



**HAL**  
open science

# Optimising punctual water sampling with an on-the-fly algorithm based on multiparameter high-frequency measurements

Jérémy Mougin, Pierre-Jean Superville, Cyril Ruckebusch, Gabriel Billon

► **To cite this version:**

Jérémy Mougin, Pierre-Jean Superville, Cyril Ruckebusch, Gabriel Billon. Optimising punctual water sampling with an on-the-fly algorithm based on multiparameter high-frequency measurements. *Water Research*, 2022, 221, pp.118750. 10.1016/j.watres.2022.118750 . hal-03715742

**HAL Id: hal-03715742**

**<https://hal.science/hal-03715742v1>**

Submitted on 5 Feb 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

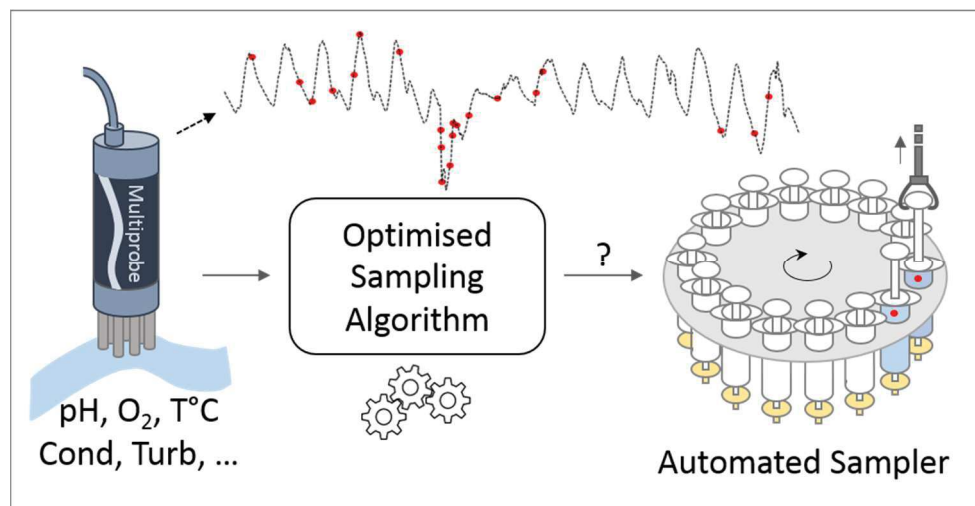
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Optimising punctual water sampling with an on-the-fly algorithm based on multiparameter high-frequency measurements

Jérémy Mougin<sup>1</sup>, Pierre-Jean Superville<sup>1\*</sup>, Cyril Ruckebusch<sup>1</sup>, Gabriel Billon<sup>1</sup>

<sup>1</sup>Université Lille, CNRS, UMR 8516 - LASIRE, Laboratoire de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000 Lille, France.

\* corresponding author



## Highlights :

- Optimised sampling to replace time-consuming high frequency sampling
- On-the-fly decision algorithm based on multiparameter measurements
- Reduction of the number of samples while retaining the dataset variability
- Systematic sampling of short events with a strong impact on the environment

17

18 **Abstract:**

19 The way in which aquatic systems is sampled has a strong influence on our understanding of  
20 them, especially when they are highly dynamic. High frequency sampling has the advantage  
21 over spot sampling for representativeness but leads to a high amount of analysis. This study  
22 proposes a new methodology to choose when sampling accurately with an automated sampler  
23 coupled with a high frequency (HF) multiparameter probe. After each HF measurement, an  
24 optimised sampling algorithm (OSA) determines on-the-fly the relevance of taking a new  
25 sample in relation to previous waters already collected. Once the OSA was optimised,  
26 considering the number of HF parameters and their variabilities, it was demonstrated through a  
27 study case that the number of samples could be significantly reduced, while still covering  
28 periods of low and high variabilities. The comparison between the total HF dataset and the  
29 sampled subdataset shows that physicochemical parameter variability is preserved (Pearson  
30 correlations  $> 0.96$ ) as well as the multiparameter variability (PCA axes remained similar with  
31 Tucker congruence  $> 0.99$ ). This algorithm simplifies HF studies by making it easier to take  
32 samples during brief phenomena such as storms or accidental spills that are often poorly  
33 monitored. In addition, it optimises the number of samples to be taken to correctly describe a  
34 system and thus reduce the human and financial costs of these environmental studies.

35 **Keywords:** river, monitoring, on line, high frequency, algorithm, sampling

36

37

38

## 39 **Introduction**

40 The composition and quality of aquatic systems are well known to be highly dynamic due to  
41 physical, biological, chemical, meteorological, and climatic factors. All these pressures have an  
42 impact on the environment quality and take place on very different scales, from minutes to  
43 years, and few square meters up to full catchments (Aguilera et al. 2016; Halliday et al. 2014;  
44 Meyer et al. 2021; Rode et al. 2016). Daily cycles resulting from recurring environmental  
45 phenomena, such as photosynthesis or temperature fluctuations (Halliday et al. 2014; Nimick  
46 et al. 2011; Shultz et al. 2018; Superville et al. 2014) are the most usually observed. Quick  
47 punctual events can also be recorded such as heavy rainfalls or industrial discharges, which are  
48 hardly predictable (Khamis et al. 2020; Seifert-Dähnn et al. 2021; Vaughan et al. 2019). At a  
49 much larger time scale, seasonal effects can be critical, as for algal bloom and the resulting  
50 organic matter decomposition, which can dramatically affect water quality (Seifert et al. 2016).

51 Monitoring aquatic ecosystems with an inadequate measurement frequency may lead to missing  
52 information and/or misinterpretation of the observed data (Marcé et al. 2016). It is therefore  
53 necessary to implement monitoring methods adapted to the environments studied and their  
54 dynamics. To address this scientific challenge, automated HF monitoring is a growing trend for  
55 operational and research purposes (Bieroza and Heathwaite 2015; Gunatilaka and Diehl 2001;  
56 Halliday et al. 2014; Ivanovsky et al. 2016; Khamis et al. 2020; Rode et al. 2016; Seifert-Dähnn  
57 et al. 2021). The capacity to measure *in situ* or *on-line* was strongly enhanced during the last  
58 decades. Miniaturization, increased power capacity and technological advances in probes have  
59 allowed new sensors to be deployed with a better stability over time (Marcé et al. 2016).

60 However, numerous parameters, *e.g.* micro-pollutants, cannot be easily monitored due to the  
61 lack of specific on line probe or analyser and/or intensive maintenance work requirement  
62 (Khamis et al. 2020; Marcé et al. 2016).

63 To overcome these limitations, most studies still rely on taking samples for off-line laboratory  
64 analysis, where a wider range of parameters can be studied. However, as the aquatic ecosystems  
65 are quickly evolving, the relevance of each sample depends strongly on when and where it was  
66 taken (Meyer et al. 2021; Piniewski et al. 2019). To be able to monitoring short-term  
67 phenomena, the increase of samples number is paramount (Khamis et al. 2020). However, some  
68 phenomena are not predictable and a strong increase in the sampling frequency (for instance  
69 several per day) is not operationally sustainable for a long term. Moreover, the probability of  
70 sampling a specific event of short duration is very low (Carstea et al. 2010) and it may lead to  
71 a misunderstanding of certain processes (Aguilera et al. 2016; Bieroza and Heathwaite 2016;  
72 Jarvie et al. 2018; Marcé et al. 2016; Reynolds et al. 2016).

73 For limnological studies, the number of samples needed to properly represents the environment  
74 can be relatively large (Aguilera et al. 2016). Some sampling strategies are based on a  
75 preliminary HF monitoring that allows for the optimisation of the sampling frequency (Aguilera  
76 et al. 2016; Ferrant et al. 2013; Piniewski et al. 2019). Another solution is to take only a few  
77 samples and rely on modelling tools to extrapolate the data (Searcy and Boehm 2021).  
78 However, these tools are not always well adapted and may provide information that is  
79 contradictory to the observations (Liu et al. 2018; Piniewski et al. 2019).

80 Our work establishes an alternative sampling solution. The main idea is to use the HF  
81 measurements data on-the-fly as a decision tool to choose when to sample next. The HF  
82 measurements are used as a visualization of the chemical status of the water body, and to trigger  
83 an automated sampler based on recorded variations. Samplers triggered by the variation of one  
84 parameter (*e.g.* conductivity or turbidity) have been previously used in environmental  
85 monitoring (Lewis and Eads 2009) but as far as we know, there are no systems based on a  
86 multivariate approach. The aims of this methodology are: (i) to minimise the number of off-line  
87 analysis without losing information from specific phenomena; and (ii) to hold data data

88 variability, statistical relevance and robustness of the off-line analysis. An application in the  
89 field illustrates critically the proof of concept of this innovative procedure.

90

## 91 **1. Material & Methods**

### 92 **1.1. Study site**

93 The study site is the Marque River, located in Northern France close to Lille. It has a length of  
94 32 km, an average flow at its confluence of  $1.2 \text{ m}^3 \text{ s}^{-1}$  and crosses an agricultural area in its  
95 upstream part and a more urban basin downstream. It is fed mainly by runoff, as well as by 8  
96 urban wastewater treatment plants (WWTP). WWTPs discharges can provide up to 30% of the  
97 Marque River flow during dry periods close to the study site. According to the Water  
98 Framework Directive criteria, its chemical and ecological quality is poor due to the presence of  
99 significant amounts of nitrogen, phosphorus, pesticides and Polycyclic Aromatic Hydrocarbons  
100 (HAP). The monitoring station ( $50^{\circ}38'43.6''\text{N}$ ,  $3^{\circ}10'54.5''\text{E}$ ) is located at the beginning of the  
101 urban part, about 1 km downstream of the Villeneuve d'Ascq WWTP (144 000 Equivalent  
102 inhabitant) and 300 m downstream of the discharge of a rainwater retention basin (Heron lake,  
103  $634\,000 \text{ m}^3$ ) (Ivanosky et al., 2016; Ivanovsky et al., 2018; Trommetter et al., submitted).

### 104 **1.2. High frequency monitoring set up**

105 *Mobile Laboratory – On line* monitoring is carried out using a mobile laboratory (ML),  
106 designed, and equipped to measure various physicochemical parameters in the field. This type  
107 of infrastructure has already been used for similar monitoring (Ivanovsky et al. 2016; Meyer et  
108 al. 2021). It is a trailer that can be towed by a commercial vehicle and can be deployed close to  
109 the water body (power supply is however necessary). It is equipped with an air conditioning  
110 system allowing to keep a relative constant temperature of  $15\text{-}25^{\circ}\text{C}$ . A submersible pump (water  
111 flow:  $10\text{-}15 \text{ m}^3 \text{ h}^{-1}$ ) supplies the ML and its various analysis devices. Briefly, most of the raw

112 water pumped is first introduced into an overflow cell in which a multiparameter probe is  
113 immersed. The overflow cell allows a measurement as close as possible to an *in-situ*  
114 measurement, allowing the constant renewal of the sample and an efficient transport of  
115 suspended matter. The second part of the hydraulic system includes an output to supply a  
116 homemade automatic filtering sampler and another output with an online filter at 100  $\mu\text{m}$  which  
117 mainly protects nutrients analysers from the biggest suspended matter. Finally, data acquired  
118 are transmitted every 30 minutes via a 4g network to a storage server, allowing the river to be  
119 monitored remotely and the whole system (pump, probes...) to be checked regularly.

120 *Multiprobe and automatic filtering sampler* - High frequency monitoring is performed with a  
121 multiprobe (*Eureka Water Probes; Manta+35*). It allows the monitoring of 7 chemical  
122 parameters: temperature, pH, conductivity, turbidity, dissolved oxygen and two fluorescence  
123 probes (*Turner Design*) for the measurement of DOM (1 sensitive to coloured dissolved organic  
124 matter (CDOM) and 1 sensitive to tryptophan-like substances). Every 10 minutes, a python  
125 script communicates with this probe, activates a wiper to clean the optical probes, and collects  
126 the average values of 10 successive measurements for each parameter (*Python Software*  
127 *Foundation. Python Language Reference, version 3.7.*). After each measurement, a decision  
128 algorithm (see section *sampling methodology*) analyses the new values and decides whether to  
129 trigger a sampling. If it is the case, a signal is sent to an automatic filtering sampler equipped  
130 with a 0.7  $\mu\text{m}$  filter (glass microfiber, Whatman) to eliminate most of the suspended matter.  
131 This homemade instrument consists of a carousel on which 24 syringes are placed and operating  
132 with a mechanical jack. As the samples are not refrigerated in the sampler, they are recovered  
133 as soon as possible (maximum 3 days) and then kept at 4°C before analyses in the laboratory  
134 (within the week).

135

136

### 137 1.3. Sampling methodology

138 *Overview* - The sample selection strategy developed in this work consists in collecting a sample  
139 on “each state” of the aquatic system that can be observed by the multi-probe. A state is defined  
140 as a combination of the values of the 7 parameters measured by the probe ( $\pm$  a certain margin).  
141 By capturing only discrete samples, the main objective of this methodology is to minimise the  
142 number of samples while preserving the variability of the data set. Specifically, a decision-  
143 making algorithm decides after each measurement made by the probe whether it should trigger  
144 a new sampling event. The mathematical formalism used in this section is constructed as  
145 follows: matrices are noted in bold with a capital letter (e.g.  $\mathbf{X}$ ), vectors are noted in bold with  
146 a lower case letter (e.g.  $\mathbf{msv}$ ), row and column indices are presented in lower case and italics  
147 (e.g.  $i, j$ ).

148 The Optimised Sampling Algorithm (OSA) is triggered after each measurement made by the  
149 mobile laboratory, *i.e.* every ten minutes. Like the ML automation, the OSA is written in python  
150 3.7.6, mainly based on the pandas 1.1.4 and numpy 1.18.3 packages (Harris et al. 2020; Reback  
151 et al. 2022). At each activation, the OSA takes as input 3 different datasets. Firstly, the  
152 measurements made by the ML since the beginning of the campaign. The corresponding data  
153 are collected in a data matrix  $\mathbf{X}$  of dimension  $n \times m$ , with  $n$  the number of measurements made  
154 since the system was launched and  $m$  the number of parameters monitored (in this study,  $m =$   
155 7). Then, a second matrix  $\mathbf{X}_s$  is defined, which groups together all the previous measurements  
156 that led to a sample being taken. This matrix  $\mathbf{X}_s$  is of dimension  $k \times m$  with  $k$  the number of  
157 samples. By construction,  $\mathbf{X}_s \subset \mathbf{X}$  and  $k < n$ . Finally, the new measure for which the OSA must  
158 decide is noted  $\mathbf{x}_{new}$  and corresponds to a vector of  $1 \times m$  dimension.

159 This algorithm works in three main steps.



160 First, a pre-processing resulting in the standardisation of the **Xs** and **xnew** data (Eqn 1) is carried  
161 out:

$$162 \quad \mathbf{X}_{i,j}^{std} = \frac{\mathbf{X}_{i,j} - \mathbf{med}_{1,j}}{\mathbf{msv}_{1,j}} \quad (\text{Equation 1})$$

163 Calculation of  $\mathbf{X}^{std}$  is done element-wise on the  $i^{\text{th}}$ -row  $j^{\text{th}}$ -column elements of **X**. **med** is a  $1 \times m$   
164  $m$  vector containing the median of each parameter over the last 1008 rows of **X**, corresponding  
165 to the last week of data. The vector **msv**, of dimension  $1 \times m$ , represents the minimum  
166 significant variation used as a standard deviation. **msv** values are set by the user, considering  
167 the quality of the sensor signal, as well as the knowledge of the variability of the observations  
168 for the river. The pre-processing step is very important as it will impact the importance of each  
169 parameter in the decision process of sampling. For some parameters, such as pH, the noise level  
170 will be used to set the values in **msv** (*e.g.* the **msv** for pH has been chosen to be 0.2 even though  
171 a variation of 0.1 could be considered relevant for the environment). For others, the  
172 corresponding value of the **msv** can be increased so that there is no oversampling for every  
173 small variation of that parameter (*e.g.* the **msv** for conductivity was set to  $25 \mu\text{S cm}^{-1}$  even  
174 though the noise is about  $2 \mu\text{S cm}^{-1}$ ). Following these recommendations, tests are performed to  
175 check if the **msv** vector is well balanced, *i.e.* variations of one parameter are not under  
176 considered compared to the others. Part D in the supplementary information presents examples  
177 of **msv** that are correct or in need of adjustment. If an unbalanced importance of a parameter is  
178 observed, **msv** can be readjusted by the user at any time during the process. Currently for this  
179 study, the **msv** vector is defined with the following values: 0.4 °C for temperature; 0.1 upH for  
180 pH;  $25 \mu\text{S/cm}$  for conductivity; 5 FNU for turbidity; 0.5 mg/L for dissolved oxygen; 2 ppb for  
181 CDOM and 5 ppb for tryptophan.

182 Once standardised, the second step is to calculate the Euclidean distances between the new  
183 measurement **xnew** and the  $k$  samples available in **Xs** (Eqn 2).

184

$$d_i = \sqrt{\sum_{j=1}^7 (\mathbf{xnew}_j - \mathbf{Xs}_{i,j})^2} \quad (\text{Equation 2})$$

185 where  $\mathbf{d}_i$  is the distance between  $\mathbf{xnew}$  and the  $i$ -th samples in  $\mathbf{Xs}$ . The main idea behind the  
 186 calculation of these distances is to determine whether the new measurement represents a new  
 187 state of the system compared to previous samples. For this purpose, all distance value  $\mathbf{d}_i$  are  
 188 compared to a threshold value, denoted  $t$ . If one of these distances is smaller than  $t$ , it is  
 189 considered that the measurement  $\mathbf{xnew}$  is already represented in the  $\mathbf{Xs}$  database. Conversely,  
 190 if all distances are greater than  $t$ , this measurement is considered to represent a new state of the  
 191 system. In this case, a sample is taken and  $\mathbf{xnew}$  is added to the  $\mathbf{Xs}$  database for the next ML  
 192 measurements.

193 Thirdly, the threshold value  $t$  is calculated with the Equation 3:

194

$$t = dto \times a + b \quad (\text{Equation 3})$$

195 where slope  $a$  and intercept  $b$  values are chosen by the users, as explained in detail in the *Results*  
 196 & *discussion* section. The coefficient  $dto$  is the distance between the origin and the new  
 197 measurement,  $\mathbf{xnew}$ .  $dto$  is defined as (Eqn 4):

198

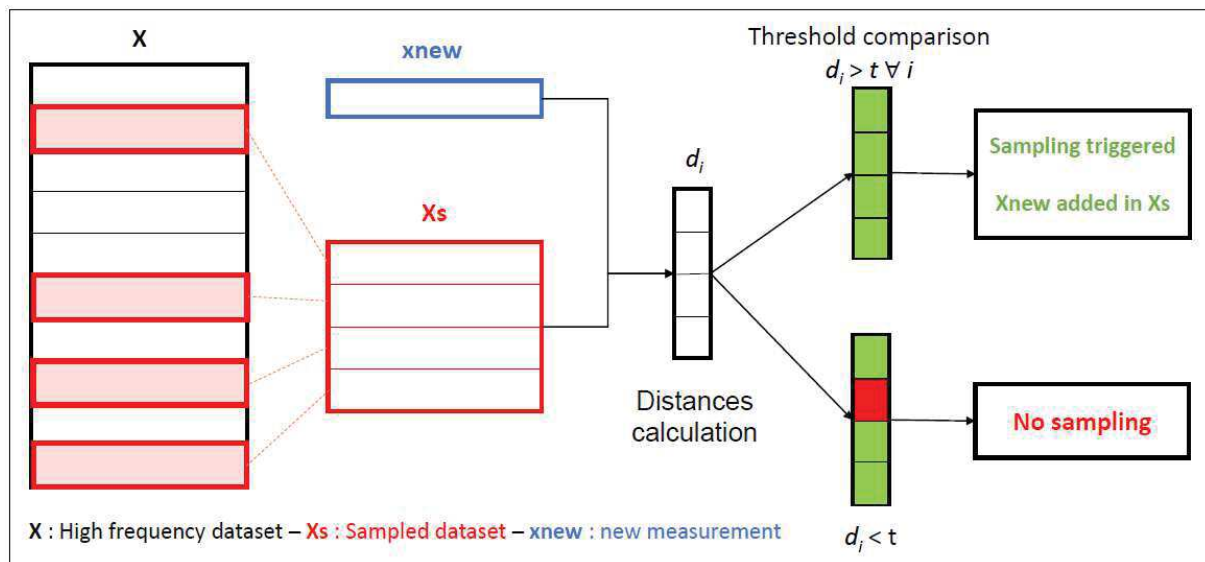
$$dto = \sqrt{\sum_{j=1}^7 (\mathbf{xnew}_j)^2} \quad (\text{Equation 4})$$

199 It should be noted that the parameters chosen for the calculation of  $t$  are of paramount  
 200 importance to extract maximum information while balancing the number of samples collected.  
 201 Making  $t$  dependent on the  $dto$  for each measurement tested by the OSA allows for a better  
 202 adaptability to the variability of the system studied. In contrast to the proposed procedure, a  
 203 fixed  $t$ -value could be chosen. However, this would make the sampling very sensitive to  
 204 extreme events. Indeed, fixing a small value for  $t$  would allow to correctly detect fine and daily

205 variations. However, during extreme events, the mean value of *e.g* turbidity can be multiplied  
 206 by 50 and, in this case, almost all measurements would lead to the decision of withdrawing a  
 207 sample. Estimating *t* using Equation 4 allows overcoming this issue, translating into low  
 208 threshold values for regular daily variations and higher values for extreme events. In this way,  
 209 good sensitivity is ensured in normal conditions while oversampling is avoided during extreme  
 210 events.

211 The overall functioning of the OSA is summarised in Figure 1.

212



213

214 **Figure 1:** Schematic view of the OSA.

215

#### 216 **1.4. Performance control**

217 Once the samples have been identified by the OSA, it is necessary to verify their relevance. A  
 218 high frequency dataset ( $X_{rebuilt}$ , dimensions  $(n \times m)$ ) is reconstructed from the sampled  
 219 dataset ( $X_s$ ). Each non-sampled point is assigned the parameter values of the closest sampled  
 220 point. To estimate the adequacy of the relevance of samples, two approaches are taken. The  
 221 first is calculating the Pearson correlation coefficient for each parameter measured by the

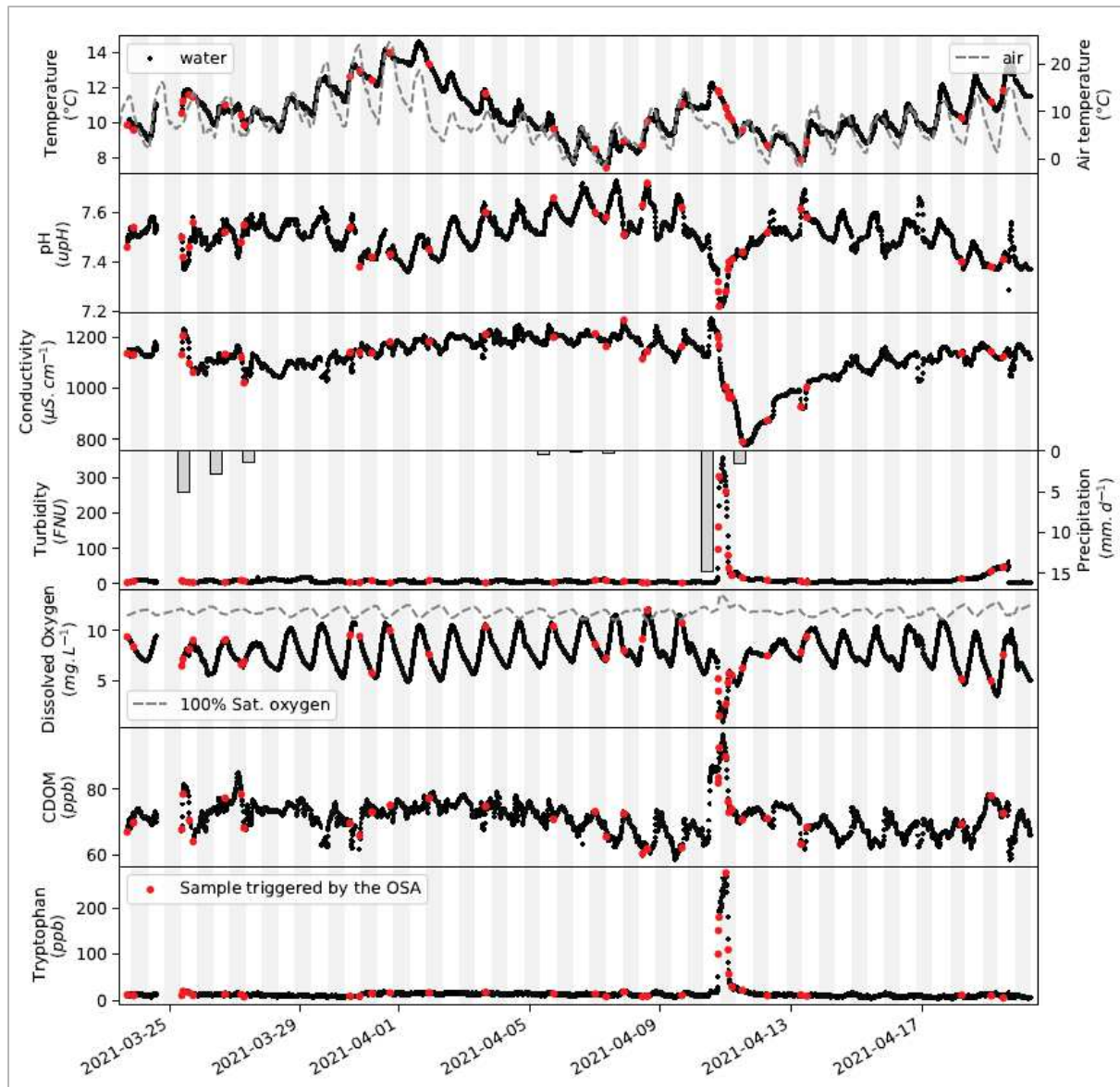
222 multiparameter probe. This evaluates the quality of the reconstruction of the complete dataset  
223 from the sampled point to a certain extent. The second is based on Principal Component  
224 Analysis (PCA). PCA is performed on both **Xrebuilt** and **X** and the loadings obtained are  
225 compared calculating Tucker Congruence Coefficient (Lorenzo-Seva and ten Berge 2006)  
226 between all principal components. In this way, we can assess the good conservation of high  
227 frequency data variability within the sampled dataset.

228

## 229 **2. Results & discussion**

### 230 **2.1. Overview of the experimental dataset acquired**

231 The first step in developing this algorithm (OSA) was to obtain a large dataset representative  
232 of the Marque River. For that purpose, a month's worth of data was collected using the  
233 multiparameter probe deployed in the ML. A total of 3754 measurements were performed from  
234 23 March to 20 April 2021, representing 26 monitoring days, with a total loss of 2 measurement  
235 days (7.7 %) due to technical issues (Figure 2). Daily cycles of temperature, pH and dissolved  
236 oxygen are clearly evidenced due to the alternance of day and night times and the development  
237 of macrophytes during this period in this highly eutrophic river (Ivanovsky and al., 2016).  
238 During this monitoring, significant meteorological evolutions also took place: (i) air daily mean  
239 temperature values evolved strongly and ranged between 2.3°C on April 7<sup>th</sup> and 17.0°C on April  
240 1<sup>st</sup>; and (ii) a heavy rainfall event was observed with 16.3 mm of water (10-11 April). The  
241 discharge of wastewaters from storm overflows was recorded during this event, leading to an  
242 important drop of dissolved oxygen and sharp peaks of dissolved organic matter. The input of  
243 rainwater in the river is also very significant as the conductivity dropped by around 40%. These  
244 events are very different (diel *vs.* punctual, small variations *vs.* plummeting/skyrocketing  
245 parameters) which makes this first dataset very relevant to optimize our algorithm.



247

248 **Figure 2:** Set of data recorded by the multiprobe and used for OSA optimisation. The red dots  
 249 represent the moments when the OSA triggers a sampling. Air temperature and daily  
 250 pluviometry have been added for information.

251

252

## 2.2.Optimisation of the OSA

The behaviour and the associated performances of the OSA have been studied from the collected dataset. The first step is to define the best combinations of  $a$  and  $b$  used in the threshold value calculation. The way in which this value is calculated affects both the number of samples and their distribution. These choices were based on preliminary tests. Different samples sets are then generated, by testing combinations of  $a$  and  $b$  over a certain range (from 0.1 to 1 for  $a$  and from 0 to 6 for  $b$ , with steps of 0.1 and 0.5 respectively). For each samples set generated, the performance control is performed as described previously by calculating the correlation coefficient between  $\mathbf{X}$  and  $\mathbf{Xrebuilt}$  and by comparing the PCA.

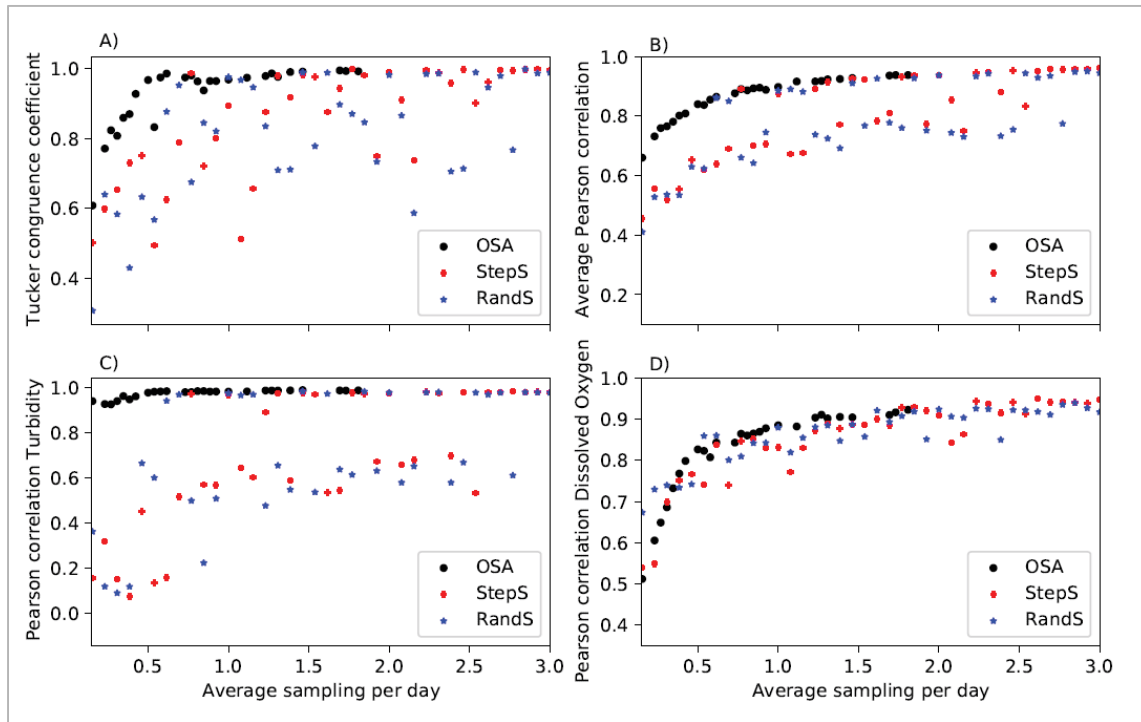
To assess the performance quality of the sampling carried out by the OSA, it is also necessary to compare these results with other sampling methods. The first comparison is made against randomly selected samples (RandS) while the second comparison is performed with a fixed step sampling method (StepS). The average sampling rates are between 0.15 and 3 samples per day for each method. This is, in our case, an operationally feasible sampling frequency range for monitoring over several months while maintaining sensitivity to one-off and daily events.

The results of these different simulations are shown in Figure 3.

Logically, whatever the methods and correlations, increasing the sampling frequency improves the description of the dataset in a non-linear way. The first observable difference between the three methods is a better stability for the OSA of the correlations with the increase of the number of samples. The RandS and StepS methods indeed show strong disparities when increasing the frequencies.

Figure 3.A shows the evolution of the average Tucker Congruence Coefficient between principal components of  $\mathbf{X}$  and  $\mathbf{Xrebuilt}$ , calculated for the different methods. This coefficient

277 has the advantage of considering all the parameters under study. OSA consistently exhibits  
 278 higher coefficients than the two other methods. Moreover, it is interesting to observe that a  
 279 ceiling seems to be reached for frequencies of the order of 1 sample per day. The gain in this  
 280 coefficient is then negligible for higher sampling frequencies.



281  
 282 **Figure 3:** Comparison between the different sampling methods: OSA, fixed step (StepS) and  
 283 random (RandS). For the OSA, each black dot represents a combination of a and b. Figure A  
 284 shows the Tucker Congruence Coefficient. Figure B shows the average Pearson correlation.  
 285 Figure C shows the Pearson correlation on turbidity. Figure D represents the Pearson correlation  
 286 on Dissolved Oxygen.

287  
 288 Figure 3.B shows the mean value of the Pearson correlations between parameters in  $\mathbf{X}$  and  
 289  $\mathbf{X}_{rebuilt}$ . The use of the average of these coefficients allows to approximate a multivariate  
 290 visualisation of the system. Again, the OSA shows both better results and greater stability  
 291 compared to the other methods.

292 Figures 3.C and 3.D display the Pearson correlations of two parameters: turbidity and dissolved  
 293 oxygen. These two correlations show very different behaviours for the OSA. For turbidity, the



294 OSA systematically gives a very high correlation where the other methods rarely manage to  
295 describe this signal correctly. The main reason is the variability of the turbidity signal (and  
296 equivalently the tryptophan). For both signals (see Figure 2), the measurements are quite stable  
297 except for a strong increase from 10 to 11 April caused by a heavy rainfall. This event has a  
298 strong impact on water quality for a very short time. For “classical” sampling methods, it is  
299 usually very difficult to take samples on this kind of short event. Conversely, the OSA makes  
300 possible in a systematic way, to consider this type of phenomenon whose impacts may be  
301 important and often poorly understood. The randomness of the ability of classical methods to  
302 sample these events is also reflected in the correlations with highly scattered values, resulting  
303 from the presence or absence of sampling during this stormy period.

304 For the dissolved oxygen (and comparably for temperature, pH, conductivity and CDOM), the  
305 behaviour of the OSA is quite different. For sampling frequencies between 0.5 and 2, OSA  
306 exhibits good results compared to other methods with high stability. For frequencies above 2,  
307 all three methods give comparable correlation values. However, for low frequencies  
308 ( $< 0.3 \text{ day}^{-1}$ ), the OSA indicates lower performance than the two other methods. This is due to  
309 the nature of the operation of the OSA and the dissolved oxygen signal. Indeed, as seen  
310 previously, the OSA systematically samples the rainfall event regardless of the sampling  
311 frequency, so that the few samples are mainly taken during this event. As a result, the dissolved  
312 oxygen values identified are not representative of the overall variability as shown in Figure 2.  
313 In other words, when only a few samples are taken, extreme events will be prioritized over  
314 small daily variations.

315 Finally, from these data, it is possible to choose a pair of values for the parameters  $a$  and  $b$   
316 corresponding to the objectives and limitations of the study under consideration. Adjustments  
317 of the **msv** values can also be made to slightly adjust the sensitivity of the OSA on the different

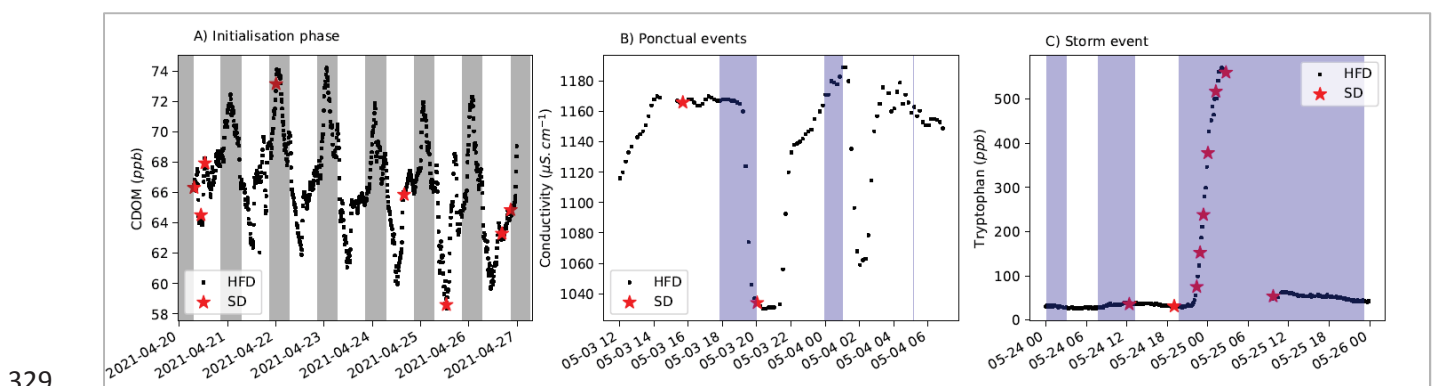


318 parameters. However, these modifications must be made with an awareness of technical  
319 limitations and environmental variations.

### 320 2.3. Application of the OSA to a monitoring campaign

321 The OSA was used for a campaign conducted from April 20<sup>th</sup> to June 28<sup>th</sup>, 2021, on the Marque  
322 River. During this period, 103 samples were taken, corresponding to an average frequency of  
323 1.6 samples per day. This frequency is higher than that predicted by the previous simulation  
324 (1.4), probably due to the high variability observed during the campaign and the strong weather  
325 changes due to the transition towards the summer season.

326 The OSA sampling system ensures a good representation of the environment, by taking samples  
327 during events that have a strong impact on the environment, regardless of their duration. Some  
328 examples of sampled events are shown Figure 4.



330 **Figure 4:** Three examples of the OSA response. Figure A shows the launch of the algorithm  
331 and its initialisation/discovery phase with day/night cycle in white/gray stripes. Figure B shows  
332 its reaction to a brief one-off phenomenon and its learning capacity. Figure C shows the ability  
333 of the algorithm to adapt its measurement frequency according to the observed variations. In B  
334 and C figures, the purple stripes correspond to the input of water from a nearby lake 500 m  
335 upstream of the station.

336

337 Figure 4.A. shows the launch of the OSA over the first 7 days, represented for only one  
338 parameter, with close sampling during the launch (initialization/discovery phase). This is

339 followed by periods without additional sampling as variability remains as low as previously.  
340 Figure 4.B. shows a one-off event of high dilution of the river by the overflow of the retention  
341 basin (the Heron lake), located just upstream. The purple areas correspond to the periods during  
342 which the water from this pond is pumped into the river; the time lag between the discharge  
343 and the impact on the conductivity is due to the distance between the discharge and the mobile  
344 laboratory. The OSA can trigger a sampling during this brief period (less than two hours), but  
345 also not to take a sample again when this event reappears some hours later. The last example  
346 (Figure 4.C.) shows the ability of the OSA to multiply samples during periods of high  
347 variability, here using the example of heavy rainfall leading to a large increase in turbidity.  
348 These periods are often critical for environmental studies and require special attention, here  
349 represented by the increase in the number of samples taken over a short period.

350 The parameters  $a$  and  $b$  identified in the test phase produced excellent results in this campaign.  
351 All correlation coefficients are above 0.96, with a Tucker congruence of 0.998. These excellent  
352 results despite different environmental changes than those observed during the simulation  
353 clearly validate the transposition of the OSA settings over different periods (HF data and the  
354 sampling points are displayed in the *Supplementary Information*, as well as the performance  
355 indicator on this period).

356

#### 357 **2.4. OSA limitation**

358 According to these findings and our experience feedback in the field, several points of vigilance  
359 must be mentioned for an optimized deployment in routine of the OSA.

360 As for any data treatment, bad data lead to bad analysis. The OSA is optimised to detect changes  
361 and will be especially sensitive to probe fouling and drift as well as recalibrations and cleaning  
362 of the instrument. For example, pH sensor re-calibration after a long period without

363 maintenance (*e.g.* several weeks) led to an over-sampling of the daily cycles, despite them  
364 having been characterized previously. To limit this kind of bias, a regular maintenance of the  
365 multiprobe have been implemented (cleaning and calibration). A weekly frequency has been  
366 chosen in this river based on the observation of the fouling, but it could be adapted depending  
367 on the characteristics of the studied water body and weather conditions (*e.g.* summer *vs.* winter).  
368 Furthermore, traceability of the maintenance and calibration must be ensured, if possible  
369 automatically, to allow an *a posteriori* understanding of the sampling by the OSA.

370 OSA is also intended to be a tool for detecting the variability occurring in a system. With good  
371 optimisation, it should be able to sample during both small and extreme phenomena. However,  
372 for lower sampling frequencies (of the order of a week, for example), only extreme events will  
373 be sampled by the OSA. The “baseline” status of the river will systematically be dismissed by  
374 the algorithm and so the information associated with it as well. Figure 3.D confirms that a  
375 misrepresentation can be observed at low sampling rate and that the OSA can become worst  
376 than random sampling in such configuration.

377 The seven parameters measured with the probe can sometimes be much correlated (*e.g.*  
378 dissolved oxygen with pH are correlated with an  $R=0.79$  over 9 months in 2021). Therefore,  
379 there is a risk that using them all can give a lot of statistical importance to the group of  
380 parameters varying together. However, there is always the chance that a decorrelation might  
381 occur, indicative of a new phenomenon happening, and the OSA should in this case be able to  
382 detect it. That is why the choice was made to keep all parameters.

383 Finally, with a more operational vision, the non-regular distribution of samples over time can  
384 be problematic. Indeed, it is possible to have no samples over several days and then 8 samples  
385 over one day during a storm. It requires flexible human resources and alert systems to grab  
386 collected samples.

## 387 **Conclusion**

388 This study was dedicated to the development, optimisation, and validation of a decision support  
389 algorithm for taking samples following multiparametric HF measurements. It allows the overall  
390 variability of the data to be maintained while reducing the number of samples collected. OSA  
391 is particularly suitable for sampling short-lived events with a high environmental impact.

392 To our best knowledge, this is the first approach of this type of sampling based on on-line  
393 multiparameter measurements. This tool is a particular response to the difficulty observed in  
394 many studies of taking samples on short and difficult to predict events. Even if it remains a  
395 perfectible tool (e.g. **msv** values could be further optimised in the future), the realisation of a  
396 campaign in spring 2021 has proved its operational applicability.

397 This type of sampling will be very useful for studies where a large variety of samples are  
398 necessary to insure a statistical robustness. Typically, it will be interesting for dissolved organic  
399 matter studies in which fluorescence excitation emission matrices are measured, as the  
400 exploitation of these matrices with the deconvolution algorithm Parafac requires some  
401 variability in the dataset to have a robust model in the end. More generally, OSA could be of  
402 interest in any environmental study that could benefit from such a system as it should improve  
403 the strength of the correlation or PCA results.

404

## 405 **Acknowledgment**

406 We acknowledge Artois-Picardie Water Agency (AEAP) and the region Haute de France for  
407 cofunding the PhD of JM. The Region Hauts de France and the French Government are also  
408 warmly acknowledged through the founding of the CPERs Climibio and ECRIN that funded  
409 part of the monitoring station. We are grateful to Sourceo for allowing us to deploy the station  
410 on their site and to the Lille European Metropolis (MEL) for providing complementary

411 information and WWTP data. The authors wish to thank the European Commission funding for  
412 the LIFE RUBIES project (LIFE20 ENV/000179). We thank the lab members Jean-Pierre  
413 Verwaerde, Viviane Blotiau and Vincent Carlucci for their help with building and maintaining  
414 the station.

415

#### 416 **References:**

- 417 Aguilera, R., Livingstone, D.M., Marcé, R., Jennings, E., Piera, J., Adrian, R., 2016. Using dynamic factor  
418 analysis to show how sampling resolution and data gaps affect the recognition of patterns in  
419 limnological time series. *Inland Waters* **6**, 284–294. <https://doi.org/10.1080/IW-6.3.948>
- 420 Bieroza, M.Z., Heathwaite, A.L., 2016. Unravelling organic matter and nutrient biogeochemistry in  
421 groundwater-fed rivers under baseflow conditions: Uncertainty in in situ high-frequency  
422 analysis. *Science of The Total Environment* **572**, 1520–1533.  
423 <https://doi.org/10.1016/j.scitotenv.2016.02.046>
- 424 Bieroza, M.Z., Heathwaite, A.L., 2015. Seasonal variation in phosphorus concentration–discharge  
425 hysteresis inferred from high-frequency in situ monitoring. *Journal of Hydrology* **524**, 333–347.  
426 <https://doi.org/10.1016/j.jhydrol.2015.02.036>
- 427 Carstea, E.M., Baker, A., Bieroza, M., Reynolds, D., 2010. Continuous fluorescence excitation–emission  
428 matrix monitoring of river organic matter. *Water Research* **44**, 5356–5366.  
429 <https://doi.org/10.1016/j.watres.2010.06.036>
- 430 Ferrant, S., Laplanche, C., Durbe, G., Probst, A., Dugast, P., Durand, P., Sanchez-Perez, J.M., Probst, J.L.,  
431 2013. Continuous measurement of nitrate concentration in a highly event-responsive  
432 agricultural catchment in south-west of France: is the gain of information useful?: HIGH-  
433 FREQUENCY SAMPLING OF NITRATE FLUSHING. *Hydrol. Process.* **27**, 1751–1763.  
434 <https://doi.org/10.1002/hyp.9324>
- 435 Gunatilaka, A., Diehl, P., 2001. A Brief Review of Chemical and Biological Continuous Monitoring of  
436 Rivers in Europe and Asia. In 'Biomonitoring and Biomarkers as Indicators of Environmental  
437 Change 2' (Eds. Butterworth, F.M., Gunatilaka, A., Gonsebatt, M.E.). Springer US, Boston, MA,  
438 pp. 9–28. [https://doi.org/10.1007/978-1-4615-1305-6\\_2](https://doi.org/10.1007/978-1-4615-1305-6_2)
- 439 Halliday, S., Skeffington, R., Bowes, M., Gozzard, E., Newman, J., Loewenthal, M., Palmer-Felgate, E.,  
440 Jarvie, H., Wade, A., 2014. The Water Quality of the River Enborne, UK: Observations from  
441 High-Frequency Monitoring in a Rural, Lowland River System. *Water* **6**, 150–180.  
442 <https://doi.org/10.3390/w6010150>
- 443 Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E.,  
444 Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M.,  
445 Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy,  
446 T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with  
447 NumPy. *Nature* **585**, 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- 448 Ivanovsky, A., Criquet, J., Dumoulin, D., Alary, C., Prygiel, J., Duponchel, L., Billon, G., 2016. Water  
449 quality assessment of a small peri-urban river using low and high frequency monitoring.  
450 *Environ. Sci.: Processes Impacts* **18**, 624–637. <https://doi.org/10.1039/C5EM00659G>
- 451 Jarvie, H.P., Sharpley, A.N., Kresse, T., Hays, P.D., Williams, R.J., King, S.M., Berry, L.G., 2018. Coupling  
452 High-Frequency Stream Metabolism and Nutrient Monitoring to Explore Biogeochemical  
453 Controls on Downstream Nitrate Delivery. *Environ. Sci. Technol.* **52**, 13708–13717.  
454 <https://doi.org/10.1021/acs.est.8b03074>

455 Khamis, K., Bradley, C., Hannah, D.M., 2020. High frequency fluorescence monitoring reveals new  
456 insights into organic matter dynamics of an urban river, Birmingham, UK. *Science of The Total*  
457 *Environment* **710**, 135668. <https://doi.org/10.1016/j.scitotenv.2019.135668>

458 Lewis, J., Eads, R., 2009. Implementation guide for turbidity threshold sampling: principles, procedures,  
459 and analysis (No. PSW-GTR-212). U.S. Department of Agriculture, Forest Service, Pacific  
460 Southwest Research Station, Albany, CA. <https://doi.org/10.2737/PSW-GTR-212>

461 Liu, X., Beusen, A.H.W., Van Beek, L.P.H., Mogollón, J.M., Ran, X., Bouwman, A.F., 2018. Exploring  
462 spatiotemporal changes of the Yangtze River (Changjiang) nitrogen and phosphorus sources,  
463 retention and export to the East China Sea and Yellow Sea. *Water Research* **142**, 246–255.  
464 <https://doi.org/10.1016/j.watres.2018.06.006>

465 Lorenzo-Seva, U., ten Berge, J.M.F., 2006. Tucker’s Congruence Coefficient as a Meaningful Index of  
466 Factor Similarity. *Methodology* **2**, 57–64. <https://doi.org/10.1027/1614-2241.2.2.57>

467 Marcé, R., George, G., Buscarinu, P., Deidda, M., Dunalska, J., de Eyto, E., Flaim, G., Grossart, H.-P.,  
468 Istvanovics, V., Lenhardt, M., Moreno-Ostos, E., Obrador, B., Ostrovsky, I., Pierson, D.C.,  
469 Potužák, J., Poikane, S., Rinke, K., Rodríguez-Mozaz, S., Staehr, P.A., Šumberová, K., Waajen,  
470 G., Weyhenmeyer, G.A., Weathers, K.C., Zion, M., Ibelings, B.W., Jennings, E., 2016. Automatic  
471 High Frequency Monitoring for Improved Lake and Reservoir Management. *Environ. Sci.*  
472 *Technol.* **50**, 10780–10794. <https://doi.org/10.1021/acs.est.6b01604>

473 Meyer, A.M., Fuenfrocken, E., Kautenburger, R., Cairault, A., Beck, H.P., 2021. Detecting pollutant  
474 sources and pathways: High-frequency automated online monitoring in a small rural  
475 French/German transborder catchment. *Journal of Environmental Management* **290**, 112619.  
476 <https://doi.org/10.1016/j.jenvman.2021.112619>

477 Nimick, D.A., Gammons, C.H., Parker, S.R., 2011. Diel biogeochemical processes and their effect on the  
478 aqueous chemistry of streams: A review. *Chemical Geology* **283**, 3–17.  
479 <https://doi.org/10.1016/j.chemgeo.2010.08.017>

480 Piniewski, M., Marcinkowski, P., Koskiaho, J., Tattari, S., 2019. The effect of sampling frequency and  
481 strategy on water quality modelling driven by high-frequency monitoring data in a boreal  
482 catchment. *Journal of Hydrology* **579**, 124186. <https://doi.org/10.1016/j.jhydrol.2019.124186>

483 Reback, J., Jbrockmendel, McKinney, W., Van Den Bossche, J., Augspurger, T., Roeschke, M., Hawkins,  
484 S., Cloud, P., Gfyoung, Sinhrks, Hoefler, P., Klein, A., Terji Petersen, Tratner, J., She, C., Ayd, W.,  
485 Naveh, S., JHM Darbyshire, Garcia, M., Shadrach, R., Schendel, J., Hayden, A., Saxton, D.,  
486 Gorelli, M.E., Fangchen Li, Zeitlin, M., Jancauskas, V., McMaster, A., Wörtwein, T., Battiston,  
487 P., 2022. pandas-dev/pandas: Pandas 1.4.2. Zenodo.  
488 <https://doi.org/10.5281/ZENODO.3509134>

489 Reynolds, K.N., Loecke, T.D., Burgin, A.J., Davis, C.A., Riveros-Iregui, D., Thomas, S.A., St. Clair, M.A.,  
490 Ward, A.S., 2016. Optimizing Sampling Strategies for Riverine Nitrate Using High-Frequency  
491 Data in Agricultural Watersheds. *Environ. Sci. Technol.* **50**, 6406–6414.  
492 <https://doi.org/10.1021/acs.est.5b05423>

493 Rode, M., Wade, A.J., Cohen, M.J., Hensley, R.T., Bowes, M.J., Kirchner, J.W., Arhonditsis, G.B., Jordan,  
494 P., Kronvang, B., Halliday, S.J., Skeffington, R.A., Rozemeijer, J.C., Aubert, A.H., Rinke, K.,  
495 Jomaa, S., 2016. Sensors in the Stream: The High-Frequency Wave of the Present. *Environ. Sci.*  
496 *Technol.* **50**, 10297–10307. <https://doi.org/10.1021/acs.est.6b02155>

497 Searcy, R.T., Boehm, A.B., 2021. A Day at the Beach: Enabling Coastal Water Quality Prediction with  
498 High-Frequency Sampling and Data-Driven Models. *Environ. Sci. Technol.* **55**, 1908–1918.  
499 <https://doi.org/10.1021/acs.est.0c06742>

500 Seifert, A.-G., Roth, V.-N., Dittmar, T., Gleixner, G., Breuer, L., Houska, T., Marxsen, J., 2016. Comparing  
501 molecular composition of dissolved organic matter in soil and stream water: Influence of land  
502 use and chemical characteristics. *Science of The Total Environment* **571**, 142–152.  
503 <https://doi.org/10.1016/j.scitotenv.2016.07.033>

504 Seifert-Dähnn, I., Furuseth, I.S., Vondolia, G.K., Gal, G., de Eyto, E., Jennings, E., Pierson, D., 2021. Costs  
505 and benefits of automated high-frequency environmental monitoring – The case of lake water

506 management. *Journal of Environmental Management* **285**, 112108.  
507 <https://doi.org/10.1016/j.jenvman.2021.112108>  
508 Shultz, M., Pellerin, B., Aiken, G., Martin, J., Raymond, P., 2018. High Frequency Data Exposes Nonlinear  
509 Seasonal Controls on Dissolved Organic Matter in a Large Watershed. *Environ. Sci. Technol.* **52**,  
510 5644–5652. <https://doi.org/10.1021/acs.est.7b04579>  
511 Superville, P.-J., Prygiel, E., Magnier, A., Lesven, L., Gao, Y., Baeyens, W., Ouddane, B., Dumoulin, D.,  
512 Billon, G., 2014. Daily variations of Zn and Pb concentrations in the Deûle River in relation to  
513 the resuspension of heavily polluted sediments. *Science of The Total Environment* **470–471**,  
514 600–607. <https://doi.org/10.1016/j.scitotenv.2013.10.015>  
515 Vaughan, M.C.H., Bowden, W.B., Shanley, J.B., Vermilyea, A., Schroth, A.W., 2019. Shining light on the  
516 storm: in-stream optics reveal hysteresis of dissolved organic matter character.  
517 *Biogeochemistry* **143**, 275–291. <https://doi.org/10.1007/s10533-019-00561-w>  
518

# Optimising punctual water sampling with an on-the-fly algorithm based on multiparameter high-frequency measurements

Supplementary information

*Jérémy Mougin<sup>1</sup>, Pierre-Jean Superville<sup>1\*</sup>, Cyril Ruckebusch<sup>1</sup>, Gabriel Billon<sup>1</sup>*

<sup>1</sup>Université Lille, CNRS, UMR 8516 - LASIRE, Laboratoire de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000 Lille, France.

## **Content**

p.2: Map of the site

p.3: First campaign results

p.4: OSA performance during the campaign

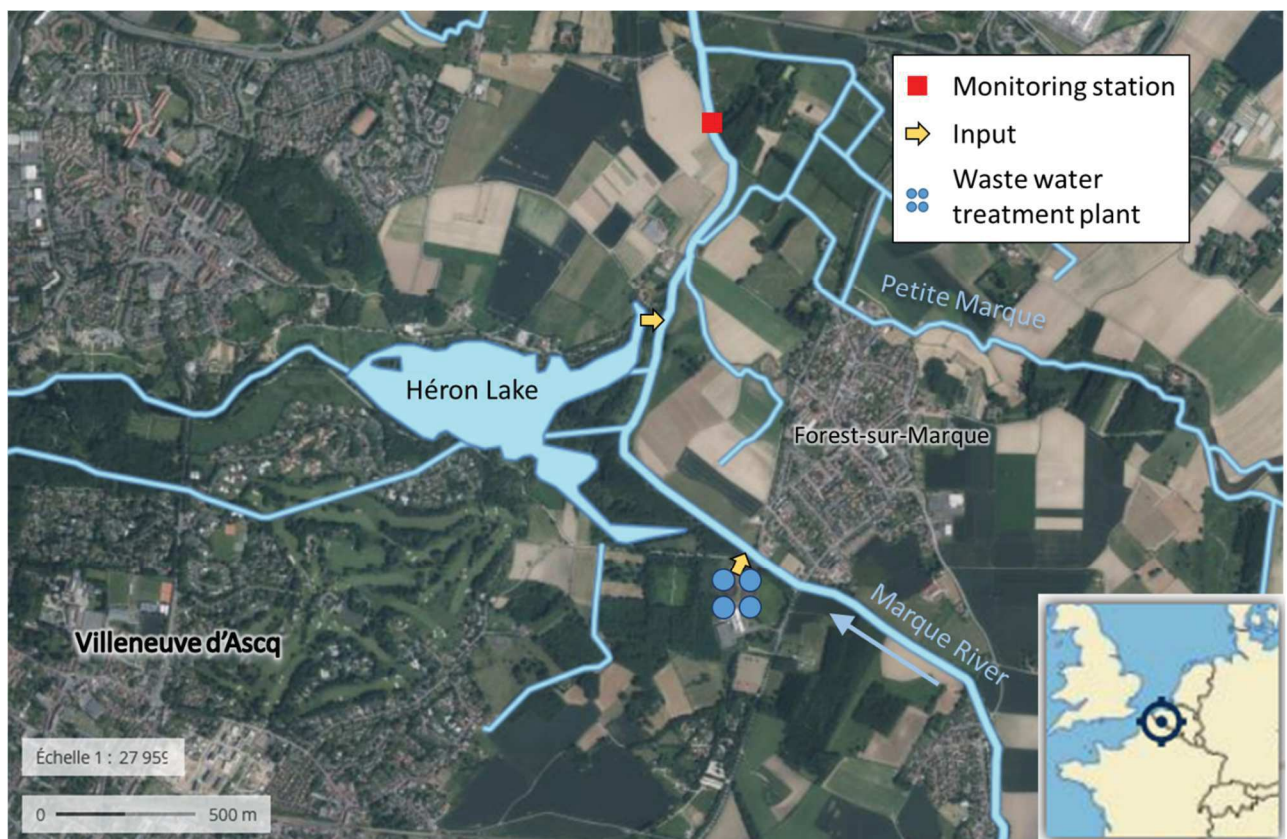
p.5: Checking and optimizing the msv



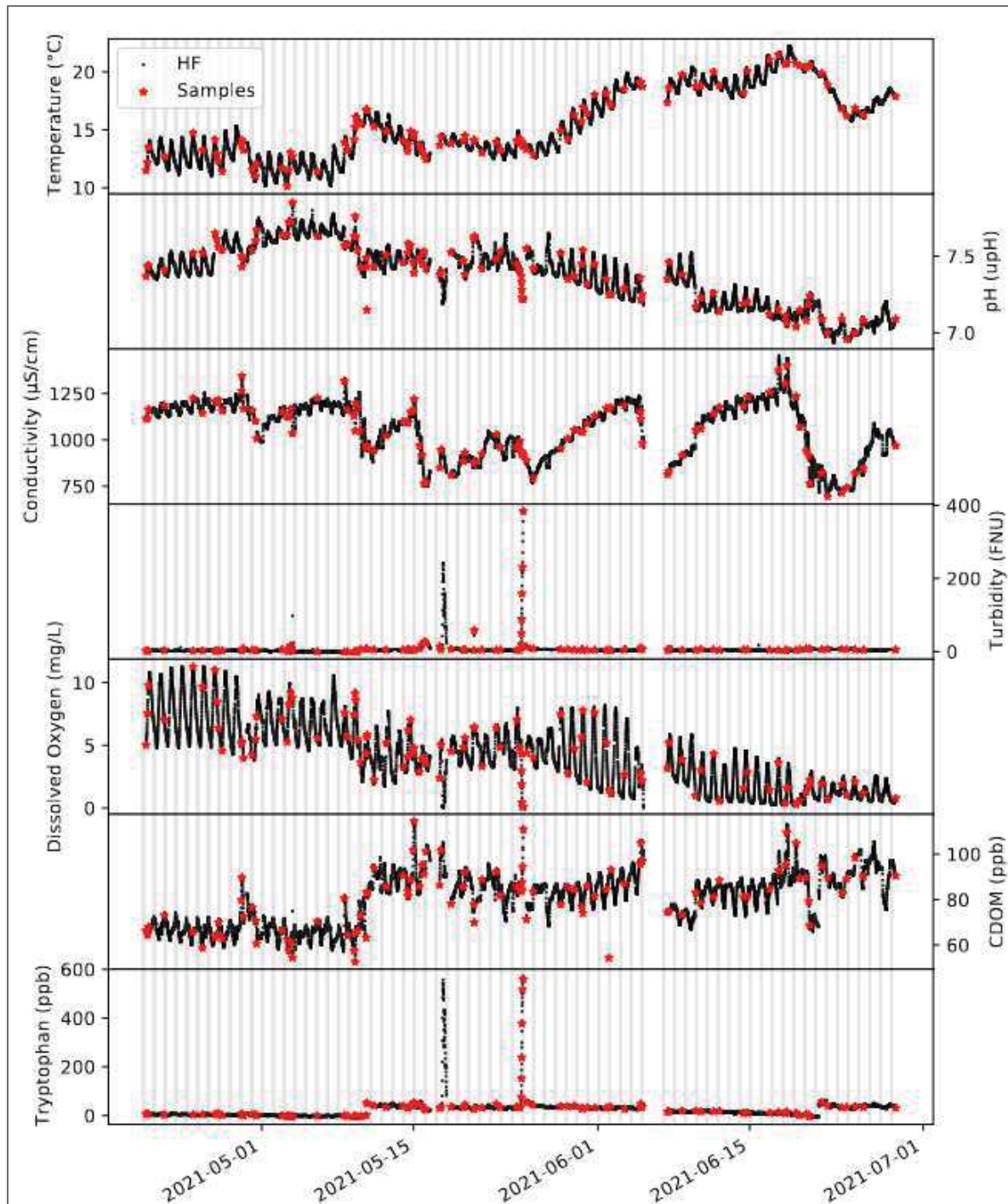
## A. Map of the site

The monitoring station is situated in Villeneuve d'Ascq, near Lille, in northern France, on the Marque River. Both rural and urban pressures impact this small river. Specifically on the site of the monitoring, 3 inputs can be highlighted:

- the wastewater treatment plant of Villeneuve d'Ascq (150 000 inhabitant equivalent) 3km upstream.
- the Heron Lake. Its waters are a mix of rainwater and some domestic untreated waters, biologically treated in a chain of lakes ending in the Heron Lake. When the level of the Lake is too high, the water is pumped into the Marque River.
- The Petite Marque is a network of ditches in the middle of fields with a few occasional farms. Agricultural and domestic impacts can be expected.



## B. First campaign results



It can be noted that the first big storm was not sampled due to a clogging of the automated sampler.

### C. OSA performance during the campaign

Parameter correlation between the original high frequency dataset and the dataset rebuilt from the samples:

	Temperature	pH	Conductivity	Turbidity	Dissolved O <sub>2</sub>	CDOM	Tryptophane
Parameter self-correlation	0.98	0.99	0.97	0.98	0.97	0.96	0.99

PCA self-correlation, i.e. Tucker congruence coefficient between the original high frequency dataset and the dataset rebuilt from the samples:

	PC1 rebuilt (47 %)	PC2 rebuilt (26 %)	PC3 rebuilt (13 %)	PC4 rebuilt (6 %)	PC5 rebuilt (4 %)	PC6 rebuilt (2 %)	PC7 rebuilt (1 %)
PC1 original (48 %)	<b>0.999</b>	0.039	0.012	0.002	0.001	0.001	0.011
PC2 original (26 %)	-0.039	<b>0.997</b>	-0.055	0.018	-0.027	0.004	0.001
PC3 original (13 %)	-0.014	0.055	<b>0.997</b>	-0.038	-0.005	-0.009	0.011
PC4 original (6 %)	0.001	0.016	-0.039	<b>0.997</b>	0.008	-0.060	-0.011
PC5 original (4 %)	0.002	-0.027	-0.004	-0.008	<b>-0.999</b>	-0.004	-0.003
PC6 original (2 %)	-0.001	-0.003	0.006	-0.061	-0.004	<b>0.996</b>	0.051
PC7 original (1 %)	-0.011	-0.002	0.012	-0.008	-0.003	-0.051	<b>0.998</b>

#### D. Checking and optimizing the *msv*

The figure D.1 presents the Pearson correlation coefficient between the raw data and the rebuilt data for different ways of sampling. It shows in particular the importance of choosing properly the values of the minimal significant variation, *msv*.

The black points in the Figure D.1 correspond to the chosen *msv* for the rest of the study. The values for each parameter can be found in the Table D.2. in the line *msv* OSA. It can be seen that most of the Pearson correlation coefficient are quite good. The lowest is for pH (R=0.79) but this parameter is very noisy so it is not expected to have a perfect correlation.

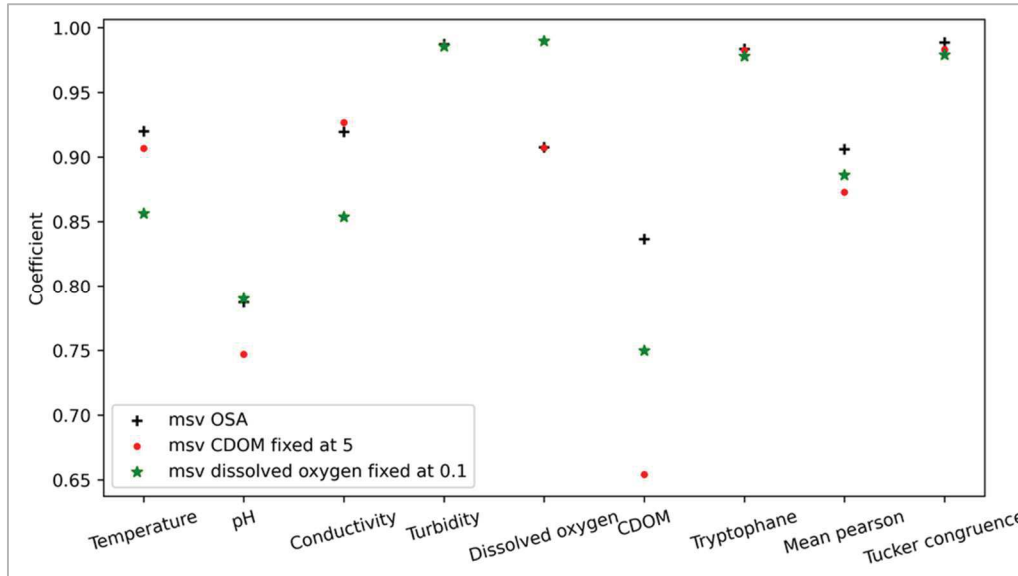


Figure D.1: Pearson correlation coefficient between the raw data and the rebuilt data for different values of *msv*.

Table D.2: values of the *msv* for the seven parameters

	Temp.	pH	Cond.	Turb.	Diss. O <sub>2</sub>	CDOM	Trypt.
<i>msv</i> OSA	0.4	0.1	25	5	<b>0.5</b>	<b>2</b>	5
<i>msv</i> 2 CDOM = 5	0.4	0.1	25	5	0.5	<b>5</b>	5
<i>msv</i> 3 DO = 0.1	0.4	0.1	25	5	<b>0.1</b>	2	5

If the *msv* for a parameter is increased, *e.g.* CDOM *msv* is changed from 2 to 5, a lower sensitivity toward this parameter should be expected. And a decrease in the correlation coefficient is indeed observed, from R=0.84 to R=0.65.

If the *msv* for a parameter is decreased, *e.g.* Dissolved oxygen *msv* is changed from 0.5 to 0.1, smaller variation of oxygen concentration will be detected as significant. Representativity of our samples are now increase in term of oxygen (R goes from 0.91 to 0.99). But it is important to note that overweighing dissolved oxygen has consequences on other parameters. Temperature (R= 0.92 to 0.86), conductivity (0.91 to 0.85) and CDOM (0.84 to 0.75) have a lower correlation coefficient, indicating a worst representativity of the sample. In a way, most of the samples are now dedicated to improve the O<sub>2</sub> signal and less are left for the other parameters. Therefore it is important to have a test phase in order to properly balance this *msv* vector.