Supplementary Materials

Statistical outlier removal in high-frequency river water-quality data: effects on total phosphorus virtual sensing and on the retention of rainfall-driven events

This document provides the supporting information referenced in the main text. Table S1a reports the sensitivity of each detection method to its principal setting, Table S1b reports the five-fold TimeSeriesSplit cross-validation results used to quantify the uncertainty of the comparison, and Table S1c lists the complete set of parameters, thresholds and software versions required to reproduce the analysis.

Table S1a. Sensitivity of the number and proportion of detected outliers to the principal setting of each detection method. All thresholds were estimated from the training period only and applied to the full record. The setting adopted in the main analysis is indicated in each group.

Method	Setting	Outliers detected (n)	Removal rate (%)
IQR	multiplier=1.0	4,809	28.19
IQR	multiplier=1.5 (adopted)	3,503	20.53
IQR	multiplier=2.0	2,619	15.35
IQR	multiplier=3.0	1,905	11.17
Robust Mahalanobis	mean+2.0SD	243	1.42
Robust Mahalanobis	mean+2.5SD	192	1.13
Robust Mahalanobis	mean+3.0SD (adopted)	158	0.93
Robust Mahalanobis	mean+3.5SD	134	0.79
Robust Mahalanobis	mean+4.0SD	120	0.70
Robust Mahalanobis	chi-square α=0.05 (3.3272)	6,246	36.61
Robust Mahalanobis	chi-square α=0.025 (3.5822)	5,851	34.30
Robust Mahalanobis	chi-square α=0.01 (3.8841)	5,360	31.42
Robust Mahalanobis	chi-square α=0.001 (4.5293)	4,634	27.16
Isolation Forest	contamination=0.03	423	2.48
Isolation Forest	contamination=0.05	704	4.13
Isolation Forest	contamination=0.1 (adopted)	1,550	9.09
Isolation Forest	contamination=0.15	2,434	14.27
Isolation Forest	contamination=0.20	3,373	19.77
Isolation Forest	contamination='auto'	1,766	10.35

The Z-score criterion is not included in this table because the $|z| > 3$ threshold was held fixed at that value in this study and was not varied in the sensitivity analysis. The robust Mahalanobis distances of the training period have a median of 2.49, close to the 2.09 expected of the square root of a chi-square variate with five degrees of freedom, but their 95th percentile is 31.05 against an expected 3.33 and their standard deviation of 46.40 is more than four times their mean of 10.34.

Table S1b. Five-fold TimeSeriesSplit cross-validation results for each combination of detection method and prediction model. The outlier detection step was refitted within each training fold. Values are the mean and standard deviation across the five folds. RF, Random Forest; GB, Gradient Boosting.

Method	Model	Ratio (mean)	Ratio (SD)	MAE mean (mg/L)	MAE SD (mg/L)	RMSE mean (mg/L)	RMSE SD (mg/L)
Raw	RF	0.850	0.451	0.0188	0.0096	0.0428	0.0212
Raw	GB	0.822	0.430	0.0187	0.0096	0.0406	0.0220
IQR	RF	0.716	0.325	0.0193	0.0102	0.0416	0.0229
IQR	GB	0.683	0.287	0.0190	0.0105	0.0407	0.0237
Z-score	RF	0.876	0.464	0.0192	0.0090	0.0443	0.0199
Z-score	GB	0.845	0.447	0.0191	0.0090	0.0423	0.0208
Robust Mahalanobis	RF	0.853	0.448	0.0188	0.0095	0.0429	0.0209
Robust Mahalanobis	GB	0.818	0.436	0.0187	0.0097	0.0408	0.0220
Isolation Forest	RF	0.931	0.536	0.0204	0.0085	0.0490	0.0198
Isolation Forest	GB	1.007	0.618	0.0223	0.0082	0.0532	0.0194

Ratio denotes the ratio of the sum of predicted values to the sum of observed values on the validation fold. The fold-to-fold standard deviations are large relative to the differences between methods; the ordering obtained from the single chronological split reported in the main text should therefore be read as indicative rather than as a statistically separated ranking.

Table S1c. Complete parameter, threshold and software information for the analysis reported in the main text.

Category	Item	Setting
Data source	Data record	Hourly sensor records from the K-water automatic water-quality monitoring station, Hoeincheon watershed (January 2023 – December 2024)
Instrumentation	Sensors	YSI EXO2 multiparameter sonde (YSI Inc., USA): TEMP, COND, TURBIDITY, DO, pH
Instrumentation	TP analyzer	RoboChem A4-TP (Centennial Technology, Republic of Korea)
Instrumentation	Calibration	Weekly routine inspection at the station**; individual calibration records are retained by the operating agency as part of routine station maintenance and were not obtained for this study**
Instrumentation	Rainfall	Rain gauge; hourly recorded amount (mm)
Preprocessing	Temporal resolution	Exact hourly records only
Preprocessing	Extreme TP filter	TP ≥ 4 mg/L removed before physical QC (48 records)
Physical QC	Sensor no-response	Records with a zero value in COND, TURBIDITY, DO or pH removed (326 records); TEMP = 0 retained as physically attainable
Physical QC	Turbidity	TURBIDITY < 0 NTU removed (111 records)
Physical QC	Rainfall	Rainfall > 150 mm removed (1 record); threshold set above the national 1-h maximum on record
Physical QC	Conductivity	No lower or upper limit imposed; low conductivity values are within the sensor measurement range and were left to the statistical detection stage
Target handling	TP = 0	Treated as below the detection limit; retained in the dataset but excluded from RF/GB fitting and evaluation
Features	Model predictors	TEMP, COND, TURBIDITY, DO, pH
Features	Rainfall	Excluded from outlier detection and from RF/GB predictors; used only for rainfall-event retention diagnostics
Split	Method	Chronological 80/20 split
Split	Cut-off	2024-08-09 19:00
Split	Training records	13,647
Split	Test records	3,412 (3,367 used after excluding TP = 0)
Outlier detection	IQR	Tukey multiplier = 1.5; thresholds from training data only
Outlier detection	Z-score	$\|z\| > 3$; mean and SD from training data only
Outlier detection	Robust Mahalanobis	Minimum Covariance Determinant; threshold = training mean robust distance + 3 SD
Outlier detection	Isolation Forest	n_estimators = 200; contamination = 0.10; random_state = 42; fitted on training data only
Prediction model	Random Forest	n_estimators = 90; method-specific max_depth (30–70) and criterion; random_state = 42; n_jobs = -1

Category	Item	Setting
Prediction model	Gradient Boosting	n_estimators = 200; max_depth = 3; learning_rate = 0.05; random_state = 42
Uncertainty	Cross-validation	5-fold TimeSeriesSplit; outlier detection refitted within each training fold
Uncertainty	Paired comparison	Circular moving block bootstrap of the paired differences in absolute error against the raw baseline; 100,000 resamples for the main analysis and 20,000 for the sensitivity checks; random_state = 42, repeated with 777; block length selected per comparison by the automatic rule of Politis and White [74] (11-141 observations, blocked in observation order); Holm adjustment over the eight comparisons
Event diagnostics	Rainfall event	24-h cumulative rainfall ≥ 5 mm; 72-h influence window
Event diagnostics	High-TP threshold	TP q95 = 0.0817 mg/L (training period only)
Event diagnostics	High-turbidity threshold	TURBIDITY q95 = 24.68 NTU (training period only)
Software	Python	3.9.7
Software	pandas / numpy	1.5.3 / 1.23.5
Software	scipy / scikit-learn	1.13.1 / 1.3.2
Software (paired comparison)	Python / numpy / pandas / scipy / arch	3.13.16 / 2.5.3 / 3.0.5 / 1.18.1 / 8.0.0

Random seeds were fixed at 42 for every stochastic component, and the paired comparison was repeated with 777 as a reproducibility check. The software versions listed above apply to the detection and prediction analysis, except for the final row, which is the separate environment in which the paired comparison was computed. The chronological split, the detection thresholds and the model fits are therefore fully reproducible from the settings listed above.

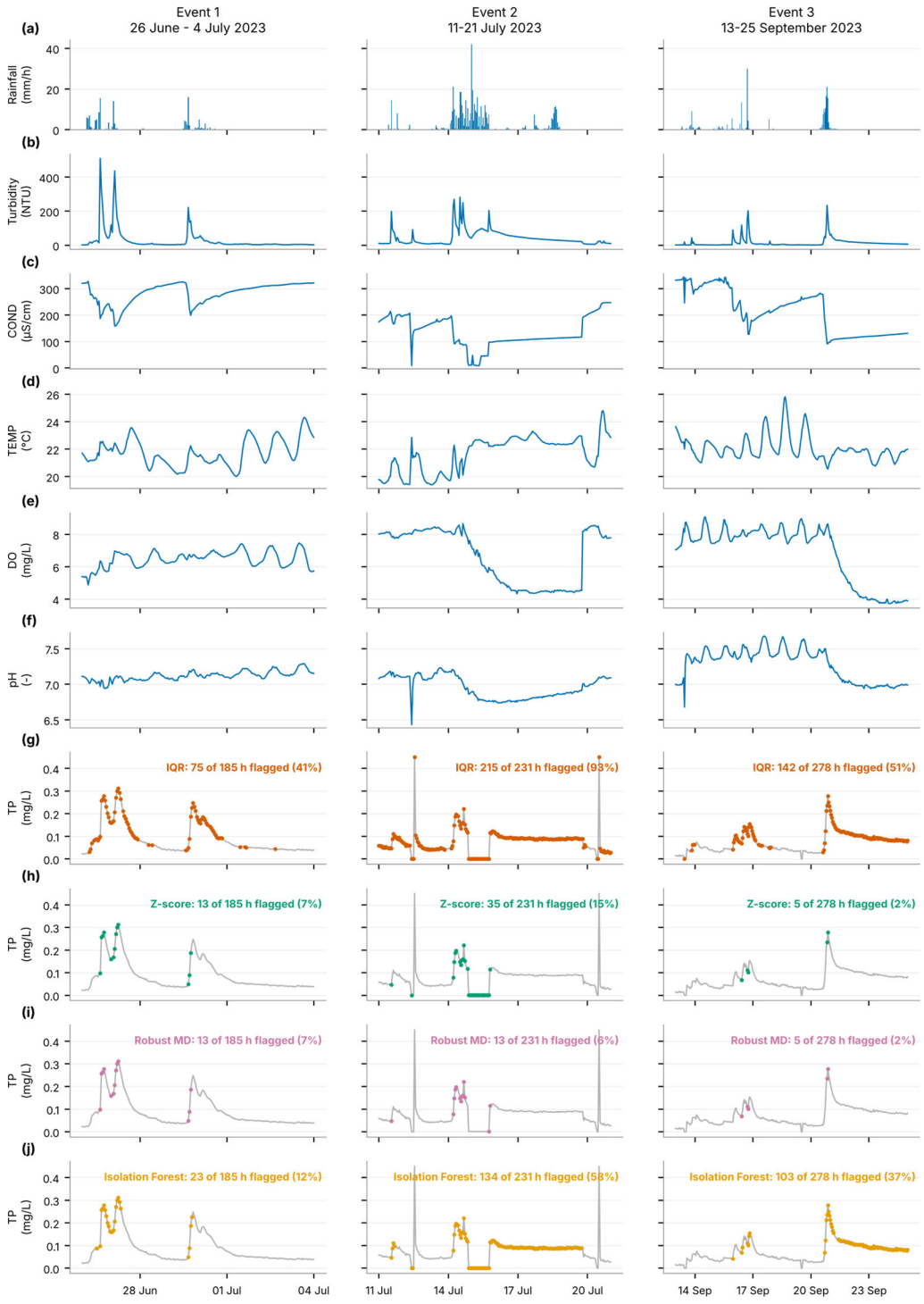


Figure S1. Observations flagged by each outlier detection criterion during the three representative rainfall events of Figure 4, shown together with all five input variables: (a) hourly rainfall; (b) turbidity; (c) electrical conductivity; (d) water temperature; (e) dissolved oxygen; (f) pH; (g-j) the TP series (gray) with the observations flagged by IQR, the Z-score, the robust Mahalanobis distance and Isolation Forest shown as colored points. Water temperature, dissolved oxygen and pH are dominated by seasonal and diurnal cycles; under the two univariate criteria they account for none of the observations flagged during these events, apart from one observation of Event 2 in which pH also lies outside its fence (Table S2).

Table S2. Variables responsible for the observations flagged during the three rainfall events of Figure 4. Each entry is the number of flagged observations for which that variable lay outside the criterion; an observation may exceed the criterion in more than one variable, so the entries can sum to more than the number flagged. The robust Mahalanobis distance and Isolation Forest evaluate the five variables jointly, so variable-specific threshold exceedances are not defined for them (n.d.).

Event	Method	Observations (h)	Flagged (n)	TEMP	COND	TURBIDITY	DO	pH
Event 1	IQR	185	75	0	5	75	0	0
Event 1	Z-score	185	13	0	0	13	0	0
Event 1	Robust MD	185	13	n.d.	n.d.	n.d.	n.d.	n.d.
Event 1	Isolation Forest	185	23	n.d.	n.d.	n.d.	n.d.	n.d.
Event 2	IQR	231	215	0	167	215	0	1
Event 2	Z-score	231	35	0	22	13	0	0
Event 2	Robust MD	231	13	n.d.	n.d.	n.d.	n.d.	n.d.
Event 2	Isolation Forest	231	134	n.d.	n.d.	n.d.	n.d.	n.d.
Event 3	IQR	278	142	0	111	133	0	0
Event 3	Z-score	278	5	0	0	5	0	0
Event 3	Robust MD	278	5	n.d.	n.d.	n.d.	n.d.	n.d.
Event 3	Isolation Forest	278	103	n.d.	n.d.	n.d.	n.d.	n.d.

Table S3. The 48 observations with TP ≥ 4 mg/L removed before the analysis, summarized by run of consecutive hours. Rainfall-influenced (n) is the number of observations in the run that fall within a rainfall-influenced period as defined in Section 3.4. The highest TP value among the 17,059 retained observations is 0.585 mg/L, and no observation falls between 0.585 and 4.561 mg/L. The record of all 48 observations is available from the corresponding author on request. When the cluster is retained, the criteria evaluated in this study flag 6 of the 48 observations under IQR and none under the Z-score, the robust Mahalanobis distance or Isolation Forest; the robust Mahalanobis distance of these observations reaches at most 32.43, against the threshold of 149.54 adopted in Section 2.2.2.

Run	Period	Duration (h)	TP (mg/L)	Turbidity (NTU)	COND (µS/cm)	Rainfall-influenced (n)
1	22 Nov 2024 18:00 – 23 Nov 12:00	19	9.000–9.363	1.60–5.72	359.2–392.9	0
2	23 Nov 15:00 – 17:00	3	4.860–4.883	3.02–3.28	382.5–455.1	0
3	24 Nov 14:00 – 25 Nov 09:00	20	4.561–9.326	1.70–7.44	357.2–445.3	0
4	25 Nov 11:00	1	9.354	1.76	352.0	0
5	25 Nov 14:00 – 15:00	2	9.360–9.406	4.09–10.24	399.3–426.5	0
6	26 Nov 11:00 – 13:00	3	9.301–9.318	15.11–25.95	303.0–337.3	3

Table S4. Prediction performance and event retention when the 48 observations with TP ≥ 4 mg/L are retained instead of removed, compared with the values reported in the main text. All 48 belong to the test period, so the training data and the detection thresholds are identical in the two cases; the common test set contains 3,367 observations when the cluster is removed and 3,415 when it is retained. The ordering of the methods is given by point estimates for this evaluation period and was not tested for statistical significance.

Method	Model	MAE removed (mg/L)	MAE retained (mg/L)	Ratio removed	Ratio retained	Event retention removed (%)	Event retention retained (%)
Raw	RF	0.0257	0.1273	0.501	0.160	100.00	100.00
Raw	GB	0.0264	0.1280	0.458	0.146	100.00	100.00
IQR	RF	0.0271	0.1287	0.467	0.149	1.99	1.99
IQR	GB	0.0281	0.1297	0.430	0.137	1.99	1.99
Z-score	RF	0.0242	0.1258	0.534	0.170	80.57	80.65
Z-score	GB	0.0260	0.1276	0.468	0.149	80.57	80.65
Robust Mahalanobis	RF	0.0257	0.1273	0.500	0.160	84.31	84.37
Robust Mahalanobis	GB	0.0263	0.1279	0.459	0.147	84.31	84.37
Isolation Forest	RF	0.0261	0.1277	0.498	0.159	28.52	28.78
Isolation Forest	GB	0.0264	0.1280	0.464	0.148	28.52	28.78

Table S5. Performance when each detection criterion is applied to the test period as well as to the training period, so that each condition is scored only on the test observations it retains. This is the operational scoring referred to in Section 4; the values in the main text are obtained on the common test set of 3,367 observations. n is the number of test observations on which each condition is scored. The ordering is given by point estimates for this evaluation period and was not tested for statistical significance.

Method	Model	n	Predicted/observed ratio	MAE (mg/L)	RMSE (mg/L)
Raw	RF	3,367	0.501	0.0257	0.0663
IQR	RF	2,787	0.493	0.0236	0.0603
Z-score	RF	3,354	0.534	0.0240	0.0658
Robust Mahalanobis	RF	3,357	0.498	0.0255	0.0663
Isolation Forest	RF	3,191	0.495	0.0254	0.0675
Raw	GB	3,367	0.458	0.0264	0.0662
IQR	GB	2,787	0.455	0.0243	0.0604
Z-score	GB	3,354	0.467	0.0257	0.0658
Robust Mahalanobis	GB	3,357	0.456	0.0262	0.0661
Isolation Forest	GB	3,191	0.456	0.0259	0.0673

Table S6. Sensitivity of the Isolation Forest results to the contamination setting, including the independently specified rule contamination = 'auto'.

Setting	Removal rate (%)	RF MAE (mg/L)	GB MAE (mg/L)	Event retention (%)	CV RF MAE (mg/L)	CV GB MAE (mg/L)
contamination = 0.03	2.48	0.0257	0.0260	73.60	0.0186 ± 0.0065	0.0168 ± 0.0068
contamination = 0.05	4.13	0.0253	0.0258	55.42	0.0184 ± 0.0068	0.0176 ± 0.0074
contamination = 0.10 (adopted)	9.09	0.0261	0.0264	28.52	0.0194 ± 0.0063	0.0188 ± 0.0063
contamination = 0.15	14.27	0.0260	0.0263	18.56	0.0218 ± 0.0089	0.0214 ± 0.0083
contamination = 0.20	19.77	0.0261	0.0263	12.70	0.0232 ± 0.0090	0.0219 ± 0.0098
contamination = 'auto'	10.35	0.0261	0.0266	25.03	0.0214 ± 0.0082	0.0210 ± 0.0076

Settings are as listed in Table S1c; n_estimators = 200 and random_state = 42 throughout. Under contamination = 'auto' the fitted offset was -0.5 and an observation was classified as an outlier when its score fell below that value, so the removal rate was determined by the observed score distribution and not prescribed by the analyst. Event retention is the percentage of the 803 rainfall-influenced high-concentration observations that remained. The two cross-validation columns are the mean and standard deviation of five TimeSeriesSplit folds computed on the training period alone, with removal applied to the fitting part of each fold only and the validation part of each fold common to all settings. Table S1b places its folds over the whole screened record, so the values in the two tables are not directly comparable.

Table S7. Paired comparison of the mean absolute error of each preprocessing condition with that of the raw baseline, on the common test set of 3,367 observations. RF, Random Forest; GB, Gradient Boosting.

Model	Method	ΔMAE (mg/L)	95% CI (mg/L)	Block length (observations)	p	Adjusted p	Adjusted p, gap-confined blocks
RF	IQR	+0.00140	+0.00050 to +0.00247	63	0.007	0.033	0.100
RF	Z-score	-0.00147	-0.00286 to -0.00027	141	0.025	0.10	0.155
RF	Robust Mahalanobis	+0.00001	-0.00022 to +0.00027	24	0.906	1.000	1.000
RF	Isolation Forest	+0.00038	+0.00014 to +0.00065	13	0.004	0.022	0.023
GB	IQR	+0.00173	+0.00086 to +0.00279	63	0.002	0.011	0.027
GB	Z-score	-0.00039	-0.00058 to -0.00019	22	<0.001	0.002	0.002
GB	Robust Mahalanobis	-0.00006	-0.00019 to +0.00007	11	0.395	1.000	1.000
GB	Isolation Forest	+0.00002	-0.00054 to +0.00047	23	0.943	1.000	1.000

ΔMAE is the mean absolute error of the condition minus that of the raw baseline, so a positive value denotes a larger error than the baseline. The paired differences in absolute error are serially correlated, with lag-1 correlations over adjacent hours between 0.24 and 0.83 across the eight comparisons, so the difference in mean absolute error was tested with a paired circular moving-block bootstrap of the difference series (100,000 resamples; random_state = 42, repeated with 777). The blocks are contiguous in the order of the retained observations and not in clock time, so the block length is a number of observations and a block can span the 27 interruptions of the hourly sequence that fall within the test period; the final column repeats the test with the blocks confined to uninterrupted stretches and is the diagnostic for that approximation. The block length was selected separately for each comparison from the estimated serial dependence of that series by the automatic rule of Politis and White, with the correction of Patton et al. [74], as implemented in the arch package, version 8.0.0. The 95% CI is the unadjusted percentile interval of the bootstrap distribution and p is the unadjusted two-sided bootstrap p-value, so an interval that excludes zero together with an adjusted p above 0.05 is not a contradiction. Repeating the analysis with the second seed left every significance decision unchanged. At fixed block lengths of 6 to 72 observations the same five comparisons are significant after adjustment, so the RF Z-score comparison is significant there although it is not at its selected length of 141; at 168 observations only the GB IQR comparison remains significant (adjusted p = 0.031). A Newey-West heteroskedasticity- and autocorrelation-consistent normal test of the same differences reproduces these decisions at bandwidths of 24 to 72 observations. The two-sided Wilcoxon signed-rank test on the same pairs gives p = 0.067 (RF) and p < 0.001 (GB) for IQR, p < 0.001 for both models under the Z-score, p < 0.001 (RF) and p = 0.003 (GB) for the robust Mahalanobis distance, and p = 0.278 (RF) and p = 0.250 (GB) for Isolation Forest; it does not account for the serial dependence of the paired differences and is listed for reference only.