Automatic optimization of temporal monitoring schemes dealing with daily water contaminant concentration patterns

Supplementary material

Number of pages: 9; number of algorithms: 2; number of figures: 11

M. Gabrielli[1], F. Trovò[2], M. Antonelli[1*]

[1] Dipartimento di Ingegneria Civile e Ambientale (DICA), Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy
[2] Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy

* Corresponding author: manuela.antonelli@polimi.it

The pseudo-code of Seq(GP-UCB-SW) is presented in Algorithm S1. It requires as input the set of possible sampling instants *{a_1, ..., a_n}*, the length of the sliding window SW, the time horizon of the monitoring campaign T and the first sampling instant a_0. At first, it initializes the GP and sets the next sampling instant *nextSample* equal to a_0. During the entire time horizon, if the available sampling instant *a* is equal to *nextSample*, then, sampling is performed, retrieving the sample concentration C_a. Afterwards, the samples collected in the days included in SW are selected and used to fit the GP. The fitted GP is, then, used to derive the upper and, eventually, lower confidence bounds (UCB and LCB) which are used to select the next sampling instant.

Algorithm S1. Pseudo-code of the Seq(GP-UCB-SW) algorithm. Pseudo-commands after "//" indicate variation of the algorithm for targeting maximum concentration variations.

Input: Possible sampling instants $\{a_1, \ldots, a_n\}$, sliding window length SW, time horizon T, first sampling instant a_0
Initialize GP
nextSample $\leftarrow a_0$
for $t \in \{1, \ldots, T\}$ **do**
  for $a \in \{a, \ldots, a_n\}$ **do**
    if $a =$ nextSample **then**
      Sample a and obtain measurement C_a
      Select samples collected after t - SW
      Fit GP
      Estimate UCB // Estimate UCB and LCB
      Select next sampling instant:

$$\begin{cases} nextSample \leftarrow argmax(UCB) \\ // \; nextSample \leftarrow argmax(UCB) \; and \; argmin(LCB) \end{cases}$$

    end if
  end for
end for

The pseudo-code of Seq(GP-UCB-CD) is presented in Algorithm S2. It requires as input the set of possible sampling instants *{a_1, ..., a_n}*, the length of the training window TW, the exploration probability α, the time horizon of the monitoring campaign T and the first sampling instant a_0. At first, it initializes the GP and the change point models (CPMs) regarding the maximum daily concentrations (CPM_{max}) and, eventually, the ones regarding the minimum daily concentrations (CPM_{min}) and the maximum daily variation (CPM_{delta}). It also sets the day of the first change point (CP) as 0 and the next sampling instant *nextSample* as a_0. At the beginning of each day, the vector *dailySamples* is initialized to store the concentrations of the samples collected during the day. Then, if the available sampling instant *a* is equal to *nextSample*, then, sampling is performed, retrieving the sample concentration C_a which is added to the *dailysamples* vector. Afterwards, the samples collected after the previous changepoint are selected and used to fit the GP. The fitted GP is, then, used to derive the upper and, eventually, lower confidence bounds (UCB and LCB). With probability 1 - α the next sampling instant is selected based on UCB and, eventually, LCB, while with probability α the next sampling instant is selected at random among the available options. At the end of each day after the training window is elapsed, maximum and, eventually, minimum concentrations and the maximum daily variation are derived from the *dailySamples* vector and used to update the CPMs. In case a CPM detects the presence of a changepoint, CP is set equal to the current day *t.* Successivley, all the active CPMs are resetted.

Algorithm S2. Pseudo-code of the Seq(GP-UCB-CD) algorithm. Pseudo-commands after "//" indicate variation of the algorithm for targeting maximum concentration variations.

Input: Possible sampling instants $\{a_1, \ldots, a_n\}$, training window length TW, exploration probability α, time horizon T, first sampling instant a_0
Initialize GP
Initialize CPM_{max} // Initialize CPM_{max}, CPM_{min}, CPM_{delta}
CP $\leftarrow 0$
nextSample $\leftarrow a_0$
for $t \in \{1, \ldots, \text{T}\}$ **do**
    dailySamples $\leftarrow \{\}$
    for $a \in \{a, \ldots, a_n\}$ **do**
        if $a =$ nextSample **then**
            Sample a and obtain measurement C_a
            dailySamples $\leftarrow$ dailySamples $\cup C_a$
            Select samples collected after CP
            Fit GP
            Estimate UCB // Estimate UCB and LCB
            Select next sampling instant:

$$\begin{cases} nextSample \leftarrow argmax(UCB) \ w.p. \ 1-\alpha \\ // \ nextSample \leftarrow argmax(UCB) \ and \ argmin(LCB) \ w.p. \ 1-\alpha \\ nextSample \leftarrow rand(\{a, \ldots, a_n\}) \ w.p. \ \alpha \end{cases}$$

        end if
    end for
    if t - CP $>$ TW **then**
        maxObs $\leftarrow$ max(dailySamples)
        // maxObs $\leftarrow$ max(dailySamples), minObs $\leftarrow$ min(dailySamples),
        // deltaObs $\leftarrow$ maxObs - minObs
        Update CPM_{max}
        // Update CPM_{max}(maxObs), CPM_{min}(minObs),
        // CPM_{delta}(deltaObs)
        if CPM_{max} detects change
        // CPM_{max}, CPM_{min} or CPM_{delta} detects change **then**
            CP $\leftarrow t$
            Reset CPM_{max} // Reset CPM_{max}, CPM_{min} or CPM_{delta}
        end if
    end if
end for

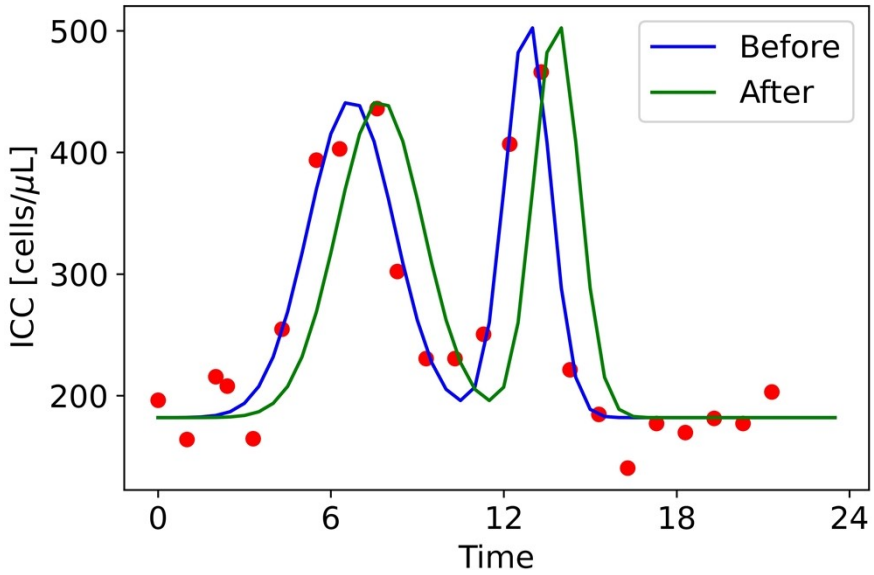

Figure S1. ICC measurements obtained from Nescerecka et al.,[1] shown as red dots, and fitted average pattern, indicated with a blue line. The green line indicates the shape of the pattern after the abrupt change.

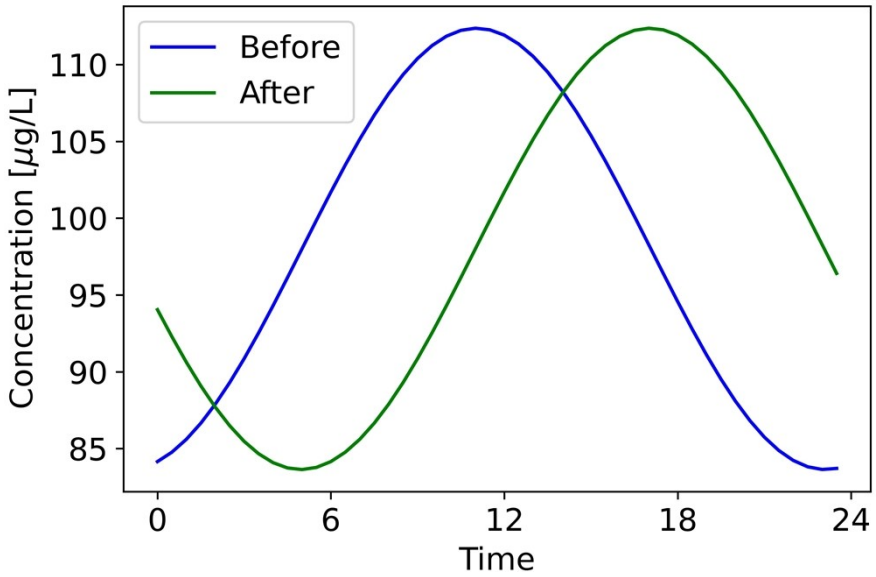

Figure S2. Average THMs concentration pattern obtained from the model developed by Chaib and Moschandreas[2], in blue, and pattern after the gradual change, in green.

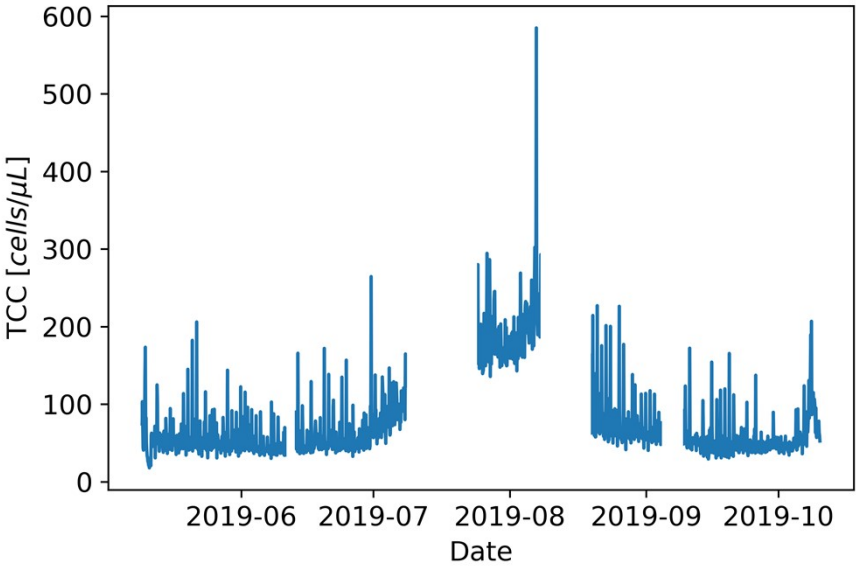

Figure S3. TCC measurements obtained from Gabrielli et al.[3].

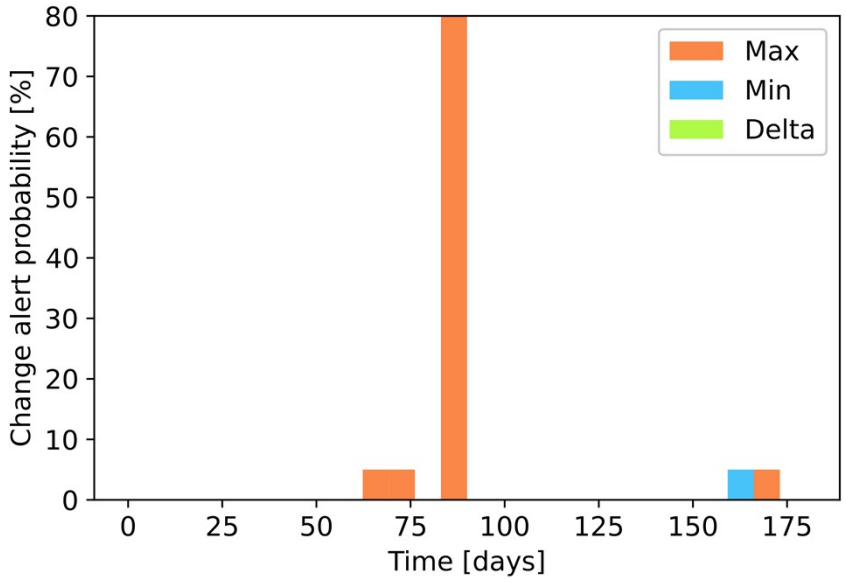

Figure S4. Histogram of the change detection alerts provided by Seq(GP-UCB-CD) in the ICC synthetic scenario.

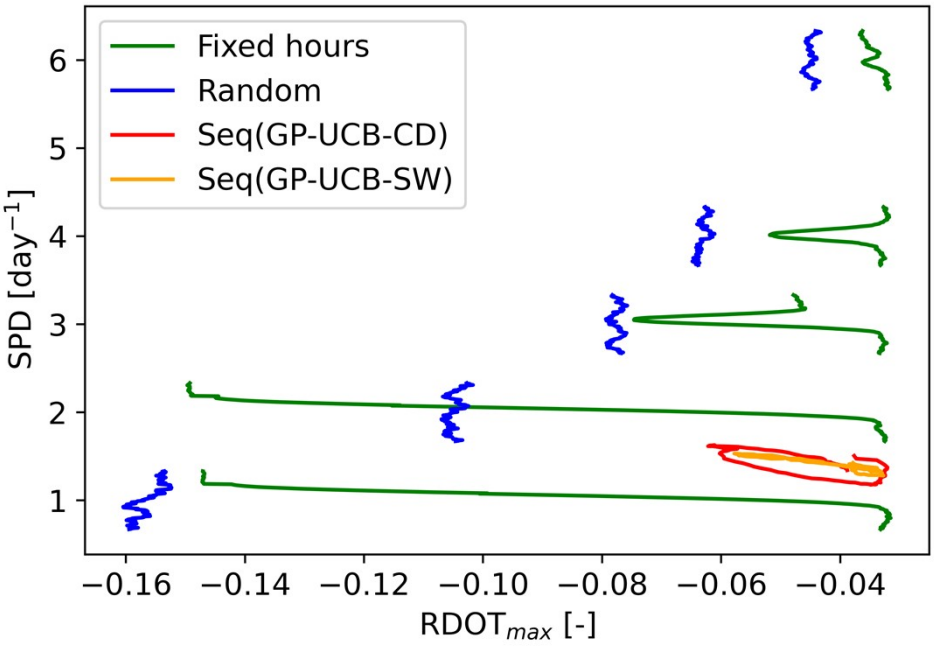

Figure S5. Average performances of tested monitoring schemes along the THMs synthetic scenario (rolling mean, n = 25). To show the temporal variation of the RDOTmax obtained by the traditional schemes along the gradual pattern change a vertical displacement was applied at each SPD value. For each SPD value, the temporal RDOTmax evolution is to be read vertically moving from the lower to the higher SPD values. To avoid clutter only the fixed-time sampling instants combination with the best performances before the pattern change was shown. The proposed algorithms' results were obtained with the following algorithms parameterization: SW = 30 d, TW = 30 d, α = 0.1.

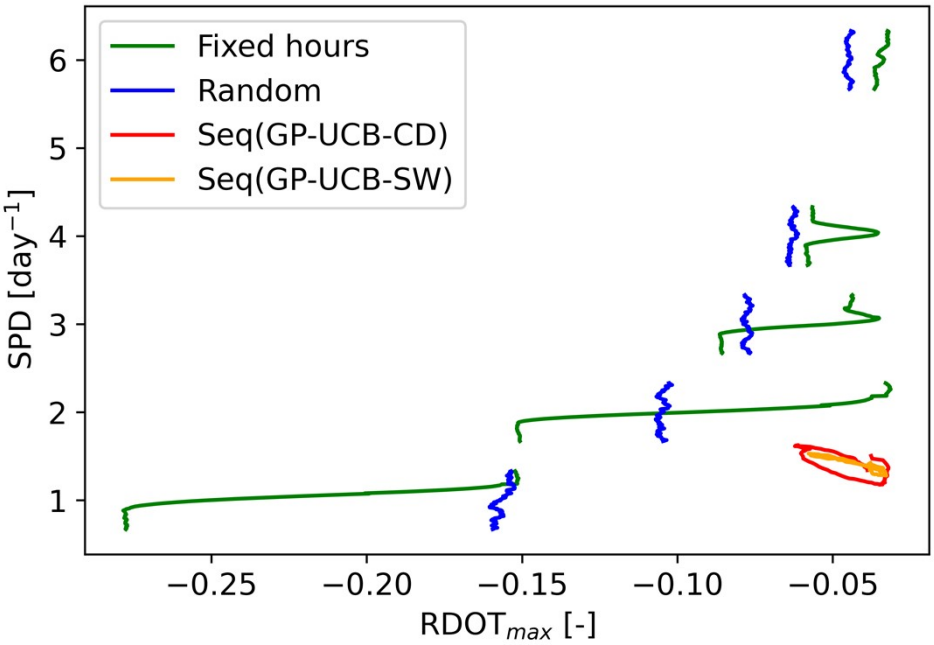

Figure S6. Average performances of tested monitoring schemes along the THMs synthetic scenario (rolling mean, n = 25). To show the temporal variation of the RDOTmax obtained by the traditional schemes along the gradual pattern change a vertical displacement was applied at each SPD value. For each SPD value, the temporal RDOTmax evolution is to be read vertically moving from the lower to the higher SPD values. To avoid clutter only the fixed-time sampling instants combination with the worst performances before the pattern change was

shown. The proposed algorithms' results were obtained with the following algorithms parameterization: SW = 30 d, TW = 30 d, α = 0.1.

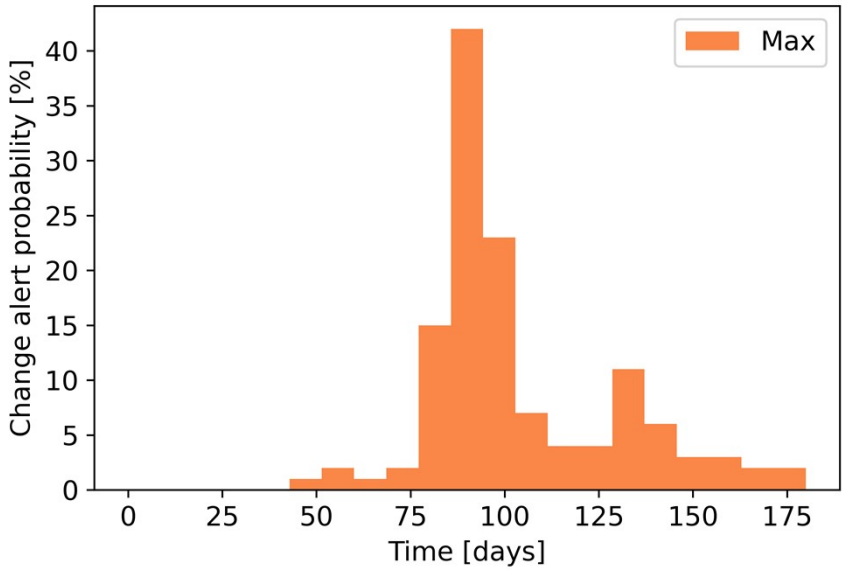

Figure S7. Histogram of the change detection alerts provided by Seq(GP-UCB-CD) in the THMs synthetic scenario.

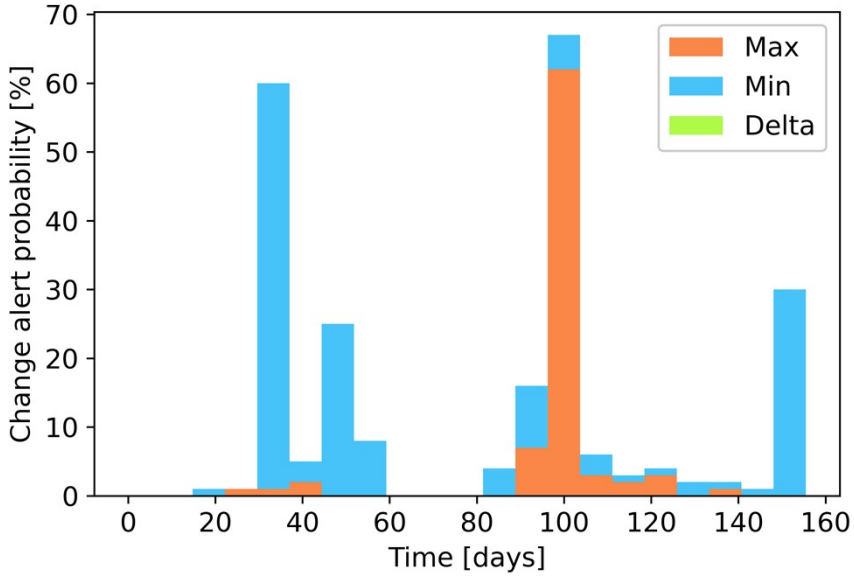

Figure S8. Histogram of the change detection alerts provided by Seq(GP-UCB-CD) in the real-world scenario.

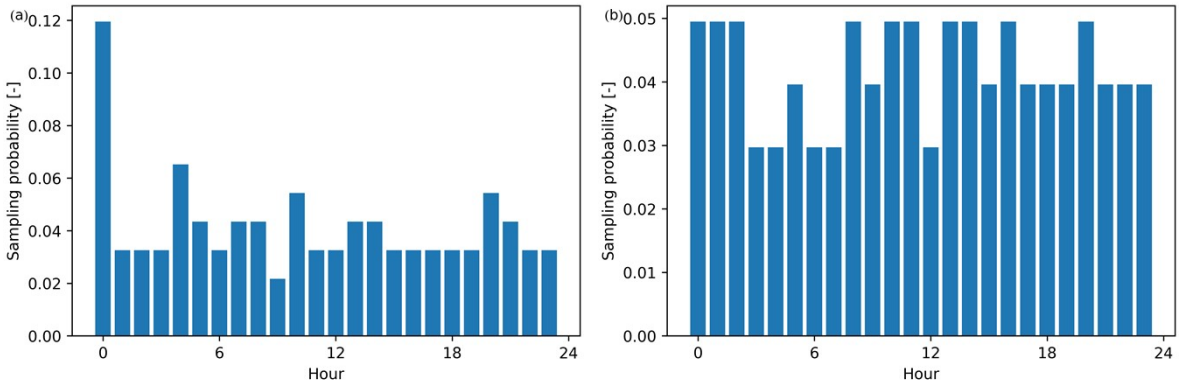

Figure S9. Sampling frequency histograms of Seq(GP-UCB-SW) in case of uniform concentration pattern characterized by minimum (a) and maximum (b) uniformity.

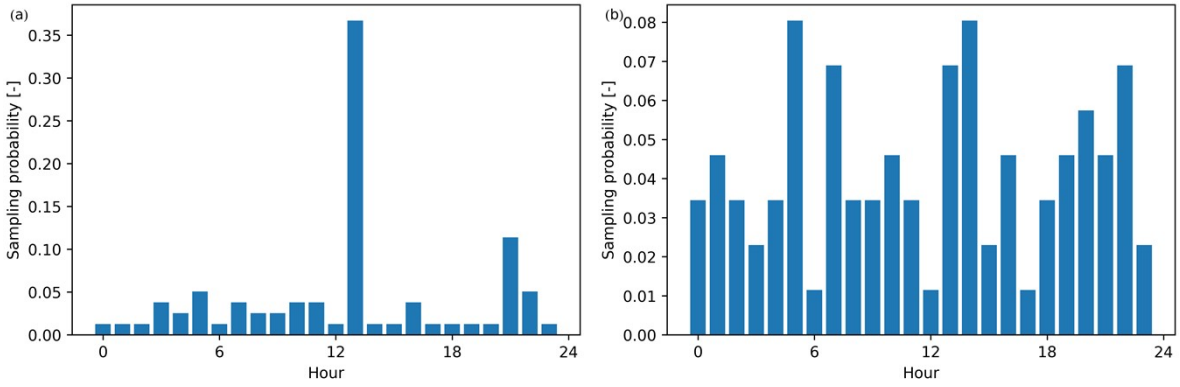

Figure S10. Sampling frequency histograms of Seq(GP-UCB-CD) in case of uniform concentration pattern characterized by minimum (a) and maximum (b) uniformity.

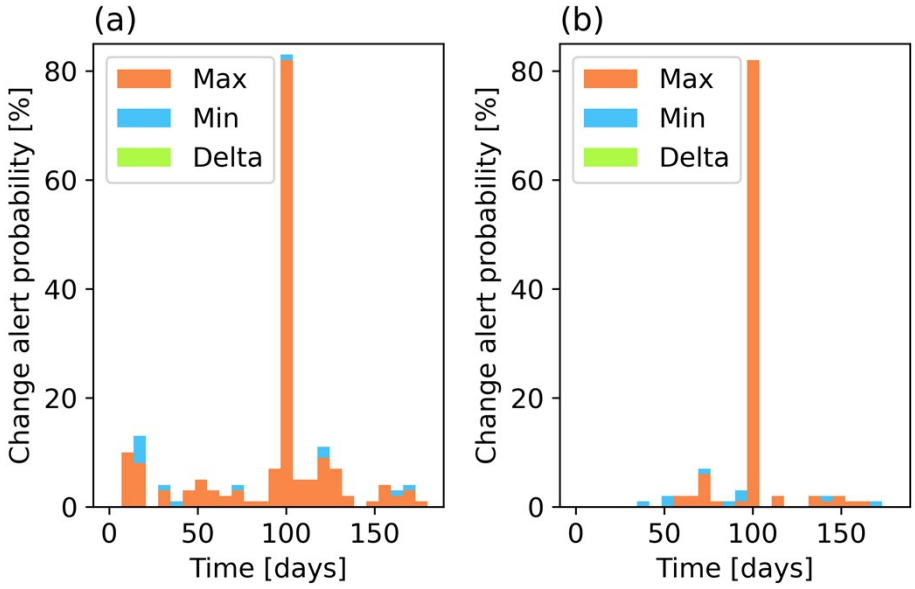

Figure S11. Histograms regarding the change detection probability obtained with a training window equal to 10 d (a) or 30 d (b).

References

1 A. Nescerecka, J. Rubulis, M. Vital, T. Juhna and F. Hammes, Biological Instability in a Chlorinated Drinking Water Distribution Network, *PLoS ONE*, 2014, **9**, e96354.
2 E. Chaib and D. Moschandreas, Modeling daily variation of trihalomethane compounds in drinking water system, Houston, Texas, *Journal of Hazardous Materials*, 2008, **151**, 662–668.
3 M. Gabrielli, A. Turolla and M. Antonelli, Bacterial dynamics in drinking water distribution systems and flow cytometry monitoring scheme optimization, *Journal of Environmental Management*, 2021, **286**, 112151.