Forecasting pollutant concentration in river to protect drinking water production

N. Cheifetz\textsuperscript{*}1, M. Laradi\textsuperscript{1}, N. Fauchon\textsuperscript{1}, F. Thouvenel\textsuperscript{2}, C. Féliers\textsuperscript{1}, and V. Heim\textsuperscript{3}

\textsuperscript{1}Veolia Eau d’Ile de France, Le Vermont, 28, Boulevard de Pesaro, Nanterre F-92751, France
\textsuperscript{2}Veolia Research & Innovation, Environment & Health Department, Chemin de la digue, Maisons-Laffitte F-78603, France
\textsuperscript{3}Syndicat des Eaux d’Ile de France (SEDIF), 120 Boulevard Saint-Germain, Paris F-75006, France
\textsuperscript{*}nicolas.cheifetz@veolia.com

\textbf{ABSTRACT}

Monitoring water quality from natural resources like lakes and rivers is very challenging due to the complexity of water sections and several environmental effects. The easy accessibility to these resources explains that any intentional or accidental pollution can hardly be prevented. However, a pollution event may be contained using some preventive simulations modeling the plume of a contaminant substance i.e. water utilities need to anticipate the substance arrival at the plant intake before taking any corrective action. Hydrodynamic simulations are usually time consuming and computationally expensive and therefore cannot be used in practice by water utilities to react efficiently in time. This paper introduces a fast data-driven approach to estimate a time-concentration curve based on a set of predefined simulations. The open-source software Telemac-2D is used to simulate numerous offline dispersion which requires a calibrated hydraulic model. Thanks to a parametric decomposition, the concentration time series is inferred quickly using only five classical characteristics of the pollution. This concentration curve is not calculated by a straightforward equation but it is estimated using some assumption on the shape of the pollution peak revealed by the hydraulic simulations. Multiple regression models are evaluated upon three realistic river sections in France around Paris and the extensive experiments show that non linear regression models outperform other linear methods and allow us to highlight the effectiveness and the rapidity of the proposed approach.

\textbf{Keywords:} plume of pollution, regression modeling, surface water, concentration-time curve

\section{INTRODUCTION}

Since a few decades, human activities have a significant impact on natural water resources e.g. oil spills or wastewater effluent \cite{1}. Several physico-chemical stressors are affecting biodiversity in agricultural and urban-industrial areas \cite{2}. In the Paris area, water quality of fluvial sectors is affected by a large population and the agricultural, domestic, and industrial activity in the watershed \cite{3}. For water utilities producing drinking water from such surface resource, the quality of raw water is an important feature used for planning and optimizing the production process. There is a need on providing water utility managers with tools that can guide them through the decision-making process of every specific use case, while being adaptable to continuous and arising situations \cite{4}.

Existing solutions to anticipate the pollution arrival at plant intake are usually based on heavy hydrodynamic simulations where the pollutant propagation can be simulated using some free surface
hydraulic modeling. Such numerical simulation tools can also be used for predicting flood inundation using 1D and 2D models of flood hydraulics [5]. In this paper, a Telemac-2D modeling is used to simulate the dispersion of a conservative tracer along river sections in the Paris area. The Telemac-2D is an open-source software solving Saint-Venant equations using the finite-element / finite-volume method and a computation mesh of triangular elements [6]. Despite effective results about estimating dynamic concentration at each node of the computational mesh, the required delay for running a set of simulations can be prohibitive. When a real pollution occurs, only uncertain and imprecise details may be reported about the pollution spreading (e.g. location of source, amount of substance,...) and numerous scenarios have to be simulated using the hydrodynamic modeling in order to contain any contamination impact. Each simulation may endure hours which makes it difficult to evaluate different scenarios of pollution in practice before the pollutant reaches the plant intake.

This article presents a new methodology to forecast rapidly the concentration time series at a particular point (i.e. water supply point of the production plant) easing the preventive simulation of multiple pollution scenarios. The formal problem can be formulated as the estimation of a time series based on a limited number of scalars characterizing the condition of pollutant injection. Note that a calibrated hydraulic model is needed for training and evaluating the proposed approach. A similar problem is studied in [7] but our contribution relies on the generic formulation of a scenario-based approach capturing the travel-time and dispersion effects in the learning phase. The next section introduces the data-driven approach as a two-step methodology to estimate concentration time-curves. An experimental study is given in section 3 using three rivers around Paris (France). Finally, the paper ends with a conclusion in section 4.

2 PROPOSED METHODOLOGY

Modeling the plume of a tracer for a river section involves complex phenomena like lateral mixing and longitudinal dispersion. In this study, existing calibrated Telemac-2D models are used to simulate free-surface flows in two dimensions of horizontal space i.e. average height and speed of water on the vertical. The resulting simulations imply extensive calculations which are not practicable in real-time. Then, a fast data-driven approach is formulated to estimate the concentration time series. The methodology adopted in this study is described in Figure 1 and it can be summarized in two consecutive steps:

1. Some pollution characteristics are gathered in a single feature vector. In this way, any pollution can be easily depicted with the same information as required in the Telemac-2D modeling. A regression model is performed to build a relation between this feature vector and a limited number of distribution parameters.

2. Based on these distribution parameters, the Telemac-2D time series of concentration can be approximated. Indeed, these concentration curves are estimated using some parametric priors about the curve patterns. This step allows to approximate the time series at the water supply point of production plants with only few parameters.

Then, the problem of estimating concentration values according with time is seen as a supervised regression problem in the feature space. This approach is much more tractable than estimating directly concentration curves with a fine granularity based on a limited number of pollution features.
The data-driven procedure relies on the assumption that each concentration time series generated by Telemac-2D simulations is specified by the following equation:

\[ C(t) = \lambda \left[ \frac{1}{\sigma_1 \sqrt{2\pi}} \exp \left( \frac{-(t - \mu_1)^2}{2\sigma_1^2} \right) + \frac{1}{\sigma_2 \sqrt{2\pi}} \exp \left( \frac{-(t - \mu_2)^2}{2\sigma_2^2} \right) \right], \tag{1} \]

where \( y = (\lambda, \mu_1, \sigma_1, \mu_2, \sigma_2) \) is the parametric vector used to approximate the concentration time series indexed with time \( t \). Note that this expression corresponds to a factor of a mixture distribution with two Gaussian components. The flexibility of the Gaussian mixture ensures a relatively good fitting to the curve and the choice of two components limits the over-fitting effect. A similar idea of mixture parametrization can be seen in [8]. As any concentration curve is no probability distribution (between 0 and 1, and sums up to 1), the cumulative distribution function of the normalized curve is used to compute the maximum likelihood estimates \( (\hat{\mu}_1, \hat{\sigma}_1, \hat{\mu}_2, \hat{\sigma}_2) \). Then, the \( \hat{\lambda} \) value is simply the ratio between concentration average and mixture distribution average.

In this paper, the only point of interest to estimate concentration curves is the water supply point of production plant. Due to lateral mixing and longitudinal dispersion, the Telemac-2D simulations may generate tracer-response curves with a peak or plateau pattern, both illustrated in Figure 2. When the time series has a single “bell curve” shape e.g. Fig. 2(a), the approximation is straightforward using the Equation 1 even if the curve is asymmetric. When the concentration is stabilizing over time e.g. Fig. 2(b), the mixture captures the rising and falling slopes but is enable to approximate correctly the constant part of the signal. A simple heuristic based on the discrimination of two distinctive peaks allows to cut the top of the peaks and to better estimate the maximum value of concentration.

![Figure 1: Block diagram describing the hydraulic modeling approach and the proposed methodology.](image)

![Figure 2: The two patterns of concentration generated by Telemac-2D simulations.](image)
The learning phase can be seen as a supervised backward process: from the concentration time series to a set of pollution features (cf. Fig. [1]. After the approximation of each concentration curve, a relation needs to be estimated between the parameter vector $\mathbf{y}$ and the pollution features, noted as $\mathbf{x}$. These pollution features have been identified as the most discriminative information used by Telemac-2D to simulate dispersion:

- Distance (km) between the injection point and the water supply point,
- Flow ($\text{m}^3/\text{s}$) of the river imposed at the section frontiers,
- Concentration (g/L) of the pollution injection,
- Flow ($\text{m}^3/\text{s}$) of the pollution injection, and
- Duration (sec) of the pollution injection.

It is worth noting that a linear relationship exists between concentration curves using several values of the river flow and for the pollutant flow as well, based on extensive simulations. This linear effect helps to reduce the feature space by exploring a single value of each flow in order to fit the best regression function $f(\mathbf{x}) = \mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are respectively the five pollution features and the parametric vector used to approximate the time series.

Various regression models have been evaluated in this study and only the most relevant ones are reported in this paper. As the main contribution of this paper is the global methodology to reproduce some Telemac-2D results, the regressors are only briefly described:

- The linear model estimates a matrix $\mathbf{\beta}$ of regression coefficients using the “normal equation” such as $\mathbf{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}$ where the matrix $(\mathbf{x}^T\mathbf{x})$ is supposed well conditioned. This result can be reached by minimizing the squared error $(\mathbf{y} - \mathbf{\beta}\mathbf{x}^T)^2$. When the product matrix is badly conditioned, the objective function may be regularized.

- The elastic net modeling aims at minimizing the regularized error function $(\mathbf{y} - \mathbf{\beta}\mathbf{x}^T)^2 + \lambda g(\mathbf{\beta})$ with a penalization term controlled by $\lambda > 0$. This term $g(\mathbf{\beta})$ is defined as a combination of the $L^1$ norm of $\mathbf{\beta}$ and the squared $L^2$ norm, thereby generalizing the lasso and ridge regression.

- The SVM (Support Vector Machine) regression is a nonparametric technique due to kernel function. This model relies on the kernel trick, implicitly mapping the inputs in a high-dimensional feature space. In this paper, a linear $\epsilon$-SVM regression model (with $L^1$ loss) is used.

- A decision tree with binary splits is performed for regression purpose. This non linear model aims to minimize the squared error $(\mathbf{y} - \hat{\mathbf{y}})^2$ where $\hat{\mathbf{y}}$ is the response estimated by the tree.

- A GPR (Gaussian Process Regression) modeling maps a non linear function $f$. This function may be seen as a kriging estimation aiming to minimize the expected squared prediction error.

- A NN (Neural Network) regression model is build with a single hidden layer and a sigmoid transfer function. This non linear model is trained by the classical back-propagation technique.

Without loss of generality, note that the input vector $\mathbf{x}$ includes a constant term for linear models.
The next section describes a case study where the proposed methodology is implemented to speed up the estimation of concentration curve at the production plant when an upstream pollution occurs. In practice, numerous scenarios are simulated to ease the reaction-making process.

3 EXPERIMENTATION

The experimental study investigates the capacity of estimating pollution plumes at drinking water production plants when located on a river. This study exhibits simulation results of the concentration curve when pollutants are released upstream of the supply point. The proposed methodology is evaluated on production plants located on different rivers in France i.e. Seine, Marne and Oise rivers. The three facilities belong to the Syndicat des Eaux d’Île-de-France (SEDIF) and SEDIF is a large association including 150 municipalities which provides drinking water for more than 4 million inhabitants of suburban Paris. The Figure 3 illustrates the three case studies with an injection source twenty kilometers from the production plants. Multiple Telemac-2D models previously calibrated are existing for each river section and are used to generate a set of reference simulations which is computationally expensive and time consuming.

This study is designed to evaluate the two steps of the methodology described in the previous section. First, offline simulations are performed by multiple Telemac-2D models based on numerous pollution configurations. A list of configurations was built up using a factorial design methodology based on the Telemac-2D model settings and the five pollution features defined in section 2. This design of experiments has the advantage to highlight the effect of several factors and even interactions but it requires to launch a large amount of simulations. Almost 3,000 Telemac-2D simulations have been generated for the three river cases with a duration ranging from 25 minutes to 15 hours per simulation. The pollution features are ranging such as river flow is from 30 to 700 $m^3/s$, injection flow is from 1 to 5 $m^3/s$, pollution concentration is from 1 to 10 $g/L$, injection duration is from 1h to 3h, and the distance to the production plant is between 5km and 60km. Let us recall that a linear relationship exists between concentration curves using several values of the river flow and for the pollutant flow, which helps to reduce the feature space by exploring a single value of each flow. Based on these simulations, the concentration curves are approximated using the 5-parameter decomposition given

![Figure 3: The three rivers around Paris supplying the production plants of SEDIF in France.](image-url)
by the Equation 1. The learning phase then consists of fitting some regression model that estimates these 5 parameters.

The proposed methodology is evaluated according to the regression performance and the curve approximation performance, respectively quantified by the two following criteria:

\[
R^2 = \frac{\sum_t (\hat{y}_t - \bar{y})^2}{\sum_t (y_t - \bar{y})^2} \quad \text{and} \quad RMSE = \sqrt{\frac{\sum_{t=1}^T (\hat{y}_t - y_t)^2}{T}}
\]

where \(R^2\) is the ratio between explained and residual variance ranging between 0 and 1, \(\hat{y}_t\) is the estimated concentration at time \(t\), \(\bar{y}\) is the average value of \(y\), and \(RMSE\) stands for Root Mean Squared Error. The higher \(R^2\) and the smaller RMSE, the better the performance.

Results are obtained using the six regression models listed in the previous section. The Table 1 shows the six method performance to estimate the 5-parameter vector for each river case. The best results for each parameter is highlighted in boldface type. The non linear models outperform the linear ones and the best models are GPR and RT depending on the river case study. The neural net is not preferred due to results below the others and a longer execution time. The poor performance of the SVM regression model may be due to its linear formulation which does not allow to capture the relationship between the pollution features and the approximation parameters. The most difficult parameter is \(\lambda\) the magnitude parameter which can be explained by its “non-mixture” nature. Then, the Table 2 gives the error to approximate concentration curves for the six regression models and again, the linear methods are outperformed by the non linear ones. Error of the Gaussian process regression model is lower for Seine and Marne rivers while the regression tree is better for the Oise case. The error medians are similar on the three river cases with a magnitude order of \(10^{-4} \text{ g/L}\).

A K-fold study (with K=10) has confirmed the results presented in tables 1 and 2. And the execution time to estimate a concentration curve with GPR or RT models was considered as reasonable, up to 20 seconds depending on the time horizon.

Table 1: Comparison of the regression performance in terms of \(R^2\) coefficient for three river cases. Regression models: linear, regularization via elastic net, support vector regression machine (SVRM), Gaussian process regression (GPR), regression tree (RT) and neural network (NN).

<table>
<thead>
<tr>
<th>Models</th>
<th>(R^2) (×100) for Seine</th>
<th>(R^2) (×100) for Marne</th>
<th>(R^2) (×100) for Oise</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\mu_1) (\mu_2) (\sigma_1) (\sigma_2) (\lambda)</td>
<td>(\mu_1) (\mu_2) (\sigma_1) (\sigma_2) (\lambda)</td>
<td>(\mu_1) (\mu_2) (\sigma_1) (\sigma_2) (\lambda)</td>
</tr>
<tr>
<td>Linear</td>
<td>31.8 31.8 34.8 32.2 26.6</td>
<td>39.7 40.7 37 38.3 25.1</td>
<td>99.1 98.8 87.3 80.3 51.7</td>
</tr>
<tr>
<td>Elastic</td>
<td>33.1 34.0 19.2 32 26.4</td>
<td>39.8 40.8 37.1 38.3 27.4</td>
<td>96.3 96.8 87.5 79 55.5</td>
</tr>
<tr>
<td>SVRM</td>
<td>16.9 16.5 14.6 15.5 9.7</td>
<td>15.1 15.1 19.8 21.2 8.7</td>
<td>85.7 87.6 89.4 80.0 38.5</td>
</tr>
<tr>
<td>GPR</td>
<td>100 100 96.1 95.2 100</td>
<td><strong>99.9</strong> 99.9 99.2 96.7 <strong>100</strong></td>
<td>99.6 99.6 94.3 94.5 11.9</td>
</tr>
<tr>
<td>RT</td>
<td><strong>100</strong> <strong>100</strong> <strong>97.4</strong> <strong>96.6</strong> 88</td>
<td><strong>99.9</strong> <strong>100</strong> <strong>99.5</strong> <strong>97.8</strong> 92.6</td>
<td><strong>99.8</strong> <strong>99.8</strong> <strong>96.2</strong> <strong>96.3</strong> <strong>99.2</strong></td>
</tr>
<tr>
<td>NN</td>
<td>99.9 99.9 93.5 92.9 65.3</td>
<td>99.7 <strong>100</strong> <strong>99.5</strong> 94.8 98.2</td>
<td>99 98.5 82.3 81.4 89.8</td>
</tr>
</tbody>
</table>
Table 2: Comparison of the approximation performance in terms of root mean squared error (RMSE) for three river cases. As the distribution of the RMSE computed for each concentration curve is asymmetric, only the median of RMSE per river is reported in this table.

<table>
<thead>
<tr>
<th>Regression models</th>
<th>Median of RMSE on concentration curves (g/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Seine</td>
</tr>
<tr>
<td>Linear</td>
<td>0.0118</td>
</tr>
<tr>
<td>Elastic</td>
<td>0.0089</td>
</tr>
<tr>
<td>SVRM</td>
<td>0.0068</td>
</tr>
<tr>
<td>GPR</td>
<td><strong>0.0004</strong></td>
</tr>
<tr>
<td>RT</td>
<td>0.0005</td>
</tr>
<tr>
<td>NN</td>
<td>0.0035</td>
</tr>
</tbody>
</table>

4 CONCLUSION

A generic methodology is introduced in this paper for automatically estimating the concentration curves of pollutant spreads injected upstream a water production plant. Considering Telemac-2D simulations as the reference for lateral mixing and longitudinal dispersion, a two step methodology is formulated to speed up the process to generate concentration time series based on few pollution features. These time series are approximated using a parametric decomposition strategy. The data-driven method is evaluated on three realistic case studies that represent the water sections upstream production plants in France. Six regression models are compared and two non linear methods are identified as the best to minimize the concentration error of about $10^{-4}$ g/L. More investigations are in progress with the water utility Veolia Eau d'Ile de France in order to refine the regression results and the proposed methodology is also being encapsulated in a GUI application.

References


