ACS ES&T Water

Open Access

pubs.acs.org/estwater

Article

Modeling of Drinking Water Biofilm Thickness

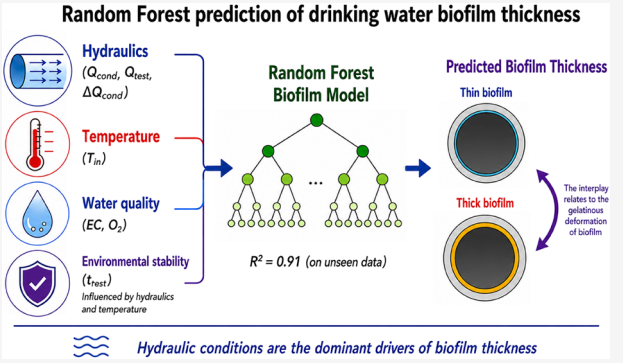
Konstantinos Glynis,* Mirjam Blokker, Zoran Kapelan, and Dragan Savić

Cite This: https://doi.org/10.1021/acsestwater.6c00651

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Biofilm development in drinking water distribution systems (DWDS) affects water quality, hydraulic performance, and microbial risk, yet its spatial distribution and structural properties remain poorly characterized. Existing assessment methods rely on microbiological or bulk water indicators that are difficult to interpret at the system scale and do not directly reflect biomass accumulation on pipe walls. This study presents the first model to predict the mean biofilm thickness in drinking water pipes using routinely measurable operational variables under controlled laboratory conditions. Biofilms were grown in a controlled pipe facility over ten months, and thickness was quantified using a hydraulic residence time method. A random forest model using seven variables describing hydraulic, thermal, and limited chemical conditions achieved high prediction accuracy (R^2 = 0.91 on unseen data) and identified flow rate, water temperature, and environmental stability as dominant factors. Feature importance and SHAP analyses highlighted the influence of conditioning shear stress and recovery time, while meta-analysis showed how operational conditions govern the formation of a stable biofilm base and a more easily removable outer layer. By linking operational conditions to biofilm thickness, this work provides a foundation for assessing and managing biofilm accumulation in drinking water systems subject to further validation.



KEYWORDS: *Drinking Water, Biofilm, Thickness, Gelatinous Deformation, Random Forest Model*

1. INTRODUCTION

Biofilms are complex microbial communities embedded in extracellular polymeric substances (EPSs) that develop on internal surfaces of drinking water distribution systems.[1] Following attachment, cells become enclosed within the EPS matrix, forming a stable microenvironment that supports persistence under varying hydraulic and physicochemical conditions.[2] This structure reduces exposure to disinfectants and environmental stresses, enabling survival even when bulk water conditions are unfavorable.[3] Biofilms influence microbial ecology in distribution systems but may compromise water quality by harboring opportunistic pathogens, accelerating disinfectant decay, and releasing biomass and corrosion products,[4] leading to operational challenges, aesthetic changes, and potential public health risks.[5] Despite their importance, biofilm spatial distribution and thickness remain poorly characterized,[6,7] limiting the identification of conditions promoting elevated accumulation.

Biofilm development is governed by interacting environmental factors. Nutrient availability influences microbial growth, while hydraulic conditions regulate attachment, stability, and detachment.[8,9] Variations in flow velocity and shear stress produce spatially distinct biofilm morphologies, affecting the thickness and detachment susceptibility.[10] Temperature further modulates growth rates and community composition, with seasonal and operational variations influencing biomass and opportunistic pathogens.[11,12] As these drivers vary across networks,[13] biofilm thickness is expected to vary spatially, particularly across zones with contrasting hydraulic conditions.[14]

Biofilm development has been assessed using biochemical and microbiological metrics such as ATP, cell counts, and 16S rRNA profiling.[15−17] While informative, these methods are costly, require specialized analysis, and produce complex data that are difficult to integrate at a network scale.[18] Physical indicators, including biofilm mass, surface roughness, and thickness, offer a complementary approach. Thickness relates directly to biovolume,[19,20] correlates with maturity and metabolic potential,[21] and influences hydraulic behavior and wall friction,[9,22] making it an interpretable descriptor of biofilm accumulation.

However, direct thickness measurements remain challenging in full-scale drinking water distribution systems[23] and are largely restricted to specialized experimental setups. Furthermore, the

Received: May 4, 2026
Revised: July 15, 2026
Accepted: July 15, 2026

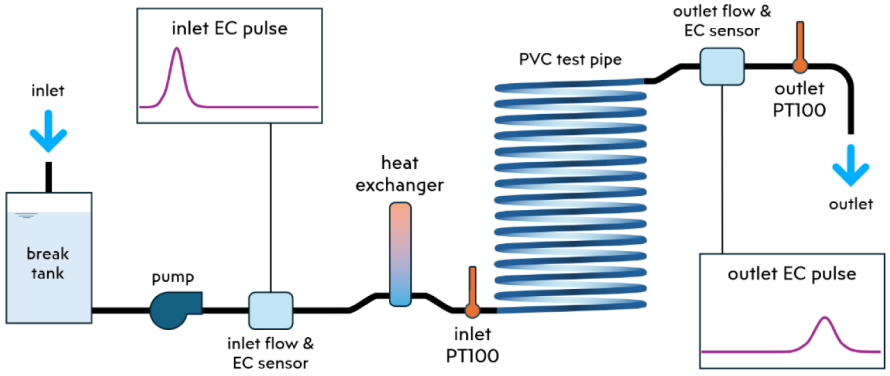


Figure 1. Schematic overview of the "Slimer" experimental setup used for drinking water biofilm growth under controlled conditions and biofilm thickness measurements.

measured thickness may depend on the hydraulic conditions under which the measurement is performed, reflecting the shear-dependent structure of the biofilm. These practical and methodological challenges limit the routine assessment of biofilm accumulation in operational systems. Consequently, predictive models are needed to infer biofilm accumulation from routinely available operational data.[24] Such models could identify locations prone to elevated biomass and microbial risk, complementing sparse field data and informing operational decision-making, including network adaptation strategies designed to improve water quality and enhance system resilience.[9,25]

Despite this potential, biofilm thickness in drinking water pipes has not yet been explicitly predicted by using such models. Existing approaches focus on microbiological proxies or bulk water variables,[26] which are difficult to quantify reliably[27] and do not directly capture the physical biofilm accumulation. Additionally, the influence of hydraulic and temperature variability on biofilm structure is rarely incorporated,[11,28] limiting the ability to predict where thicker, more mature biofilms may develop.[18]

This study addresses the lack of models explicitly quantifying biofilm thickness and its dependence on hydraulic and temperature variability. The objective of this work was primarily to isolate the influence of hydraulic and thermal conditions on the biofilm thickness under otherwise biologically stable conditions. The study is based on an 11-month laboratory experiment comprising 219 biofilm thickness measurements under systematically varied hydraulic conditions, enabling the isolation of effects not resolvable in full-scale networks. Using these data, we developed a novel data-driven model to predict biofilm thickness at the pipe scale under controlled laboratory conditions, providing a foundation for future validation in operational drinking water systems. By linking thickness to hydraulic and temperature conditions, the model enables inference of spatial variation in biofilm accumulation and evaluation of how network design or operational strategies may influence biofilm development and the associated risks.

The remainder of this paper is organized as follows. Section 2 describes the experimental setup and modeling framework, including the random forest formulation and validation. Section 3 presents results on biofilm thickness patterns, model performance, feature importance, and inferred structural behavior. Section 4 discusses findings, limitations, and implications for network management, and Section 5 summarizes conclusions and future research directions.

2. METHODOLOGY

2.1. Experimental Setup

A controlled laboratory experiment was conducted to generate the data required for the development of the biofilm thickness prediction model. The experimental design enabled systematic variation of hydraulic and environmental conditions while maintaining otherwise stable operating conditions, thereby facilitating investigation of the factors influencing biofilm thickness. To achieve this, experiments were performed using the laboratory-scale "Slimer" facility at the KWR Water Research Institute, specifically designed to reproduce drinking water pipe conditions under controlled laboratory settings.[29]

The "Slimer" facility comprises a 50 m transparent plasticized polyvinyl chloride (PVC-p) pipe (inner diameter 13.2 mm), selected for its high plasticizer content, which promotes biofilm growth.[30] As shown in Figure 1, the pipe was arranged helically to minimize the footprint and maintain homogeneous ambient conditions. To prevent the presence of preexisting biofilm that could bias the results, a new pipe was used and had no prior contact with water before the start of the experiment. No additional physical or chemical pretreatment was applied prior to commissioning. To minimize the potential for phototrophic growth, the experimental setup was operated in a dark room throughout the campaign. Flow rate and electrical conductivity (EC) were monitored at both pipe ends using magnetic induction sensors, while inlet, outlet, and ambient temperatures were recorded with standard PT100 resistance temperature detectors.[31] The inlet PT100 sensor was positioned immediately upstream of the test section and downstream of the heat exchanger; consequently, the measured inlet temperature represents the actual water temperature entering the pipe and was used as a model input parameter (Section 2.2). The heat exchanger was used to establish the desired inlet water temperature prior to entry into the test section. Steady-state flow was maintained via automated pump control.

Biofilm was grown from December 2024 to October 2025 under conditioning flow rates of 50−400 L h^{-1} (Reynolds number, Re ≈ 1,300−10,600) using residual disinfectant-free tap water supplied by the local water utility. Ambient temperature was regulated, and inlet water temperature varied between 12.2 and 24.0 °C. Biofilm thickness measurements were performed regularly at testing flow rates (Q_{test}) not exceeding the conditioning rate (Q_{cond}) to avoid shear-induced detachment of biofilm. Multiple biofilm thickness estimates may therefore be associated with the same date, as measurements were often performed sequentially at different testing flow rates,

https://doi.org/10.1021/acsestwater.6c00651
ACS EST Water XXXX, XXX, XXX−XXX

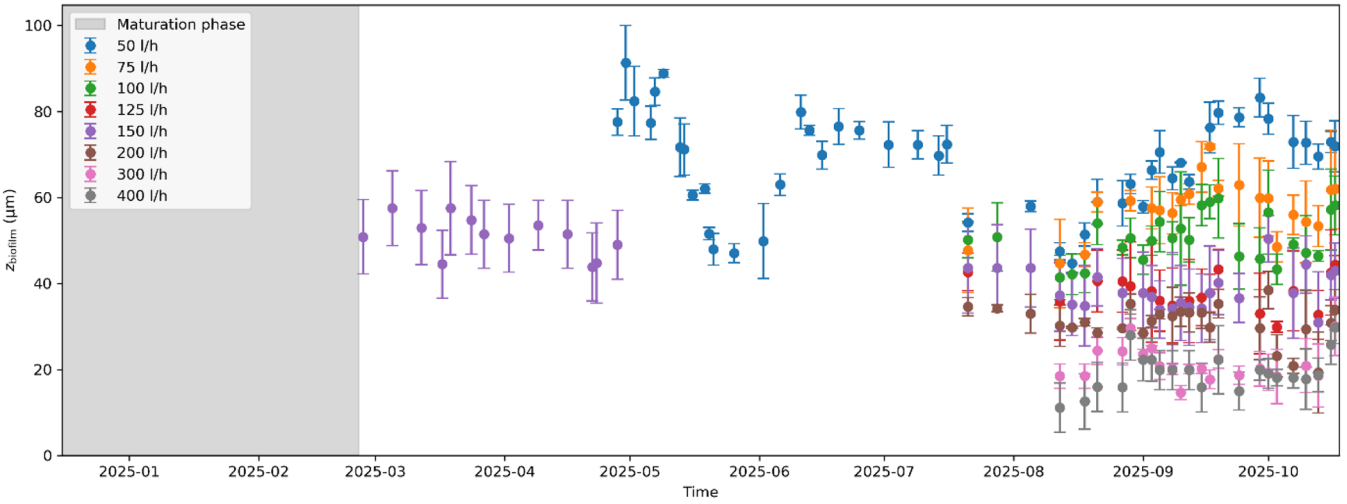


Figure 2. Time series of biofilm thickness ($z_{biofilm}$) calculated throughout the experimental campaign for all testing flow rates (Q_{test}). Colors represent different testing flow rate (Q_{test}) groups. Data points represent mean biofilm thickness values, and error bars indicate the per-pulse variability among the different thickness estimates obtained with the HRT method.

each yielding an independent thickness estimate under the corresponding hydraulic conditions.

Mean biofilm thickness $z_{biofilm}$ was quantified using the hydraulic residence time (HRT) method.[23] A low-salinity NaCl pulse was injected, producing transient EC increases detected upstream and downstream of the pipe. The underlying principle of this method is that biofilm accumulation reduces the effective cross-sectional area of the pipe, thereby increasing the hydraulic residence time relative to that of a clean-pipe reference condition. For each pulse, characteristic times (25th, 50th, and 75th percentiles and EC-weighted mean) were extracted, and residence times were computed as the difference between corresponding inlet and outlet signals. The mean biofilm thickness per pulse was inferred from deviations of the measured residence times relative to a clean-pipe reference, following a calibration procedure.[23] The mean of these estimates was reported as the biofilm thickness for each pulse, while their variability was retained as a measure of the uncertainty of the HRT-based method and is shown as error bars in Figure 2. The HRT-based biofilm thickness method and experimental setup were previously evaluated against destructive biovolume measurements and complementary ATP analyses. While destructive biovolume measurements provided direct physical evidence of biomass accumulation, ATP measurements served as an independent biological indicator, supporting the presence and development of biofilm within the system.[23]

Electrical conductivity (EC) was characterized at two scales: (i) background (bulk) EC values provided by the water utility at the outlet of the water treatment plant, representing long-term influent water quality conditions, and (ii) transient EC signals measured locally at the inlet and outlet of the experimental pipe during NaCl pulse injections for hydraulic residence time determination. Water temperature was logged continuously at the setup, while bulk EC (weekly) and O_2 (weekly) were obtained from the supplying water utility and are representative of the influent water entering the "Slimer" experimental system. No additional disinfection was applied, and the supplied water was biologically stable with low organic carbon content. An overview of the basic water quality parameters is provided in Appendix A.

Representative optical images of the pipe wall before and after biofilm formation are provided in Appendix B, illustrating the visible accumulation of biomass at the same locations where HRT measurements indicated an increased biofilm thickness.

2.2. Experimental Data Processing

Experimental data were processed to derive the model inputs and outputs. Biofilm thickness $z_{biofilm}$ served as the target variable to be predicted. Input variables comprised the testing flow rate (Q_{test}), conditioning flow rate (Q_{cond}), change in conditioning flow rate (ΔQ_{cond}), inlet water temperature (T_{in}), electrical conductivity (EC), dissolved oxygen concentration (O_2), and time since the most recent abrupt change in hydraulic or thermal conditions (t_{rest}).

Among the model inputs, ΔQ_{cond} and t_{rest} are engineered features. ΔQ_{cond} represents the step change in the conditioning flow rate and serves as a proxy for abrupt changes in the maximum shear stress experienced by the biofilm. In contrast, t_{rest} quantifies the time elapsed since the most recent disturbance to the system, whether hydraulic (captured through ΔQ_{cond}) or thermal. Together, these variables characterize the stability or instability of the conditions governing biofilm adaptation and regrowth. Bulk EC and O_2 values obtained from the water utility did not always coincide with experimental timestamps; therefore, linear interpolation was used to align these variables with the biofilm measurement series.

A summary of all input variables, including their ranges and the number of data points available, is provided in Table 1.

2.3. Model Inputs and Rationale

The data-driven model developed in this study predicts mean, i.e., pipe-average, biofilm thickness as a function of seven variables: testing flow rate (Q_{test}), conditioning flow rate (Q_{cond}), change in conditioning flow rate (ΔQ_{cond}), inlet water temperature (T_{in}), electrical conductivity (EC), dissolved oxygen concentration (O_2), and time since the most recent abrupt change in hydraulic or thermal conditions (t_{rest}). The selection of model input variables was guided by two criteria: (i) mechanistic relevance to biofilm accumulation and (ii) feasibility of measurement or estimation in real-life systems.

These variables represent the combined influence of hydraulic, thermal, chemical, and stability-related drivers of biofilm development. Hydraulic conditions were characterized by the testing flow rate, which determines the shear stress acting on the biofilm during thickness measurements, and the conditioning flow rate, reflecting the long-term shear history

https://doi.org/10.1021/acsestwater.6c00651
ACS EST Water XXXX, XXX, XXX−XXX

Table 1. Summary of Model Input Variables, Including Their Description, Observed Range, Units, and Number of Data Points

Symbol	Description	Range	Units	Data points
Q_{test}	Testing flow rate	[50, 400]	L h^{-1}	219
Q_{cond}	Conditioning flow rate	[50, 400]	L h^{-1}	219
ΔQ_{cond}	Change in conditioning flow rate	[−100, 200]	L h^{-1}	219
T_{in}	Inlet water temperature	[12.2, 24.0]	°C	219
EC	Bulk electrical conductivity (utility data)	[230, 260]	μS/cm	46
O_2	Dissolved oxygen concentration (utility data)	[9.9, 11.6]	mg/L	46
t_{rest}	Time since the most recent abrupt change in hydraulic or thermal conditions	[0, 79]	days	219

under which the biofilm developed. Sudden hydraulic changes were captured using an engineered variable quantifying step changes in the conditioning flow rate, representing abrupt shifts in maximum shear stress. Thermal influences were incorporated via inlet water temperature, which is a key regulator of microbial activity and biofilm growth.

Chemical conditions were described using electrical conductivity (EC), readily measured using low-cost sensors,[32] and dissolved oxygen concentration (O_2), which is an important regulator of aerobic microbial activity and biofilm development in drinking water systems.[33,34]

A second engineered variable quantified the time since the most recent hydraulic or thermal disturbance, defined as a change greater than 15% in the conditioning flow rate or inlet water temperature relative to the preceding measurement. This threshold was selected to represent meaningful operational shifts while excluding minor fluctuations or measurement noise; an exploratory analysis indicated limited sensitivity to its exact value. This metric reflects environmental stability and can be estimated in operational systems by using SCADA data or hydraulic model outputs.[35]

The above seven variables capture both long-term operating conditions and short-term perturbations, providing a mechanistically grounded basis for the data-driven prediction of biofilm thickness.

2.4. Random Forest Model

A random forest (RF) regression model[36] was selected to predict biofilm thickness from the measured hydraulic, thermal, and chemical quality variables. This approach was chosen because it handles nonlinear relationships and interactions effectively,[37] which is essential given the complex influence of flow rate, temperature, and water chemical composition on biofilm growth.[9] It is also robust to noise and multicollinearity[38] and performs well with moderate-sized datasets[39] typical of controlled laboratory studies. Furthermore, RFs provide straightforward variable-importance metrics and are compatible with the SHapley Additive exPlanation (SHAP) analysis framework,[40] allowing transparent interpretation of predictor influence.[41]

Finally, unlike parametric models that require explicit assumptions regarding the functional form of biofilm growth, random forests support a flexible, data-driven representation of the system, making them particularly suitable in cases, such as biofilm thickness, where no established mechanistic models currently exist.[42] For these reasons, the RF regression framework was selected as a robust and interpretable basis for the development of our biofilm thickness predictive model.

The RF model was trained and evaluated by using a structured procedure designed to prevent information leakage, enable robust hyperparameter selection, and quantify uncertainty arising from data partitioning. All analyses were performed in Python 3.11 using Scikit-learn.[43]

To ensure that the independent test set was fully isolated from model development, 20% of observations ($n = 44$) were selected a priori using an equally spaced sampling strategy across the full time series. This approach is justified because each biofilm thickness measurement represents a quasi-steady-state condition established under fixed hydraulic and thermal regimes; therefore, temporal ordering does not contain additional dynamic information, and successive observations are not temporally autocorrelated. Selecting test samples uniformly throughout the experiment ensured broad coverage of hydraulic and thermal conditions while avoiding temporal clustering, thereby providing a representative and unbiased assessment of model generalizability. The remaining 80% of observations ($n = 175$) constituted the training−validation pool used for all model-development activities, including hyperparameter optimization and model training. The independent test subset was excluded from these activities and reserved exclusively for the final assessment of model performance.

Hyperparameter optimization was performed using repeated single-split evaluations applied only to this training−validation pool. In each of 20 tuning iterations, the pool was randomly divided into 60% training and 20% validation subsets, corresponding to an overall 60/20/20 structure relative to the full dataset. This partitioning was selected as a compromise between providing sufficient data for model training while retaining adequately sized validation and independent test subsets for hyperparameter optimization and unbiased performance assessment. A grid search was conducted over combinations of tree number, tree depth, minimum samples per split, minimum samples per leaf, and the number of features considered at each split. Each configuration was scored using a harmonic-mean criterion combining shifted training and validation R^2 values (the coefficient of determination), favoring models with balanced predictive performance. The best configuration from each iteration was recorded, and the final hyperparameter set was selected as the statistical mode across the 20 tuning runs.

Using these globally selected hyperparameters, model performance was assessed across 100 independent training−validation splits. In each run, the 80% training−validation pool was randomly divided again, the model was fitted, and predictions were generated for the training, validation, and, most importantly, the fixed, untouched test set. This procedure quantified the variability in predictive performance attributable solely to data partitioning, ensuring a robust characterization of model uncertainty.

2.5. Model Performance Evaluation

Model performance was assessed across the 100 evaluation runs by comparing the predicted and observed biofilm thickness values for the training, validation, and test subsets. As in the hyperparameter-tuning stage (Section 2.4), the coefficient of determination (R^2) was used as a key performance indicator of predictive skill, quantifying the proportion of variance in measured biofilm thickness explained by the model.

https://doi.org/10.1021/acsestwater.6c00651
ACS EST Water XXXX, XXX, XXX−XXX

Table 2. Chronological Overview of the Hydraulic Conditioning Program and Biofilm Thickness Measurements

Period	Dates	Duration (days)	Q_{cond} (L h^{-1})	Q_{test} (L h^{-1})	Purpose
Initial maturation	Dec 2024−Feb 2025	71	200	—	Initial biofilm growth and maturation
Hydraulic conditioning	Feb 2025−Apr 2025	61	150	150	Intermediate-shear conditioning
	Apr 2025−Jul 2025	84	50	50	Long-term low shear growth
	Jul 2025−Aug 2025	22	200	[50, 200]	Intermediate-shear conditioning
	Aug 2025−Oct 2025	66	400	[50, 400]	Biofilm adaptation to increased shear

To complement R^2, two error-based metrics were included: the root mean squared error (RMSE), which expresses absolute prediction error in μm, and the normalized RMSE (RMSE normalized by the mean biofilm thickness of the corresponding subset), which quantifies prediction errors relative to typical biofilm levels. Together, these indicators capture explanatory power, absolute accuracy, and proportional accuracy, providing a comprehensive assessment of model performance.

For each run, all three metrics were computed for the training, validation, and test subsets. Final performance was summarized as the mean ± standard deviation, also shown as boxplots across the 100 runs, providing an uncertainty-aware assessment of accuracy and robustness.

2.6. Feature Importance Analysis

Predictor influence on the RF model was assessed using global feature importance scores and SHapley Additive exPlanations (SHAP) values.[40] Prior to interpretation, potential multicollinearity among the input features was evaluated using pairwise correlation analysis (Appendix D). This analysis revealed moderate to strong correlations among some predictors, particularly among hydraulic descriptors, reflecting physical and operational dependencies in the experimental design.

Impurity-based feature importance from scikit-learn provides an initial ranking of variables by quantifying their average contribution to variance reduction across the ensemble. Although efficient, these scores reflect only global effects and are less reliable when the predictors interact or are correlated.

To obtain a more interpretable and locally faithful assessment, SHAP values were computed using the TreeExplainer algorithm applied to the independent test set.[44,45] SHAP summary plots reveal both the magnitude and direction of each feature's effect on predicted biofilm thickness, capturing nonlinear relationships while providing more robust insights into individual predictor contributions under multicollinearity.

Using these two complementary methods allowed us to characterize model behavior transparently: impurity-based importances offered a quick global ranking of predictors, while SHAP analysis provided a more detailed explanation of how different individual inputs influenced the biofilm thickness prediction.

2.7. Derivation of Structural Biofilm Descriptors

Beyond mean thickness, biofilms exhibit a shear-dependent mechanical structure, meaning that their thickness varies with the applied shear stress as protrusions compress or detach and the near-wall boundary layer changes.[46] Because no universally accepted definition of biofilm thickness exists, and measured thickness depends inherently on the hydraulic conditions under which it is assessed, we introduce two structural descriptors that explicitly characterize how the apparent thickness responds to shear. In this study, the term *gelatinous deformation* is used to describe the shear-dependent reduction in apparent biofilm thickness resulting from compression, deformation, and even detachment of loosely attached biofilm structures when exposed to increasing hydraulic shear stress.

To quantify this response, we first defined a parametric model describing how the predicted biofilm thickness changes with the testing flow rate Q_{test}. For each combination of environmental and hydraulic conditions, the RF model was evaluated over a sweep of testing flow rates spanning the experimental range (50−400 L h^{-1}), yielding a predicted thickness curve $z_{biofilm}(Q_{test})$.

An exponential decay function was used to describe the relationship between predicted biofilm thickness and testing flow rate, capturing the expected shear-induced thinning: a rapid initial decrease followed by an asymptotic approach to a shear-resistant core. This provides a parsimonious representation of both thinning behavior and residual thickness.

The function was fitted as

$$z_{biofilm}(Q_{test}) = A_{residual}(1 + e^{-k_{thinning}Q_{test}}) \quad (1)$$

In eq 1, $A_{residual}$ represents the asymptotic thickness as $Q_{test} \to \infty$, corresponding to the portion of the biofilm that remains under very high shear (i.e., the shear-resistant core). The parameter $k_{thinning}$ defines the rate of thickness reduction with increasing Q_{test}, with larger values indicating greater sensitivity to shear and values approaching zero indicating a more compact, shear-resistant structure. These parameters, $A_{residual}$ and $k_{thinning}$, constitute the structural biofilm descriptors used in the subsequent analysis.

To generate a sufficiently broad representation of environmental−operational conditions, we first defined a multidimensional feature space using all model predictors except the testing flow rate Q_{test}. For each feature, the minimum and maximum observed values were used to establish sampling bounds and a corresponding uniform distribution. A total of 100,000 synthetic environmental−operational conditions were then generated by drawing random combinations of all non-Q_{test} predictors, i.e., Q_{cond}, ΔQ_{cond}, T_{in}, EC, t_{rest}, O_2.

For each synthetic sample, the random forest model was used to generate $z_{biofilm}(Q_{test})$ across the full sweep of testing flow rates. The exponential decay model was fitted using nonlinear least squares; samples for which the fit did not converge were discarded. This procedure yielded a meta-dataset containing, for each environmental state, the corresponding values of $A_{residual}$, $k_{thinning}$, and all non-Q_{test} predictors.

Two independent random forest regressors were then trained to predict $A_{residual}$ and $k_{thinning}$ from the environmental predictors. This meta-modeling approach enabled us to investigate how hydraulic history, disturbance time scales, temperature, and nutrient availability influence the shear-dependent mechanical structure of the biofilm, as implied by the original thickness model.

The resulting $A_{residual}$ and $k_{thinning}$ parameters were examined in Section 3.4 to characterize how the biofilm shear-dependent mechanical structure behavior varies across the hydraulic and environmental feature space.

https://doi.org/10.1021/acsestwater.6c00651
ACS EST Water XXXX, XXX, XXX−XXX

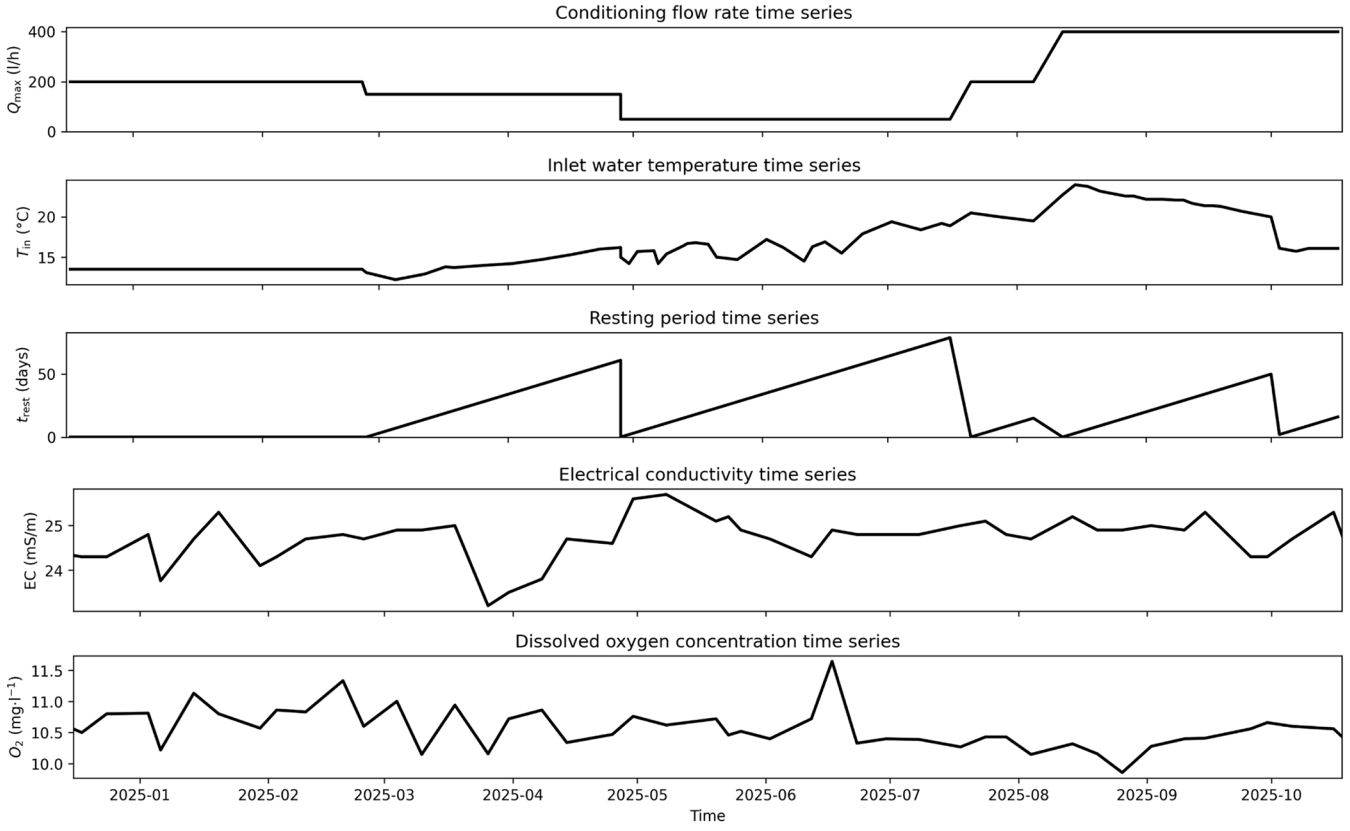


Figure 3. Time series of conditioning flow rate (Q_{cond}), inlet water temperature (T_{in}), and resting period (t_{rest}) measured in the setup, and bulk electrical conductivity (EC) and dissolved oxygen concentration (O_2) measured at the production location in the influent drinking water during the experimental campaign.

3. RESULTS

3.1. Descriptive Analysis of Experimental Data

Biofilm thickness exhibited substantial temporal and hydraulic variability over the 10-month experimental campaign (see Figure 2). The chronology of the hydraulic conditioning program and associated biofilm thickness measurements is summarized in Table 2. During the initial maturation period (Dec 2024−Feb 2025), biofilm thickness stabilized around 50−60 μm. Once regular testing began, thickness values ranged from approximately 10 to 100 μm, with clear dependence on the testing flow rate and the preceding hydraulic history.

The time series data shown in Figure 2 illustrate how biofilm thickness evolved in response to the successive conditioning regimes described in Table 2. Periods of prolonged stable conditioning flow rates, particularly during the low-shear conditioning phase between April and July 2025, corresponded to gradual increases in the biofilm thickness. In contrast, transitions to higher conditioning flow rates were generally associated with reductions in measured thickness, indicating the influence of hydraulic shear on the biofilm accumulation and structure. The denser clustering of measurements during the final months of the experiment reflects the dedicated model-development phase, during which biofilm thickness was assessed under a broader range of testing flow rates and hydraulic conditions to expand the calibration dataset.

Summary statistics for the full dataset (Appendix C) reveal a systematic decrease in calculated mean biofilm thickness with increasing testing flow rate. The thickest biofilms were present at 50 L h^{-1} (mean 68.2 μm, maximum 91.4 μm), while the thinnest biofilms occurred at 400 L h^{-1} (mean 19.4 μm). This monotonic decline reflects the sensitivity of the HRT-based measurement to hydraulic conditions, with lower testing flow rates associated with thicker biofilms, consistent with previous observations that biomass accumulation increases at lower flow velocities.[47]

Median values follow the same trend, confirming robust differences across flow conditions (e.g., 70.9 μm at 50 L h^{-1} vs 19.1 μm at 400 L h^{-1}). The spread of values (min−max range) also narrows with increasing flow rate, suggesting reduced structural heterogeneity at higher shear stresses.

The environmental and operational time series shown in Figure 3 provide further context for the observed biofilm thickness evolution. Periods with prolonged stable conditioning flow rates (e.g., April−June) correspond to gradual increases in thickness, whereas abrupt changes in the conditioning flow rate (visible as step transitions in Q_{cond}) coincide with marked reductions in thickness or resets in structure, as reflected by the resting-time variable t_{rest}.

Temperature increased steadily from early spring to midsummer (13−22 °C), producing parallel increases in biofilm thickness within corresponding steady-state periods, consistent with known temperature−growth relationships. Later in the campaign, when temperature declined and dissolved oxygen increased, measured thicknesses exhibited a downward shift, particularly at higher testing flow rates

3.2. Model Performance

The RF model demonstrated strong predictive performance across the 100 evaluation runs, with consistently high accuracy on both validation and independent test subsets. Across all runs, the mean test-set coefficient of determination was R^2 = 0.905 ±

https://doi.org/10.1021/acsestwater.6c00651
ACS EST Water XXXX, XXX, XXX−XXX

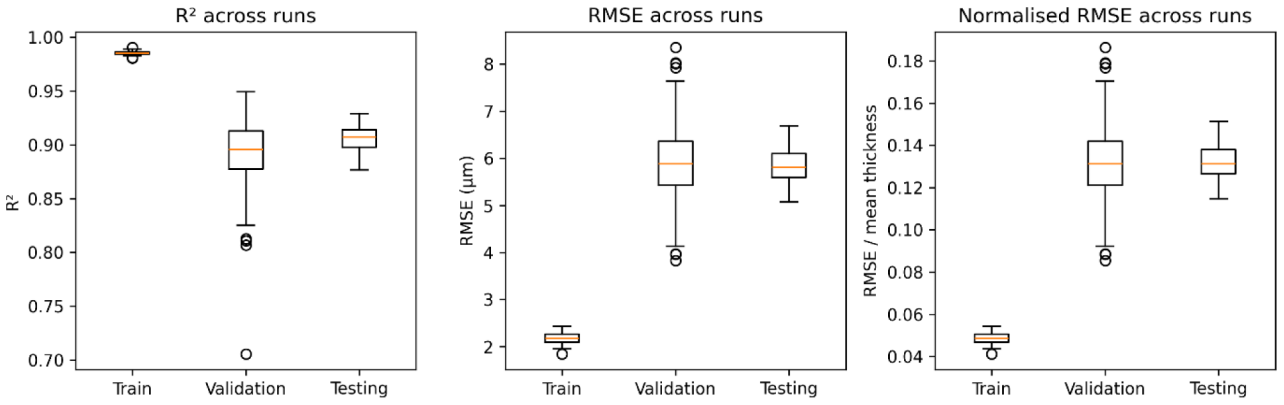


Figure 4. Performance of the final RF configuration across 100 evaluation runs. Boxplots show the distribution of R^2 (left), absolute RMSE (middle), and normalized RMSE (right) for the training, validation, and test subsets.

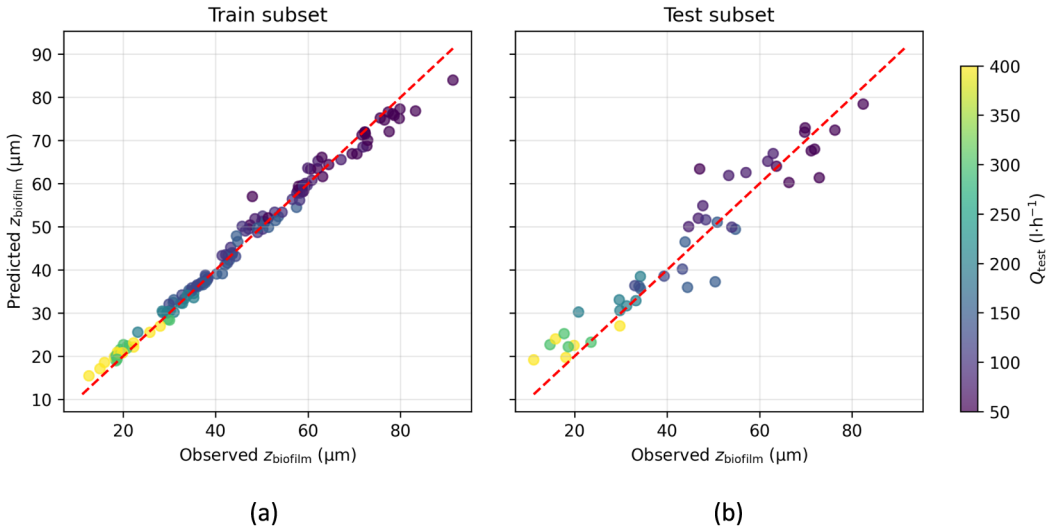


Figure 5. Predicted versus observed biofilm thickness for the training (a) and test (b) subsets. Colors denote the testing flow rate Q_{test}. The dashed line represents the 1:1 correspondence.

0.012, indicating that the model explained approximately 90% of the variance in measured biofilm thickness. Test-set prediction errors were low relative to the absolute magnitude of the measurements, with an average RMSE of 5.863 ± 0.355 μm, corresponding to a normalized RMSE of 0.133 ± 0.008. These values are consistent with what is typically regarded as good performance in environmental modeling with biological variability.

Performance distributions across runs are summarized in Figure 4 (boxplots). Training-set metrics were very high ($R^2 \approx$ 0.985; normalized RMSE ≈ 0.049), which is expected for ensemble tree models fitted on moderate sample sizes, while validation and test subsets showed closely aligned error distributions, indicating that the model generalized reliably without signs of overfitting. The slightly larger variability observed in validation performance (R^2 = 0.892 ± 0.035) reflects sensitivity to random train/validation splits within the reduced 80% training–validation pool. Test-set performance, by contrast, remained stable because the same fixed set of test instances was used for all 100 runs.

Predicted versus observed biofilm thickness is shown in Figure 5 for both the training and test subsets. In the training subset, predicted values align closely with the 1:1 line, consistent with the low training RMSE. In the test subset, the model reproduces the overall magnitude and spread of measured thickness values across the full range (≈15–90 μm), with no systematic bias toward over- or under-prediction. The color gradient indicates that prediction quality is comparable across flow regimes, suggesting that the model captures hydraulic influences on biofilm structure without being dominated by a single operating condition.

The test-set R^2 of 0.905 ± 0.012 and normalized RMSE of 0.133 ± 0.008 confirm that the model provides accurate and reliable predictions of biofilm thickness under unseen operational conditions. Overall, these results confirm that the model provides accurate and robust predictions of biofilm thickness under diverse hydraulic and thermal conditions, with stable performance across repeated data partitions and strong agreement with observed measurements.

3.3. Importance of Model Inputs

Prior to interpreting predictor importance, potential multicollinearity among the input features was assessed using pairwise correlation analysis (Appendix D). This analysis revealed moderate to strong correlations among some predictors, particularly among hydraulic descriptors (Q_{cond} and ΔQ_{cond}) and between hydraulic and water quality variables, reflecting physical and operational dependencies in the experimental design. Consequently, impurity-based feature importance values

https://doi.org/10.1021/acsestwater.6c00651
ACS EST Water XXXX, XXX, XXX–XXX

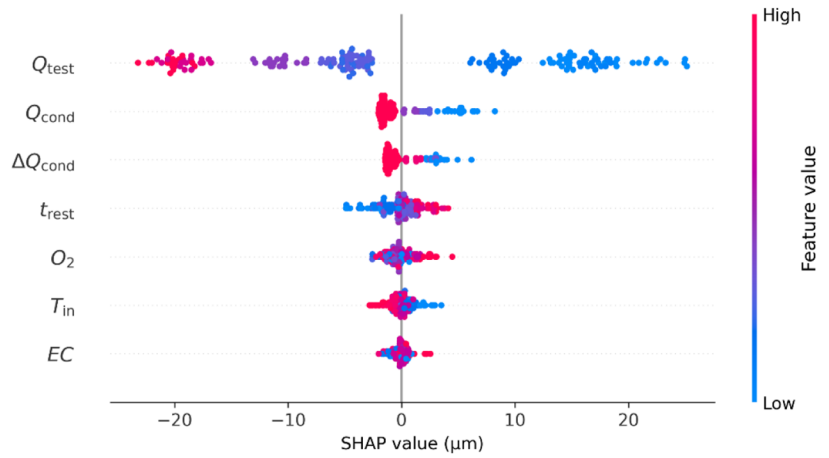


Figure 6. SHAP summary plot illustrating the magnitude and direction of each input feature's influence on predicted biofilm thickness across the test set.

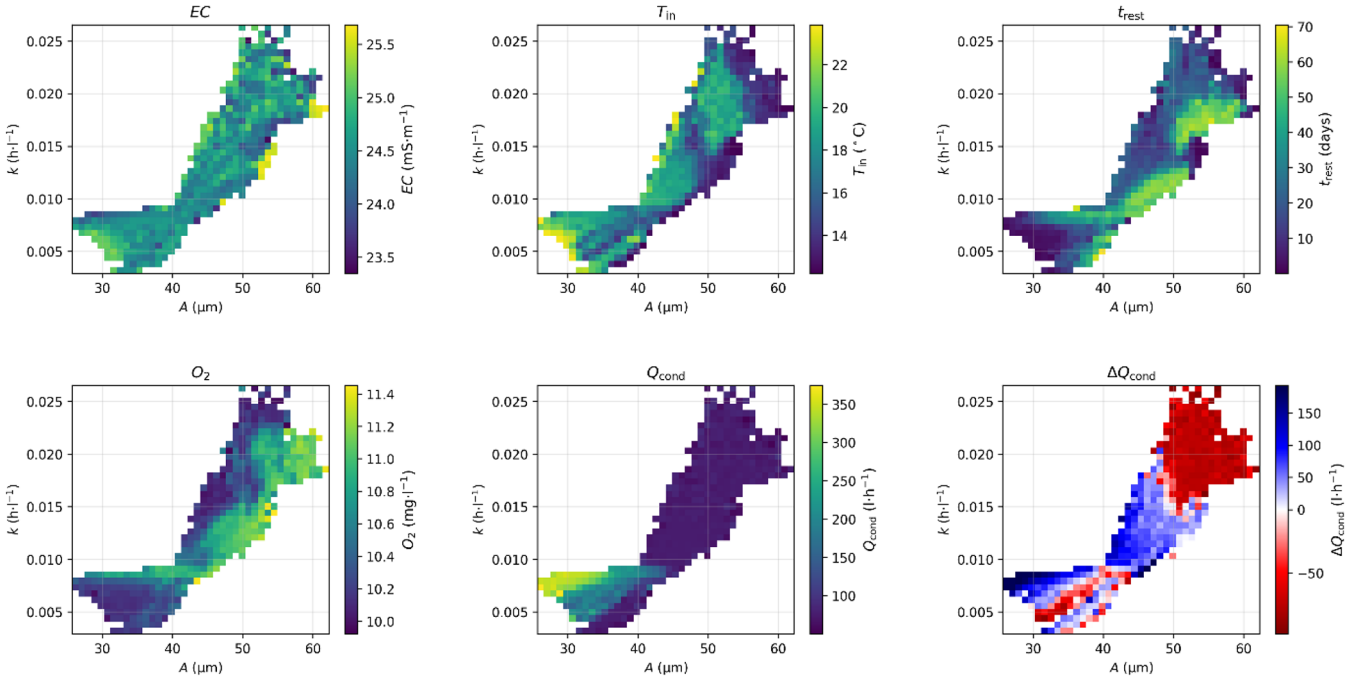


Figure 7. $A-k$ behavioral maps showing how residual thickness (A) and thinning rate (k) predicted by the meta-model vary across the sampled hydraulic, thermal, and chemical feature space. Colors indicate the mean value of each predictor within the (A, k) bins, highlighting how electrical conductivity, temperature, resting time, dissolved oxygen concentration, and hydraulic history shape the structural robustness and shear-response of the biofilm.

are interpreted with caution and primarily as an indication of overall predictor relevance rather than as definitive measures of causal influence.

The RF model revealed a clear hierarchy in feature importance. Q_{test} was the dominant predictor of biofilm thickness, accounting for 65% of the total importance, likely reflecting its direct mechanistic link to shear stress during measurement and its influence on biofilm compression or mobilization during HRT-based thickness estimation. A second tier of predictors included Q_{cond}, the engineered ΔQ_{cond}, O_2, and t_{rest}, each contributing 6−10%, while T_{in} and bulk EC contributed marginally (<5%).

The SHAP summary plot (Figure 6) provides deeper insight into how each predictor shaped individual predictions. For Q_{test}, high values (red) consistently produced strongly negative SHAP values, indicating a reduction in predicted thickness as the testing flow rate increased. Conversely, low testing flow rates (blue) yielded positive SHAP values, consistent with thicker biofilms measured under gentle hydraulic loading. This directional pattern aligns with established links between hydrodynamic shear and biofilm structure and further confirms that the HRT-based measurement responds systematically to the imposed test conditions.

Q_{cond} also showed a coherent directional influence: high conditioning flow rates produced negative SHAP values, indicating that biofilms subjected to sustained high shear were predicted to be thinner, while lower conditioning flow rates favored thicker biofilm accumulation. The ΔQ_{cond} variable displayed a more dispersed distribution, with both positive and negative contributions. Large positive changes, representing a

https://doi.org/10.1021/acsestwater.6c00651
ACS EST Water XXXX, XXX, XXX−XXX

sudden increase in the conditioning hydraulic stress, were associated with negative SHAP values, reflecting likely detachment or compaction events, whereas negative changes (reductions in shear) tended to favor positive SHAP contributions and thicker biofilms.

The stabilization variable t_{rest} showed a consistent positive influence on biofilm thickness, with longer undisturbed periods promoting larger predicted thicknesses. This suggests that the model captures the gradual recovery and accumulation of biofilm following hydraulic disturbances. The dissolved oxygen concentration (O_2) exhibited a weaker but generally positive effect, indicating that oxygen availability may also contribute to biofilm development within the experimental system.

Temperature (T_{in}) and EC contributed less overall, but their SHAP distributions revealed an asymmetric influence. Elevated temperatures occasionally produced negative SHAP values, consistent with an enhanced microbial metabolic rhythm during warmer periods, which, coupled with limited nutrient availability, led to partial biofilm decay. EC showed only weak, nondirectional effects. The effects of O_2 and EC should be interpreted cautiously, given the limited variability and sampling frequency of the measurements.

Together, these patterns demonstrate that the model not only identifies the dominant hydraulic controls on biofilm thickness but also captures subtler interactions among shear history, temperature conditions, and nutrient availability. By interpreting feature importance in conjunction with SHAP analysis and accounting for correlations among predictors (Appendix D), the results provide confidence that the model is learning physically meaningful relationships rather than relying on incidental statistical associations.

3.4. Model-Derived Structural Behavior of Biofilms

To investigate how environmental and hydraulic conditions shape biofilm structure beyond absolute thickness, we analyzed the relationship between the RF-derived residual thickness $A_{residual}$ and thinning rate $k_{thinning}$ across the globally sampled meta-dataset. For clarity of presentation, these parameters are denoted as A and k, respectively, in Figure 7. The behavioral maps of the $A-k$ analysis in Figure 7 reveal coherent patterns that reflect distinct structural responses of the biofilm to different drivers.

Across all subplots in Figure 7, two distinct behavioral regimes are apparent. The low-A/low-k region in the lower left reflects thin, compact biofilms with limited sensitivity to shear, whereas the high-A/high-k region in the upper right denotes thicker and more compliant structures that are more readily deformed by hydraulic forces.

EC shows no clear systematic trend, suggesting only a weak relationship between the bulk water composition and biofilm structure. In contrast, the dissolved oxygen concentration (O_2) exhibits a distinct positive gradient, with higher O_2 concentrations associated with increases in both residual thickness and thinning rate. This indicates that conditions characterized by elevated dissolved oxygen tend to coincide with thicker biofilms possessing more shear-responsive outer layers. Temperature displays the opposite behavior, with higher T_{in} values generally associated with lower residual thickness and, to a lesser extent, lower thinning rate, as points corresponding to warmer conditions cluster toward the lower-left portion of the behavioral space.

The disturbance time period t_{rest} influences both $A_{residual}$ and $k_{thinning}$ positively, with longer undisturbed periods favoring the development of thicker and more shear-responsive biofilm structures. Hydraulic history exerts the strongest influence on biofilm structure. Higher conditioning flow rates (Q_{cond}) are associated with lower $A_{residual}$ and $k_{thinning}$ values, whereas lower conditioning flow rates promote thicker and more compliant biofilms. Similarly, negative changes in conditioning flow rate (ΔQ_{cond}), corresponding to transitions toward lower hydraulic stress, tend to align with the high-A/high-k regime, while positive changes are associated with thinner and less deformable structures.

Together, these maps provide a structural interpretation of the RF model, showing that hydraulic conditions primarily govern biofilm architecture, while environmental variables such as dissolved oxygen concentration, temperature, and hydraulic stability further influence the balance between compact, shear-resistant biofilms and thicker, more shear-responsive structures.

4. DISCUSSION

This study presents the first explicit model capable of predicting biofilm thickness in drinking water pipes under controlled laboratory conditions using routinely measurable operational variables. Hydraulic conditions were the dominant drivers of biofilm thickness within the controlled experimental conditions of this study, where water quality and biological factors were relatively stable. Both the instantaneous shear imposed during measurement (Q_{test}) and the long-term hydraulic conditioning (Q_{cond} and ΔQ_{cond}) strongly reduced predicted thickness. This is consistent with established theory, in which shear stress regulates detachment and constrains vertical growth, producing thinner and more compact biofilms under high-shear regimes,[14,28,48] with implications for particle mobilization and discoloration processes observed in PVC systems.[49] Periods of undisturbed operation (captured by t_{rest}) also promoted thicker biofilms, reflecting the well-documented tendency of biofilms to consolidate, stratify, and accumulate biomass under stable environmental conditions.[47,50]

Chemical and thermal influences on biofilm thickness were also captured, albeit with weaker effect sizes than those of the hydraulic variables. Higher dissolved oxygen concentrations were generally associated with an increased predicted biofilm thickness, indicating that variations in oxygen availability may contribute to differences in biofilm development within the experimental system. In contrast, higher temperatures decreased the thickness, consistent with the experimental observations and highlighting the secondary role of environmental conditions relative to hydraulic drivers.

The extended $A-k$ structural analysis further revealed how operational drivers influence not only thickness but also the shear-dependent mechanical structure of the biofilm. The residual thickness parameter $A_{residual}$ increased with dissolved oxygen concentration (O_2) and with long resting periods (t_{rest}), indicating the formation of more substantial biofilm base layers under conditions characterized by elevated oxygen availability and prolonged environmental stability. The thinning rate $k_{thinning}$, interpreted as the shear-dependent mechanical structure of the biofilm or its susceptibility to shear, also increased with O_2 and, to a lesser extent, EC, suggesting that these conditions are associated with the development of more shear-responsive biofilm. In contrast, higher temperatures consistently reduced both $A_{residual}$ and $k_{thinning}$, indicating that warmer, oligotrophic conditions favor thinner, more compact, and less shear-sensitive biofilms. Hydraulic history (Q_{cond} and ΔQ_{cond}) contributed strongly to $A_{residual}$ but showed no clear association with $k_{thinning}$.

https://doi.org/10.1021/acsestwater.6c00651
ACS EST Water XXXX, XXX, XXX−XXX

This suggests that long-term shear primarily influences the size of the biofilm's structural core, whereas elevated O_2 concentrations and prolonged environmental stability are associated with the development of softer, more shear-responsive outer layers.

Together, these findings demonstrate that the combined effects of shear history, environmental stability, oxygen availability, and temperature can be distilled into moderately predictable patterns of biofilm structure. By modeling physical biofilm thickness and, through the $A-k$ analysis, its shear-dependent mechanical structure, this work provides system-level insight into biofilm development that is difficult to achieve using microbiological indicators alone and establishes a foundation for operationally relevant biofilm prediction.

4.1. Comparison with Previous Studies

Previous biofilm modeling efforts in drinking water systems have primarily focused on microbiological indicators, such as ATP, cell counts, or 16S rRNA profiles, or on bulk-water quality parameters such as nutrient concentrations.[51] While these proxies remain valuable, their deployment is often limited by high capital and operational costs, sensor data complexity, and difficulties in interpretation at the whole-system scale, particularly because they do not directly measure biomass attached to the pipe walls.[24,27]

The present work advances beyond these approaches by explicitly modeling the physical outcome of biofilm processes, i.e., thickness. The modeled relationships are strongly aligned with established experimental and theoretical findings: increased shear reduces biofilm height,[2,52] stable low-shear conditions promote thickening,[46,53] temperature increase affects the biofilm negatively in oligotrophic environments,[54] and oxygen availability modulates microbial activity, biofilm metabolism, and the development of biofilm structure.[33] However, unlike previous studies, the RF model integrates these effects simultaneously and quantifies their relative contributions in a form that can be applied to real-life systems.

The added $A-k$ behavioral analysis further extends the literature by demonstrating that biofilm structural behavior, specifically the balance between a shear-resistant core and loosely attached protrusions, can also be inferred from operational data. To our knowledge, this is the first demonstration that biofilm "gelatinous deformation", i.e., its shear-dependent mechanical structure changes, can be approximated from data-driven modeling and linked to operational conditions such as dissolved oxygen concentration, temperature, and hydraulic stability. This contributes new mechanistic insight into how the structure of drinking water biofilms evolves, complementing prior microbiological or mechanistic models and enabling interpretation at a whole-system scale.

4.2. Limitations of This Work

Several factors influence the interpretation of the results obtained. First, the variability of EC and O_2 measurements was limited by the intermittent sampling frequency and the relative biological stability of the provided water, reducing the model's ability to learn intricate relationships between water chemistry and biofilm thickness. This is an experimental rather than a modeling limitation; higher-frequency chemical monitoring may reveal additional chemical controls on thickness, such as hysteresis phenomena in biofilm growth dynamics induced by chemical water quality perturbations. Moreover, because these measurements were taken at the utility's production site rather than in situ at the point of entry to the setup, the model may not capture changes occurring along the distribution network, such as biologically or chemically induced variations in dissolved oxygen or mineral content. As a result, the model should not be viewed as a comprehensive representation of chemical water-quality controls on biofilm development, as it primarily captures hydraulic and thermal effects under biologically stable water-quality conditions.

Second, the experimental setup differs from full-scale drinking water distribution systems in several important respects. The use of a single pipe material (PVC-p), small diameter, and controlled laboratory conditions does not reflect the diversity of pipe materials (e.g., unplasticized PVC and asbestos cement pipes), geometries, aging processes, and complex hydraulics encountered in operational networks. Moreover, the indirect hydraulic method used to quantify biofilm thickness relies on the controlled laboratory conditions employed here and may not be readily transferable to operational DWDSs; however, this limitation primarily affects the validation dataset rather than the proposed model structure, which requires further validation using independent datasets and operational DWDS conditions. A more detailed discussion of the assumptions and limitations of the measurement method is provided by Glynis et al.[23] At the same time, these controlled conditions enabled isolation of hydraulic and temperature effects on biofilm thickness, which would be difficult to disentangle in more heterogeneous systems. Furthermore, while flow rate was an appropriate predictor within the experimental setup, velocity (v) or wall shear stress (τ) may represent more transferable descriptors for systems with different pipe diameters and hydraulic regimes; therefore, extrapolation beyond the experimental domain should be undertaken with caution.

Third, biofilm development is also governed by water quality characteristics and microbial community composition, including nutrient availability and community-dependent cohesion and detachment resistance. These biological and chemical drivers were not explicitly resolved in the present study and may contribute to the variability in biofilm thickness under similar hydraulic conditions. In this work, water quality and pipe material were intentionally kept constant in the experimental facility in order to isolate and systematically quantify the influence of hydraulic and temperature conditions on biofilm thickness, an effect that would be difficult to disentangle in full-scale distribution systems, where multiple drivers vary simultaneously.

Fourth, the experiments were conducted using biologically stable (Appendix A), unchlorinated drinking water and therefore do not account for the effects of disinfectant residuals present in many full-scale networks. Finally, while biofilm thickness provides a useful structural indicator of biofilm accumulation, it does not directly reflect microbial activity or functional characteristics such as metabolic rates or community composition.

Fifth, the model is trained on quasi-steady-state biofilm thickness values obtained at discrete measurement times and not on continuously varying biofilm thickness. As such, it captures the direction and magnitude of distinct steady-state responses to operational drivers but does not explicitly model dynamic biofilm regrowth processes. Additionally, the current formulation does not account for hydraulic complexities such as flow reversals, intermittent supply, or local hydraulic anomalies near pumps and valves, all of which can substantially influence biofilm development in real networks.

https://doi.org/10.1021/acsestwater.6c00651
ACS EST Water XXXX, XXX, XXX−XXX

Sixth, the random forest framework, while robust and interpretable, has some inherent limitations, including difficulty in extrapolating outside the training domain and limited representation of finely structured temporal dependencies. The repeated-sampling evaluation strategy mitigates risks of overfitting but does not eliminate them entirely.

4.3. Recommendations for Future Research

The findings highlight several promising directions for the continued development of biofilm thickness modeling. A key priority is the development of dynamic models that incorporate temporal growth and decay processes rather than relying solely on the steady-state thickness. Such models would provide a richer representation of biofilm dynamics and facilitate better forecasting of responses to operational changes. Equally important is full-scale validation through the application of the model to pilot or real distribution systems, to assess its transferability across a wider range of pipe materials, diameters, hydraulic conditions, and disinfection regimes.

The present study demonstrated that hydraulic variables were the dominant predictors of biofilm thickness and structure. Future research should therefore investigate whether a hydraulics-focused modeling framework can provide comparable predictive performance while reducing data requirements. In parallel, dedicated studies with higher-resolution water-quality datasets are needed to better quantify the influence of environmental variables, including dissolved oxygen, organic carbon, and disinfectant residuals, on biofilm development. Such studies would help clarify the extent to which water quality controls can improve predictive performance beyond that achieved using hydraulic variables alone. The inclusion of complementary biofilm proxies, capturing microbial activity or community shifts, also holds promise. While thickness reflects structural outcomes, combining it with targeted microbiological proxies could enhance the mechanistic insight and strengthen future model formulations.

Finally, methodological developments offer additional opportunities. Integrating data-driven approaches with mechanistic biofilm models may enable hybrid frameworks that capture both underlying physical principles and complex interactions. Comparative evaluations of alternative machine learning algorithms or physics-based models would further clarify how best to represent biofilm behavior at the system scale.

4.4. Implications for Practice

This work represents an initial step toward operational biofilm thickness modeling, and its practical implications for water utilities ultimately depend on further development and validation. Before such a model can be deployed on a large scale, it will need to be extended to accommodate a wider range of pipe materials, diameters, environmental conditions, and operational settings. Nevertheless, the results presented here demonstrate the feasibility of predicting biofilm thickness directly from operational data typically available to utilities, such as flow rate, temperature, and bulk chemical water quality measurements.

Such models could support several practical applications. They may enable the identification of zones prone to thick or highly shear-responsive biofilms, which are often associated with more diverse microbial communities and exhibit greater niche heterogeneity, conditions that may increase the likelihood of opportunistic pathogen establishment. They could be integrated with hydraulic models or SCADA systems to generate real-time or scenario-based estimates of biofilm accumulation throughout the network. The model could also inform targeted cleaning or flushing strategies by highlighting locations where biofilms are predicted to be thicker or more structurally unstable, and it may assist in network design and asset management by indicating how changes in flow regimes or system configuration could influence long-term biofilm development.

The incorporation of the $A-k$ structural analysis could provide additional diagnostic value by distinguishing between compact, shear-resistant biofilms and more filamentous, easily detached structures. This capability may assist utilities in prioritizing areas at greater risk of aesthetic issues, mobilization events, or pathogen retention.

5. CONCLUSIONS

This paper presents a new model for predicting biofilm thickness in drinking water pipes. The model predicts biofilm thickness as a function of readily measurable variables representing different hydraulic, thermal, and chemical conditions in the pipe. This model was developed using random forest and experimentally collected data.

Using long-term experimental measurements, the random forest model achieved high predictive accuracy (testing R^2 = 0.905) and identified hydraulic forcing, especially instantaneous shear and, to a lesser extent, long-term conditioning, as the dominant determinant of biofilm thickness, with temperature and oxygen availability exerting secondary influences.

To move beyond a single notion of "thickness", two structural descriptors, the residual thickness $A_{residual}$ and thinning rate $k_{thinning}$, were introduced to characterize the biofilm's shear-dependent mechanical behavior. These descriptors revealed links between environmental conditions and biofilm structure, distinguishing compact, shear-resistant biofilms from thicker, more deformable layers.

While promising, the model remains constrained by the controlled laboratory setting, limited chemical variability, and the use of a single pipe diameter and material. It also reflects quasi-steady-state responses rather than dynamic growth. These limitations highlight the need for further development and validation under more diverse hydraulic and water-quality conditions.

Even so, this framework provides a foundation for the operational biofilm thickness assessment. With continued refinement and full-scale verification, such models could support hotspot identification, scenario-based risk analysis, and optimization of network cleaning strategies, contributing to improved biofilm management in drinking water distribution systems.

APPENDIX A

Overview of Water Quality Parameters

Tables A1 and C1 summarizes the basic influent water quality parameters used in the experiments (December 2024−October 2025), based on data provided by the supplying water utility.

The consistently low TOC and stable pH, dissolved oxygen, and mineral composition reported in Table A1 indicate biologically stable, oligotrophic conditions typical of unchlorinated drinking water distribution systems. Accordingly, the influent water was used directly in the experimental facility without additional chemical treatment or disinfection.

https://doi.org/10.1021/acsestwater.6c00651
ACS EST Water XXXX, XXX, XXX−XXX

Table A1. Summary of Basic Influent Water Quality Parameters during the Experimental Period (December 2024−October 2025), Based on Measurements Provided by the Supplying Water Utility

parameter	mean	min	max	units
pH	8.00	7.85	8.10	
EC	247.2	232.0	257.0	μS/cm
O_2	10.56	9.90	11.60	mg/L
bicarbonate (HCO_3^-)	154.20	146.00	163.00	mg/L
calcium (Ca)	37.60	33.80	41.00	mg/L
magnesium (Mg)	6.32	5.71	6.80	mg/L
TOC	1.60	1.50	1.70	mg/L

Table C1. Summary Statistics of Measured Biofilm Thickness ($z_{biofilm}$, μm) Grouped by Testing Flow Rate (Q_{test}, L h^{-1})

Q_{test}	mean	median	min	max
50	68.2	70.9	44.7	91.4
75	57.7	59.1	44.7	71.9
100	50.4	50.2	41.4	59.8
125	37.9	38.0	29.9	44.3
150	43.0	43.0	31.0	57.5
200	30.9	31.1	19.3	38.5
300	21.2	20.2	14.7	30.0
400	19.4	19.1	11.2	29.8

APPENDIX B

Optical Verification of Biofilm Development (Figure 8)

APPENDIX C

Biofilm Thickness Dataset Summary (Table C1)

APPENDIX D

Assessment of Multicollinearity among Input Features

To evaluate potential multicollinearity among the input features used in the random forest model, pairwise Pearson correlation coefficients (r) were calculated for all predictors (Q_{test}, Q_{cond}, ΔQ_{cond}, T_{in}, O_2, EC, and t_{rest}). The resulting correlation matrix is shown in Figure 9.

The strongest correlation was observed between the conditioning flow rate (Q_{cond}) and the engineered change in conditioning flow rate (ΔQ_{cond}) (r = 0.96), reflecting the fact that large changes in hydraulic conditions were typically associated with higher conditioning flow rates. Moderate positive correlations were also present between T_{in} and both Q_{cond} and ΔQ_{cond} (r = 0.52−0.55). Dissolved oxygen concentration (O_2) exhibited a strong negative correlation with T_{in} (r = −0.70) and weaker negative correlations with Q_{cond} and ΔQ_{cond} (r = −0.35 to −0.40). All remaining feature pairs showed weak correlations ($|r| < 0.4$).

These relationships largely reflect physical and operational dependencies within the experimental design rather than statistical artifacts. For example, seasonal variations in water temperature were accompanied by changes in dissolved oxygen concentration and, to some extent, by adjustments to the hydraulic conditioning program. While random forest models are relatively robust to correlated predictors, multicollinearity can influence impurity-based feature-importance scores. Therefore, feature importance in this study was interpreted in conjunction with SHAP analysis, which provides more reliable insight into predictor influence in the presence of correlated inputs.

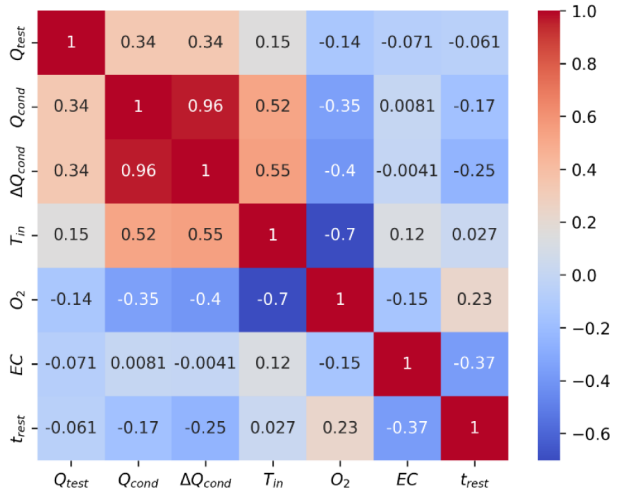


Figure 9. Pairwise Pearson correlation coefficients between all input features (Q_{test}, Q_{cond}, ΔQ_{cond}, T_{in}, O_2, EC, and t_{rest}) used in the random forest model for biofilm thickness prediction.

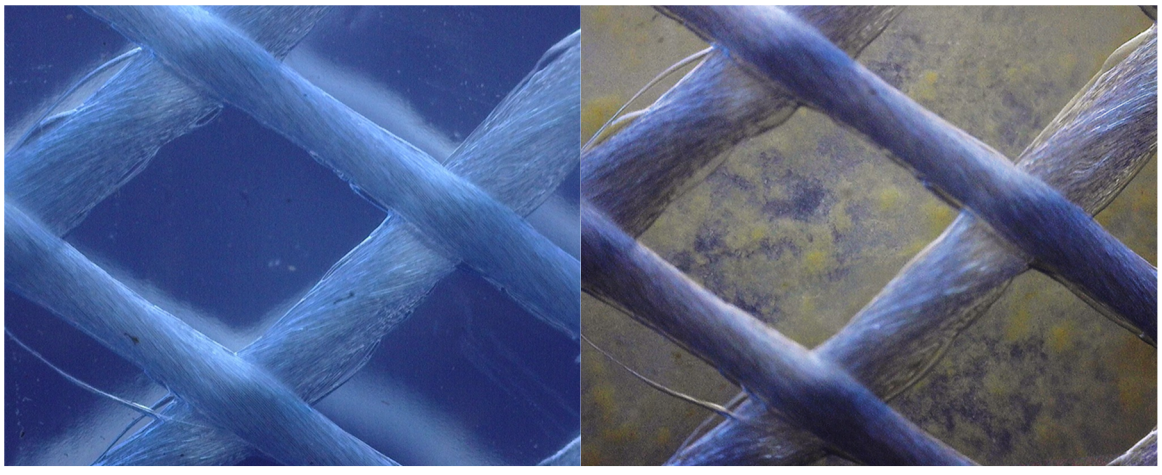


Figure 8. Optical images of the "Slimer" setup pipe wall before (left) and after (right) biofilm development.

https://doi.org/10.1021/acsestwater.6c00651
ACS EST Water XXXX, XXX, XXX−XXX

■ ASSOCIATED CONTENT

Data Availability Statement

Data can be made available on request.

■ AUTHOR INFORMATION

Corresponding Author

Konstantinos Glynis − *Department of Water Management, Delft University of Technology, Delft 2628 CN, Netherlands; KWR Water Research Institute, Nieuwegein 3433 PE, Netherlands;* orcid.org/0000-0002-0911-6793; Email: K.G.Glynis@tudelft.nl

Authors

Mirjam Blokker − *Department of Water Management, Delft University of Technology, Delft 2628 CN, Netherlands; KWR Water Research Institute, Nieuwegein 3433 PE, Netherlands*

Zoran Kapelan − *Department of Water Management, Delft University of Technology, Delft 2628 CN, Netherlands*

Dragan Savić − *KWR Water Research Institute, Nieuwegein 3433 PE, Netherlands; Centre for Water Systems, University of Exeter, Exeter EX4 4PY, U.K.;* orcid.org/0000-0001-9567-9041

Complete contact information is available at: https://pubs.acs.org/10.1021/acsestwater.6c00651

Author Contributions

K.G.: Writing − original draft, Methodology, Investigation, Formal analysis, Data curation. M.B.: Methodology, Investigation, Supervision. Z.K.: Methodology, Supervision. D.S.: Methodology, Supervision.

Notes

During the preparation of this work, the authors used ChatGPT (OpenAI) to support tasks such as improving phrasing, enhancing clarity, and refining the structure of the manuscript. After using this tool, the authors reviewed and edited all generated content as needed and took full responsibility for the content of the published article.

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the "Smart Water Futures" project which has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, Grant agreement No. 951424. The authors thank Paul van der Wielen (KWR & Radboud University) and Anke Brouwer−Hanzens (KWR) for their support and advice on the microbiological aspects of the research and Robbie van Pelt (KWR) for his technical advice and assistance in realizing and operating the experiments.

■ REFERENCES

(1) Bachmann, R. T.; Edyvean, R. G. J. Biofouling: An Historic and Contemporary Review of Its Causes, Consequences and Control in Drinking Water Distribution Systems. *Biofilms* **2005**, *2* (3), 197−227.

(2) Douterelo, I.; Husband, S.; Loza, V.; Boxall, J. Dynamics of Biofilm Regrowth in Drinking Water Distribution Systems. *Appl. Environ. Microbiol.* **2016**, *82* (14), 4155−4168.

(3) Abberton, C. L.; Bereschenko, L.; van der Wielen, P. W. J. J.; Smith, C. J. Survival, Biofilm Formation, and Growth Potential of Environmental and Enteric Escherichia Coli Strains in Drinking Water Microcosms. *Appl. Environ. Microbiol.* **2016**, *82* (17), 5320−5331.

(4) Husband, P. S.; Boxall, J. B. Asset Deterioration and Discolouration in Water Distribution Systems. *Water Res.* **2011**, *45* (1), 113−124.

(5) EPA *Health Risks from Microbial Growth and Biofilms in Drinking Water Distribution Systems*; EPA, 2002.

(6) Cowle, M. W.; Babatunde, A. O.; Rauen, W. B.; Bockelmann-Evans, B. N.; Barton, A. F. Biofilm Development in Water Distribution and Drainage Systems: Dynamics and Implications for Hydraulic Efficiency. *Environ. Technol. Rev.* **2014**, *3* (1), 31−47.

(7) Fish, K. E.; Collins, R.; Green, N. H.; Sharpe, R. L.; Douterelo, I.; Osborn, A. M.; Boxall, J. B. Characterisation of the Physical Composition and Microbial Community Structure of Biofilms within a Model Full-Scale Drinking Water Distribution System. *PLoS One* **2015**, *10* (2), No. e0115824.

(8) Braga, A. S.; Filion, Y. Examining the Conditioning Factors That Influence Material Shear Strength of Particle Deposits in a Full-Scale Drinking Water Distribution Laboratory. *Environ. Sci.: Water Res. Technol.* **2023**, *9* (10), 2619−2630.

(9) Liu, S.; Gunawan, C.; Barraud, N.; Rice, S. A.; Harry, E. J.; Amal, R. Understanding Monitoring, and Controlling Biofilm Growth in Drinking Water Distribution Systems. *Environ. Sci. Technol.* **2016**, *50* (17), 8954−8976.

(10) Cowle, M. W.; Webster, G.; Babatunde, A. O.; Bockelmann-Evans, B. N.; Weightman, A. J. Impact of Flow Hydrodynamics and Pipe Material Properties on Biofilm Development within Drinking Water Systems. *Environ. Technol.* **2020**, *41* (28), 3732−3744.

(11) Calero Preciado, C.; Boxall, J.; Soria-Carrasco, V.; Martínez, S.; Douterelo, I. Implications of Climate Change: How Does Increased Water Temperature Influence Biofilm and Water Quality of Chlorinated Drinking Water Distribution Systems? *Front. Microbiol.* **2021**, *12*, 12.

(12) van der Wielen, P. W. J. J.; Dignum, M.; Donocik, A.; Prest, E. I. Influence of Temperature on Growth of Four Different Opportunistic Pathogens in Drinking Water Biofilms. *Microorganisms* **2023**, *11* (6), 1574.

(13) Agudelo-Vera, C.; Avvedimento, S.; Boxall, J.; Creaco, E.; de Kater, H.; Di Nardo, A.; Djukic, A.; Douterelo, I.; Fish, K. E.; Iglesias Rey, P. L.; Jacimovic, N.; Jacobs, H. E.; Kapelan, Z.; Martinez Solano, J.; Montoya Pachongo, C.; Piller, O.; Quintiliani, C.; Ručka, J.; Tuhovčák, L.; Blokker, M. Drinking Water Temperature around the Globe: Understanding, Policies, Challenges and Opportunities. *Water* **2020**, *12* (4), 1049.

(14) Tsagkari, E.; Sloan, W. T. Turbulence Accelerates the Growth of Drinking Water Biofilms. *Bioprocess Biosyst. Eng.* **2018**, *41* (6), 757−770.

(15) Boubetra, A.; Le Nestour, F.; Allaert, C.; Feinberg, M. Validation of Alternative Methods for the Analysis of Drinking Water and Their Application to Escherichia Coli. *Appl. Environ. Microbiol.* **2011**, *77* (10), 3360−3367.

(16) Douterelo, I.; Boxall, J. B.; Deines, P.; Sekar, R.; Fish, K. E.; Biggs, C. A. Methodological Approaches for Studying the Microbial Ecology of Drinking Water Distribution Systems. *Water Res.* **2014**, *65*, 134−156.

(17) van der Wielen, P. W. J. J.; van der Kooij, D. Effect of Water Composition, Distance and Season on the Adenosine Triphosphate Concentration in Unchlorinated Drinking Water in the Netherlands. *Water Res.* **2010**, *44* (17), 4860−4867.

(18) Douterelo, I.; Sharpe, R. L.; Husband, S.; Fish, K. E.; Boxall, J. B. Understanding Microbial Ecology to Improve Management of Drinking Water Distribution Systems. *Wires Water* **2019**, *6* (1), No. e01325.

(19) Dreszer, C.; Vrouwenvelder, J. S.; Paulitsch-Fuchs, A. H.; Zwijnenburg, A.; Kruithof, J. C.; Flemming, H.-C. Hydraulic Resistance of Biofilms. *J. Membr. Sci.* **2013**, *429*, 436−447.

(20) Gomes, I. B.; Simões, M.; Simões, L. C. An Overview on the Reactors to Study Drinking Water Biofilms. *Water Res.* **2014**, *62*, 63−87.

(21) Wagner, M.; Horn, H. Optical Coherence Tomography in Biofilm Research: A Comprehensive Review. *Biotechnol. Bioeng.* **2017**, *114* (7), 1386−1402.

(22) Türetgen, I. Comparison of the Efficacy of Free Residual Chlorine and Monochloramine against Biofilms in Model and Full Scale Cooling Towers. *Biofouling* **2004**, *20* (2), 81−85.
(23) Glynis, K.; Blokker, M.; Kapelan, Z.; Savić, D. Experimental Measurements of Biofilm Thickness in Drinking Water Pipes. *Urban Water J.* **2026**, 1−18.
(24) Pick, F. C.; Fish, K. E.; Husband, S.; Boxall, J. B. Non-Invasive Biofouling Monitori.ng to Assess Drinking Water Distribution System Performance. *Front. Microbiol.* **2021**, *12*, 730344.
(25) Goliopoulos, N.; Makropoulos, C. Near Real-Time Topological Adaptation in Looped Water Distribution Networks: Enhancing Resilience and Water Quality. *J. Hydroinf.* **2025**, *27* (6), 1030−1046.
(26) Erdei-Tombor, P.; Kiskó, G.; Taczman-Brückner, A. Biofilm Formation in Water Distribution Systems. *Processes* **2024**, *12* (2), 280.
(27) Burnet, J. B.; Demeter, K.; Dorner, S.; Farnleitner, A. H.; Hammes, F.; Pinto, A. J.; Prest, E. I.; Prévost, M.; Stott, R.; van Bel, N. Automation of On-Site Microbial Water Quality Monitoring from Source to Tap: Challenges and Perspectives. *Water Res.* **2025**, *274*, 123121.
(28) Fish, K. E.; Sharpe, R. L.; Biggs, C. A.; Boxall, J. B. Impacts of Temperature and Hydraulic Regime on Discolouration and Biofilm Fouling in Drinking Water Distribution Systems. *PLOS Water* **2022**, *1* (8), No. e0000033.
(29) Glynis, K.; Blokker, M.; Kapelan, Z.; Savić, D. Experimental Setup for Measuring the Effect of Biofilm Build-up on Heat Transfer in Drinking Water Pipes. *Eng. Proc.* **2024**, *69* (1), 169.
(30) Learbuch, K. L. G.; Smidt, H.; Van Der Wielen, P. W. J. J. Influence of Pipe Materials on the Microbial Community in Unchlorinated Drinking Water and Biofilm. *Water Res.* **2021**, *194*, 116922.
(31) Małek, M. T.; Koczyk, H. Influence of Temperature Sensor (Pt100) Accuracy on the Interpretation of Experimental Results of Measuring Temperature on the Surface. *Civ. Environ. Eng. Rep.* **2024**, *34* (4), 1.
(32) European Commission *Review Of Sensors To Monitor Water Quality: eRNCIP Thematic Area "Chemical Biological Risks In The Water Sector", Deliverable D1 Task 1*; Publications Office of the European Union: Luxembourg, 2013.
(33) Wang, Y.; Xu, H.; Shen, Z.; Ding, M.; Xue, R.; Lin, T. Effects of Dissolved Oxygen on Water Quality and Biofilms in the Raw Water Distribution Pipeline for a Reserved Water Source. *Environ. Sci.: Water Res. Technol.* **2022**, *8* (10), 2277−2288.
(34) Braissant, O.; Astasov-Frauenhoffer, M.; Waltimo, T.; Bonkat, G. A Review of Methods to Determine Viability, Vitality, and Metabolic Rates in Microbiology. *Front. Microbiol.* **2020**, *11*, 547458.
(35) Creaco, E.; Campisano, A.; Fontana, N.; Marini, G.; Page, P. R.; Walski, T. Real Time Control of Water Distribution Networks: A State-of-the-Art Review. *Water Res.* **2019**, *161*, 517−530.
(36) Breiman, L. Random Forests. *Mach. Learn* **2001**, *45* (1), 5−32.
(37) Cheng, L.; De Vos, J.; Zhao, P.; Yang, M.; Witlox, F. Examining Non-Linear Built Environment Effects on Elderly's Walking: A Random Forest Approach. *Transp. Res. D: Transp. Environ.* **2020**, *88*, 102552.
(38) Lindner, T.; Puck, J.; Verbeke, A. Beyond Addressing Multicollinearity: Robust Quantitative Analysis and Machine Learning in International Business Research. *J. Int. Bus Stud.* **2022**, *53* (7), 1307−1314.
(39) Kemei, E. K.; Laerhoven, K. V.; Karuri, N. W.; Tewo, R. K. Multivariate Prediction of Total Organic Carbon in River Water Using Random Forest and Deep Learning Regression Algorithms. *Appl. Comput. Inform.* **2025**, *5*, 264−285.
(40) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc, 2017; Vol.: *30*.
(41) Ponce-Bobadilla, A. V.; Schmitt, V.; Maier, C. S.; Mensing, S.; Stodtmann, S. Practical Guide to SHAP Analysis: Explaining Supervised Machine Learning Model Predictions in Drug Development. *Clin. Transl. Sci.* **2024**, *17* (11), No. e70056.
(42) Tabib, S.; Larocque, D. Non-Parametric Individual Treatment Effect Estimation for Survival Data with Random Forests. *Bioinformatics* **2020**, *36* (2), 629−636.
(43) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(44) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* **2020**, *2* (1), 56−67.
(45) Sharma, P.; Mirzan, S. R.; Bhandari, A.; Pimpley, A.; Eswaran, A.; Srinivasan, S.; Shao, L. Evaluating Tree Explanation Methods for Anomaly Reasoning: A Case Study of SHAP TreeExplainer and TreeInterpreter. In *Advances in Conceptual Modeling*; Grossmann, G.; Ram, S.; Eds.; Springer International Publishing: Cham, 2020; pp. 35−45. DOI: .
(46) Paul, E.; Ochoa, J. C.; Pechaud, Y.; Liu, Y.; Liné, A. Effect of Shear Stress and Growth Conditions on Detachment and Physical Properties of Biofilms. *Water Res.* **2012**, *46* (17), 5499−5508.
(47) Learbuch, K. *Influence of Flow Velocity on the Distribution of Bacteria in Biofilm and Water*; BTO 2018.090, KWR Water Research Institute, 2018. https://library.kwrwater.nl/publication/102149/influence-of-flow-velocity-on-the-distribution-of-bacteria-in-biofilm-and-water-executive-summary-flow-velocity-not-an-influence-on-the-distribution-of-microorganisms-in-biofilm-and-water/.
(48) Stoodley, P.; Cargo, R.; Rupp, C. J.; Wilson, S.; Klapper, I. Biofilm Material Properties as Related to Shear-Induced Deformation and Detachment Phenomena. *J. Ind. Microbiol. Biotechnol.* **2002**, *29* (6), 361−367.
(49) Braga, A. S.; Filion, Y.; Anderson, B. The Synergy between Particles and Biofilms That Drives Drinking Water Discolouration Processes in PVC Pipes. *Environ. Sci.: Water Res. Technol.* **2026**, *12* (1), 168−184.
(50) Telgmann, U.; Horn, H.; Morgenroth, E. Influence of Growth History on Sloughing and Erosion from Biofilms. *Water Res.* **2004**, *38* (17), 3671−3684.
(51) Waegenaar, F.; García-Timermans, C.; Van Landuyt, J.; De Gusseme, B.; Boon, N. Impact of Operational Conditions on Drinking Water Biofilm Dynamics and Coliform Invasion Potential. *Appl. Environ. Microbiol.* **2024**, *90* (5), No. e0004224.
(52) Mathieu, L.; Bertrand, I.; Abe, Y.; Angel, E.; Block, J. C.; Skali-Lami, S.; Francius, G. Drinking Water Biofilm Cohesiveness Changes under Chlorination or Hydrodynamic Stress. *Water Res.* **2014**, *55*, 175−184.
(53) Tsagkari, E.; Connelly, S.; Liu, Z.; McBride, A.; Sloan, W. T. The Role of Shear Dynamics in Biofilm Formation. *Npj Biofilms Microbiomes* **2022**, *8* (1), 33.
(54) Blokker, M. *Modelling Growth in the Distribution Network*; BTO Report 2017.077, KWR Water Research Institute: Nieuwegein, Netherlands, 2017, p 38.

https://doi.org/10.1021/acsestwater.6c00651
ACS EST Water XXXX, XXX, XXX−XXX