

# 1 Supplementary information

## A. Laboratory setup

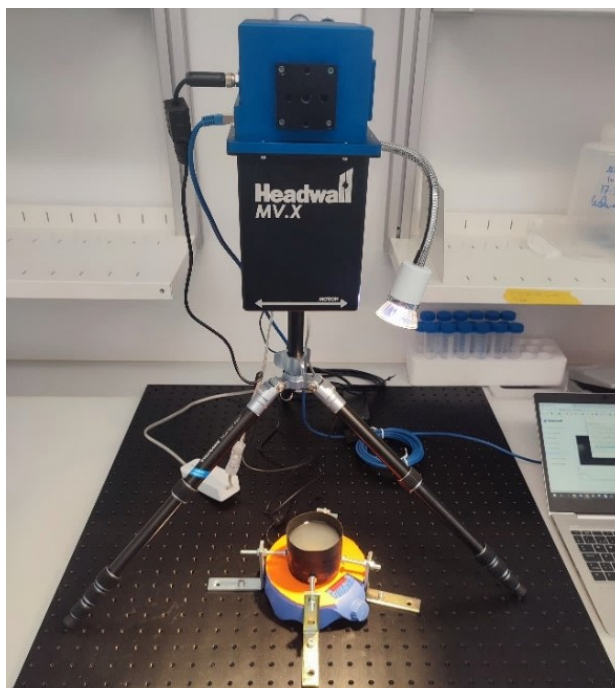


Figure 7: Picture of the laboratory setup

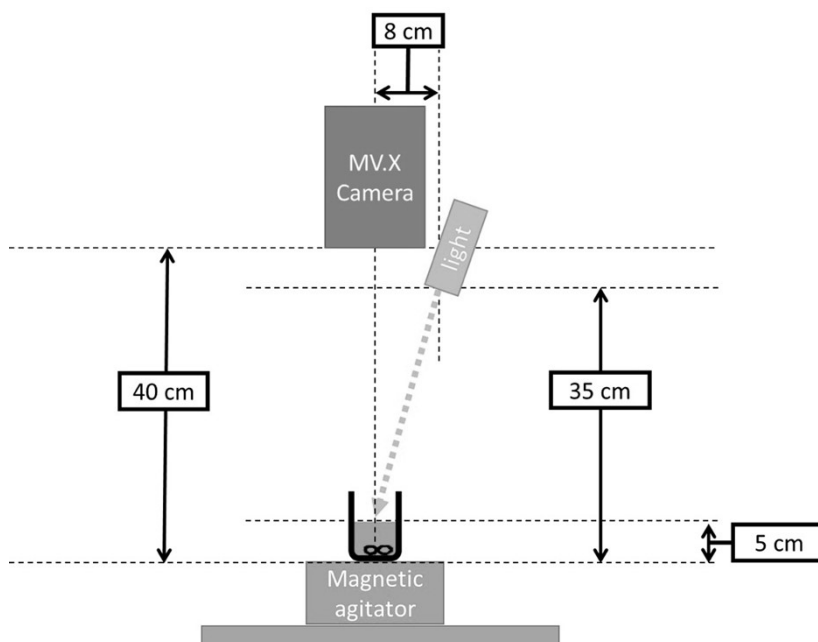


Figure 8: Dimensions of the laboratory setup

B. Scatterplot and range, mean and median values of the seven pollution indicators measured in this study

Table 5: Range and metrics for the pollution in the wastewater mixtures generated.

Water quality variable	Unit	Min	Max	Mean	Median
COD	mg/L	91.2	379.0	227.0	233.7
Turbidity	NTU	21.6	267.3	147.4	132.9
DOC	mg/L	45.1	302.9	132.3	124.2
TDN	mg/L	13.5	44.6	29.9	30.6
PO <sub>4</sub> -P	mg/L	0.8	5.0	2.9	3.1
SO <sub>4</sub> -S	mg/L	27.8	74.7	53.3	58.7
NH <sub>4</sub> -N	mg/L	5.4	26.6	18.0	19.4

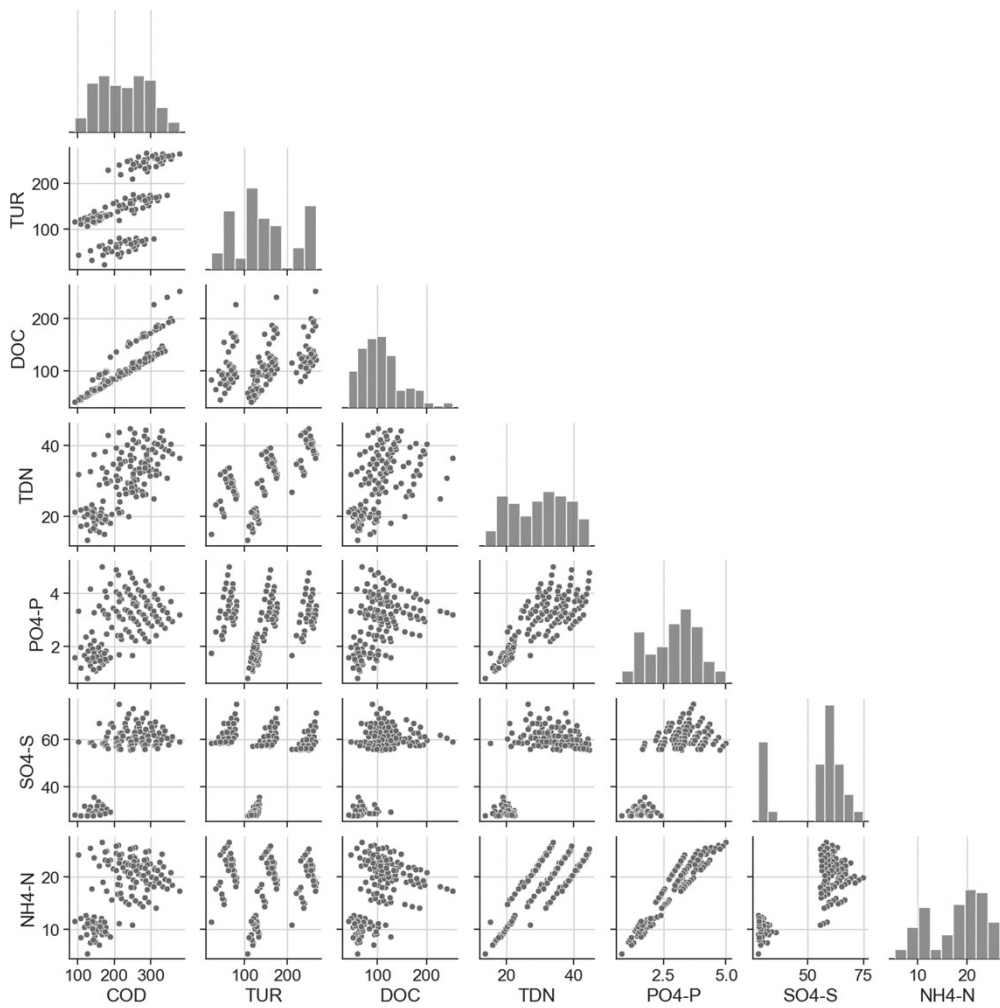


Figure 9: Scatterplot of the water quality indicators

### C. Mathematical equations for: hyperspectral data-cubes pre-processing and spectra extraction

The hyperspectral data-cube of a given water sample is organized into a 3D reflectance array  $\theta$  with dimensions 1020 x 51 x 300. We use the following indices to represent the dimensions of the array:

- $i$  represents the pixel number, ranging from 1 to 1020.
- $j$  represents the line number, ranging from 1 to 51.
- $k$  represents the wavelength number, ranging from 1 to 300 and corresponding to the wavelength range between 400 and 1000nm with a 2nm resolution

Substep 1.1 Normalization to of the raw data-cube  $\theta$  with dark and white reference

The normalized reflectance array  $\theta_{i,j,k}^-$  is calculated with:

$$\theta_{i,j,k}^- = \frac{\theta_{i,j,k} - D_{i,j,k}}{W_{i,j,k} - D_{i,j,k}} \quad (i = 1 \dots I; j = 1 \dots J; k = 1 \dots K) \quad (EQ. 1)$$

where  $D_{i,j,k}$  is the dark reference data-cube and  $W_{i,j,k}$  is the white reference data-cube.

Substep 1.2 Data-cube reframing

The reframed data-cube  $\theta_{i,j,k}^{-*}$  is obtained with the formula:

$$\theta_{i,j,k}^{-*} = \theta_{i+110,j,k+10}^- \quad (i = 1 \dots I^*; j = 1 \dots J^*; k = 1 \dots K^*) \quad (EQ. 2)$$

with:  $I^* \times J^* \times K^* = 750 \times 45 \times 280$

Substep 1.3 Pixel selection:

First, the light reflection intensity  $R_{i,j}^{int}$  of each pixel ( $i, j$ ) is calculated as the average reflectance across all wavelengths:

$$R_{i,j}^{int} = \frac{1}{K^*} * \sum_k \theta_{i,j,k}^{-*} \quad (i = 1 \dots I^*; j = 1 \dots J^*) \quad (EQ. 3)$$

Second, the mask  $M_{i,j}$  is calculated as an  $I^* \times J^*$  array of boolean values:

$$M_{i,j} = \begin{cases} 1 & \text{if } L \leq R_{i,j}^{int} \leq U, \quad (i = 1 \dots I^*; j = 1 \dots J^*) \\ 0 & \text{otherwise} \end{cases} \quad (EQ. 4)$$

where: L and U are the 20% and 80% percentiles of the array  $R^{int}$ .

Substep 1.4 Calculation of the mean reflectance spectra

After mask application, the total number of remaining pixels is:

$$0.6 \times I^* \times J^* = 20250 .$$

With these pixels, the mean ( $m$ ) and standard deviation ( $S$ ) of reflectance is calculated for every wavelength ( $k$ ):

$$m_k = \frac{\sum_{i,j} M_{i,j} \cdot \theta_{i,j,k}^-}{\sum_{i,j} M_{i,j}} \quad (k = 1 \dots K^*) \quad (EQ. 5)$$

$$s_k = \sqrt{\frac{\sum_{i,j} M_{i,j} \cdot (\theta_{i,j,k}^- - m_k)^2}{\sum_{i,j} M_{i,j}}} \quad (k = 1 \dots K^*) \quad (EQ. 6)$$

**The array  $m$  is representing the raw reflectance spectra extracted from a given data-cube**

Substep 1.5 Spectra pre-processing

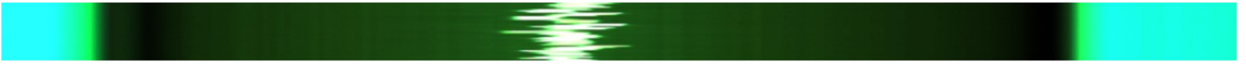
The processed reflectance spectrum R is finally obtained with:

$$R_k = \log_{10}(P_k(x = k)) \quad (k = 1 \dots K^*) \quad (EQ. 7)$$

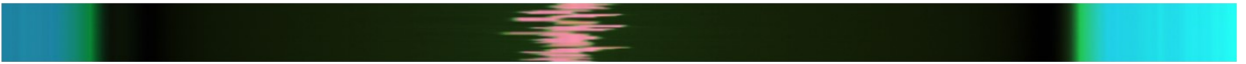
Where:  $P_k(x)$  is a second-order polynomial function fitted with the 17 reflectance values centered on the  $k^{\text{th}}$  value of the reflectance spectra  $m$ .

The following figure gives a visualization of the first three substeps, starting from a raw data-cubes.

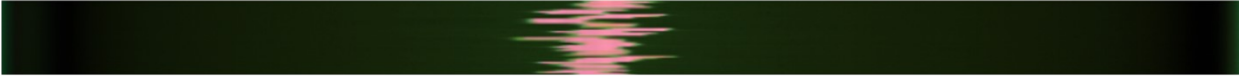
a. Raw hyperspectral data-cube



b. Step 1.1 conversion of the raw image to reflectance



c. Step 1.2 re-framing of the image



d. Step 1.3 mask to remove the first and last 20% of reflectance to the previous image



*Figure 10: Step by step modification of the hyperspectral images to extract a representative spectrum*

## D. Mathematical equations for: Partial least squares parameter optimization and model evaluation

For this section, we are using  $s$  as the indices representing a wastewater sample ( $s=1\dots 144$ ) and  $v$  as the indices representing the water quality variable as defined in Table 1 ( $v=1\dots 7$ ).

The reflectance spectra  $R$  of each sample  $s$  (defined in EQ7) are combined into a  $S \times K^*$ -dimensional matrix  $X$  to be used as model features:

$$\text{model features: } X_{s,k} (s = 1\dots S = 144)(k = 1\dots K^* = 280)$$

The seven reference measurements of each sample  $s$  are combined in an  $S \times 7$ -dimensional matrix  $Y$ .

$$\text{model labels: } Y_{s,v} (s = 1\dots S = 144)(v = 1\dots 7)$$

### Substep 2.1 Classification of the wavelength

For each number of latent variables  $n_l$  and each water quality variable  $v$ , a PLS model was fitted with the full dataset to retrieve the wavelength classification  $C_{n_l, v}$ :

$$C_{n_l, v} = \text{sort}(\text{abs}(\text{coeff}(\text{PLS}_{n_l} \cdot \text{fit}(Y_v, X)))) \quad (\text{EQ. 8})$$

where  $\text{PLS}_{n_l}$  is a PLS model with  $n_l$  latent variables and  $Y_v$  is the  $v$ -th column of the matrix  $Y$ .

### Substep 2.2 Optimization of the PLS parameters

The PLS model performance was measured with the root mean squared error (RMSE) between the reference pollution values  $Y_v$  and the predictions  $Y_v^{\text{pred}}$  obtained with leave-one-out cross-validation (LOOCV). Therefore, the optimal parameters are defined by:

$$(n_l, n_w)$$

$$\text{where: } \text{RMSE}(Y_v^{\text{pred}}(n_l, n_w), Y_v) = \min_{a:1 \rightarrow 20, b:0 \rightarrow 279} (\text{RMSE}(Y_v^{\text{pred}}(a, b), Y_v)) \quad (\text{EQ. 9})$$

$$\text{with: } RMSE(Y_v^{pred}, Y_v) = \sqrt{\frac{1}{S} * \sum_{s=1}^{144} (Y_{s,v}^{pred} - Y_{s,v})^2} \quad (EQ. 10)$$

Step 2.3: Detailed optimal model evaluation

$$R^2(Y_v^{pred}, Y_v) = 1 - \frac{\sum_s (Y_{s,v}^{pred} - \bar{Y}_v)^2}{\sum_s (Y_{s,v} - \bar{Y}_v)^2} \quad (EQ. 11)$$

$$RMSE_{relative}(Y_v^{pred}, Y_v) = 100 * \frac{RMSE(Y_v^{pred}, Y_v)}{\bar{Y}_v} \quad (EQ. 12)$$

$$\text{with } \bar{Y}_v = \frac{1}{S} * \sum_{s=1}^{144} Y_{s,v} \quad (EQ. 13)$$

E. Two additional approaches to estimate turbidity

### Model based on the light reflectance intensity

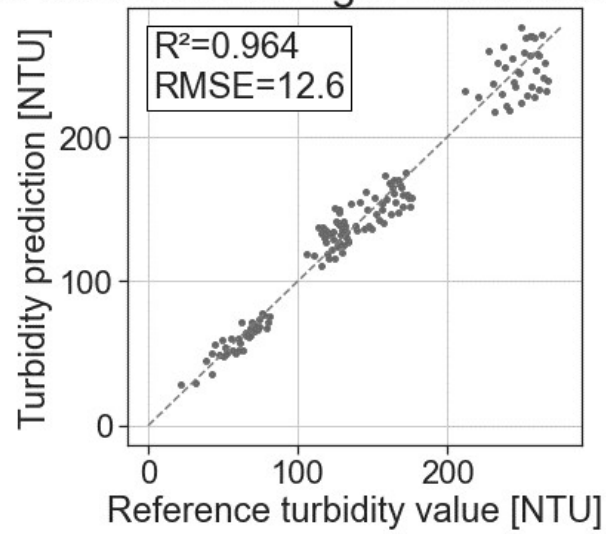


Figure 11: Estimation of turbidity from the light reflection intensity

### Model based on the three RGB wavelengths

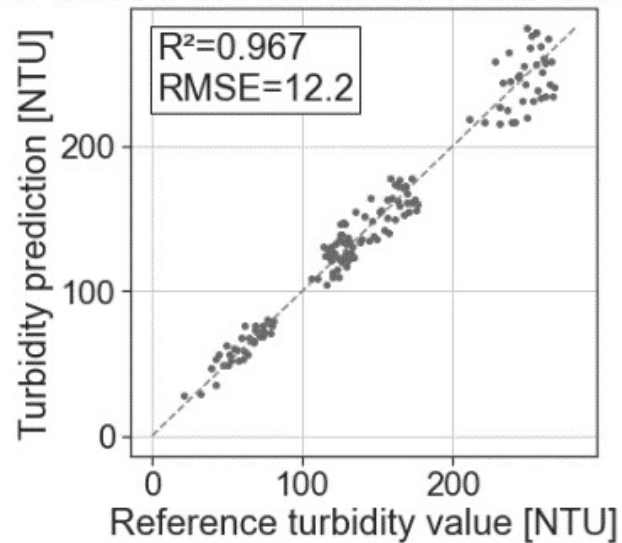


Figure 12: Estimation of the turbidity with a PLS model trained with wavelength at 464 (blue), 532 (green) and 630nm (red)



F. Results of the optimal PLS-models trained with the 36 samples composed of wastewater without formazine

Table 6: Detailed results of the model

Water quality variable	Unit	Min	Max	Optimal number of latent variables	Optimal number of wave-length	R <sup>2</sup>	RMSE	RMSE (relative)
COD	mg/L	101.6	308.0	20	56	0.88	16.0	7.4%
Turbidity	NTU	21.6	80.7	20	143	0.94	3.6	5.9%
DOC	mg/L	45.1	228.0	20	60	0.92	10.7	10.2%
TDN	mg/L	15.1	33.7	20	100	0.97	0.7	2.6%
PO <sub>4</sub> -P	mg/L	1.7	5.0	20	99	0.94	0.2	4.9%
SO <sub>4</sub> -S	mg/L	58.3	74.7	20	85	0.77	1.9	3.0%
NH <sub>4</sub> -N	mg/L	11.4	26.6	20	107	0.97	0.6	2.7%

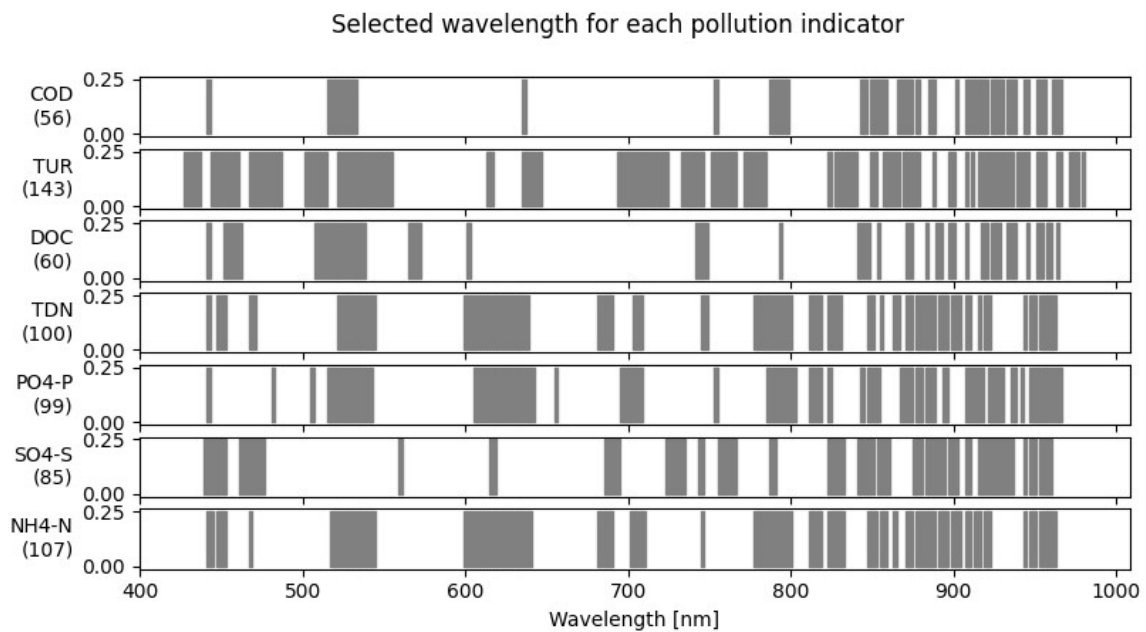


Figure 13: Model selected wavelengths