality.

The regression coefficients for the six predictors were 2.994×10^{-3} (temperature), 9.452×10^{-4} (turbidity), -2.584×10^{-2} (pH), -4.186×10^{-4} (EC), 2.213×10^{-3} (color), and 8.279×10^{-1} (previous-day dosing), with an intercept of 0.380. These coefficients align with the correlation structure, confirming that turbidity and historical dosing exert the strongest influence on current dosing requirements.

This strong influence is mechanistically reasonable. Turbidity directly reflects the concentration of colloidal particles requiring charge neutralization and is therefore the primary determinant of coagulant demand. The previous-day dosing rate captures a residual carry-over effect: hydrolyzed aluminum species (e.g., $Al(OH)^{2+}$, $Al(OH)_2^{+}$, and polymeric Al_{13} clusters) persist for several hours in full-scale DWTPs and influence particle stability in subsequent periods. This persistence explains why historical dosing remains a dominant predictor. In addition, dosing adjustments in full-scale DWTPs are typically made on a daily basis through jar-test-based decision-making combined with flow-paced control, creating day-to-day continuity in operator-driven dosing patterns. The previous-day dosing rate therefore reflects both physicochemical carry-over and operational behavior, explaining its consistently high importance in the model.

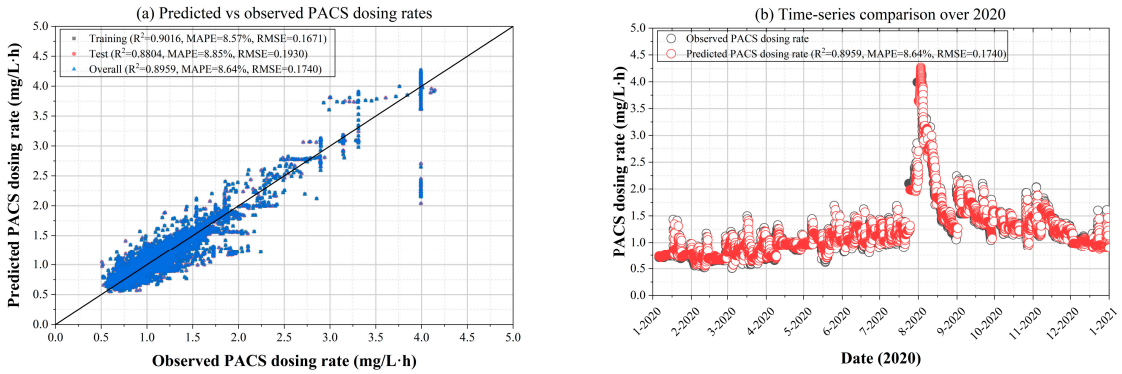


Figure 1. Baseline multiple linear regression (MLR) model performance for PACS dosing rate prediction. (**a**) Parity plot of observed vs. predicted dosing rates for training (75%), testing (25%), and overall datasets, with the 1:1 reference line (y = x) included for comparison. (**b**) Time-series comparison of observed and predicted dosing rates for the entire study period (2020). The MLR achieved R^2 = 0.896 and MAPE = 8.64%, indicating reasonable predictive accuracy but limited ability to capture nonlinear behavior.

3.2. Polynomial Multiple Linear Regression

To better capture nonlinear interactions among influent water-quality parameters, polynomial multiple linear regression (PMLR) was applied. By introducing higher-order and interaction terms, PMLR can represent complex physicochemical relationships—for example, the combined effects of turbidity and pH on coagulant demand.

Polynomial expansion increased the number of predictors from 6 (degree = 1) to 8007 (degree = 10). This exponential growth highlights the risk of overfitting as model complexity approaches or exceeds the number of available samples. With 8007 predictors and 8303 samples, the model operates in a near $p \approx n$ regime ($p/n \approx 0.96$). Under such high-dimensional conditions, the design matrix becomes nearly rank-deficient because of strong multicollinearity among polynomial interaction terms. This inflates coefficient variance, destabilizes ordinary least squares (OLS) estimation, and ultimately leads to severe overfitting. These dimensionality constraints justify the need for regularization to ensure generalization and numerical stability.

Predictive performance for polynomial degrees 1–4 is shown in Figure 2. As the polynomial degree increased, model accuracy improved: R^2 increased from 0.896 (degree = 1, equivalent to baseline MLR) to 0.958 (degree = 4), while MAPE decreased from 8.64% to 6.68%. These results demonstrate the benefit of polynomial expansion in capturing nonlinear dynamics of the coagulation process.

However, when the degree increased further (degrees 5–10), overfitting became pronounced (Figure 3). Training R^2 values exceeded 0.97, yet test-set performance deteriorated sharply, with error values exceeding MAPE > 100% at degrees ≥ 7. Predictions diverged unrealistically from observed values, indicating numerical instability. From an operational standpoint, such instability would result in erratic or unsafe dosing recommendations that cannot be used in real-time DWTP operation.

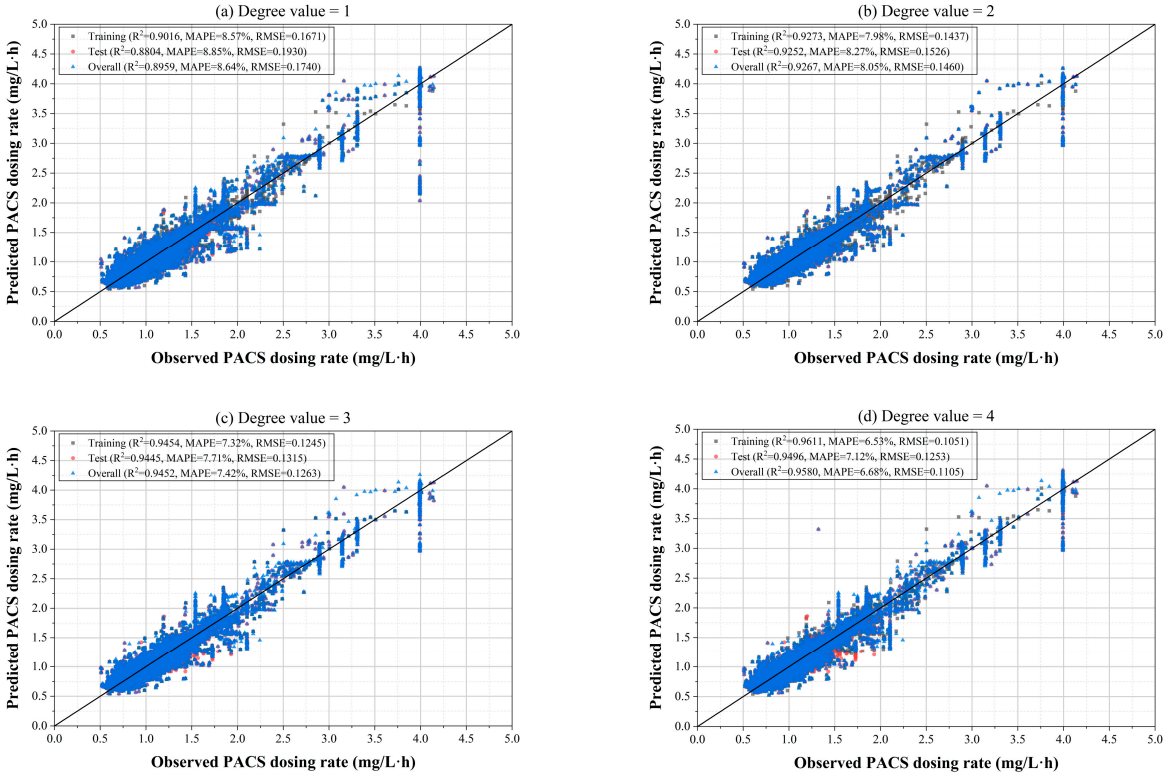


Figure 2. PMLR without regularization (degrees 1–4). Observed vs. predicted PACS dosing rates across polynomial degrees 1–4. As the degree increased, predictive accuracy improved (R^2 up to 0.958; MAPE reduced to 6.68%), demonstrating the benefit of polynomial expansion in capturing nonlinear interactions. The 1:1 reference line (y = x) is included to indicate ideal agreement between observed and predicted values.

Overall, these findings indicate that moderate polynomial expansion (degrees 2–4) improves predictive accuracy by capturing essential nonlinear effects, whereas higher-degree expansions introduce excessive model complexity and poor generalization. This trade-off underscores the necessity of robust regularization techniques to balance accuracy and stability when employing polynomial features—motivating the application of Ridge and Lasso methods examined in Section 3.3.

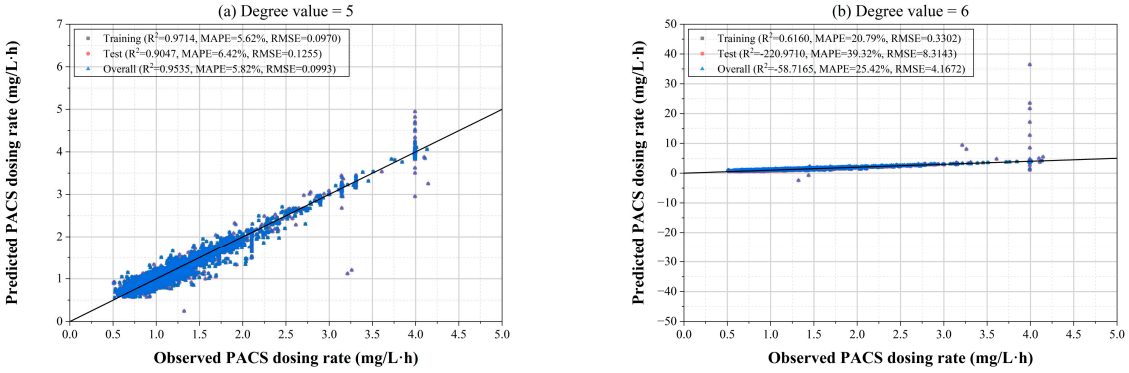


Figure 3. *Cont.*

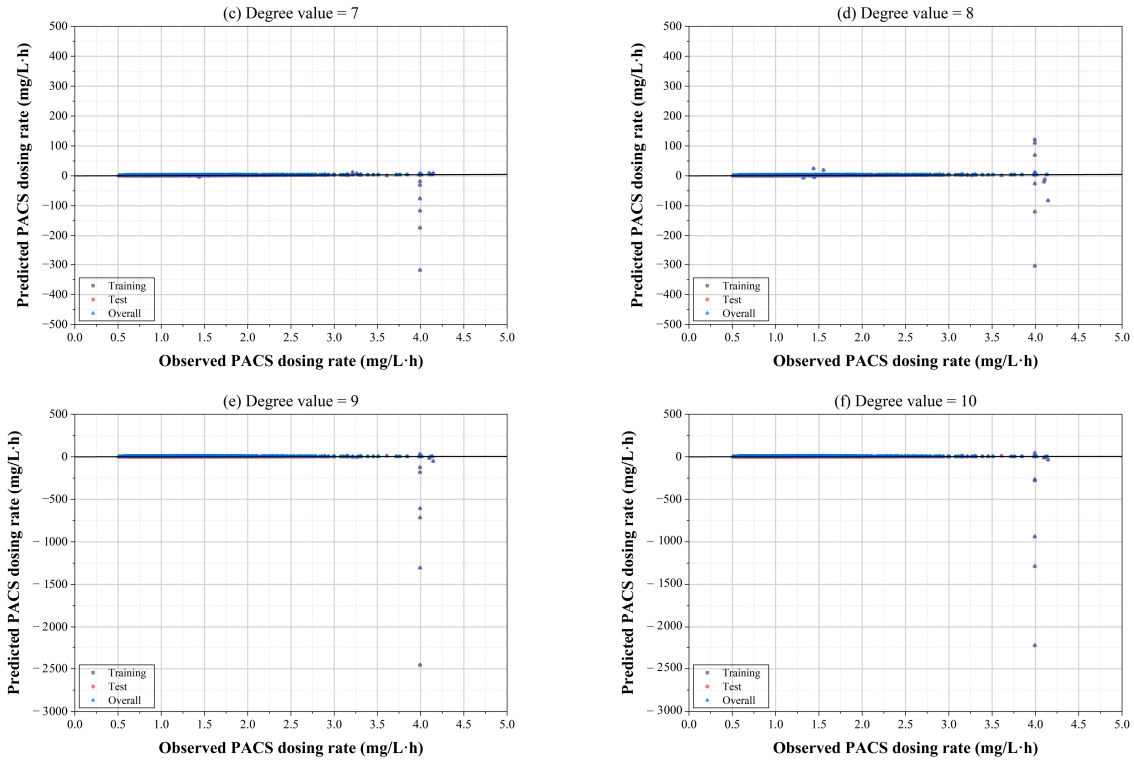


Figure 3. PMLR without regularization (degrees 5–10). Observed vs. predicted PACS dosing rates across polynomial degrees 5–10. Training accuracy increased, but testing performance deteriorated (e.g., MAPE > 100% at degree $\geq$ 7), highlighting severe overfitting.

3.3. Polynomial Multiple Linear Regression with Lasso and Ridge Regularizations

Polynomial expansion substantially increased the expressive capacity of the regression model, but also amplified the risk of overfitting. As shown in Figure 4, unregularized PMLR exhibited steadily increasing training R^2 (exceeding 0.97 at degree 10), while test errors diverged sharply beyond degree 4. This divergence reflects a fundamental limitation of ordinary least squares under high-dimensional expansion: increased model complexity enhances in-sample fit but undermines predictive reliability under unseen conditions. Such instability is unacceptable for real-time coagulant dosing in DWTPs. Therefore, model performance was evaluated using the chronological 75:25 train–test split described in Section 2.2 to prevent temporal information leakage.

To mitigate overfitting, Ridge and Lasso regularizations were applied (Figures 5 and 6). Ridge regression reduced variance relative to unregularized PMLR; however, residual overfitting persisted at higher degrees, as test errors failed to converge and correlated coefficients grew in magnitude. In contrast, Lasso regularization demonstrated consistent generalization, with training and testing metrics converging smoothly across polynomial degrees. R^2 values increased gradually, while both RMSE and MAPE decreased for the training, testing, and overall datasets.

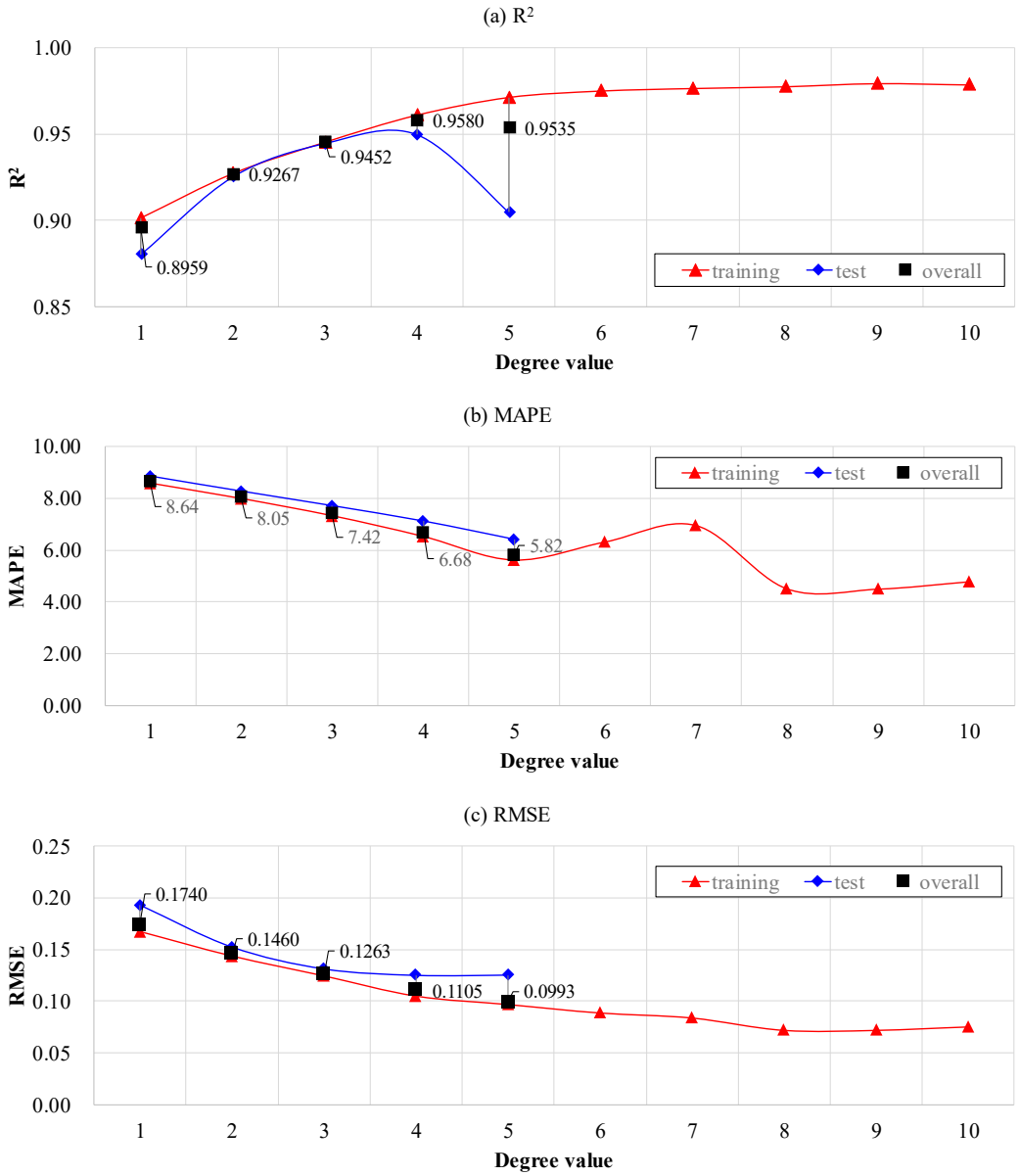


Figure 4. Predictive performance of unregularized PMLR across degrees 1–10. Comparison of R^2, MAPE, and RMSE for training and testing datasets. While training R^2 exceeded 0.97 at degree 10, test errors increased, confirming overfitting beyond degree 4.

The superior performance of Lasso arises from its ability to shrink less influential coefficients exactly to zero, performing embedded variable selection. For example, variables such as color, which exhibited weak correlation with coagulant demand, were effectively excluded, whereas turbidity and pH retained strong influence. This sparsity reduces unnecessary polynomial terms, enhances interpretability, and prevents coefficient inflation within the expanded feature space.

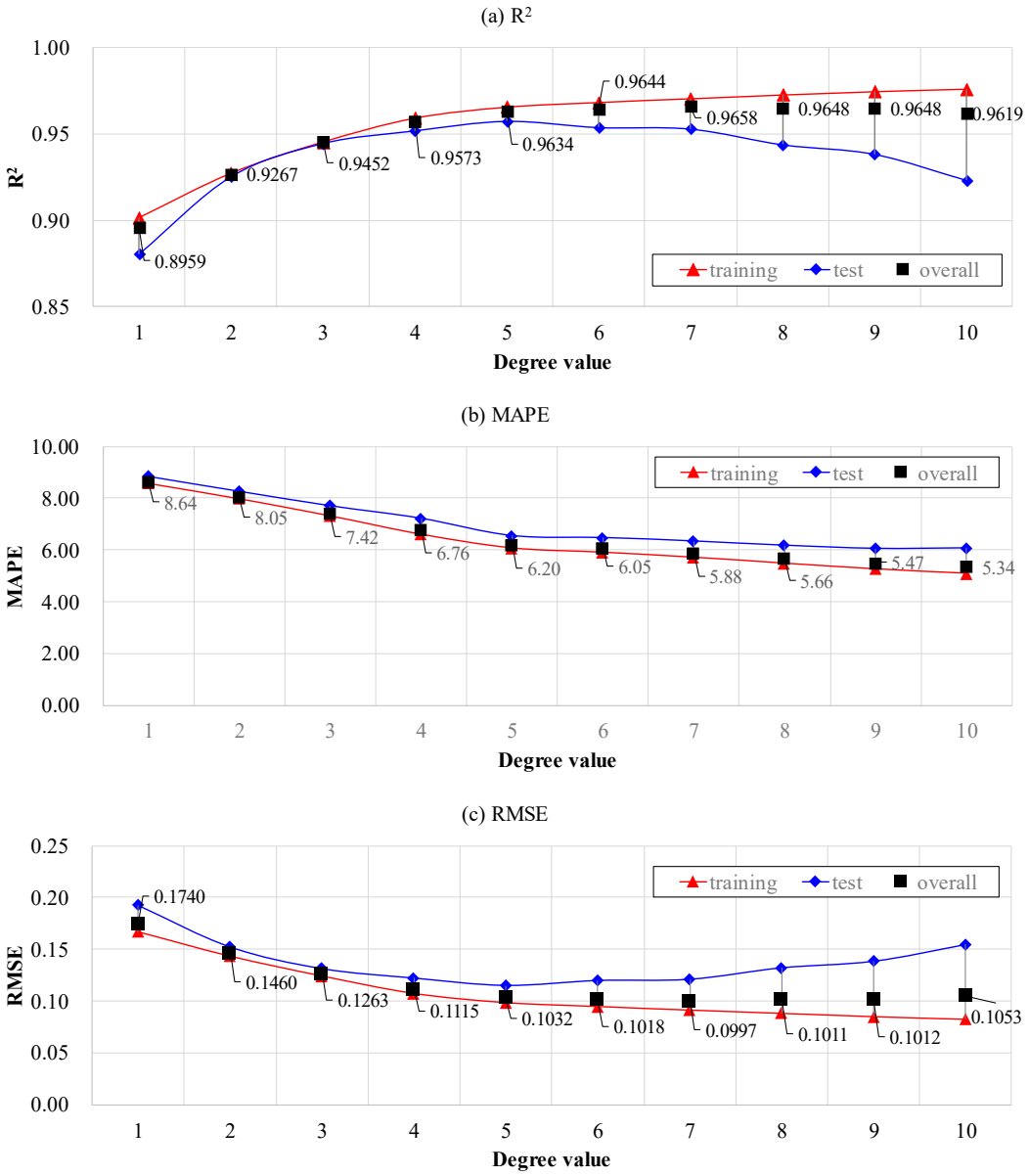


Figure 5. Ridge-regularized PMLR performance across polynomial degrees. Comparison of R^2 (**a**), MAPE (**b**), and RMSE (**c**) for training and testing datasets. Ridge reduced variance compared with unregularized PMLR, but residual overfitting persisted beyond degree 4.

Ridge and Lasso were further compared in terms of their stability as the polynomial degree (D) increased. Ridge regression provided initial variance reduction but showed progressive overfitting at higher degrees, particularly when multicollinearity intensified among interaction terms. In contrast, Lasso maintained stable generalization by suppressing redundant predictors and preserving only the most physically meaningful contributors. This indicates that the L_1 penalty is more robust than the L_2 penalty for managing the severe collinearity present in high-degree polynomial expansions. Consequently, the Lasso-regularized PMLR was selected as the most suitable configuration, achieving the best balance between predictive accuracy, numerical stability, and scientific transparency.

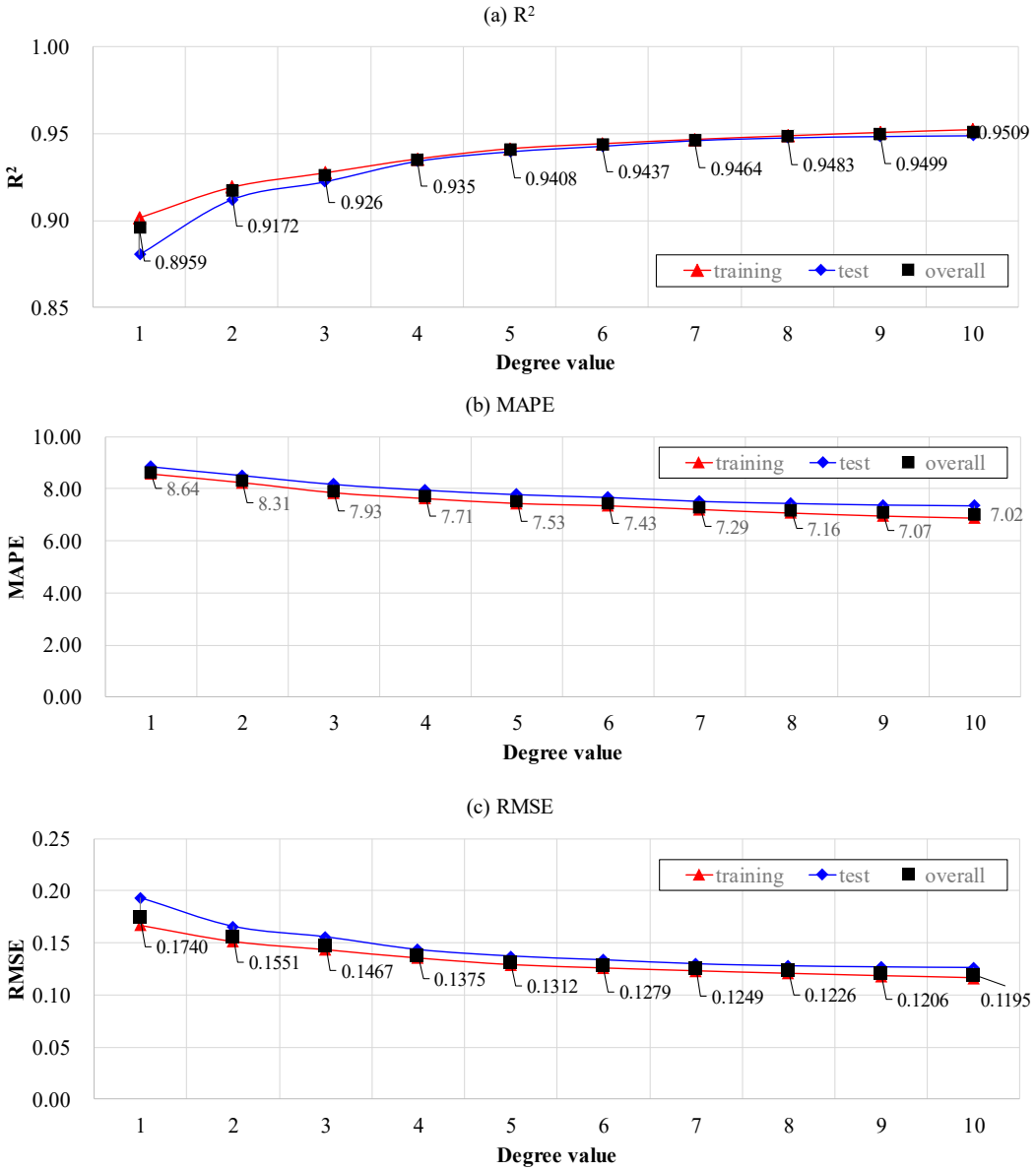


Figure 6. Lasso-regularized PMLR performance across polynomial degrees. Comparison of R^2 (**a**), MAPE (**b**), and RMSE (**c**) for training and testing datasets. Lasso provided superior generalization by shrinking irrelevant coefficients, ensuring convergence of training and testing performance.

The robustness of the Lasso-regularized PMLR is further illustrated in Figure 7, which presents observed versus predicted dosing for the degree-10 model. Unlike the unregularized degree-10 model, which exhibited unstable divergence, the Lasso-regularized degree-10 model achieved highly stable predictions (R^2 = 0.951, RMSE = 0.120, MAPE = 7.02%). This confirms that even with thousands of predictors, Lasso can accommodate complex nonlinearities while maintaining generalization. For DWTP operation, this stability translates into reliable and interpretable dosing recommendations across diverse water quality conditions, including atypical turbidity events.

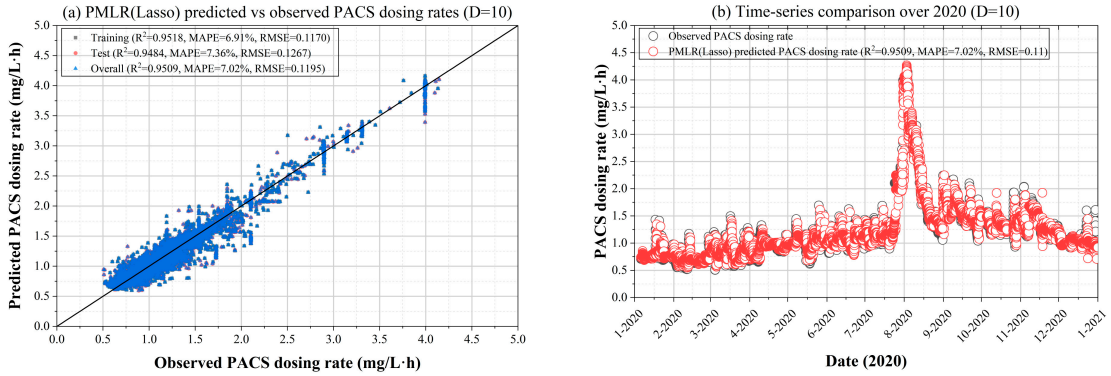


Figure 7. Lasso-regularized PMLR at degree 10. Parity plot of observed vs. predicted PACS dosing rates. Despite the high polynomial degree, Lasso maintained robust prediction accuracy (R^2 = 0.951; MAPE = 7.02%; RMSE = 0.120) and eliminated divergence observed in unregularized models. The 1:1 reference line (y = x) is included to indicate ideal agreement between observed and predicted values.

Although both Ridge and Lasso were examined, Elastic Net regularization—which blends L_1 and L_2 penalties—was not applied in this study. The six original influent variables exhibited limited multicollinearity, and the high-dimensional polynomial terms were sufficiently stabilized by the L_1 penalty alone. Therefore, Lasso provided the necessary sparsity and interpretability without the added complexity of Elastic Net. In comparison with ANN or tree-based ensemble models (e.g., GBM, XGBoost), which require post hoc techniques such as SHAP to interpret variable importance, the Lasso-regularized PMLR offers direct coefficient-based interpretability with minimal computational overhead—a key advantage for real-time SCADA integration.

Temporal dependence in DWTP operation was addressed through the previous-day dosing feature. More complex sequential models (e.g., ARIMA, LSTM) were not adopted because they provide limited interpretability and are not easily transferable to full-scale PLC–SCADA systems. Future work will extend the framework to include Elastic Net and ensemble models to quantitatively compare interpretability, numerical stability, and operational applicability across regularization strategies.

4. Conclusions

Accurate determination of coagulant dosing is essential for maintaining stable and efficient operation of drinking water treatment plants (DWTPs). This study developed a scientifically interpretable and operationally transferable regression framework for predicting coagulant demand using routinely monitored water-quality data. By combining polynomial multiple linear regression (PMLR) with Lasso regularization, the proposed approach provides both high predictive accuracy and coefficient-level transparency, enabling operators to understand, validate, and adjust dosing behavior without specialized machine learning expertise.

Among the evaluated models, the Lasso-regularized PMLR achieved the most balanced and robust performance ($R^2 \approx 0.95$, RMSE = 0.12, MAPE = 7.0%). Its ability to maintain stable generalization across increasing polynomial degrees demonstrates that the L_1 penalty effectively controls the severe multicollinearity generated by high-degree expansions—an important requirement for real-time deployment in DWTPs. The resulting sparse coefficient structure offers mechanistic interpretability, revealing how turbidity, pH, and conductivity jointly influence coagulant demand and linking data-driven predictions to well-established physicochemical coagulation mechanisms.

Importantly, the one-year operational dataset included high-turbidity periods (>30 NTU) and winter low-temperature conditions (<10 °C), conditions under which many ML models

struggle to generalize. The Lasso-regularized PMLR maintained stable performance across these challenging conditions, indicating that its predictive structure is resilient to seasonal and episodic fluctuations in raw-water quality.

From a practical standpoint, the proposed framework can be seamlessly integrated into existing PLC–SCADA architectures, providing transparent and real-time dosing recommendations without the tuning burden associated with complex black-box models. This transparency enhances operator trust and supports mechanism-aware decision-making, which is essential for sustainable and resilient DWTP operation.

Future research should validate the model across multiple treatment plants, coagulant types, and seasonal regimes to assess its transferability. Extending the framework to include Elastic Net or hybrid explainable ensemble approaches may further strengthen its robustness across diverse influent conditions. Additionally, incorporating treated-water quality indicators (e.g., settled turbidity, filtered turbidity, and residual aluminum) would enable direct assessment of how prediction errors translate into downstream water-quality outcomes, thereby enhancing the operational reliability and real-time applicability of the framework.

Author Contributions: J.A.: formal analysis, investigation, and writing—original draft; S.Y.: conceptualization, and writing—review and editing; C.O.: conceptualization, and writing—review and editing; B.L.: conceptualization and writing—review and editing; W.J.: conceptualization and writing—review and editing; J.Y.: investigation and writing—review and editing; H.J.O.: funding acquisition and writing—review and editing; J.P.: formal analysis, investigation, and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Korea Environment Industry & Technology Institute (KEITI) through Project for developing innovative drinking water and wastewater technologies Project, funded by Korea Ministry of Environment (MOE) (2020002700003) and the KICT Research Program (project no. 20250195-001, Development of low-energy/high-efficiency MCDI process and application to Malaysia) funded by the Ministry of Science and ICT.

Data Availability Statement: All data supporting the findings of this study are provided in this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Shi, Z.; Chow, C.W.K.; Fabris, R.; Liu, J.; Sawade, E.; Jin, B. Determination of coagulant dosages for process control using online UV-vis spectra of raw water. *J. Water Process Eng.* **2022**, *45*, 102526. [CrossRef]
2. Gagnon, C.; Grandjean, B.P.; Thibault, J. Modelling of coagulant dosage in a water treatment plant. *Artif. Intell. Eng.* **1997**, *11*, 401–404. [CrossRef]
3. Jayaweera, C.D.; Othman, M.R.; Aziz, N. Improved predictive capability of coagulation process by extreme learning machine with radial basis function. *J. Water Process Eng.* **2019**, *32*, 100977. [CrossRef]
4. Li, L.; Rong, S.; Wang, R.; Yu, S. Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. *Chem. Eng. J.* **2021**, *405*, 126673. [CrossRef]
5. Jang, J.-S.R. ANFIS: Adaptive-network-based fuzzy inference system. *IEEE Trans. Syst. Man Cybern.* **1993**, *23*, 665–685. [CrossRef]
6. Chiu, S.L. Fuzzy model identification based on cluster estimation. *J. Intell. Fuzzy Syst.* **1994**, *2*, 267–278. [CrossRef]
7. Chun, M.-G.; Kwak, K.-C.; Ryu, J.-W. Application of ANFIS for coagulant dosing process in a water purification plant. In Proceedings of the IEEE International Fuzzy Systems. Conference Proceedings (Cat. No.99CH36315), Seoul, Republic of Korea, 22–25 August 1999. [CrossRef]
8. Chen, C.-L.; Hou, P.-L. Fuzzy model identification and control system design for coagulation chemical dosing of potable water. *J. Water Supply Res. Technol.* **2006**, *6*, 97–104. [CrossRef]
9. Heddam, S.; Bermad, A.; Dechemi, N. ANFIS-based modeling for coagulant dosage in drinking water treatment plant: A case study. *Environ. Monit. Assess.* **2012**, *184*, 1953–1971. [CrossRef]
10. Baxter, C.W.; Stanley, S.J.; Zhang, Q. Development of a full-scale artificial neural network model for the removal of natural organic matter by enhanced coagulation. *J. Water Supply Res. Technol. AQUA* **1999**, *48*, 129–136. [CrossRef]

11. Joo, D.-S.; Choi, D.-J.; Park, H. Determination of optimal coagulant dosing rate using an artificial neural network. *J. Water Supply Res. Technol. AQUA* **2000**, *49*, 49–55. [CrossRef]
12. Deveughele, S.; Do-Quang, Z. Neural networks: An efficient approach to predict on-line the optimal coagulant dose. *Water Supply* **2004**, *4*, 87–94. [CrossRef]
13. Wu, G.-D.; Lo, S.-L. Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system. *Eng. Appl. Artif. Intell.* **2008**, *21*, 1189–1195. [CrossRef]
14. Griffiths, K.A.; Andrews, R.C. The application of artificial neural networks for the optimization of coagulant dosage. *Water Supply* **2011**, *11*, 605–611. [CrossRef]
15. Heddam, S.; Bermad, A.; Dechemi, N. Applications of radial-basis function and generalized regression neural networks for modeling of coagulant dosage in a drinking water treatment plant: Comparative study. *J. Environ. Eng. ASCE* **2011**, *137*, 1209–1214. [CrossRef]
16. Zangooei, H.; Delnavaz, M.; Asadollahfardi, G. Prediction of coagulation and flocculation processes using ANN models and fuzzy regression. *Water Sci. Technol.* **2016**, *74*, 1296–1311. [CrossRef] [PubMed]
17. Baxter, C.W.; Stanley, S.J.; Zhang, Q.; Smith, D.W. Developing artificial neural network models of water treatment processes: A guide for utilities. *J. Environ. Eng. Sci.* **2002**, *1*, 201–211. [CrossRef]
18. Kim, C.M.; Parnichkun, M. MLP, ANFIS, and GRNN based real-time coagulant dosage determination and accuracy comparison using full-scale data of a water treatment plant. *J. Water Supply Res. Technol. Aqua* **2017**, *66*, 49–61. [CrossRef]
19. Peixoto, J.L. A property of well-formulated polynomial regression models. *Am. Stat.* **1990**, *44*, 26–30. [CrossRef]
20. Kolluri, J.; Kotte, V.K.; Phridviraj, M.S.B.; Razia, S. Reducing overfitting problem in machine learning using novel l1/4 regularization method. In Proceedings of the 2020 4th International Conference on Trends in Electronics and Informatics, Tirunelveli, India, 15–17 June 2020. [CrossRef]
21. Joo, D.-S.; Choi, D.-J.; Park, H. The effects of data preprocessing in the determination of coagulant dosing rate. *Water Res.* **2000**, *34*, 3295–3302. [CrossRef]
22. Kim, J.; Hua, C.; Kim, K.; Lin, S.; Oh, G.; Park, M.-H.; Kang, S. Optimizing coagulant dosage using deep learning models with large-scale data. *Chemosphere* **2023**, *329*, 140989. [CrossRef]
23. Bachri, A.; Mahamad, A.K.; Utomo, W.M.; Saon, S.; Laksono, A.B. Optimization of coagulation process in water treatment plant using artificial intelligence: A systematic literature review and framework. *J. Robot. Control (JRC)* **2025**, *6*, 2471–2488. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.