Electronic Supplementary Material (ESI) for Environmental Science: Water Research & Technology.
This journal is © The Royal Society of Chemistry 2023

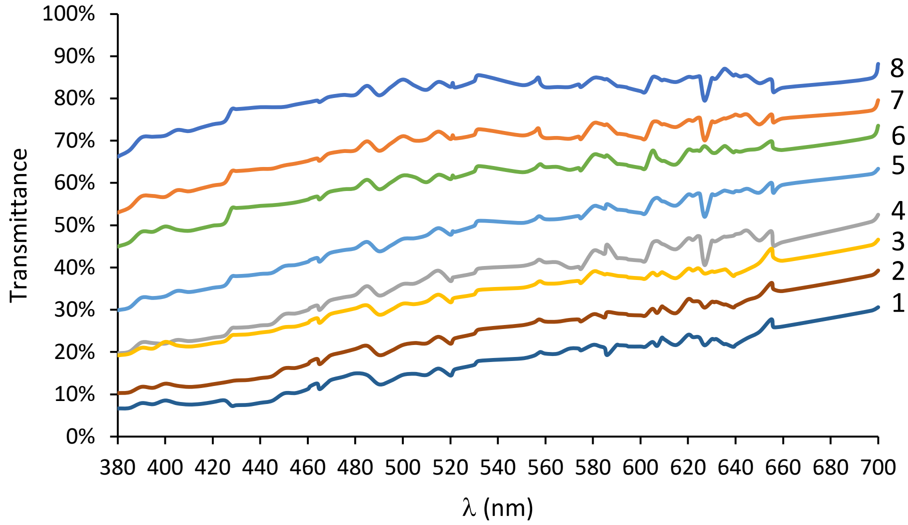
Supplementary Materials

Optimization of indirect wastewater characterization: A hybrid approach based on decision trees, genetic algorithms and spectroscopy

Daniel Carreres-Prieto[1*], Juan T. García[2*], *, José M. Carrillo[2] and Antonio Vigueras-Rodríguez[2]

1. Center for Technological Innovation in Construction and Civil Engineering (CITEEC), Universidade da Coruña, 15008, A Coruña, Spain
2. Department of Mining and Civil Engineering, Universidad Politécnica de Cartagena, 30202 Cartagena, Spain

* Correspondence: daniel.carreres@udc.es (D.C.-P.); juan.gbermejo@upct.es (J.T.G.)



Polluting parameters	Samples 1	2	3	4	5	6	7	8
COD (mg/l)	1842	1574	849	981	672	394	252	11
BOD_5 (mg/l)	150	340	540	660	400	280	215	4.3
TSS (mg/l)	1032	844	320	286	368	196	73	2.8
Phosphorus (P) (mg/l)	12.5	11.8	10.5	10.5	5.84	2.8	3.11	8.19
Total Nitrogen (TN) (mg/l)	86	86	86	96.3	42.9	41	31	15.9
NO_3-N (mg/l)	0.8	0.7	0.7	0	0	0.3	0	9.8
PH	7.6	7.38	7.6	7.96	7.95	7.73	8.17	7.61
Conductivity (μS/cm)	2500	2480	2400	1746	965	2200	979	878

Figure S1. Spectral response of eigth samples taken at random, accompanied by their laboratory characterization, relating to both raw wastewater samples and treated wastewater.

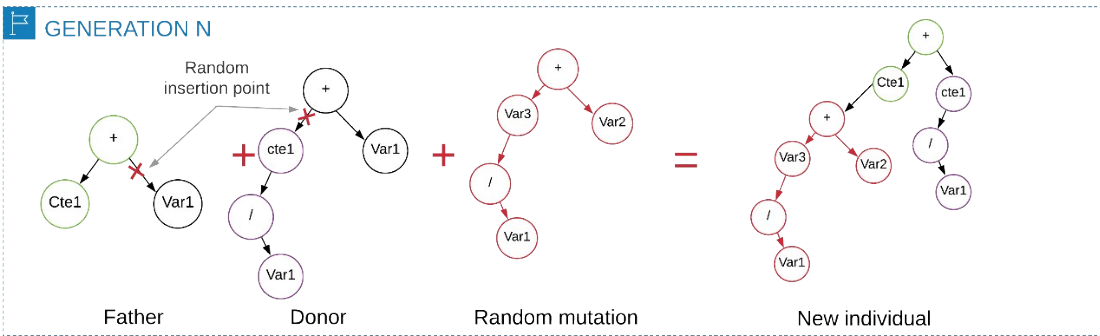


Figure S2. Simplified diagram of model generation based on genetic algorithms.

S1. Specific estimation models for raw wastewater

S1.2. Biochemical Oxygen Demand estimation models (BOD_5)

S.1.2.1. GA model based on global trend line

Equation (S1) shows the model for estimating BOD_5 from the overall trend line of the spectral response, calculated from 328 raw water samples, after removing outliers. The model presents a Pearson's coefficient of 66.36% for training and 51.47% for testing, reaching the optimum at generation 17 of 25, with a mutation rate of 20%.

$$BOD_{5(mg/l)} = \left(\frac{c_0}{c_1 * M_{Global} + c_2 * N_{Global}} - \left((c_3 * N_{Global})^2 - (c_4 * N_{Global} - c_5)\right)\right) + c_6$$
$$c_0 = 655{,}06;\ c_1 =- 149.35;\ c_2 =- 7.519;\ c_3 =- 5{,}434.49;\ c_4 =- 2{,}921.78; \quad \text{(S1)}$$
$$c_5 =- 1446{,}39;\ c_6 = 1.5566$$

S.1.2.2. GA model based on multiple individual trend lines for each colour group

Similar estimation results are achieved with the model of Equation (S2), relative to the trend lines of the different colour groups of the visible spectrum, with a Pearson's coefficient of 67.86% for training and 54.17% for testing, achieving the optimum in generation 16 of 25 and a mutation rate of 20%.

$$BOD_{5(mg/l)} = \frac{\left((c_0 * M_{Cyan} - c_1 * N_{Violet}) - (c_2 * N_{Blue} + c_3 * N_{Orange})\right)}{\left((c_4 * N_{Cyan} + c_5 * M_{Yellow}) - (c_6 + c_7 * M_{Violet})\right)} + c_8$$
$$c_0 = 3{,}120{,}040.98;\ c_1 = 196{,}000.47;\ c_2 =- 35{,}647.17;\ c_3 = 7{,}614.70; \quad \text{(S2)}$$
$$c_4 = 92.42;\ c_5 = 41.46;\ c_6 =- 5.10;\ c_7 = 1759.30;\ c_8 = 1753.2'$$

S.1.2.3. GA model based on point values of the spectral response

Equation (S3) shows the model for estimating BOD_5 from point values of the spectral response, achieving a Pearson's Coefficient of 70.77% for training and 53.02% for test.

$$BOD_{5(mg/l)} = \left(\frac{c_0 * A_{627} + c_1 * T_{627}}{c_2 * T_{450} - c_3 * T_{425}} - \left(\frac{c_4}{A_{627}} + (c_5 * T_{420} + c_6 * A_{627})\right)\right) + c_7$$
$$c_0 =- 942.30;\ c_1 = 1{,}384.47;\ c_2 =- 24.04;\ c_3 =- 35.44;\ c_4 =- 15.34; \quad \text{(S3)}$$
$$c_5 = 593.14;\ c_6 =- 1{,}347.66;\ c_7 = 8.20$$

S.1.2.4. Hybrid characterization model based on decision trees

Figure S3 shows the classification tree for the hybrid model of combined water characterization for BOD_5, with a R^2 of 61.05% for training and 67.86% for test.

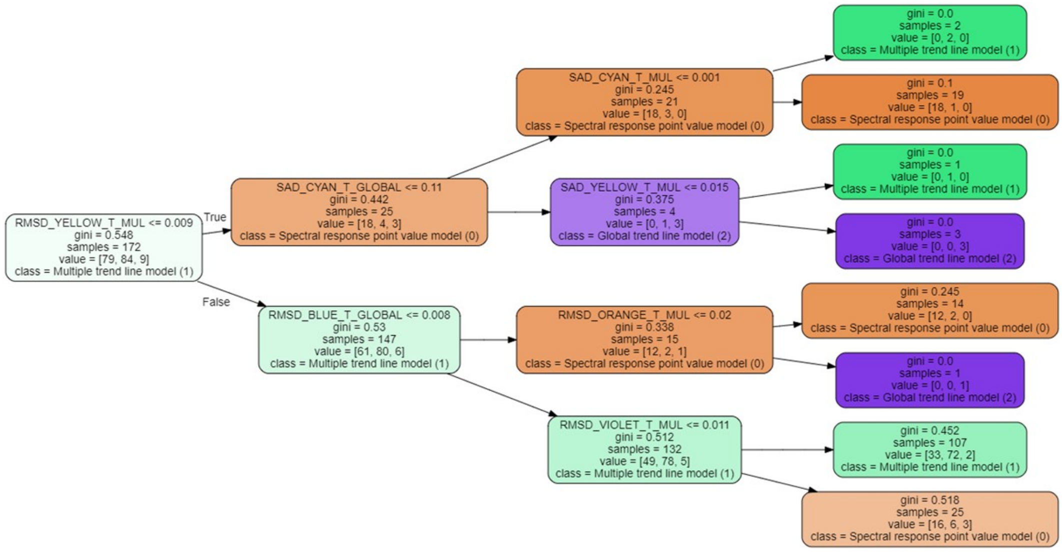
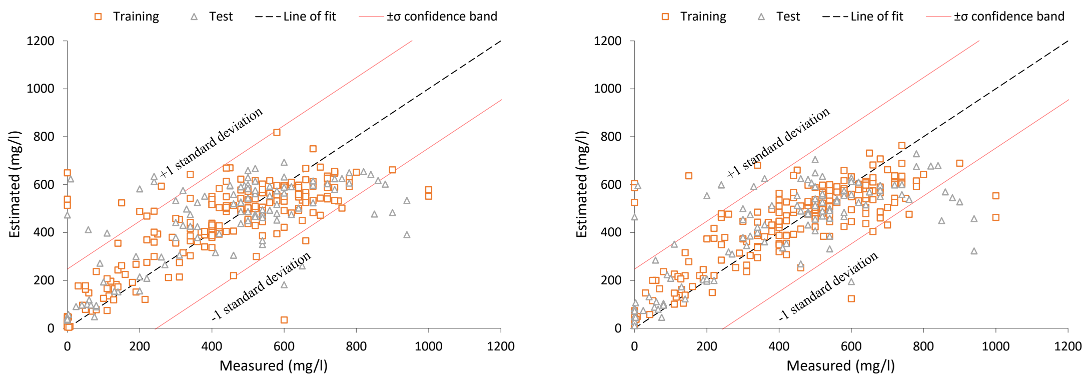
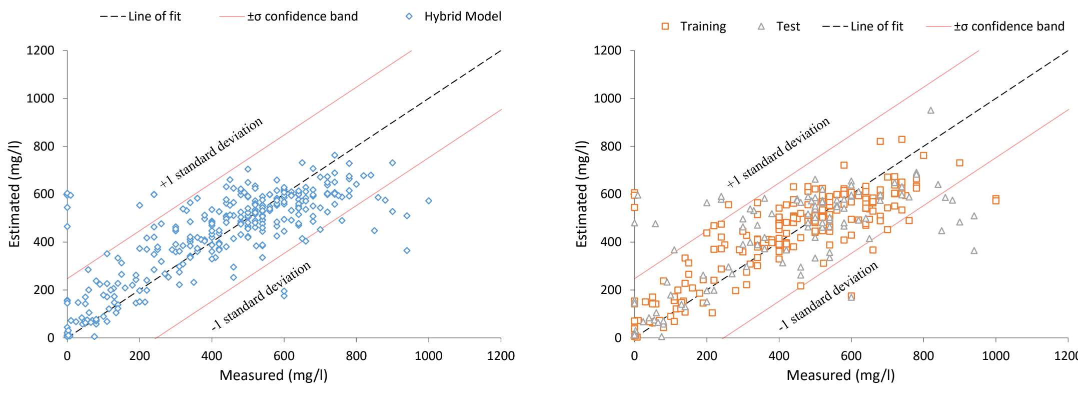


Figure S3. Classification tree for hybrid model of raw wastewater characterization for BOD_5.

Figure S4 shows the scatter plot of the estimation models based on trend lines (Equation (S1 and S2)), as well as the hybrid model of Figure S5. As can be seen, the scatter plot in Figure S4 C shows a lower dispersion of the data, which denotes an improvement in the ability to characterize the sample with respect to the exclusive use of other techniques.



A **B**



C **D**

Figure S4. Scatter plot between laboratory measured BOD_5 values (Measured) and those estimated by: (A) Global Model, Equation (S1). (B) Individual trend model, Equation (S2). (C) Hybrid model. (D) Model based on spectral point values by Offspring Selection technique, Equation (S3).

S1.3. Total Suspended Solids estimation models (TSS)

S.1.3.1. GA model based on global trend line

Equation (S4) shows the model for estimating the TSS from the global trend line, with a Pearson's coefficient of 61.90% for training and 70.37% for testing, from 294 samples after removing outliers.

$$TSS_{(mg/l)} = \left(\left((c_0 * N_{Global} + c_1) - (c_2 * N_{Global})^2 \right) - \frac{(c_3 * N_{Global})^2}{c_4 * N_{Global} + c_5 * M_{Global}} \right) + \quad \text{(S4)}$$

$$c_0 = 1099.27;\ c_1 = 450.29;\ c_2 = -3{,}369.09;\ c_3 = -8{,}385.07;\ c_4 = -3.0189;$$
$$c_5 = -110.05;\ c_6 = 3.4933$$

S.1.3.2. GA model based on multiple individual trend lines for each colour group

Similar estimation levels are achieved with the model of Equation (S5), to estimate TSS from the trend line of the different colour groups, with a Pearson's coefficient of 67.42% and 71.95% for training and test, respectively.

$$TSS_{(mg/l)} = \frac{\left((c_0 * N_{Blue} - c_1 * N_{Orange}) - (c_2 * N_{Green} + c_3) \right)}{(c_4 * N_{Green} + c_5) + (c_6 * N_{Green})^2} + c_7 \quad \text{(S5)}$$

$$c_0 = 132{,}974.44;\ c_1 = -44{,}412.66;\ c_2 = -379{,}809.60;\ c_3 = -70{,}644.03;$$
$$c_4 = 555.67;\ c_5 = 81.545;\ c_6 = -26.872;\ c_7 = -441.9$$

S.1.3.3. GA model based on point values of the spectral response

Equation (S6), shows the model for estimating TSS from point values of the spectral response, achieving a Pearson's Coefficient of 70.39% for training and 74.09% for test.

$$TSS_{(mg/l)} = \left(\frac{c_0 * T_{521} + c_1 * T_{630} + c_2 * T_{500} + c_3 * T_{642}}{(c_4 * T_{521} + c_6 * T_{630}) - (c_6 - c_7 * A_{574})} \right) + c_8$$

$$c_0 = 160{,}880.01;\ c_1 =- 714{,}652.13;\ c_2 = 171{,}926.03;\ c_3 = 424{,}116.73; \quad \text{(S6)}$$

$$c_4 =- 237.47;\ c_5 =- 519.34;\ c_6 =- 561.32;\ c_7 =- 864.11;\ c_8 = 275.01$$

S.1.3.4. Hybrid characterization model based on decision trees

Figure S5 shows the classification tree for the hybrid model of combined water characterization for TSS, with an R^2 for training of 59% and 72% for test, respectively.

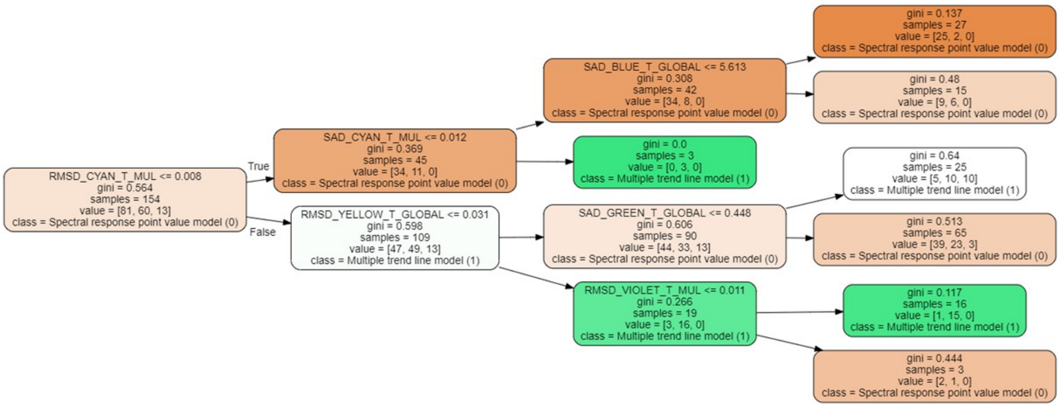
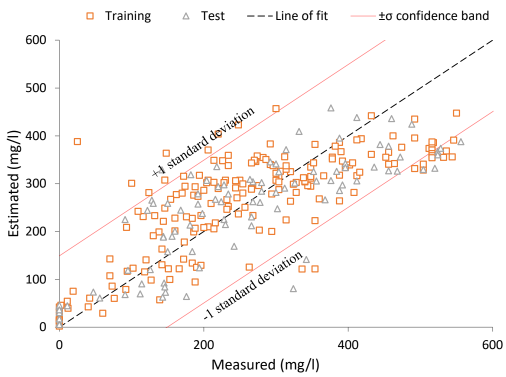
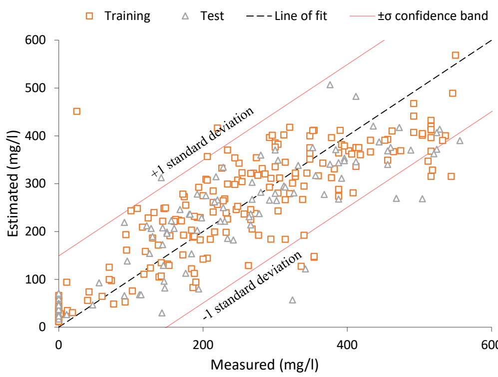


Figure S5. Classification tree for hybrid model of raw wastewater characterization for TSS.

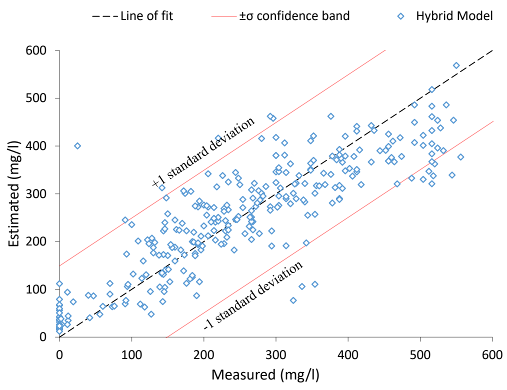
Figure S6 shows the scatter plot for each of the calculated models.



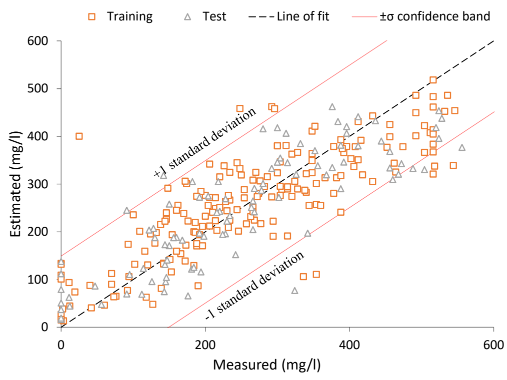
A



B



C



D

Figure S6. Scatter plot between laboratory measured TSS values (Measured) and those estimated by: (A) Global Model, Equation (S4). (B) Individual trend model, Equation (S5). (C) Hybrid model. (D) Model based on spectral point values by Offspring Selection technique, Equation (S6).

S1.4. Total Nitrogen estimation models (TN)

S.1.4.1. GA model based on global trend line

Equation (S7) shows the model for estimating TN estimates from the global trend line of the spectral response, calculated from 299 raw water samples, after removing outliers. The model shows a Pearson's coefficient of 60.48% for training and 52.08% for testing.

$$TN_{(mg/l)} = \frac{(c_0 * N_{Global})^2 + \frac{c_1}{N_{Global}}}{(c_3 * N_{Global})^2 - c_4 * M_{Global} * c_5} * c_6 + c_7 \quad \text{(S7)}$$

$$c_0 =- 12.85;\ c_1 = 0.60;\ c_3 = 4.85;\ c_4 = 58.70;\ c_5 =- 57.80;\ c_6 =- 14.315;$$

$$c_7 = 83.97$$

S.1.4.2. GA model based on multiple individual trend lines for each colour group

The model of Equation (S8), to estimate TN from the trend line of the different colour groups, achieves similar estimation levels, with a Pearson's coefficient of 68.12% and 53.86 % for training and test, respectively.

$$TN_{(mg/l)} = \frac{(c_0 * N_{Orange} * M_{Green} + (c_1 * N_{Cyan} + c_2))}{(c_3 * M_{Green} + (c_4 * N_{Orange} - c_5 * N_{Cyan}))} + c_6 \quad \text{(S8)}$$

$$c_0 =- 1{,}311{,}112.59;\ c_1 =- 34{,}075.30;\ c_2 = 11{,}736.60;\ c_3 = 19{,}704.98;$$
$$c_4 = 166.7;\ c_5 =- 67.63;\ c_6 = 86.407$$

S.1.4.3. GA model based on point values of the spectral response

Equation (S9) shows the model for estimating TN from point values of the spectral response, achieving a Pearson's Coefficient of 67.40% for training and 57.23% for test.

$$TN_{(mg/l)} = \left(\frac{(c_0 * T_{425} + c_1 * T_{468}) - (c_2 * T_{490} + c_3)}{(c_4 * A_{435} + c_5) - (c_6 * T_{558} + c_7 * T_{490})}\right) + c_8 \tag{S9}$$
$$c_0 = 4{,}918.08;\ c_1 = 40{,}612.12;\ c_2 = 41{,}828.26;\ c_3 = 834.67;$$
$$c_4 = 29.39;\ c_5 =- 34.63;\ c_6 = 132.51;\ c_7 =- 126.01;\ c_8 = 54.34$$

S.1.4.4. Hybrid characterization model based on decision trees

Figure S7 shows the classification tree for the hybrid model of combined water characterization for TN, with an R^2 for training of 66.67% and 73.61% for test, respectively.

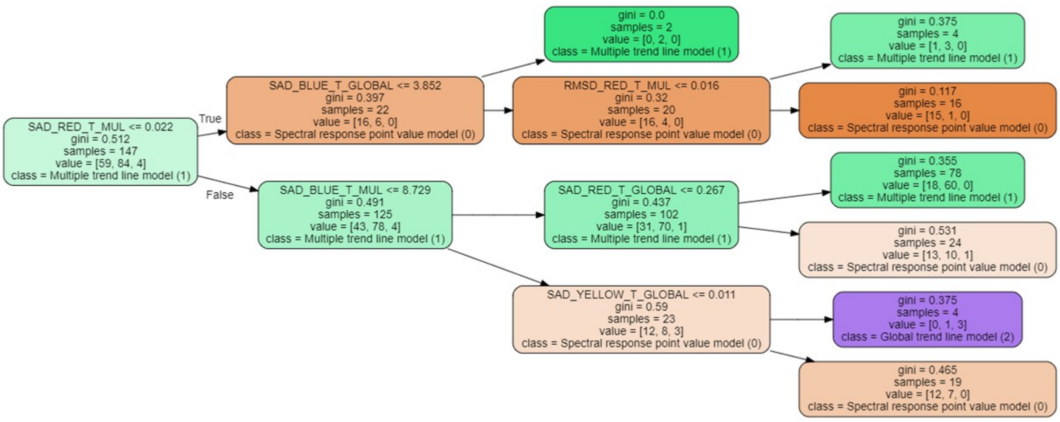


Figure S7. Classification tree for hybrid model of raw wastewater characterization for TN.

Figure S8 shows the scatter plot for each of the calculated models, where less dispersion is observed in the hybrid model (Figure S9 C).

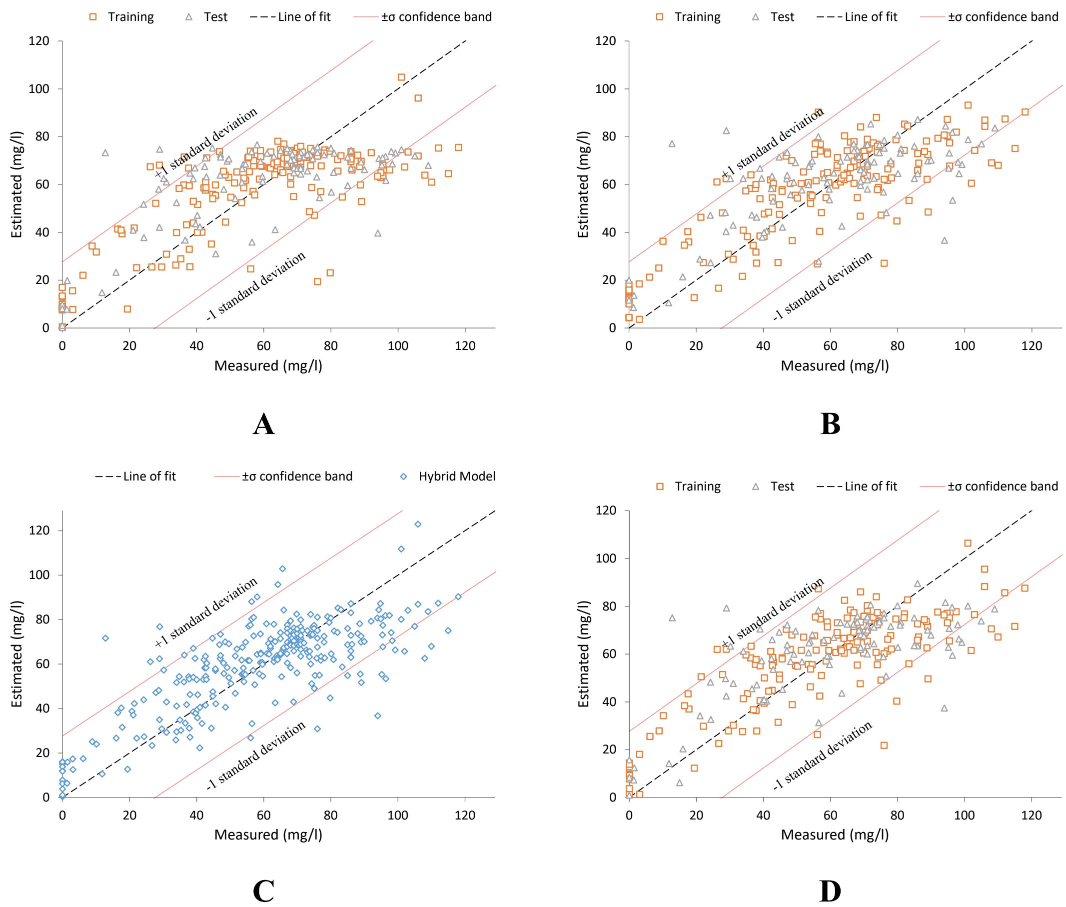


Figure S8. Scatter plot between laboratory measured TN values (Measured) and those estimated by: (A) Global Model, Equation (S7). (B) Individual trend model, Equation (S8). (C) Hybrid model. (D) Model based on spectral point values by Offspring Selection technique, Equation (S9).

S1.5. Total Phosphorus estimation models (TP)

S.1.5.1. GA model based on global trend line

The model to estimate TP from the global trend line of the spectral response, based on 304 raw water samples, after eliminating the outliers, is shown in equation (S10). The model shows a Pearson's coefficient of 54.77% for training and 61.07% for testing.

$$TP_{(mg/l)} = \left(\left(c_0 * N_{Global}\right)^2 * \left(c_1 * N_{Global} - c_2 * M_{Globa} \right.\right. * c_6 + c_7 \quad \text{(S10)}$$

$c_0 = 8.6897$; $c_1 = 0.2788$; $c_2 =- 22.37$; $c_3 = 4.7403$; $c_4 = 25.167$; $c_5 =- 26.233$; $c_6 = 2.5068$; $c_7 = 5.2719$

S.1.5.2. GA model based on multiple individual trend lines for each colour group

The similar estimation levels are achieved with the model of Equation (S10), for estimating TP using the trend line of the different colour groups, with a Pearson's coefficient of 59.05% and 57.46 % for training and test, respectively.

$$TP_{(mg/l)} = \frac{(c_0 + c_1 * N_{Green}) * (c_2 - c_3 * N_{Cyan})}{(c_4 * N_{Yellow} - c_5 * N_{Blue}) - (c_6 * N_{Cyan} - c_7 * N_{Orange})} + c_8 \quad \text{(S11)}$$

$$c_0 =- 12.00;\ c_1 =- 55.31;\ c_2 = 14.70;\ c_3 = 37.56;\ c_4 =- 7.37;\ c_5 = 45.13;$$
$$c_6 = 10.69;\ c_7 =- 8.71;\ c_8 = 8.47$$

S.1.5.3. GA model based on point values of the spectral response

Equation (S12) shows the model for estimating TP from point values of the spectral response, achieving a Pearson's Coefficient of 63.24% for training and 54.52% for test.

$$TP_{(mg/l)} = ((c_0 * T_{480} * T_{627} + (c_1 * T_{440} + c_2 * T_{385})) + c_6 \quad \text{(S12)}$$

$$c_0 = 25.47;\ c_1 = 50.94;\ c_2 = 88.31;\ c_3 = 84.46;\ c_4 =- 63.91;\ c_5 =- 195.47;$$
$$c_6 = 12.87$$

S.1.5.4. Hybrid characterization model based on decision trees

Figure S9 shows the classification tree for the hybrid model of combined water characterization for TP, with an R^2 for training of 65.82% and 72.73% for test, respectively.

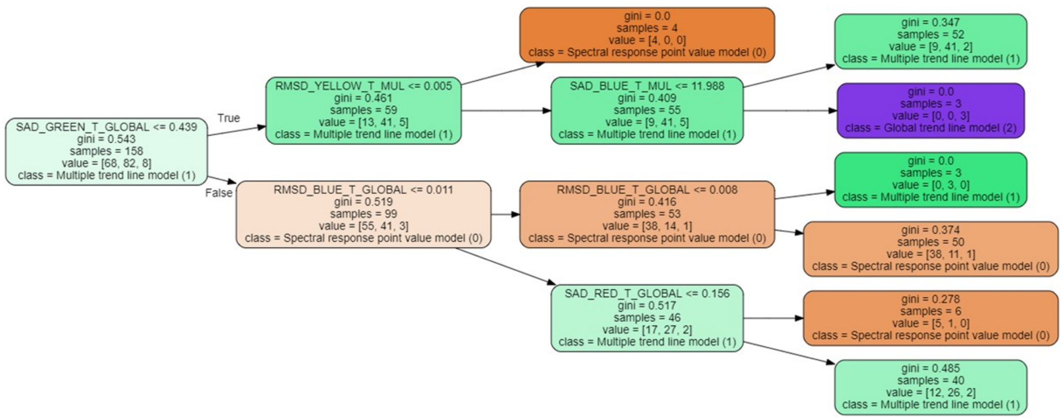


Figure S9. Classification tree for hybrid model of raw wastewater characterization for TP.

Figure S10 shows the scatter plot for each of the calculated models, where less dispersion is observed in the hybrid model (Figure S10 C).

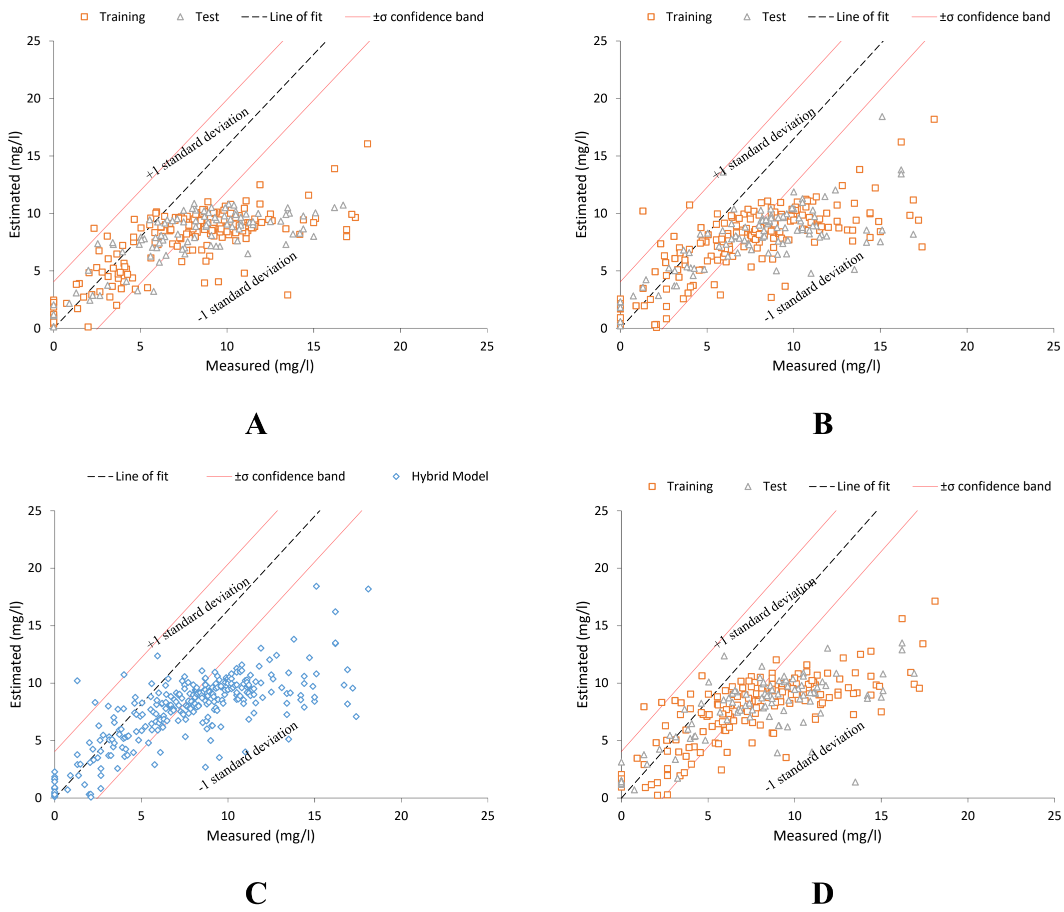


Figure S10. Scatter plot between laboratory measured TP values (Measured) and those estimated by: (A) Global Model, Equation (S10). (B) Individual trend model, Equation (S11). (C) Hybrid model. (D) Model based on spectral point values by Offspring Selection technique, Equation (S12).

S2. Specific estimation models for treated wastewater

S2.2. Biochemical Oxygen Demand estimation models (BOD_5)

S.2.2.1. GA model based on global trend line

Equation (S13) shows the model for estimating BOD_5 from the overall trend line of the spectral response, calculated from 279 treated water samples, after removing outliers. The model presents a Pearson's coefficient of 23.84% for training and 20.77% for testing.

$$BOD_{5(mg/l)} = \left(\left(c_0 * N_{Global}\right)^2 - \left(c_1 * N_{Global} + c_2\right)\right) * c_3 * + c_8 \quad \text{(S13)}$$

$$c_0 = -4.9436;\ c_1 = 25.73;\ c_2 = -6.8428;\ c_3 = 11.958;\ c_4 = -11.218;$$
$$c_5 = -14.478;\ c_6 = 14.461;\ c_7 = 63.438;\ c_8 = 1.8826$$

S.2.2.2. GA model based on multiple individual trend lines for each colour group

Similar estimation results are achieved with the model of Equation (S14), relative to the trend lines of the different colour groups of the visible spectrum, with a Pearson's coefficient of 23.13% for training and 41.74% for testing.

$$BOD_{5(mg/l)} = \left(c_0 * N_{Green} - c_1 * N_{Yelow}\right) * \left(c_2 * N_{Violet} - \right. + c_8 \quad \text{(S14)}$$

$$c_0 = 4.16;\ c_1 = -8.99;\ c_2 = 31.82;\ c_3 = 24.09;\ c_4 = -0.39;\ c_5 = -18.64;$$
$$c_6 = 3.24;\ c_7 = -51.89;\ c_8 = 3.3892$$

S.2.2.3. GA model based on point values of the spectral response

Equation (S15) shows the model for estimating BOD_5 from point values of the spectral response, achieving a Pearson's Coefficient of 45.68% for training and 26.29% for test.

$$BOD_{5(mg/l)} = \frac{\left(c_0 * T_{558} - c_1 * T_{555}\right) * \dfrac{c_2 * A_{586}}{A_{642}}}{\left(c_3 * T_{660} - c_4\right) + \left(c_5 * A_{586} - c_6 * T_{624}\right)} + c_7 \quad \text{(S15)}$$

$$c_0 = 38.46;\ c_1 = 29.26;\ c_2 = 4.10;\ c_3 = 3.50;\ c_4 = -11.94;\ c_5 = -18.59;$$
$$c_6 = 17.06$$

$$c_7 = -3.86;$$

S.2.2.4. Hybrid characterization model based on decision trees

Figure S11 shows the classification tree for the hybrid characterization model for treated water for BOD_5, with an R^2 for training of 61.27% and 66.21% for test, respectively.

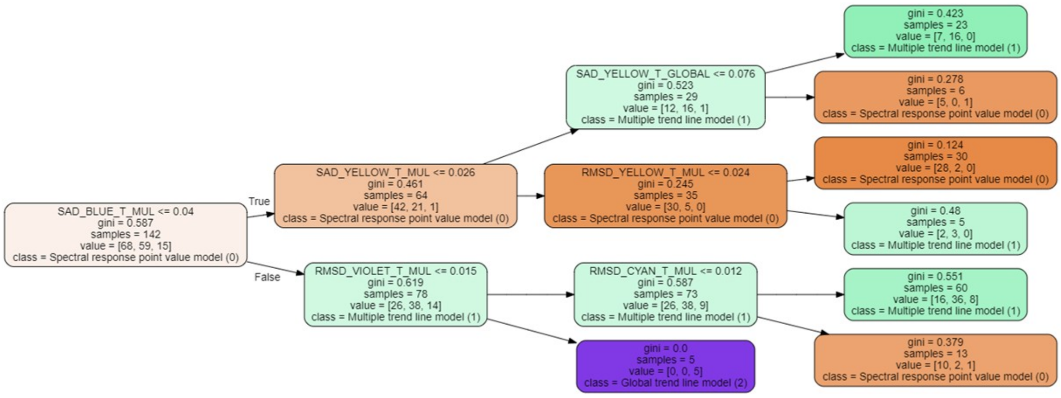
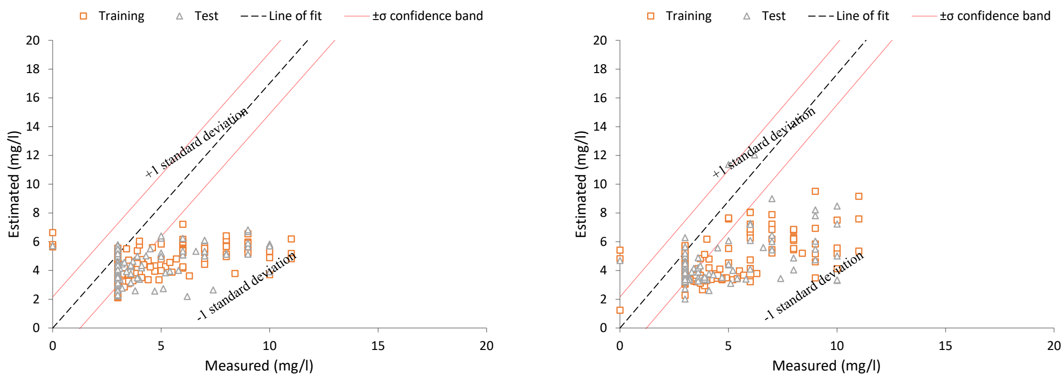
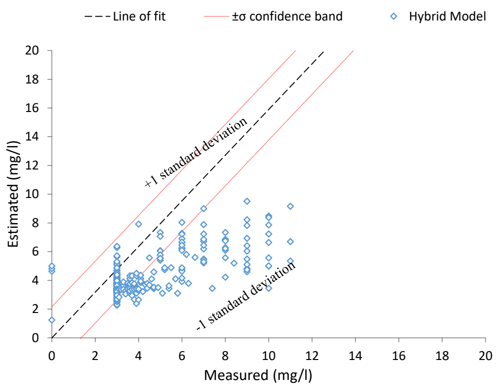


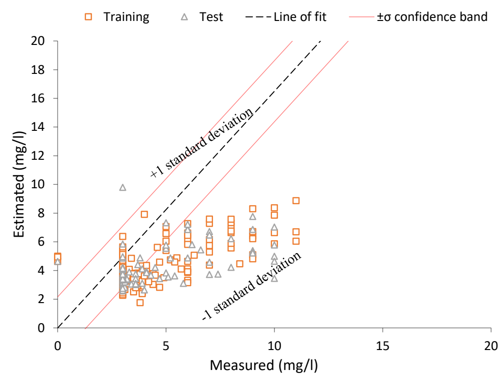
Figure S11. Classification tree for hybrid model of treated water characterization for BOD_5.

Figure S12 shows the scatter plot of the estimation models based on trend lines (Equation (S13 and S14)), as well as the hybrid model of Figure S11. As can be seen, the scatter plot in Figure S12 C shows a lower dispersion of the data, which denotes an improvement in the ability to characterize the sample with respect to the exclusive use of other techniques.



A **B**





C D

Figure S12. Scatter plot between laboratory measured BOD_5 values (Measured) and those estimated by: (A) Global Model, Equation (S13). (B) Individual trend model, Equation (S14). (C) Hybrid model. (D) Model based on spectral point values by Offspring Selection technique, Equation (S15).

S2.3. Total Suspended Solids estimation models (TSS)

S.2.3.1. GA model based on global trend line

Equation (S16) shows the model for estimating the TSS from the global trend line, with a Pearson's coefficient of 28.85% for training and 29.45% for testing.

$$TSS_{(mg/l)} = \left(c_0 * M_{Global}^2 * c_1 * \left(c_2 * N_{Global} - c_3\right) * c_4 * \right. \quad \text{(S16)}$$

$$c_0 = 33.118;\ c_1 = 33.411;\ c_2 = 61.383;\ c_3 = 33.389;\ c_4 = 29.123;$$
$$c_5 = -29.122;\ c_6 = 61.383;\ c_7 = 33.389;\ c_8 = -14.749;\ c_9 = 1.7179$$

S.2.3.2. GA model based on multiple individual trend lines for each colour group

Similar estimation levels are achieved with the model of Equation (S17), to estimate TSS from the trend line of the different colour groups, with a Pearson's coefficient of 36.04% and 27.82% for training and test, respectively.

$$TSS_{(mg/l)} = \left(c_0 + c_1 * N_{Violet}\right) * M_{Red} * N_{Blue} * \left(\frac{c_2 * N_O}{M_{Vio}}\right. \quad \text{(S17)}$$
$$c_5$$

$c_0 = 6{,}496.11$; $c_1 =- 8{,}459.44$; $c_2 = 0{,}81$; $c_3 =- 66.03$; $c_4 = 12.86$; $c_5 = 5.3638$

S.2.3.3. GA model based on point values of the spectral response

Equation (S18) shows the model for estimating TSS from point values of the spectral response, achieving a Pearson's Coefficient of 42.42% for training and 21.73% for test.

$$TSS_{(mg/l)} = \frac{\frac{(c_0 * T_{631} - c_1 * T_{461}) * (c_2 * T_{627} + c_3)}{c_4 * A_{700}}}{(c_5 * A_{435} - c_6 * A_{410})} * c_7 + c_8 \quad \text{(S18)}$$

$c_0 = 4.30$; $c_1 = 4.41$; $c_2 =- 16.15$; $c_3 = 11.16$; $c_4 = 1.32$; $c_5 = 15.07$; $c_6 = 14.81$

$c_7 = 2.31$; $c_8 = 2.74$;

S.2.3.4. Hybrid characterization model based on decision trees

Figure S13 shows the classification tree for the hybrid characterization model for treated water for TSS, with an R^2 for training of 67.67% and 60% for test, respectively.

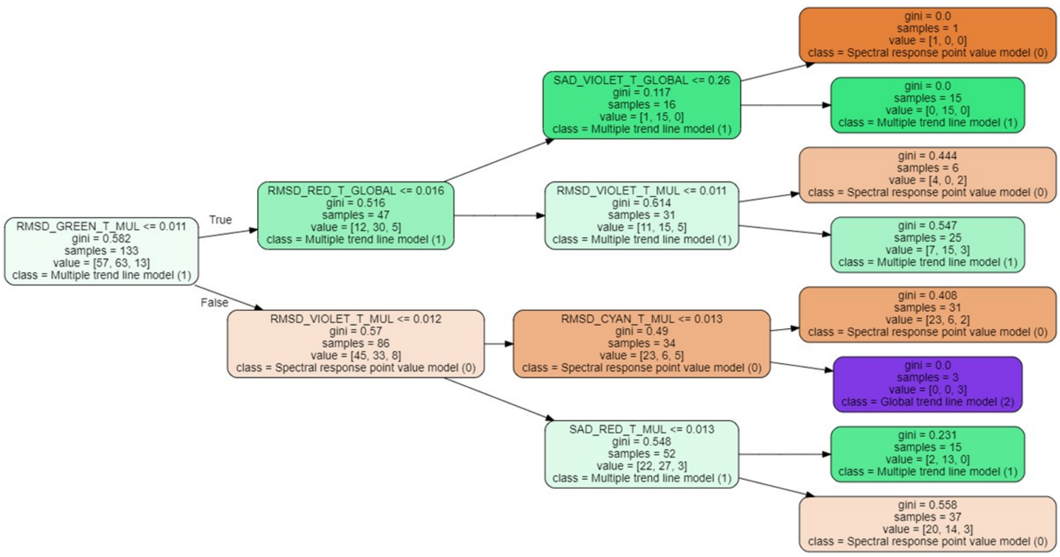
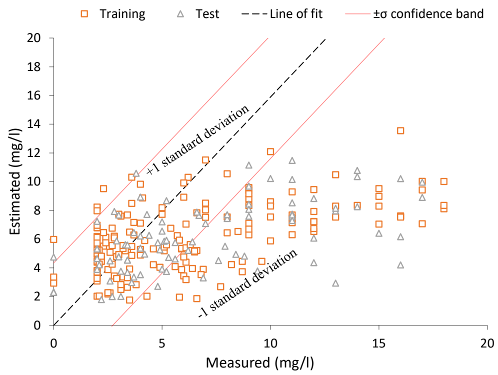
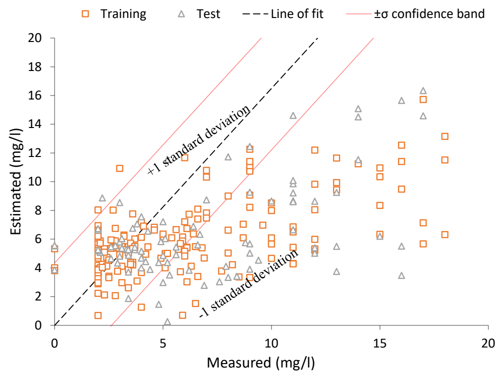


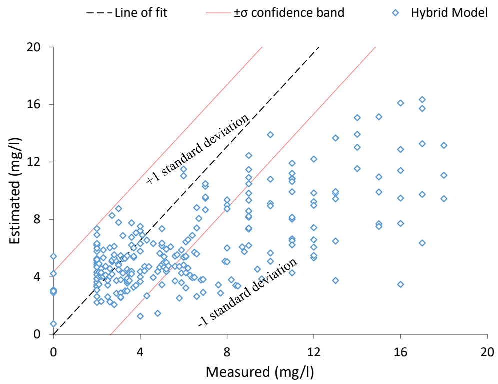
Figure S13. Classification tree for hybrid model of treated water characterization for TSS.



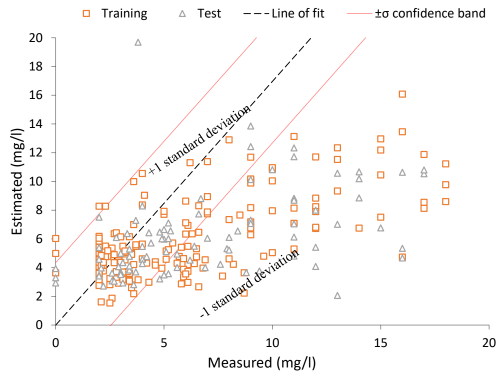
A



B



C



D

Figure S14. Scatter plot between laboratory measured TSS values (Measured) and those estimated by: (A) Global Model, Equation (S16). (B) Individual trend model, Equation (S17). (C) Hybrid model. (D) Model based on spectral point values by Offspring Selection technique, Equation (S18).

S2.4. Total Nitrogen estimation models (TN)

S.2.4.1. GA model based on global trend line

Equation (S19) shows the model for estimating TN estimates from the global trend line of the spectral response, calculated from 264 treated water samples, after removing outliers. The model shows a Pearson's coefficient of 32.86% for training and 13.46% for testing.

$$TN_{(mg/l)} = (c_0 * N_{Global} - c_1) * (c_2 * N_{Global} - c_3) * c_4 * \ast c_8 + c_9 \quad \text{(S19)}$$

$c_0 =- 52.47$, $c_1 =- 15.796$, $c_2 =- 47.398$, $c_3 =- 24.975$, $c_4 =- 18.905$,

$c_5 =- 19.566$, $c_6 =- 63.091$, $c_7 = 33.246$, $c_8 =- 5.1155$, $c_9 =- 1.0713$

S.2.4.2. GA model based on multiple individual trend lines for each colour group

The model of Equation (S20), to estimate TN from the trend line of the different colour groups, achieves similar estimation levels, with a Pearson's coefficient of 56.98% and 31.04% for training and test, respectively.

$$TN_{(mg/l)} = \left(\left(c_0 * M_{Orange} - c_1 * M_{Violet}\right) + c_2 * M_{Red}\right) + c_7 \quad \text{(S20)}$$

$$c_0 = 136.66;\ c_1 =- 98.15;\ c_2 =- 388.17;\ c_3 = 186.51;\ c_4 = 201.12;\ c_5 = 54.93;$$
$$c_6 =- 358.76;\ c_7 = 20.105$$

S.2.4.3. GA model based on point values of the spectral response

Equation (S21) shows the model for estimating TN from point values of the spectral response, achieving a Pearson's Coefficient of 61.65% for training and 20.06% for test.

$$TN_{(mg/l)} = \left(\left(c_0 * T_{660} + c_1 * A_{650}\right) + \left(c_2 * T_{615} + c_3 *\right.\right. \quad \text{(S21)}$$

$$c_0 = 26.93;\ c_1 =- 3.16;\ c_2 =- 34.54;\ c_3 = 8.97;\ c_4 =- 27.83;\ c_5 =- 25.43$$
$$c_6 =- 44.05;\ c_7 = 6.09;\ c_8 = 2.60;\ c_9 = 6.7$$

S.2.4.4. Hybrid characterization model based on decision trees

Figure S15 shows the classification tree for the hybrid characterization model for treated water for TN, with an R^2 of 100% both training and test.

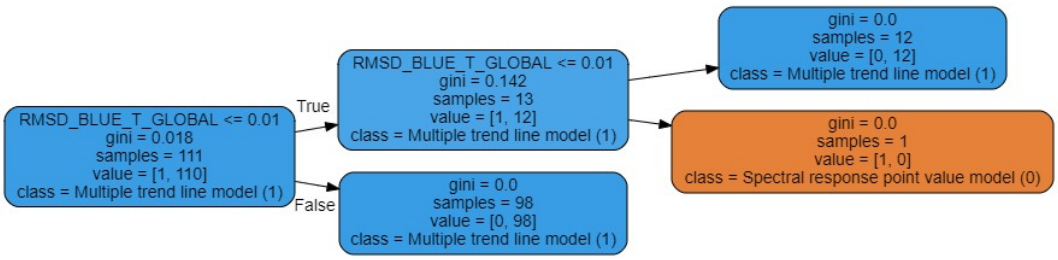


Figure S15. Classification tree for hybrid model of treated water characterization for TN.

Figure S16 shows the scatter plot of the estimation models based on trend lines (Equation (25 and 26)), as well as the hybrid model of Figure S15.

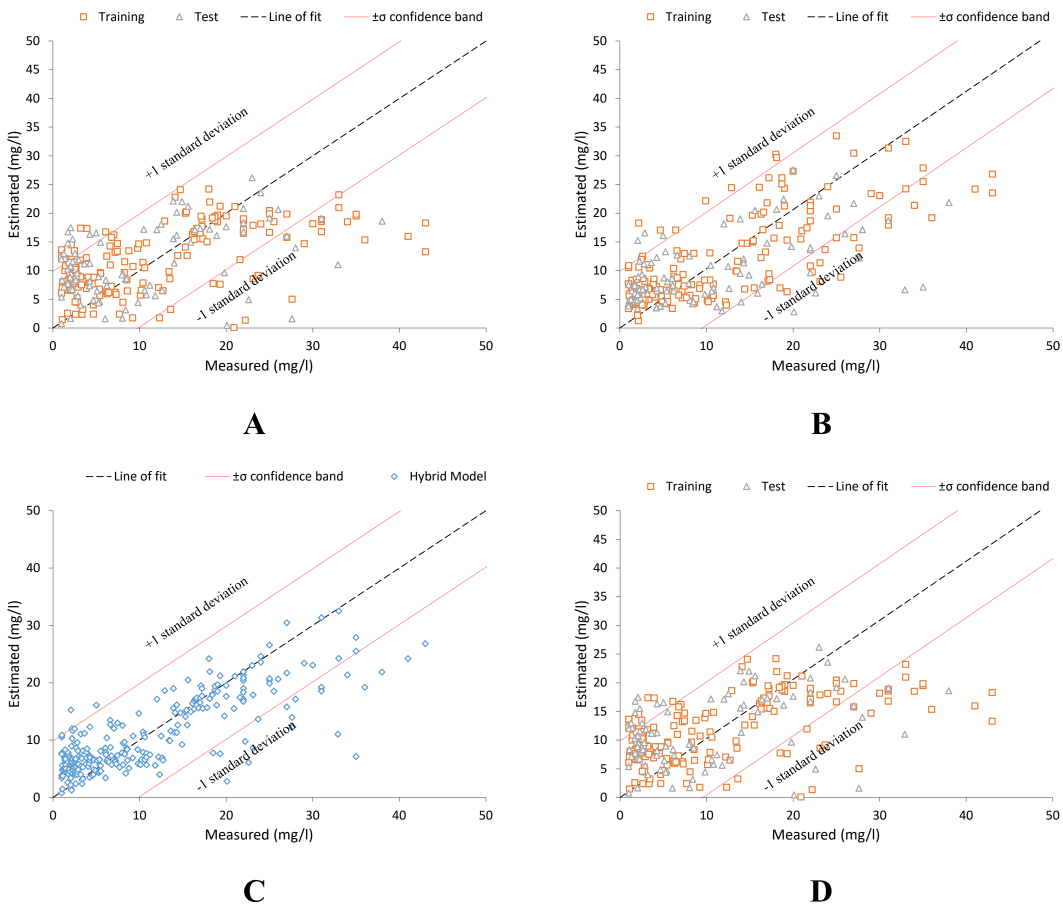


Figure S16. Scatter plot between laboratory measured TN values (Measured) and those estimated by: (A) Global Model, Equation (S19). (B) Individual trend model, Equation (S20). (C) Hybrid model. (D) Model based on spectral point values by Offspring Selection technique, Equation (21).

Table S1. Wastewater treatment plants used during the study.

	WWTP	Province	Population		Capacity (m3/a) In average		SST In average			COD In average			BOD_5 In average		
			Served	Equivalent	Design (m3/a)	Current (m3/a)	In (mg/l)	Out (mg/l)	Perf (%)	In (mg/l)	Out (mg/l)	Perf (%)	In (mg/l)	Out (mg/l)	Perf (%)
1	**Abanilla**	Murcia	3.626	15.711	547.500	779.051	294	4	98.6	739	18	97.6	442	4	99.1
2	**Abarán**	Murcia	13.371	12.626	1.642.500	726.065	257	5	98.1	596	28	95.3	381	3	99.2
3	**Albudeite**	Murcia	1.296	1.043	365.000	45.738	205	8	96.1	756	31	95.9	499	4	99.2
4	**Alcantarilla**	Murcia	41.447	62.342	4.745.000	2.588.649	301	6	98.0	835	33	96.0	527	4	99.2
5	**Alguazas**	Murcia	9.102	37.629	5.475.000	1.076.650	371	4	98.9	1.208	22	98.2	765	3	99.6
6	**Archena**	Murcia	24.413	54.425	2.737.500	1.792.326	459	6	98.7	1.104	27	97.6	665	3	99.5
7	**Baños y Mendigo**	Murcia	218	344	173.375	21.521	352	10	97.2	591	36	93.9	350	3	98.9
8	**Barinas**	Murcia	756	1.982	197.100	73.884	390	4	99.0	906	21	97.7	588	4	99.3
9	**Barqueros**	Murcia	1.030	1.872	109.500	60.376	443	17	96.2	1.245	58	95.3	679	6	99.1
10	**Beniel Nueva**	Murcia	11.900	25.818	1.825.000	1.245.618	659	4	99.4	944	26	97.2	454	3	99.3
11	**Blanca**	Murcia	5.184	5.636	730.000	356.464	271	4	98.5	559	19	96.6	346	3	99.1
12	**Cabezo Beaza**	Murcia	176.223	173.924	12.775.000	9.031.284	470	18	96.2	924	52	94.4	422	12	97.2
13	**Cabezo de la Plata**	Murcia	104	358	44.165	44.165	248	11	95.6	1.024	30	97.1	702	3	99.6
14	**Calasparra**	Murcia	9.505	26.938	2.190.000	659.778	408	3	99.3	1.426	23	98.4	894	3	99.7
15	**Campos del Río**	Murcia	1.998	1.635	547.500	88.252	212	5	97.6	649	21	96.8	406	3	99.3
16	**Cañada de la leña**	Murcia	93	28	21.900	6.166	68	19	72.1	172	56	67.4	99	6	93.9
17	**Cañares / Bronchos**	Murcia	442	195	1.350.500	54.371	805	2	99,7	3.253	2.751	37.5	156	3	98,0
18	**Casas Nuevas**	Murcia	152	220	73.000	8.170	847	6	99.3	1.190	28	97.6	590	4	99.3
19	**Ceutí Nueva**	Murcia	11.774	36.311	2.920.000	1.052.685	448	11	97.5	1.274	33	97.4	755	3	99.6
20	**Cieza**	Murcia	33.797	63.567	3.650.000	2.485.914	362	5	98.6	872	23	97.4	560	3	99.5

21	**Corvera**	Murcia	2.443	2.464	109.500	133.534	284	2	99.3	682	24	96.5	404	3	99.3
22	**El Cantón**	Murcia	66	506	18.250	18.250	324	18	94.4	1.058	34	96.8	608	5	99.2
23	**El Raal**	Murcia	15.940	23.706	2.737.500	3.950.557	151	7	95.4	240	21	91.3	131	4	96.9
24	**El Valle**	Murcia	194	464	511.000	58.089	378	5	98.7	343	17	95.0	175	3	98.3
25	**Fortuna**	Murcia	7.557	11.544	912.500	423.126	445	9	98.0	975	34	96.5	598	4	99.3
26	**Fuente Librilla**	Murcia	579	1.418	146.000	44.776	266	19	92.9	1.122	38	96.6	694	4	99.4
27	**Hacienda Riquelme**	Murcia	224	658	574.875	64.366	145	5	96.6	313	24	92.3	159	3	98.1
28	**Jumilla Nueva**	Murcia	24.588	70.595	4.380.000	1.739.564	825	3	99.6	1.761	24	98.6	889	3	99.7
29	**La Murta**	Murcia	91	545	44.165	15.378	353	4	98.9	1.271	28	97.8	776	3	99.6
30	**Lorqui**	Murcia	6.622	26.108	1.825.000	1.221.497	376	4	98.9	835	19	97.7	468	3	99.4
31	**Macisvenda**	Murcia	504	557	41.975	26.219	242	5	97.9	730	30	95.9	465	3	99.4
32	**Molina Norte**	Murcia	68.296	218.823	9.125.000	6.093.740	490	6	98.8	1.456	36	97.5	786	3	99.6
33	**Mosa Trajectum**	Murcia	144	285	642.400	42.568	177	3	98.3	270	15	94.4	147	3	98.0
34	**Mula Nueva**	Murcia	15.496	17.210	2.190.000	672.031	335	2	99.4	892	18	98.0	561	3	99.5
35	**Murcia Este**	Murcia	375.775	553.451	36.500.000	36.952.999	277	9	96.8	577	32	94.5	328	5	98.5
36	**Pliego**	Murcia	3.631	4.490	547.500	162.769	583	3	99.5	1.150	23	98.0	604	3	99.5
37	**Pol. Ind. Fortuna**	Murcia	0	584	65.700	22.945	797	30	96.2	855	68	92.0	557	13	97.7
38	**Santomera Norte**	Murcia	14.956	16.139	2.190.000	1.137.404	242	5	97.9	526	33	93.7	311	4	98.7
39	**Sucina Nueva**	Murcia	1.924	3.634	1.825.000	173.650	212	3	98.6	681	22	96.8	458	3	99.3
40	**Torres de Cotillas N.**	Murcia	19.996	53.597	4.380.000	1.602.051	641	9	98.6	1.281	21	98.4	733	3	99.6
41	**El Trampolín**	Murcia	149	158	73.000	13.930	329	37	88.8	396	33	91.7	248	4	98.4
42	**Yecla**	Murcia	31.876	43.586	2.920.000	1.648.354	490	7	98.6	1.015	20	98.0	579	3	99.5
43	**Yecla Raspay**	Murcia	97	109	18.250	8.385	147	5	96.6	477	18	96.2	286	4	98.6

Table S2. Example of optimal model selection for treated water samples from the decision tree in Figure 5.

Ref (mg/l)	G.T.M* (mg/l)	M.T.M** (mg/l)	S.P.M*** (mg/l)	Best model	RMSD_VIOLET_T GLOBAL	RMSD_BLUE_T GLOBAL	RMSD_CYAN_T GLOBAL	RMSD_GREEN_T GLOBAL	RMSD_YELLOW GLOBAL	RMSD_ORANGE GLOBAL	RMSD_RED_T GLOBAL	RMSD_VIOLET_T MUL	RMSD_BLUE_T MUL	RMSD_CYAN_T MUL	RMSD_GREEN_T MUL	RMSD_YELLOW_T MUL	RMSD_ORANGE_T MUL	RMSD_RED_T MUL	Tree's model	Tree's best estimation
20	19.43	16.98	33.07	G.T.M	0.041	0.013	0.029	0.029	0.009	0.017	0.037	0.012	0.006	0.012	0.011	0.011	0.015	0.037	G.T.M	19.43
31	21.71	18.15	30.41	S.P.M	0.037	0.011	0.029	0.035	0.024	0.021	0.038	0.012	0.005	0.011	0.016	0.024	0.018	0.036	M.T.M	18.15
17	18.27	18.06	27.90	M.T.M	0.039	0.011	0.029	0.034	0.019	0.020	0.038	0.011	0.004	0.010	0.014	0.021	0.018	0.037	M.T.M	18.06
10	11.68	15.03	10.84	S.P.M	0.031	0.010	0.035	0.037	0.051	0.035	0.028	0.015	0.004	0.013	0.018	0.037	0.030	0.024	S.P.M	10.84
10	13.50	14,71	11.03	S.P.M	0.035	0.012	0.038	0.039	0.049	0.036	0.032	0.015	0.005	0.012	0.017	0.037	0.031	0.029	S.P.M	11.03
30	24.09	23,87	26.03	S.P.M	0.039	0.010	0.029	0.024	0.015	0.013	0.021	0.013	0.005	0.012	0.012	0.006	0.012	0.016	S.P.M	26.03
23	16.95	20.38	18.28	M.T.M	0.040	0.014	0.034	0.038	0.034	0.026	0.041	0.012	0.006	0.011	0.016	0.032	0.023	0.038	M.T.M	20.38
23	17.79	15.67	34.69	G.T.M	0.042	0.010	0.032	0.035	0.020	0.021	0.040	0.013	0.004	0.014	0.013	0.022	0.018	0.038	M.T.M	15.67
24	21.52	16.31	28.97	Any	0.041	0.013	0.030	0.032	0.018	0.019	0.037	0.012	0.004	0.011	0.013	0.019	0.016	0.035	Any	28.97
10	15.23	14.04	20.11	M.T.M	0.034	0.010	0.026	0.031	0.033	0.024	0.040	0.011	0.003	0.010	0.014	0.024	0.021	0.038	M.T.M	14.04
10	16.39	13.42	23.67	M.T.M	0.038	0.010	0.027	0.032	0.024	0.021	0.038	0.013	0.004	0.011	0.014	0.020	0.016	0.036	M.T.M	13.42
10	13.78	13.40	19.54	M.T.M	0.035	0.008	0.026	0.033	0.031	0.024	0.041	0.012	0.003	0.011	0.015	0.024	0.021	0.039	M.T.M	13.40
10	16.31	15.90	28.06	M.T.M	0.035	0.010	0.022	0.027	0.012	0.019	0.041	0.012	0.004	0.010	0.013	0.011	0.016	0.039	M.T.M	15.90
20	22.08	18.12	18.80	S.P.M	0.043	0.013	0.028	0.029	0.010	0.017	0.036	0.015	0.004	0.010	0.013	0.011	0.015	0.034	M.T.M	18.12
10	18.06	15.69	20.37	M.T.M	0.035	0.011	0.027	0.035	0.034	0.025	0.039	0.011	0.004	0.011	0.016	0.029	0.022	0.037	M.T.M	15.69
11	17.23	11.48	24.60	M.T.M	0.043	0.012	0.030	0.036	0.029	0.024	0.041	0.014	0.005	0.013	0.015	0.025	0.020	0.039	M.T.M	11.48
0	17.61	13.88	22.47	M.T.M	0.039	0.013	0.028	0.034	0.028	0.024	0.040	0.013	0.005	0.011	0.014	0.024	0.020	0.038	M.T.M	13.88
13	20.81	13.40	-1.52	M.T.M	0.038	0.011	0.024	0.035	0.020	0.023	0.038	0.014	0.006	0.011	0.014	0.023	0.021	0.037	M.T.M	13.40
13	15.33	10.43	20.52	Any	0.041	0.016	0.037	0.040	0.045	0.034	0.045	0.013	0.007	0.013	0.017	0.033	0.028	0.042	Any	20.52
14	14.04	14.26	16.54	Any	0.032	0.010	0.025	0.034	0.039	0.026	0.041	0.012	0.006	0.011	0.016	0.029	0.023	0.038	Any	16.54

*G.T.M: Global trend line model **M.T.M: Multiple trend line model *** S.P.M: Spectral point value model

Table S2 in Supplementary Information shows an example application of the hybrid characterization model based on the decision tree shown in Figure 7 for 20 treated wastewater samples taken at random, where the reference values measured in the laboratory are indicated, as well as the estimates made by each of the models, and the RSMD of each colour group that will be used to determine which model is more appropriate to apply.

This table shows which estimation model is the most adequate ('Best model') according to the relation between the reference value and those provided by the different estimation models, as well as which is the model proposed by the tree in Figure 7 based on the RSMD values shown in Table S2. As can be seen, the classification performed by the tree match in most cases with the best possible model (marked in green), allowing to achieve better estimates.

Likewise, in those models, whose estimates are similar to each other (but not necessarily to the reference value), they have been designated with the label 'Any' to indicate that any of the models can be equally valid, i.e., all models are equally good/bad at characterizing the sample

Table S3: Performance analysis of the models presented for different levels of random perturbation in terms of RMSE.

Parameter	Type of wastewater	Perturbation	RMSE			
			Punctual model	Global straight trend model	Multiple trend lines model	Hybrid model
COD	Raw	0%	194.03	212.91	205.63	187.59
		2%	235.50	212.70	231.27	186.63
		5%	406.94	214.93	294.90	201.75
		10%	1371.01	219.37	464.09	210.59
		15%	2753.27	222.38	659.93	238.45
		20%	4250.74	360.44	895.74	353.55
BOD_5		0%	148.00	154.67	150.54	143.36
		2%	342.08	154.95	193.96	144.80
		5%	1242.98	155.05	1331.52	150.60
		10%	927.05	157.62	1864.17	146.07
		15%	1178.66	155.10	1864.39	185.03
		20%	934.26	165.75	1404.90	174.69
TSS		0%	79.22	88.39	83.17	75.73
		2%	128.73	88.51	88.53	82.78
		5%	292.30	89.41	112.22	83.47
		10%	1292.38	89.54	226.21	93.95
		15%	1548.40	92.52	473.32	139.07
		20%	1934.73	97.31	800.57	196.26

TN		0%	16.63	18.01	16.89	16.48
		2%	37.61	18.00	34.17	16.50
		5%	98.71	18.00	116.12	17.08
		10%	176.35	18.70	298.13	18.39
		15%	323.57	18.16	434.46	26.10
		20%	403.79	71.53	524.02	60.54
TP		0%	2.56	2.66	2.61	2.49
		2%	4.05	2.66	4.77	2.60
		5%	7.83	2.65	23.92	2.68
		10%	14.66	2.76	43.18	2.75
		15%	22.70	2.86	60.90	3.72
		20%	29.41	3.06	55.32	4.15
COD	Treated	0%	10.66	12.74	10.75	10.30
		2%	12.28	12.51	15.36	10.33
		5%	18.25	13.89	59.55	13.42
		10%	77.87	17.34	99.47	17.35
		15%	186.27	22.42	104.61	19.96
		20%	206.98	24.11	132.15	21.55
BOD_5		0%	1.70	1.90	1.78	1.56
		2%	2.97	1.90	4.52	1.83
		5%	26.47	1.93	6.44	1.91
		10%	30.43	2.02	25.39	1.98
		15%	34.66	2.12	46.98	2.42
		20%	32.97	2.37	63.16	3.25
TSS		0%	3.54	3.66	3.59	3.16
		2%	4.61	3.67	4.80	3.55
		5%	15.55	3.75	25.87	3.78
		10%	49.61	4.19	48.42	3.83
		15%	70.94	4.81	59.33	7.82
		20%	83.81	6.29	71.58	12.57
TN		0%	7.70	8.56	7.04	5.86
		2%	10.12	8.50	8.23	7.18
		5%	20.66	8.62	13.75	7.99
		10%	57.15	9.78	34.95	9.91
		15%	120.64	10.17	75.51	17.88
		20%	167.30	11.57	113.47	34.86