

Article

Leakage-Controlled Comparative Prediction of Microplastics in Sewage Treatment Plant Effluents: Season- and Morphology-Aware Modelling at Erandwane STP, Pune, India

Supriya B. Shinde * and Moni U. Khobragade

Department of Civil Engineering, COEP Technological University, Pune 411005, Maharashtra, India;
muk.civil@coeptech.ac.in

* Correspondence: sbs21.civil@coeptech.ac.in

Abstract

Microplastics (MPs) discharged in sewage treatment plant (STP) effluents vary with influent concentration, particle morphology and temporal conditions, but predictive studies can be weakened by data leakage, limited benchmarking and mechanistic overinterpretation. This study evaluates within-site prediction of final-effluent MP concentration at the Erandwane STP, Pune, India, using 792 contemporaneously paired influent–effluent observations from 396 dates between 1 August 2024 and 31 August 2025. Predictors were total influent MPs, influent fibres, fragments, films and beads, categorical hydrological season and sampling session. Morning and afternoon records from the same date were kept in the same modelling subset to prevent same-day leakage, and five-fold date-grouped cross-validation compared multivariable linear regression, support vector regression, random forest, XG-Boost and CatBoost. On the independent grouped test set, CatBoost achieved a coefficient of determination (R^2) of 0.498, a root mean square error (RMSE) of 35.52 MPs L⁻¹ and a mean absolute error (MAE) of 28.58 MPs L⁻¹; however, its mean grouped cross-validation R^2 (0.433) did not exceed multivariable linear regression (0.473). The principal finding is therefore moderate site-level predictive information in the available multivariable structure rather than superiority of a specific algorithm. Season was the leading CatBoost importance variable but is interpreted only as a temporal proxy because direct rainfall, flow, hydraulic retention time, clarifier loading and total suspended solids were not continuously available. Quantitative recovery, duplicate-precision and detection-limit data were also unavailable, and the same validation subset was used for tuning, calibration and conformal-residual estimation. Consequently, the uncertainty results and model interpretation should be considered exploratory. The framework may support prioritisation of confirmatory monitoring at this site, but it cannot replace laboratory measurement or establish transferable treatment mechanisms without prospective, multi-site validation and stronger analytical and operational covariates.



Academic Editor: Daniel Rittschof

Received: 9 September 2026

Revised: 25 September 2026

Accepted: 3 October 2026

Published: 22 October 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: microplastics; sewage treatment plant; seasonal variability; morphology-resolved prediction; grouped cross-validation; wastewater monitoring

1. Introduction

Microplastics (MPs), generally defined as synthetic polymer particles smaller than 5 mm, are persistent contaminants of aquatic, terrestrial, atmospheric and engineered systems [1,2]. Their small size, variable density, hydrophobic surfaces and affinity for

co-contaminants facilitate transport through wastewater networks and receiving waters. MPs can be ingested by aquatic organisms and transferred through food webs, while also acting as carriers for other contaminants [3,4]. Recent international studies continue to show that plastic sources and human activity shape environmental MP occurrence, including tourism-related inputs in coastal systems [5].

Sewage treatment plants (STPs) are important control points in the urban MP cycle because they receive fibres and particles from laundry, personal-care products, industrial discharges, urban runoff, packaging waste and tyre wear [6–8]. Screening, sedimentation, biological treatment and advanced polishing can remove a large fraction of incoming MPs, yet residual particles remain in final effluent and retained particles can accumulate in sewage sludge [9,10]. STPs should therefore be viewed both as removal barriers and as residual discharge pathways, rather than simply as sources of MPs. Broader research on plastic substitution and circular-economy pathways also demonstrates the continuing need to reduce plastic inputs at source [11]. Related wastewater and microplastic studies further document treatment removal, sludge retention, analytical recovery, polymer identification, and variability across treatment configurations [12–16].

Prediction of effluent MP concentrations is challenging because influent concentration, removal performance and particle composition vary with season, runoff, sewer flushing, hydraulic conditions, sampling time and morphology. Wet-weather conditions may mobilise previously deposited particles and alter hydraulic residence and settling behaviour, whereas dry-weather conditions are generally more stable [17,18]. Studies of variable-quality wastewater treatment likewise demonstrate that flow and residence conditions can materially influence treatment performance [19]. These observations provide a practical rationale for evaluating season-aware and nonlinear prediction rather than assuming a fixed influent–effluent proportionality. Work on hydrodynamic transport, weather-related mobilisation, inflow variability, and environmental microplastic transfer supports treating temporal context cautiously rather than assigning a single causal driver [20–22].

Particle morphology adds further complexity because fibres, fragments, films and beads differ in shape, density, buoyancy and interactions with flocs and aerated treatment systems [23–25]. Morphology-resolved predictors may therefore preserve information that is lost when only total influent concentration is considered. Polymer identity can add further mechanistic information through density and surface-chemistry differences; however, polymer-specific FTIR/Raman confirmation was not available as a predictive feature in the present dataset.

The primary methodological question in this study is whether the available season- and morphology-resolved influent information contains reproducible within-site predictive signal under leakage-controlled evaluation, rather than whether any single algorithm is intrinsically superior. Environmental prediction can appear stronger than it is when repeated measurements from the same sampling context are split across development and test data, when correlated predictors are interpreted independently, or when a nonlinear model is compared only with a weak baseline. The present analysis therefore treats validation design, model-family benchmarking and interpretation discipline as part of the scientific contribution. CatBoost is examined as one candidate nonlinear model within this comparative framework. Recent machine-learning studies in wastewater, microplastic detection, environmental transport, spectroscopy, and model interpretability likewise show that predictive performance depends on validation design, data quality, and the measurement domain [26–31].

This study develops a site-specific comparative prediction framework for final-effluent MP concentration at the Erandwane STP, Pune. The contribution lies in integrating a full-year paired influent–effluent dataset with morphology-resolved candidate predictors, date-

grouped leakage control, an independent grouped test set, five-fold grouped benchmarking across linear and nonlinear models, redundancy assessment, interpretable diagnostics and split-conformal uncertainty intervals. CatBoost is evaluated as one candidate model rather than presented as the source of novelty. The objectives were to: (i) characterise seasonal and morphology-specific variability; (ii) quantify within-site predictive performance on unseen sampling dates; (iii) compare multivariable linear regression, support vector regression, random forest, XGBoost and CatBoost under the same grouped validation framework; (iv) interpret season and morphology as descriptive predictive variables without assigning unsupported treatment mechanisms; and (v) define the generalisability and uncertainty limits relevant to monitoring prioritisation.

2. Methodology

2.1. Study Site and Experimental–Computational Framework

Data collection was conducted at the Erandwane Sewage Treatment Plant (STP), adjacent to Mhatre Bridge on the Mutha River in Pune, India (18.5037° N, 73.8349° E). The site receives municipal wastewater from a highly urbanised catchment. Sampling was performed downstream of a 20 mm coarse screen and at the final treated-effluent point. The study documentation confirms screening and primary/secondary treatment stages, represented schematically in Figure 1, but does not provide verified unit-by-unit design and operating records for grit removal, equalisation, clarifier configuration, tertiary filtration or related treatment-unit performance. Verified plant-specific values for hydraulic retention time, sludge age/sludge-management practice, clarifier loading and tertiary-unit operation were also unavailable. Consequently, treatment-process variables were not used as model predictors, and no unit-specific removal mechanism is inferred.

The dataset contains 792 paired influent–final effluent observations from 396 sampling dates between 1 August 2024 and 31 August 2025, with two sessions per date (morning and afternoon). Season was represented by the categorical labels recorded for the observations: dry, pre-monsoon and monsoon. Exact calendar boundaries and site-specific rainfall or meteorological time series corresponding to these labels were not available in the analysis materials; season is therefore treated as a study-period temporal category rather than a directly measured hydrological state. The analytical unit is one influent–final effluent pair from the same date and session. Keeping both sessions from a date in the same modelling subset prevents contextual leakage between development and evaluation data.

Figure 1 summarises the experimental and computational workflow. Candidate variables included total influent MP concentration, morphology-resolved influent counts, derived subclass/total ratios, categorical season and sampling session. Direct rainfall intensity, air temperature, humidity, influent flow, chemical oxygen demand (COD), biochemical oxygen demand (BOD₅), total suspended solids (TSS) and other treatment-process measurements were not available as continuous model inputs and were not introduced by imputation or reconstruction. Because no measured rainfall/flow series or defensible independent proxy series were available, a sensitivity analysis that adds proxy rainfall or flow was not performed; constructing such proxies directly from the season label would be circular and would not provide an independent test of season importance. Derived ratios were evaluated during feature engineering but excluded from the final predictive fit because they are algebraically dependent on total and subclass counts. The final CatBoost predictors were total influent MPs, influent fibres, fragments, films and beads, season and session; the target was measured final-effluent MP concentration.

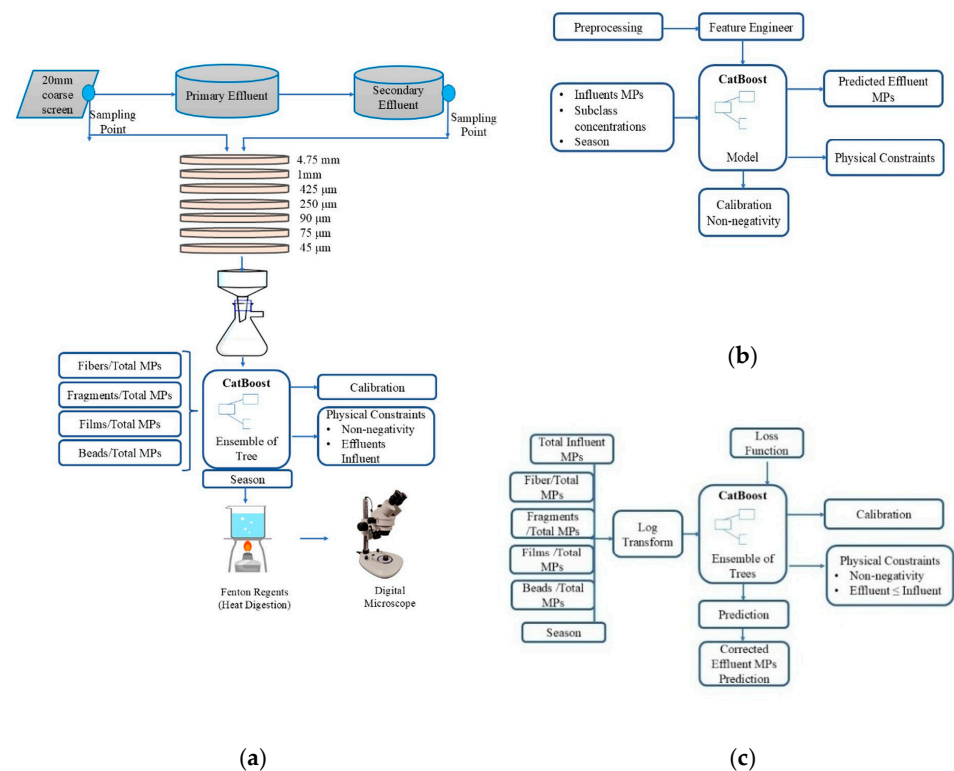


Figure 1. Integrated workflow for microplastic sampling and prediction: (a) laboratory processing and identification; (b) linkage of influent and temporal predictors to CatBoost; and (c) preprocessing, target transformation, model fitting, calibration and post-prediction plausibility checks. Season is a model input and a stratification variable; it is not a separate preprocessing outcome. (a) Sampling and microplastic-isolation workflow. The burner, beaker and microscope symbols denote laboratory digestion/processing and microscopic particle counting/identification. (b) Experimental-computational framework linking influent measurements and season/session information to final-effluent prediction. (c) Preprocessing and CatBoost workflow. Ratio variables shown here represent engineered candidate features; the final model excluded the ratios after redundancy screening. Only the effluent target, not all predictors, was log-transformed.

Field/procedural blanks, duplicate counting, spiked-sample recovery checks, filtered deionised rinsing water, covered glassware and contamination-control practices were documented as part of the analytical workflow. However, quantitative recovery percentages, duplicate relative variability, analytical detection/quantification limits, blank concentrations and observer-classification uncertainty were not available in the study documentation and cannot be recovered from the concentration values alone. The analytical response variable therefore has unquantified measurement uncertainty, which limits the strength of conclusions about model accuracy. Established work on analytical recovery correction and polymer identification illustrates why missing quantitative QA/QC and spectroscopic confirmation remain important limitations [13,14].

2.2. Wastewater Sampling and Fractionation

Wastewater was sampled at two hydraulic points: influent downstream of the 20 mm coarse screen and final treated effluent immediately before discharge to the Mutha River. Two composite sampling sessions were used per date: morning (09:00–12:00 h) and afternoon (12:00–15:00 h), with hourly grab subsamples composited within each session. The study dataset contains 792 paired observations across 396 dates from August 2024 through August 2025. Influent and effluent composites were collected within corresponding day-time sessions and were not time-shifted by a measured plant hydraulic retention time. Thus,

“paired” denotes a same-date/session association rather than tracking the same wastewater parcel through the treatment train; calculated removal efficiency is descriptive and may not represent parcel-specific treatment removal. The fixed daytime window improves protocol consistency but does not represent overnight or early-morning peak-flow conditions.

Sampling vessels and sieves were pre-cleaned and rinsed with filtered deionised water. Glass or stainless-steel equipment was used wherever practicable to reduce plastic contamination. The laboratory workflow shown in Figure 1a used stainless-steel sieves of 4.75 mm, 1 mm, 425 µm, 250 µm, 90 µm, 75 µm and 45 µm; the nominal lower sieve size was therefore 45 µm. Material retained on the sieves was transferred to pre-cleaned glass beakers using filtered deionised water. The morphology-based dataset records particle-abundance categories (fibres, fragments, films and beads); no mass-based or weight-ratio measurements by morphology were available, so the modelling analysis uses particle counts/concentrations rather than mass fractions. The study documentation does not specify the complete digestion reagent and conditions, density-separation medium, filtration membrane and pore size, microscope make and magnification, exact analysed sample volume, image-analysis software or all visual-identification criteria sufficiently for full protocol reproduction. These unreported procedural details are therefore acknowledged as a methodological limitation rather than reconstructed.

The final predictor definitions, variable types, and modelling roles are summarised in Table 1.

Table 1. Summary of Input Features and Target Variable for Machine Learning Dataset.

Feature Category	Feature Name	Type	Description/Notes
Influent measurements	Total MPs	Continuous	Total microplastic concentration (MPs L ⁻¹)
Influent measurements	Fibres, fragments, films, beads	Continuous	Subclass-specific MP counts (MPs L ⁻¹)
Derived ratios	Fibres/Total; Fragments/Total; Films/Total; Beads/Total	Derived continuous	Exploratory engineered features; excluded from final model because of algebraic redundancy with total/subclass counts
Seasonal/hydrological	Season	Categorical	Dry, pre-monsoon and monsoon study-period labels; exact meteorological boundaries were not available as model inputs
Temporal	Session	Categorical	Morning or afternoon sampling session; no continuous rainfall/flow/TSS/COD/BOD ₅ field was available in the study dataset
Target variable	Effluent MPs	Continuous	Measured final-effluent MP concentration (MPs L ⁻¹)

2.3. Quality Assurance, Quality Control, and Analytical Validation

Field blanks accompanied wastewater samples to estimate contamination introduced during sampling and transport, and procedural blanks were carried through sieving, digestion, filtration and counting. Blank correction was completed before feature construction. The blank-correction equation and quantitative blank concentrations, recovery percentages, duplicate-count coefficient of variation/relative per cent difference, method detection limit and method quantification limit were not available in the study documentation; therefore, no numerical QA/QC performance claim is made.

A subset of samples was independently or repeatedly counted, and spiked samples containing known MP particles were processed through the isolation workflow to check recovery and potential particle losses. Samples showing substantial disagreement were re-evaluated. These controls confirm that repeat-counting and recovery checks were part of the workflow, but the underlying replicate and spike logs were unavailable; consequently,

they do not provide a quantitative estimate of analytical precision or recovery for the present modelling analysis.

Samples were kept covered when not being processed; work surfaces and equipment were cleaned with filtered deionised water and ethanol, with ethanol used only as a routine surface-cleaning agent rather than for microbial analysis. Non-plastic equipment and cotton laboratory clothing were preferred where practicable. These measures reduce contamination but cannot eliminate uncertainty associated with visual/microscopic identification. Polymer-specific confirmation by FTIR or Raman spectroscopy was not included. Because polymer identity was not independently verified, misclassification of some nonplastic particles cannot be excluded; the response variable should therefore be interpreted as morphology-based particle counts under the applied optical identification workflow. Published reviews of spectroscopic microplastic analysis likewise emphasise the distinction between morphology-based identification and polymer confirmation [14].

2.4. Train–Validation–Test Splitting and Leakage Control

The 792 observations were split at the unique-date level so that morning and afternoon records from the same date could never appear in different subsets. Using a fixed random seed (42), 396 dates were stratified by hydrological season and allocated to training (237 dates; $n = 474$ observations), validation (79 dates; $n = 158$) and independent test (80 dates; $n = 160$) subsets. This grouped random split was chosen instead of a purely sequential split to reduce ordering bias while preserving contextual independence at the date level.

Seasonal representation was maintained across the subsets. The training subset contained 90 dry, 55 pre-monsoon and 92 monsoon dates; validation contained 30 dry, 19 pre-monsoon and 30 monsoon dates; and the test subset contained 31 dry, 18 pre-monsoon and 31 monsoon dates. All subsets therefore span the monitoring period rather than representing only one contiguous season. The validation set was used for hyperparameter selection, early stopping and calibration; the test set was reserved for final evaluation.

In addition to the single grouped holdout split, five-fold stratified group cross-validation was performed with sampling date as the group variable and season as the stratification variable. This provides a more robust estimate of performance for a moderate-sized single-site dataset and directly tests sensitivity to the particular holdout allocation. The grouped design estimates within-site generalisation to unseen dates; it is not external validation across plants, treatment technologies or years.

The grouped random split does not constitute prospective forecasting because temporally adjacent dates from the same annual period can occur in different subsets. Date grouping prevents only same-date leakage; it does not remove possible dependence between neighbouring days. A blocked chronological holdout and a formal assessment of temporal autocorrelation were not available in the retained analysis outputs, so adjacent-day dependence could not be quantified or eliminated retrospectively. Accordingly, conventional inferential statistics are treated as descriptive, and the reported model performance estimates only within-site generalisation to unseen dates under the observed monitoring period. No claim of prospective or temporally independent forecasting performance is made. A blocked time-series evaluation with explicit autocorrelation diagnostics is required in future work. The evaluation is accordingly restricted to grouped within-site prediction, rather than forecasting independently observed future periods.

2.5. CatBoost as a Candidate Nonlinear Model

Final-effluent MP concentration was modelled using CatBoost regression, a gradient-boosting decision-tree method suitable for nonlinear interactions and categorical

predictors [32,33]. The final predictors were total influent MPs, influent fibres, fragments, films and beads, hydrological season and sampling session. Feature inclusion was process-driven rather than target-driven. The derived subclass/total ratios were excluded from the final fit because of algebraic redundancy. A VIF diagnostic confirmed the expected severe redundancy when total MPs and their component counts are simultaneously treated as linear predictors ($VIFs > 10^3$ for several variables); this does not invalidate tree-based prediction but means individual importance values must be interpreted cautiously rather than causally.

The target was transformed as $\log(1 + y)$, where y is measured effluent concentration, to reduce heteroscedasticity. CatBoost was tuned on the validation subset over tree depth, learning rate, iteration number and L2 regularisation. For the grouped split used in this analysis, the selected configuration used depth = 6, learning rate = 0.03 and L2 leaf regularisation = 3. Early stopping monitored validation RMSE with a patience of 50 iterations; the minimum validation log-RMSE (0.2085) occurred at iteration 126, and training terminated after no further improvement within the patience window. A fixed random seed of 42 was used for data splitting and reproducibility where applicable. The exact candidate hyperparameter ranges, software/package versions and computational hardware settings were not retained in the study documentation and could not be reconstructed reliably. We therefore report only the settings that are verifiably available and explicitly acknowledge that exact computational replication is limited. Future archival practice should deposit the environment file, package versions, hardware specification and complete tuning grid together with the modelling code.

A fixed seed was used for reproducibility. CatBoost handled season and session as categorical predictors. The independent test targets were not used for feature screening, hyperparameter selection, calibration, early stopping, model interpretation or uncertainty calibration.

2.6. Calibration and Post-Prediction Plausibility Checks

Raw CatBoost predictions were back-transformed to $MPs L^{-1}$. A linear calibration model was fitted using validation predictions only (slope 0.997; intercept $3.07 MPs L^{-1}$) and then applied unchanged to the independent test predictions. The term calibration is used here only for this validation-fitted linear rescaling; it is distinct from the subsequent deterministic plausibility checks.

The same validation subset was used for hyperparameter selection/early stopping, linear calibration and construction of split-conformal residuals. This reuse can introduce optimistic bias because tuning and uncertainty calibration are not fully independent. The conformal coverage results are therefore reported as exploratory internal uncertainty diagnostics rather than as formally unbiased prospective intervals; nested or separately partitioned tuning and calibration should be used in future validation.

After calibration, two transparent plausibility checks were specified: negative predictions would be set to zero, and any prediction exceeding its paired influent concentration would be capped at that influent value. The upper bound was exactly $1.00 \times$ influent. In the independent test subset, neither rule altered a prediction, so the reported test metrics are identical before and after plausibility checking.

These checks are not physical constraints embedded in CatBoost and do not model settling, buoyancy, clarifier loading or hydraulic transport. They simply prevent impossible concentration bounds under the assumptions of this paired dataset. Because no test prediction was altered, the checks had zero measurable effect on test R^2 , RMSE or MAE.

2.7. Model Interpretation

Global CatBoost feature importance was used to rank predictive contribution. Partial dependence plots (PDPs) were generated for total influent MPs and morphology-resolved influent variables to visualise the fitted marginal response after averaging over the remaining predictors. Because total concentration and morphology counts are strongly correlated, PDPs can evaluate predictor combinations that are sparse or absent in the observed data and are not interpreted as isolated physical effects. Grouped permutation importance, fold-wise SHAP analysis, accumulated local effects and conditional-dependence diagnostics were not available; feature-importance and PDP results are therefore treated only as descriptive diagnostics of the fitted model. Related studies in wastewater modelling and microplastic identification underscore the need to keep interpretability separate from mechanistic verification [26,27,30,31].

2.8. Comparative Model-Family Benchmarking and Evaluation Metrics

Performance was assessed using R^2 , RMSE and MAE. R^2 measures explained variance, RMSE emphasises larger errors while remaining in MPs L^{-1} , and MAE represents the average absolute error in the same physical unit. MSE was not reported separately because it contains the same squared-error information as RMSE but in less directly interpretable squared units. MAPE was removed because it was not required for the principal interpretation and would add a redundant percentage metric.

Benchmarking included a single-predictor influent–effluent linear reference and, for the full available predictor set, multivariable linear regression, support vector regression (SVR), random forest, XGBoost and CatBoost. The multivariable and nonlinear models were evaluated using the same five-fold date-grouped cross-validation framework; this model-family comparison is the principal basis for algorithmic interpretation. Training-set mean, seasonal-mean and previous-date persistence benchmarks were not available in the retained analysis outputs, and a matched set of total-only, morphology-only and season/session-only ablation fits was not available. Therefore, the study does not quantify the independent incremental value of morphology or season, and differences between the single-predictor reference and full models are not presented as a formal ablation effect. Feature importance and PDPs are used only to describe fitted-model behaviour.

Feature importance and PDPs were generated only after the predictive workflow was specified and were not used to tune the independent test performance.

2.9. Robustness and Uncertainty Quantification

Split-conformal uncertainty intervals were constructed from absolute validation residuals after model calibration. The 90% interval used a residual quantile of 64.04 MPs L^{-1} and achieved 95.0% empirical coverage on the independent test set; the 95% interval used 70.10 MPs L^{-1} and achieved 96.9% coverage. Because a single absolute-residual quantile was used for each nominal level, the corresponding untruncated full interval widths were 128.08 and 140.20 MPs L^{-1} , respectively. Season-specific coverage, concentration-conditional coverage and comparisons with alternative uncertainty methods were not available. Together with the validation-reuse caveat described above, these intervals should be interpreted as internal exploratory uncertainty estimates for this site and monitoring period rather than external uncertainty bounds for other STPs.

3. Results

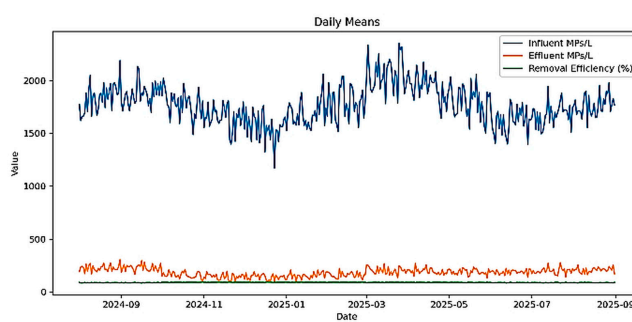
3.1. Temporal and Seasonal Behaviour of Microplastic Concentrations

The dataset contains 792 paired observations from 396 dates. Influent MP concentration averaged $1773.4 \pm 228.7 \text{ MPs L}^{-1}$ (median 1772; range 1095–2589), whereas effluent

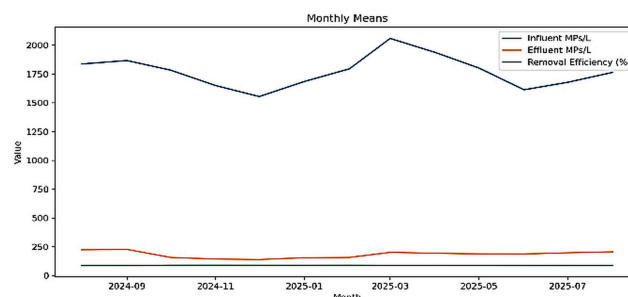
concentration averaged 184.6 ± 49.5 MPs L⁻¹ (median 182; range 59–332). Mean removal efficiency was $89.61 \pm 2.38\%$. These distributions establish substantial variability despite consistently lower effluent than influent concentrations.

Seasonal summaries were based on 151 dry-period dates (302 observations), 92 pre-monsoon dates (184 observations) and 153 monsoon dates (306 observations). Mean effluent concentrations were 151.8 MPs L⁻¹ in the dry period, 195.9 MPs L⁻¹ in the pre-monsoon period and 210.1 MPs L⁻¹ in the monsoon period. Because the seasonal labels were not accompanied by continuous site-specific meteorological data, and rainfall intensity, flow rate, hydraulic retention time, clarifier loading and TSS were not directly measured in the modelling dataset, the seasonal pattern is reported as a temporal association rather than evidence of rainfall-driven hydraulic overloading or a specific treatment mechanism.

Figure 2 presents daily and monthly temporal summaries of influent concentration, effluent concentration and removal efficiency. Because concentration variables and removal efficiency have different numerical scales, interpretation is based on their reported values and trends rather than on direct comparison of visual amplitude. The seasonal distributional statistics reported below provide the quantitative basis for interpretation.



(a) Daily mean influent and effluent MP concentrations and removal efficiency.

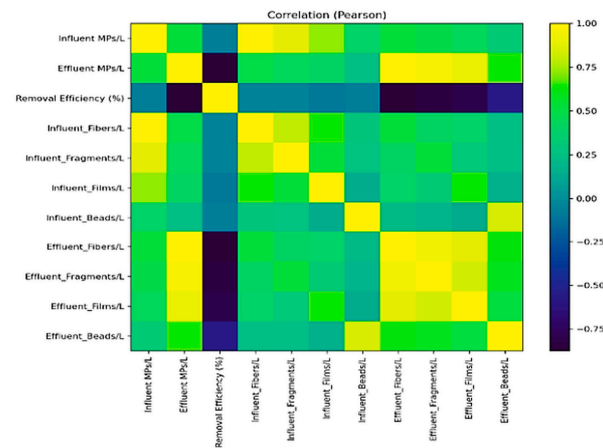


(b) Monthly mean influent and effluent MP concentrations and removal efficiency.

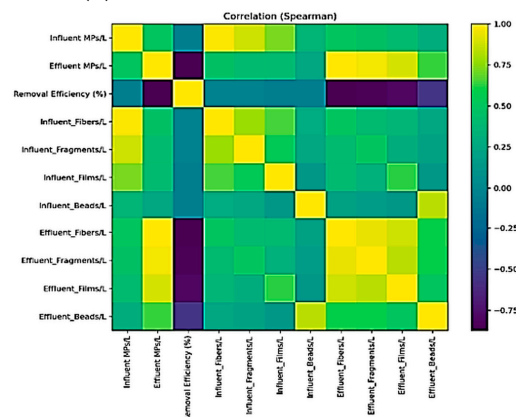
Figure 2. Temporal variation in influent and effluent microplastic concentrations and removal efficiency: (a) daily means and (b) monthly means. Concentration (MPs L⁻¹) and removal efficiency (%) differ in scale; quantitative comparison therefore relies on the reported numerical summaries.

3.2. Linear Baseline and Correlation Structure

Pearson and Spearman analyses were used to describe associations among influent characteristics, morphology-resolved counts, effluent concentration and removal efficiency (Figure 3). Total influent concentration was moderately positively correlated with effluent concentration (Pearson $r = 0.522$, $p < 0.001$). Influent fibres ($r = 0.494$), fragments ($r = 0.446$), films ($r = 0.422$) and beads ($r = 0.249$) were also positively correlated with effluent concentration (all $p < 0.001$). Because temporal autocorrelation was not formally modelled, the p -values are treated as descriptive rather than as definitive evidence of temporally independent effects.



(a) Pearson correlation matrix.



(b) Spearman rank correlation matrix.

Figure 3. Pearson and Spearman correlation matrices for influent characteristics, morphology resolved counts, effluent concentration and removal efficiency. Statistical significance for the principal influent–effluent associations is reported in the text.

Removal efficiency was strongly negatively correlated with effluent concentration ($r = -0.876$, $p < 0.001$). However, removal efficiency is calculated directly from influent and effluent concentrations, so this association is mathematically coupled and is not interpreted as an independent scientific relationship. Other influent–effluent correlations quantify co-variation but do not identify hydraulic or source mechanisms because flow, retention time and source-specific tracers were not measured.

The high correlation between total influent MPs and fibres ($r = 0.972$), together with strong correlations among several subclass counts, confirms predictor redundancy. The derived composition ratios were therefore excluded from the final CatBoost fit, and feature-importance/PDP findings are interpreted as distributed predictive information among correlated variables rather than as independent effects. No inference about household-textile sources is made from correlation alone.

The observed single-predictor influent–effluent scatter has $R^2 = 0.272$ (Figure 4), showing that total influent concentration alone explains only about 27% of effluent variability. This baseline motivates the evaluation of additional seasonal and morphology-resolved information, but it does not by itself prove that a nonlinear algorithm is necessary.

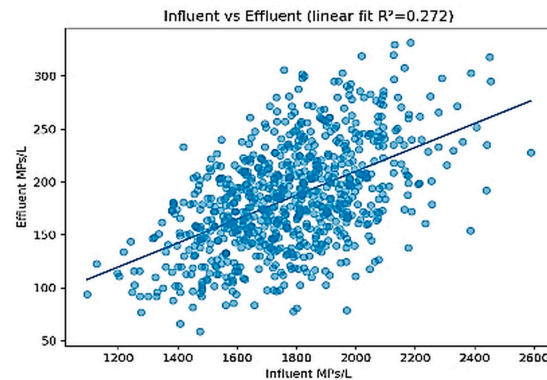


Figure 4. Observed influent versus observed effluent MP concentration with the single-predictor linear baseline. Blue circles represent individual paired observations, and the solid blue line represents the fitted least-squares linear baseline.

3.3. Seasonal and Sampling-Session Descriptive Comparisons

Observed effluent MP concentration differed among the three seasonal categories (one-way ANOVA: $F = 155.38$, $p < 0.001$; Kruskal–Wallis $H = 230.58$, $p < 0.001$), although these tests do not account for temporal autocorrelation and are therefore interpreted descriptively. Effluent medians were 151 MPs L^{-1} in the dry period (IQR $120\text{--}178$), 192.5 MPs L^{-1} in pre-monsoon (IQR $161\text{--}225.5$) and 205 MPs L^{-1} in monsoon (IQR $179\text{--}240.75$). Corresponding influent medians were $1703.5 \text{ MPs L}^{-1}$ in the dry period (IQR $1535.75\text{--}1854.75$), $1944.5 \text{ MPs L}^{-1}$ in pre-monsoon (IQR $1794\text{--}2082$) and 1757 MPs L^{-1} in monsoon (IQR $1629\text{--}1888.5$). Influent concentration was highest in pre-monsoon, whereas effluent concentration was highest in monsoon, reinforcing that the season label captures broader temporal context rather than a demonstrated hydraulic mechanism.

Morning and afternoon observations form a paired within-date structure. Removal efficiency was nearly identical between morning and afternoon samples (means 89.62% and 89.60% ; medians 89.72% and 89.70% , respectively); the paired afternoon-minus-morning difference was -0.028 percentage points (paired $t = -0.20$, $p = 0.843$). Measured effluent concentration averaged $190.24 \text{ MPs L}^{-1}$ in the afternoon and $178.92 \text{ MPs L}^{-1}$ in the morning, a mean paired difference of 11.32 MPs L^{-1} (95% CI $5.56\text{--}17.09$; paired $t = 3.86$, $p < 0.001$). These paired comparisons account for the repeated morning/afternoon structure within a date, but they do not address longer-range temporal autocorrelation.

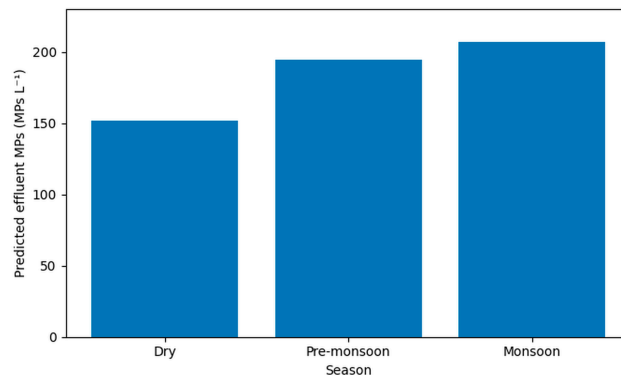
Figure 4 shows the observed influent–effluent relationship ($R^2 = 0.272$) and is distinct from multivariable model evaluation. It provides a direct graphical baseline showing considerable scatter around a simple proportional relationship.

Taken together, the seasonal summaries, paired session comparison and weak single-predictor baseline justify evaluating season, session and morphology-resolved influent information as candidate predictors, while not establishing any of them as causal process drivers.

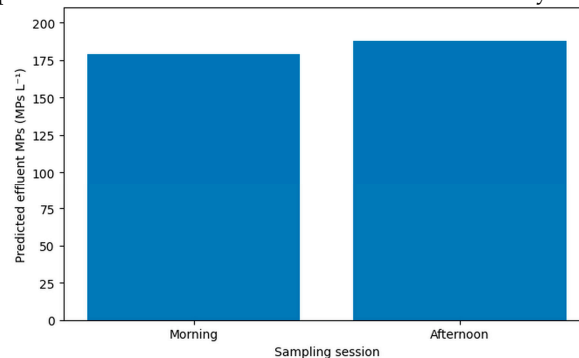
3.4. CatBoost-Predicted Category Patterns

The fitted CatBoost model reproduced systematic differences among hydrological seasons (Figure 5a). Mean model-predicted effluent concentration was 152.0 MPs L^{-1} in dry conditions, 194.8 MPs L^{-1} in pre-monsoon conditions and 207.2 MPs L^{-1} during monsoon.

Predicted afternoon concentration averaged 187.6 MPs L^{-1} compared with 178.9 MPs L^{-1} in the morning (Figure 5b). These plots describe conditional behaviour learned by CatBoost and are not independent causal estimates of season or sampling time.



(a) CatBoost-predicted mean effluent MP concentration across hydrological seasons.

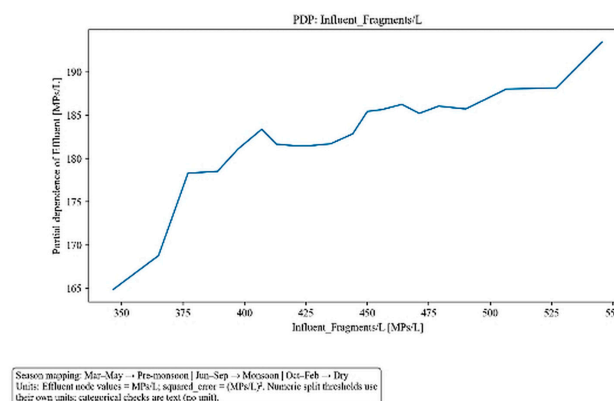


(b) CatBoost-predicted mean effluent MP concentration by sampling session.

Figure 5. Fitted CatBoost predictions summarised by (a) hydrological season and (b) sampling session. These summaries describe model behaviour rather than causal treatment effects.

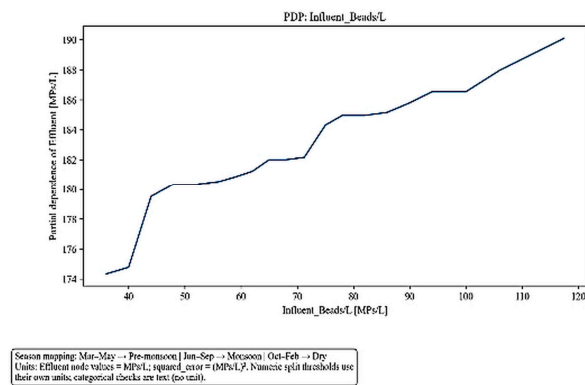
3.5. Partial Dependence Analysis

PDPs were used to visualise the fitted CatBoost response to total influent concentration and morphology-resolved influent counts (Figure 6). The curves show generally increasing marginal predictions with influent concentration, while some morphology-specific profiles contain local plateaus or changes in slope. Because the predictors are strongly correlated and the observed-data distribution is not shown beneath each curve, these shapes may include marginalised combinations that are sparse in the dataset and should be interpreted cautiously.

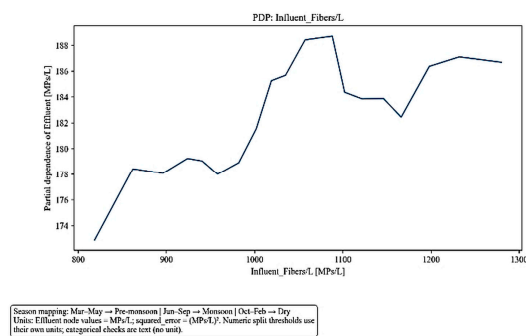


(a) PDP for influent fragments.

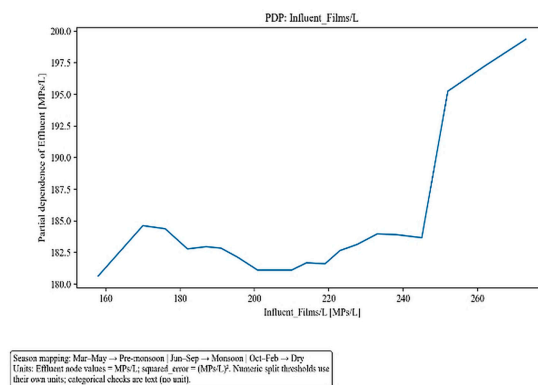
Figure 6. Cont.



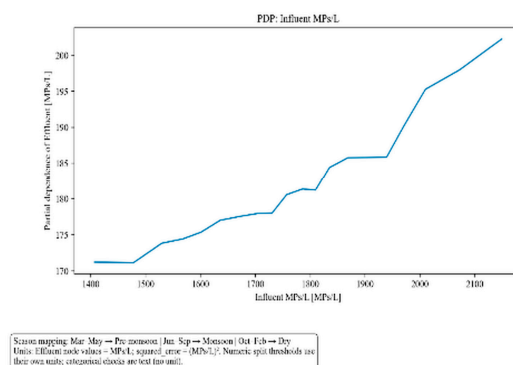
(b) PDP for influent beads.



(c) PDP for influent fibres.



(d) PDP for influent films.



(e) PDP for total influent MPs.

Figure 6. Partial dependence plots for CatBoost-predicted effluent concentration as a function of (a) influent fragments, (b) beads, (c) fibres, (d) films and (e) total influent MPs. Curves are model-specific marginal associations and are not causal or universal treatment thresholds.

Because the predictors are correlated, these PDP shapes are not interpreted as isolated physical response functions. Apparent thresholds may reflect the joint distribution of predictors and the fitted tree structure, and should not be equated with floc-capture, settling or buoyancy thresholds without independent process measurements.

The five PDP panels are used only to visualise model associations for the available predictors. Their interpretation is unified through the caption and surrounding text; no panel is treated as evidence of an experimentally validated treatment mechanism.

3.6. Feature Importance

Within the fitted CatBoost model, feature-importance scores ranked season highest (36.42%), followed by influent fibres (15.51%), fragments (14.86%), total MPs (12.01%), films (11.92%) and beads (9.28%); session had negligible importance (Figure 7). Because total MPs and subclass counts are strongly correlated, these percentages represent distributed model weight rather than independent causal contributions. The high rank of season may absorb unmeasured meteorological, hydraulic or operational variation. No sensitivity analysis with independent rainfall/flow proxies was possible because such measured or externally aligned proxy data were unavailable; proxies constructed from the season category itself would be circular. Because grouped permutation importance and fold-wise SHAP analyses were also unavailable, this ranking is treated as model-specific and descriptive rather than as a stable effect hierarchy.

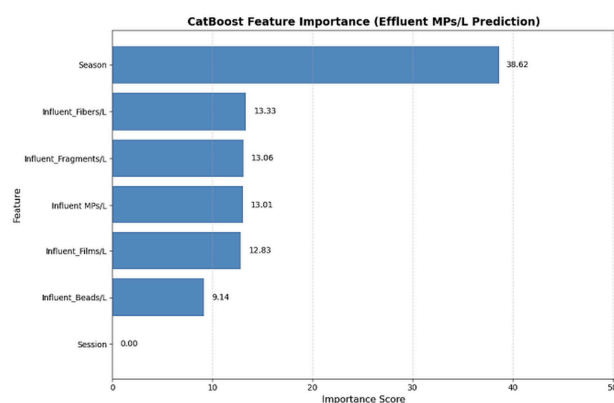


Figure 7. CatBoost feature importance for final-effluent prediction. Values describe predictive contribution within the fitted model and should not be interpreted as causal effect sizes.

3.7. Independent Test Performance and Algorithm Benchmarking

Using the date-grouped, season-stratified split described in Section 2.4, the independent test set contained 160 observations from 80 previously unseen sampling dates. The calibrated CatBoost model achieved $R^2 = 0.498$, $RMSE = 35.52 \text{ MPs L}^{-1}$ and $MAE = 28.58 \text{ MPs L}^{-1}$. Relative to the overall mean effluent concentration (184.6 MPs L^{-1}), the RMSE corresponds to approximately 19.2% and the MAE to 15.5%; however, these ratios are descriptive only. Because quantitative laboratory recovery, duplicate precision, blank variability and detection/quantification limits were unavailable, the observed errors cannot be separated into modelling error and analytical measurement uncertainty or judged against a validated laboratory-precision benchmark. The model is therefore not characterised as having high or measurement-grade predictive accuracy and should not replace direct laboratory measurement.

Five-fold date-grouped cross-validation was used to benchmark model families on the same predictor set. Multivariable linear regression achieved the highest mean cross-validated R^2 (0.473 ± 0.048), followed by CatBoost (0.433 ± 0.052), random forest (0.415 ± 0.057), XGBoost (0.403 ± 0.058) and SVR (0.275 ± 0.126). The comparison there-

fore does not support a claim of CatBoost superiority. Instead, it shows that moderate within-site predictability is present across more than one model family and that model complexity should be justified empirically. The single-predictor influent–effluent relationship ($R^2 = 0.272$) provides a descriptive reference, but because it was not generated through the identical grouped multivariable benchmark, the contrast is not presented as a formal feature-ablation effect. Season-specific model-performance metrics were not available, so aggregate performance may conceal weaker prediction within smaller subgroups, particularly the pre-monsoon category.

The five-fold date-grouped cross-validation results for all benchmarked model families are summarised in Table 2. These interpretation limits are consistent with the broader literature on ML applications and microplastic measurement [25–27].

Table 2. Five-fold date-grouped cross-validation benchmark (mean $\pm$ SD).

Model	R^2	RMSE (MPs L ^{−1})	MAE (MPs L ^{−1})
Multivariable linear regression	0.473 $\pm$ 0.048	35.62 $\pm$ 1.95	28.55 $\pm$ 2.18
Support vector regression	0.275 $\pm$ 0.126	41.53 $\pm$ 1.12	32.64 $\pm$ 1.38
Random forest	0.415 $\pm$ 0.057	37.50 $\pm$ 2.14	30.07 $\pm$ 1.73
XGBoost	0.403 $\pm$ 0.058	37.88 $\pm$ 2.07	30.38 $\pm$ 1.57
CatBoost	0.433 $\pm$ 0.052	36.95 $\pm$ 2.09	29.58 $\pm$ 1.85

3.8. Uncertainty and Scope of Plausibility Checks

Split-conformal intervals based on validation residuals provided an explicit but exploratory measure of prediction uncertainty. The nominal 90% interval achieved 95.0% test coverage with a fixed full width of 128.08 MPs L^{−1}, and the nominal 95% interval achieved 96.9% coverage with a fixed full width of 140.20 MPs L^{−1}. These widths are substantial relative to the mean effluent concentration and reinforce that predictions should be used only to flag periods for enhanced monitoring. Because the same validation subset was also used for tuning and calibration, and because season-specific/conditional coverage was not evaluated, interval performance should not be interpreted as a fully independent uncertainty assessment.

The higher monsoon effluent concentrations are statistically supported within the observed dataset, but direct hydraulic measurements were not available to assign a causal mechanism. Increased runoff, reduced retention time or clarifier stress remain plausible hypotheses supported by literature, not mechanisms demonstrated here. Similarly, correlations among MP subclasses cannot identify household or industrial sources without source-tracing evidence. Because quantitative analytical uncertainty and regulatory MP concentration thresholds were not available, the RMSE and MAE cannot be interpreted against a validated compliance threshold or laboratory precision benchmark.

The non-negativity and influent upper-bound rules were applied only after calibration. No test prediction was negative or exceeded its paired influent concentration, so neither rule changed any of the 160 test predictions and the test metrics were unchanged. No performance improvement is therefore attributed to post-processing.

The observed single-predictor influent–effluent relationship ($R^2 = 0.272$) is weaker than the performance of models fitted to the full available predictor set, but this contrast was not generated through a matched feature-ablation experiment. It therefore cannot establish the independent incremental contribution of season, session or morphology-resolved variables. Grouped cross-validation further shows that the full predictor structure contains within-site signal across more than one model family and that this signal is not uniquely attributable to CatBoost. Algorithm choice and feature-specific contributions require re-evaluation under controlled ablation and external validation.

4. Discussion

The monitoring dataset shows a clear seasonal pattern: monsoon effluent concentrations were highest (mean 210.1; median 205 MPs L⁻¹), pre-monsoon values were intermediate (mean 195.9; median 192.5 MPs L⁻¹), and dry-period concentrations were lowest (mean 151.8; median 151 MPs L⁻¹). Similar seasonal changes have been reported in wastewater and receiving-water studies [17,18,34]. However, the categorical season labels used here were not paired with continuous site-specific rainfall, temperature, humidity or flow measurements. In addition, influent and effluent samples were collected within the same daytime sessions rather than being offset by a measured hydraulic residence time. The observed mean removal efficiency of 89.61% should therefore be interpreted as a descriptive same-session concentration contrast, not as parcel-specific process removal. Season remains a temporal proxy rather than proof of hydraulic overloading or a causal treatment mechanism.

Morphology-resolved influent variables were included in the full predictor set and received non-zero importance within CatBoost, but their independent incremental predictive contribution cannot be established from the present analysis because matched total-only, morphology-only and season/session-only ablations were not retained. The strong correlation between total MPs and subclass counts also demonstrates substantial redundancy. Derived subclass ratios were therefore excluded, and feature importance and PDPs are interpreted cautiously. Effluent morphology-specific counts were not available in the modelling dataset, so morphology-specific removal efficiencies or effluent distributions for fibres, fragments, films and beads cannot be reported. Polymer-specific identification could add mechanistic information through density and surface-chemistry differences, but the magnitude of any predictive improvement cannot be quantified without paired FTIR/Raman-confirmed polymer data.

The comparative benchmark changes the interpretation of the modelling results. CatBoost achieved $R^2 = 0.498$ on the independent grouped test set, but its mean grouped cross-validation R^2 (0.433) did not exceed multivariable linear regression (0.473). Accordingly, the study does not identify CatBoost as the optimal model for STP microplastic data. The defensible conclusion is that the full available predictor structure contains reproducible within-site signal across several model families, while algorithm choice remains secondary and site-specific. Because analytical measurement uncertainty is not quantified, predictive error should be interpreted as combined model-and-measurement discrepancy rather than pure model error.

4.1. Contribution Beyond Algorithm Choice

The originality of this work lies in the evaluation framework rather than in the use of a particular machine-learning algorithm. For a full year of paired influent–effluent monitoring at one STP, the study combines morphology-resolved predictors, date-level leakage control, an independent grouped test set, five-fold grouped comparison of linear and nonlinear models, explicit treatment of predictor redundancy, interpretable diagnostics and uncertainty intervals. This combination allows the analysis to distinguish three questions that are often conflated: whether the site contains predictable signal, whether a complex nonlinear model is needed, and whether model associations can be interpreted as treatment mechanisms. The results answer the first question moderately, do not support the second as a general claim, and deliberately avoid the third without direct process measurements. Additional studies of wastewater and microplastic prediction reinforce the importance of comparing model families rather than assuming a complex learner will perform best [26,27,29,31].

4.2. Practical Implications

For Erandwane STP, the modelling framework can be considered only as an exploratory screening approach for prioritising sampling or laboratory confirmation when predicted effluent concentrations are elevated. The CatBoost test RMSE of 35.52 MPs L⁻¹ and the wide conformal intervals are substantial, and the absence of quantitative laboratory-precision data prevents determination of how much of this discrepancy arises from model error versus analytical uncertainty. No validated regulatory MP threshold or analytical-uncertainty benchmark was available for direct decision calibration. Any operational use would therefore require independent validation and confirmatory microscopy or spectroscopy rather than reliance on model output as a compliance measurement.

4.3. Generalisability and Limitations

The study is limited to one municipal STP and one August 2024–August 2025 monitoring campaign. Date-grouped splitting and grouped cross-validation reduce same-day leakage but do not remove possible short-term temporal dependence because adjacent dates can occur across subsets and temporal autocorrelation was not explicitly modelled. Thus, the evaluation should not be interpreted as a blocked prospective forecast. The categorical season labels were not supported by continuous site-specific meteorological records, and no independent rainfall/flow proxy sensitivity analysis could be performed without introducing circular assumptions. The dataset lacks continuous rainfall intensity, temperature, humidity, influent flow, COD, BOD₅, TSS, hydraulic retention time, clarifier loading and verified unit-level operating records. Influent and effluent samples were contemporaneous rather than retention-time matched. Quantitative recovery, blank, duplicate-precision and detection-limit data were unavailable; the full digestion, filtration and microscopy protocol was incompletely documented; polymer identity was not spectroscopically confirmed; and effluent morphology-specific data were unavailable. The same validation subset was reused for tuning, calibration and conformal residual estimation. Exact tuning ranges, package versions and hardware settings were not retained. These limitations restrict analytical reproducibility, mechanistic interpretation, uncertainty calibration, temporal generalisability and transferability beyond the observed site and period.

4.4. Future Work

Future work should prioritise prospective and external validation at multiple STPs with contrasting treatment configurations, catchments and climates. A chronological holdout or blocked time-series design should accompany grouped cross-validation, with explicit autocorrelation diagnostics such as lag plots, autocorrelation functions and residual temporal-dependence testing. Direct rainfall, temperature, humidity, flow, COD, BOD₅, TSS, hydraulic retention time, clarifier loading and treatment-unit performance should be recorded so that season can be decomposed into measured hydrometeorological and operational drivers rather than used as a broad categorical proxy. Influent/effluent sampling should be aligned to measured hydraulic residence time where feasible. The analytical protocol should report sample volume, digestion and density-separation conditions, filter type and pore size, microscope specifications, counting/classification rules, blank-correction equation, recovery, duplicate precision and detection/quantification limits, with polymer confirmation on a representative subset. Future modelling should include pre-specified feature ablations, training-mean/seasonal-mean/persistence baselines, season-specific performance, nested or separately partitioned tuning and calibration, grouped permutation importance, SHAP or accumulated-local-effect diagnostics, and complete computational environment archiving. Public deposition of the processed dataset, code, split indices, data dictionary and model settings would further improve reproducibility.

5. Conclusions

This study provides a single-site, leakage-controlled comparison of model families for predicting final-effluent microplastic concentration from a full year of paired influent measurements, morphology-resolved counts and categorical temporal information at Erandwane STP, Pune. CatBoost achieved $R^2 = 0.498$, RMSE = 35.52 MP_s L^{−1} and MAE = 28.58 MP_s L^{−1} on the independent grouped test set, but multivariable linear regression achieved a higher mean five-fold grouped cross-validated R^2 (0.473 versus 0.433). The central result is therefore evidence of moderate within-site predictive signal in the full available multivariable structure, not superiority of a nonlinear algorithm. Because matched feature ablations were unavailable, the independent incremental contribution of morphology and season cannot be quantified from the present analysis.

The findings remain exploratory because the study is limited to one site and monitoring period; season is a categorical proxy rather than a measured hydrological variable; key operational and quantitative QA/QC data are absent; influent and effluent samples were not retention-time matched; short-term temporal autocorrelation was not formally assessed; and prospective chronological validation was not performed. Consequently, RMSE and MAE represent combined model-and-measurement discrepancy rather than pure model error. The framework may help prioritise confirmatory monitoring after further validation, but it should not replace laboratory measurement or be transferred to other STPs without stronger analytical documentation, measured operational predictors, independent uncertainty calibration and external validation.

Author Contributions: S.B.S.: investigation, data curation, formal analysis, visualisation, and original draft. M.U.K.: conceptualisation, supervision, methodology, review, and editing. All authors have read and agreed to the published version of the manuscript.

Funding: The authors declare that no specific external funding was received for this study.

Data Availability Statement: The processed monitoring dataset, modelling code, date-grouped split indices, data dictionary and model settings are available from the corresponding author on reasonable request. Public deposition in a persistent repository had not been completed at the time of manuscript preparation; such deposition is recommended, subject to institutional approval, to support independent reproduction of the computational analysis.

Acknowledgments: The authors gratefully acknowledge the institutional and laboratory facilities provided by COEP Technological University, Pune, which supported the successful completion of this study. The authors sincerely express their gratitude to their research guide for valuable guidance, encouragement, and technical support throughout the course of this research work. The authors also extend their sincere thanks to the Head of the Department, Civil Engineering, COEP Technological University, Pune, for continuous support, encouragement, and cooperation throughout the research work. The authors further acknowledge the library facilities and academic resources provided by COEP Technological University, Pune, which were valuable in carrying out this research work. During the preparation of this manuscript, the authors used ChatGPT (OpenAI; Pro 100) to assist in preparing the graphical abstract. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

References

1. Zhang, S.; Wang, J.; Liu, X.; Qu, F.; Wang, X.; Wang, X.; Li, Y.; Sun, Y. Microplastics in the environment: A review of analytical methods, distribution and biological effects. *TrAC Trends Anal. Chem.* **2019**, *111*, 62–72. [CrossRef]
2. Hale, R.C.; Seeley, M.E.; La Guardia, M.J.; Mai, L.; Zeng, E.Y. A global perspective on microplastics. *J. Geophys. Res. Ocean.* **2020**, *125*, e2018JC014719. [CrossRef]

3. Benson, N.U.; Agboola, O.D.; Fred-Ahmadu, O.H.; De-la-Torre, G.E.; Oluwalana, A.; Williams, A.B. Micro (nano) plastics prevalence, food web interactions and toxicity assessment in aquatic organisms: A review. *Front. Mar. Sci.* **2022**, *9*, 851281. [\[CrossRef\]](#)
4. Provencher, J.F.; Ammendolia, J.; Rochman, C.M.; Mallory, M.L. Assessing plastic debris in aquatic food webs: What we know and don't know about uptake and trophic transfer. *Environ. Rev.* **2019**, *27*, 304–317. [\[CrossRef\]](#)
5. Chairunnisa, N.K.; Adam, M.A.; Kristianto, S.; Candri, D.A.; Talbia, H.S.; Aprilia, M.; Mutia, T.; Masruroh, H.; Iswara, A.P.; Prayogo, W. Linking the Tourism Activity to the Occurrence and Distribution of Microplastics. *Civ. Eng. J.* **2025**, *11*, 2811–2825. [\[CrossRef\]](#)
6. Carr, S.A.; Liu, J.; Tesoro, A.G. Transport and fate of microplastic particles in wastewater treatment plants. *Water Res.* **2016**, *91*, 174–182. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Gatidou, G.; Arvaniti, O.S.; Stasinakis, A.S. Review on the occurrence and fate of microplastics in Sewage Treatment Plants. *J. Hazard. Mater.* **2019**, *367*, 504–512. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Iyare, P.U.; Ouki, S.K.; Bond, T. Microplastics removal in wastewater treatment plants: A critical review. *Environ. Sci. Water Res. Technol.* **2020**, *6*, 2664–2675. [\[CrossRef\]](#)
9. Li, X.; Chen, L.; Mei, Q.; Dong, B.; Dai, X.; Ding, G.; Zeng, E.Y. Microplastics in sewage sludge from the wastewater treatment plants in China. *Water Res.* **2018**, *142*, 75–85. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Rolsky, C.; Kelkar, V.; Driver, E.; Halden, R.U. Municipal sewage sludge as a source of microplastics in the environment. *Curr. Opin. Environ. Sci. Health* **2020**, *14*, 16–22. [\[CrossRef\]](#)
11. Erdiwansyah, N.S.; Masyudi, S.; Irhamni, D.; Hidayat, F. Global Trends in Agricultural Waste-Based Bioplastic Research: A Scientometric Review. *J. Hum. Earth Future* **2025**, *6*, 234–246. [\[CrossRef\]](#)
12. Bayo, J.; Olmos, S.; López-Castellanos, J. Assessment of microplastics in a municipal wastewater treatment plant with tertiary treatment: Removal efficiencies and loading per day into the environment. *Water* **2021**, *13*, 1339. [\[CrossRef\]](#)
13. Zhu, Z.; Parker, W.J. Implementing differential recovery corrections enhances accuracy of mass balances on microplastics in wastewater treatment. *Water Res.* **2025**, *271*, 122912. [\[CrossRef\]](#) [\[PubMed\]](#)
14. K  ppler, A.; Fischer, D.; Oberbeckmann, S.; Schernewski, G.; Labrenz, M.; Eichhornand, K.-J.; Voit, B. Analysis of environmental microplastics by vibrational microspectroscopy: FTIR, Raman or both? *Anal. Bioanal. Chem.* **2016**, *408*, 8377–8391. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Cao, Y.; Wang, Q.; Ruan, Y.; Wu, R.; Chen, L.; Zhang, K.; Lam, P.K. Intra-day microplastic variations in wastewater: A case study of a sewage treatment plant in Hong Kong. *Mar. Pollut. Bull.* **2020**, *160*, 111535. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Jani, V.; Wu, S.; Venkiteshwaran, K. Advancements and regulatory situation in microplastics removal from wastewater and drinking water: A comprehensive review. *Microplastics* **2024**, *3*, 98–123. [\[CrossRef\]](#)
17. Xia, F.; Wang, Y.; Wang, D.; Cai, Y.; Zhang, J. Seasonal pulse effect of microplastics in the river catchment-From tributary catchment to mainstream. *J. Environ. Manag.* **2023**, *342*, 118316. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Sun, X.; Jia, Q.; Ye, J.; Zhu, Y.; Song, Z.; Guo, Y.; Chen, H. Real-time variabilities in microplastic abundance and characteristics of urban surface runoff and sewer overflow in wet weather as impacted by land use and storm factors. *Sci. Total Environ.* **2023**, *859*, 160148. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Ahmed, M.E.; Khajah, M.; Abdullah, H.; Al-Yaseen, R. Treatment of Industrial Wastewater of Variable Quality Using Ultra-sound Irradiation. *Civ. Eng. J.* **2025**, *11*, 1476–1491. [\[CrossRef\]](#)
20. Chen, H.; Jia, Q.; Sun, X.; Zhou, X.; Zhu, Y.; Guo, Y.; Ye, J. Quantifying microplastic stocks and flows in the urban agglomeration based on the mass balance model and source-pathway-receptor framework: Revealing the role of pollution sources, weather patterns and environmental management practices. *Water Res.* **2022**, *224*, 119045. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Bentes, I.; Silva, D.; Vieira, C.; Matos, C. Inflow quantification in urban sewer networks. *Hydrology* **2022**, *9*, 52. [\[CrossRef\]](#)
22. Kumar, R.; Sharma, P.; Verma, A.; Jha, P.K.; Singh, P.; Gupta, P.K.; Chandra, R.; Prasad, P.V. Effect of physical characteristics and hydrodynamic conditions on transport and deposition of microplastics in riverine ecosystem. *Water* **2021**, *13*, 2710. [\[CrossRef\]](#)
23. Zhang, B.; Huang, S.; Wu, L.; Guo, Y.; Shi, W.; Lens, P.N. Micro (nano) plastic size and concentration co-differentiate the treatment performance and toxicity mechanism in aerobic granular sludge systems. *Chem. Eng. J.* **2023**, *457*, 141212. [\[CrossRef\]](#)
24. Ahmadi, P.; Elagami, H.; Dichgans, F.; Schmidt, C.; Gilfedder, B.S.; Frei, S.; Peiffer, S.; Fleckenstein, J.H. Systematic evaluation of physical parameters affecting the terminal settling velocity of microplastic particles in lakes using CFD. *Front. Environ. Sci.* **2022**, *10*, 875220. [\[CrossRef\]](#)
25. Rosal, R. Morphological description of microplastic particles for environmental fate studies. *Mar. Pollut. Bull.* **2021**, *171*, 112716. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Hosseinzadeh, A.; Amirkhani, F.; Azizi, N.; Dashti, A.; Zhou, J.L.; Altaee, A. Machine learning modeling of microplastics removal by coagulation in water and wastewater treatment. *J. Water Process Eng.* **2025**, *76*, 108108. [\[CrossRef\]](#)
27. Spogis, N.; Neves, J.; Santos, S.L.B.S.; Padoin, N.; Soares, C. Machine Learning and Deep Learning Applied to Wastewater Treatment. In *Wastewater Treatment Plants: Processes, Assessment, Design and Operation*; Springer Nature: Cham, Switzerland, 2025; pp. 607–637. [\[CrossRef\]](#)

28. Meyers, N.; Everaert, G.; Hostens, K.; Schmidt, N.; Herzke, D.; Fuda, J.-L.; Janssen, C.R.; De Witte, B. Towards reliable data: Validation of a machine learning-based approach for microplastics analysis in marine organisms using Nile red staining. *Mar. Pollut. Bull.* **2024**, *207*, 116804. [[CrossRef](#)] [[PubMed](#)]
29. Chen, L.; Wang, J.; Zhu, M.; He, R.; Mu, H.; Ren, H.; Wu, B. Quality evaluation parameter and classification model for efflu-ents of wastewater treatment plant based on machine learning. *Water Res.* **2025**, *268*, 122696. [[CrossRef](#)] [[PubMed](#)]
30. Zhen, Y.; Wang, L.; Sun, H.; Liu, C. Prediction of microplastic abundance in surface water of the ocean and influencing fac-tors based on ensemble learning. *Environ. Pollut.* **2023**, *331*, 121834. [[CrossRef](#)] [[PubMed](#)]
31. Fazil, A.Z.; Gomes, P.I.; Sandamal, R.K. Applicability of machine learning techniques to analyze Microplastic transportation in open channels with different hydro-environmental factors. *Environ. Pollut.* **2024**, *357*, 124389. [[CrossRef](#)] [[PubMed](#)]
32. Yang, C.; Xie, J.; Gowen, A.; Xu, J.-L. Machine learning driven methodology for enhanced nylon microplastic detection and characterization. *Sci. Rep.* **2024**, *14*, 3464. [[CrossRef](#)] [[PubMed](#)]
33. Khanam, M.M.; Uddin, M.K.; Kazi, J.U. Advances in machine learning for the detection and characterization of microplas-tics in the environment. *Front. Environ. Sci.* **2025**, *13*, 1573579. [[CrossRef](#)]
34. Huang, J.; Tuo, J.; Wang, L.; Liu, J. Abundance, characteristics and seasonal variation of microplastics in a domestic sewage treatment plant in Nanjing, China. *J. Water Process Eng.* **2023**, *55*, 104200. [[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.