Environmental Modelling & Software 24 (2009) 969–981

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

# Design of a sampling strategy to optimally calibrate a reactive transport model: Exploring the potential for *Escherichia coli* in the Scheldt Estuary

Anouk de Brauwere [a,c,*], Fjo De Ridder [b], Olivier Gourgue [c], Jonathan Lambrechts [c], Richard Comblen [c], Rik Pintelon [d], Julien Passerat [e], Pierre Servais [e], Marc Elskens [a], Willy Baeyens [a], Tuomas Kärnä [c], Benjamin de Brye [c], Eric Deleersnijder [c]

[a] *Vrije Universiteit Brussel, Analytical and Environmental Chemistry, Pleinlaan 2, B-1050 Brussels, Belgium*
[b] *Vlaamse Instelling voor Technologisch Onderzoek (VITO), Boeretang 200, B-2400 Mol, Belgium*
[c] *Université catholique de Louvain, Centre for Systems Engineering and Applied Mechanics (CESAME), 4 Avenue G. Lemaître, B-1348 Louvain-la-Neuve, Belgium*
[d] *Vrije Universiteit Brussel, Department of Electricity and Instrumentation, Pleinlaan 2, B-1050 Brussels, Belgium*
[e] *Université Libre de Bruxelles, Ecologie des Systèmes Aquatiques (ESA), Campus de la Plaine CP221, Boulevard du Triomphe, B-1050 Brussels, Belgium*

ARTICLE INFO

*Article history:*
Received 5 September 2008
Received in revised form
22 January 2009
Accepted 11 February 2009
Available online 17 March 2009

*Keywords:*
Optimal experimental design
Parameter estimation
Parameter uncertainty
Reactive tracer model
Scheldt, Fisher information matrix

ABSTRACT

For the calibration of any model, measurements are necessary. As measurements are expensive, it is of interest to determine beforehand which kind of samples will provide maximal information. Using a criterion related to the Fisher information matrix as a measure for information content, it is possible to design a sampling scheme that will enable the most precise parameter estimates. This approach was applied to a reactive transport model (based on the Second-generation Louvain-la-Neuve Ice-ocean Model, SLIM) of *Escherichia coli* concentrations in the Scheldt Estuary. As this estuary is highly influenced by the tide, it is expected that careful timing of the samples with respect to the tidal cycle can have an effect on the quality of the data. The timing and also the positioning of samples were optimised according to the proposed criterion. In the investigated case studies the precision of the estimated parameters could be improved by up to a factor of ten, confirming the usefulness of this approach to maximize the amount of information that can be retrieved from a fixed number of samples. Precise parameter values will result in more reliable model simulations, which can be used for interpretation, or can in turn serve to plan subsequent sampling campaigns to further constrain the model parameters.


## 1. Introduction

Taking environmental samples and subsequently analysing them is often an expensive and time-consuming business. In particular, when the study concerns trace elements or biological species, the sampling and analysis cannot usually be automated; instead delicate and expert handling is required. It is therefore of obvious interest to know beforehand which and how samples should be taken such that a maximum of information will be gathered, or such that a predetermined level of information can be achieved with a minimum of resources. Usually this step is performed in a more or less intuitive way, based on previous experiences or other prior subjective knowledge. This article uses a more rigorous criterion to determine which samples will be most informative. Using this criterion, different sampling designs can be compared a priori in order to find the optimal one.

This sampling design strategy was applied to *Escherichia coli* (*E. coli*) concentrations in the Scheldt Estuary. *E. coli* is one of the most common bacteria present in the intestines of mammals. Huge numbers are released to the environment every day by human and animal excrements. Therefore, the abundance of *E. coli* in water is generally used as an indicator of faecal pollution. Although most of *E. coli* strains are not pathogenic themselves, the *E. coli* concentration indicates the level of potential presence of other pathogenic micro-organisms from faecal origin and thus the sanitary risk associated with various water utilisations (bathing, shellfish harvesting, production of drinking water,…) (Edberg et al., 2000; Fewtrell and Bartram, 2001).

Within the framework of a Belgian interuniversity research project (http://www.climate.be/TIMOTHY/), we are interested in the spatial and temporal variability of *E. coli* abundance in the Scheldt Estuary. It is an illusion to try to answer this question by measurements alone unless huge resources are invested. Therefore, a coupled hydrodynamical – reactive tracer model was constructed

* Corresponding author. Vrije Universiteit Brussel, Analytical and Environmental Chemistry, Pleinlaan 2, B-1050 Brussels, Belgium. Tel.: +32 2 629 32 64; fax: +32 2 629 32 74.
*E-mail address:* adebrauw@vub.ac.be (A. de Brauwere).

1364-8152/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2009.02.004

to simulate the dynamics of *E. coli* in the domain of interest. This model will provide high resolution simulations of temporally and spatially varying *E. coli* abundance. Although the structure (i.e. the equations) is assumed to be correct, this model still needs to be calibrated. This means that measurements of *E. coli* are needed and the question of the sampling design is relevant. In this particular case, very little is known about the distribution of *E. coli* in the Scheldt. Based on previous studies in other areas (e.g. Steets and Holden, 2003; Garcia-Armisen et al., 2006; Servais et al., 2007a,b) some general features can be expected (e.g. average disappearance rate and general model structure) but it is clear that extrapolation of this knowledge to the macrotidal Scheldt basin is not straightforward. These facts were the actual motivations to find a useful criterion to guide the planning of future sampling campaigns.

The criterion to design an optimal sampling scheme is related to the Fisher information matrix (Fedorov, 1972). Optimality here refers to maximal information content of measurements, in terms of their ability to deliver precise parameter estimates. In other words, guided by the information criterion, the experimental setup is selected which will reduce the uncertainty associated with the parameter estimates most. This uncertainty is the result of measurement uncertainties propagated through the model. Even if the measurement uncertainty is independent of space and time, measurements taken at different locations and times will deliver different parameter estimates with different uncertainties. Using the information criterion approach, the parameter uncertainty obtained from any measurement set can be predicted, and thus minimized – resulting in the identification of the optimal sampling setup. More particularly, in this study we focused on optimising the sampling setup in terms of the location and timing of a sampling campaign. The objective of the article is to apply this strategy to *E. coli* concentrations in the Scheldt, as an investigation of its potential utility for this real application. Since the results appear promising, the next step will be the application to a more realistic model setup, where the results will actually be combined with field constraints to eventually derive the optimal realistic sampling strategy.

Experimental design is an important issue for all experimental studies, although may be not equally recognized in all fields. The broad area of water quality studies is one of the fields where considerable work has been done on this subject (see Whitfield, 1988; Dixon and Chiswell, 1996 for reviews). Several criteria or procedures have been proposed to find the optimal sampling distribution. One approach is to distribute the samples or experiments such that the design space is covered as uniformly as possible, e.g. using a procedure placing experimental points such that their distance is maximized (Kennard and Stone, 1969; Marengo and Todeschini, 1992). Alternatively, Sanders (1982) used analysis of variance to determine how many and where samples should be taken along a river's cross-section to obtain representative mean water quality concentrations. Both these approaches have the advantage not to require the assumption of a particular model for the system under study. Most other methods do use this assumption. For instance, Sanders and Adrian (1978) used a (simple) model to determine the station sampling frequencies such that a uniform variance of the water quality variance would be achieved. Lo et al. (1996) used kriging to select the monitoring points which ensured to produce an average water quality closest to the true (modelled) one. Alvarez-Vázquez et al. (2006) choose samples to be most representative in their river section, by minimizing the difference between the point measurement and the average section value (both estimated using a model). Apart from obtaining uniform or representative information, an important problem in water quality studies is the identification of pollution sources. For this purpose, Sharp (1971) proposed a topological strategy (not requiring a model), but many other model-based (inverse) methods have been proposed since (Sun, 2007 and references therein). Yet another objective for experimental design can be the discrimination between hypothesized models (Steinberg and Hunter, 1984 and references therein), or simply to reduce the cost as much as possible while satisfying some minimum requirements (Vandenberghe et al., 2002). Knopman and Voss (1989) proposed a multiobjective strategy combining the two last objectives with the objective to reduce the variance in model parameter estimates. This last objective is the focus of the present study, searching those experiments (samples) which will enable a most reliable model calibration. If the model is precisely calibrated, the model itself will be able to reliably reproduce the spatiotemporal variations of the system, which is eventually what we desire. A sampling strategy can be optimised to achieve this objective by using the Fisher information matrix as a criterion expressing the precision of the estimated model parameters is. Vandenberghe et al. (2002) applied this same criterion to investigate the optimal measuring points for water quality variables in a river. Based again on the same criterion, Vanrolleghem and Coen (1995) proposed a procedure to gather maximum on-line information on processes in a biosensor, enabling both optimal model selection and parameter estimation. Wagner (1995) and Catania and Paladino (2009) also used a similar criterion, respectively for a groundwater modelling application and the estimation of dispersion coefficients in laboratory experiments. However, none of these literature results can be extrapolated to *E. coli* concentrations in the Scheldt with its typical dynamics due to its different domain shape and important tidal influence.

The article is structured as follows. In the Methods section, the information criterion is introduced (Section 2.1), followed by a description of how to use this criterion in practice to design an optimal sampling scheme (Section 2.2). Next, some information is given on the application: some background on the Scheldt Estuary (Section 3.1), the model used (Section 3.2), the model inputs (Section 3.3), the important model parameters (Section 3.4) and the measurement of *E. coli* (Section 3.5). After the tools have been presented, in Section 4 the results are shown and discussed for several conceptual applications. Finally, some concluding remarks are given, reminding the major points of the results and giving an outlook to future opportunities.

## 2. Methods

### 2.1. Information criterion

The fundamental idea is that a model will produce the most reliable output if its parameters were estimated using the most informative measurements. If we can find a formal expression for "information content", this can be used as a criterion to design an optimal sampling distribution. The criterion often used is based on the Fisher information matrix (Fedorov, 1972). It has already been applied in several distinct areas (e.g. Vandenberghe et al., 2002; Rensfeld et al., 2008). For those not familiar with this approach, a basic background is given in this section.

Consider that a model $f$ is able to accurately simulate a variable $y$ given some inputs $x$ and parameters $p$, i.e. the measured variable $y$ equals the modelled value $f(x,p)$ plus some error term $e$:

$$y = f(x,p) + e. \tag{1}$$

$y$ and $p$ are vectors of length $N$ and $n_p$ respectively. In order to have an accurate model, i.e. being close to the measurements ($y$), the error term should be small. Therefore, the fit between model and measurements is usually optimised by varying the parameter values until the sum of squared (weighted) errors $((y-f(x,p))^T cov_e{}^{-1}(y-f(x,p)))$ is minimal. The use of this weighting, with the inverse of the measurement covariance matrix $cov_e{}^{-1}$, has some convenient consequences.

The uncertainty associated with the optimal parameter values can be estimated, even before any samples are taken, because it simply results from the propagation of the measurement uncertainties through the model. Indeed, the parameter covariance matrix can be estimated only knowing the measurement uncertainties and the model by

$$cov_p = \left(J(R,T)^T cov_e(R,T)^{-1} J(R,T)\right)^{-1}, \quad (2)$$

where $J$ is the $(N \times n_p)$ Jacobian or sensitivity matrix, containing all first derivatives of the model output with respect to the model parameters. $J$ and $cov_e$ are explicitly said to be dependent on $R$ and $T$, symbolizing respectively the locations and times of the considered model outputs and samples. Using Equation (2) the uncertainty can be estimated that would be associated with parameters if they were estimated using measurements taken at $(R, T)$. In other words, the actual measurements are only needed to estimate the parameter *values*; the associated parameter *uncertainties* can be estimated even before any measurements exist, as long as it is known where and when the samples will be taken and which would be the covariance associated with these measurements. The latter can usually be assumed based on previous studies, or else it may be a reasonable approximation to say that the measurements will be independent and all have the same (possibly unknown) variance. To be correct, Equation (2) is only exact in the case that the model $f$ is linear in the parameters. In the nonlinear case, Equation (2) is only asymptotically valid, i.e. for the sample size tending to infinity. In other words, in that case Equation (2) may not represent the actual parameter covariance matrix but it is still a measure of the best achievable parameter uncertainty. Further, note that these properties are independent of the distribution of the measurement errors $(e)$. Under some additional assumptions, the inverse of the parameter covariance matrix is also called the Fisher information matrix (Fedorov, 1972), but to avoid restricting ourselves to those assumptions, we won't use this term further. If the model is nonlinear, the Jacobian matrix $J$ also depends on the parameter values. Therefore, some prior values for the model parameters must be assumed for the construction of $cov_p$

If there is only one parameter $cov_p$ is actually a scalar, quantifying the variance of $p$. In that case we define the best sampling distribution $(\hat{R}, \hat{T})$ as the one resulting in the lowest parameter variance:

$$\begin{aligned}\left(\hat{R}, \hat{T}\right) &= \arg\min_{R,T} cov_p = \arg\min_{R,T}\left(J(R,T)^T cov_e^{-1} J(R,T)\right)^{-1} \\ &= \arg\max_{R,T}\left(J(R,T)^T cov_e^{-1} J(R,T)\right), \end{aligned} \quad (3)$$

or in words: $(\hat{R}, \hat{T})$ are those values of $(R,T)$ for which $cov_p$ has its minimal value.

So, for all realistic combinations of $(R, T)$, $J^T cov_e J$ can be computed. The spatial and temporal distribution that gives the highest value for $J^T cov_e J$ can be designated as the optimal sampling scheme.

In the more general situation where more than one parameter are estimated, $J^T cov_e J$ is not a scalar anymore but a matrix. To rank the different sampling distributions according to the "total" parameter uncertainty, a scalar function has to be applied to $J^T cov_e J$ first. A common choice is to maximize the determinant of $J^T cov_e J$ (Jacquez, 1998)

$$\left(\hat{R}, \hat{T}\right) = \arg\max_{R,T}\left(\det\left[J(R,T)^T cov_e^{-1} J(R,T)\right]\right). \quad (4)$$

The experimental setup found this way is called *D-optimal* (from determinant). D-optimal experiments have the advantage to be invariant with respect to any rescaling of the parameters (Pukelsheim, 2006). Other criteria expressing the magnitude of $cov_p^{-1}$ with a scalar can be proposed as well, e.g. the trace, or even criteria that emphasize the importance of some parameters more than others (Fedorov, 1972). Of course, this kind of strategy can be used to optimise any aspect of the experiment; yet in this study we focus on the sampling distribution in space and time, as in our application these are the major controllable factors.

In practice, the assumption of independent measurement uncertainties being constant (i.e. independent of location and time) will often be the only reasonable approximation available. In that case, $cov_e$ equals a constant times the identity matrix $(\sigma\, I_N)$, which can be omitted without changing the ranking of the different $(R, T)$:

$$\begin{aligned}\left(\hat{R}, \hat{T}\right) &= \arg\max \det\left(J(R,T)^T cov_e^{-1} J(R,T)\right) \\ &= \arg\max \det\left(J(R,T)^T (\sigma \cdot I_N) J(R,T)\right) \\ &= \arg\max \det\left(J(R,T)^T J(R,T)\right) \\ &= \arg\max \det(F_S). \end{aligned} \quad (5)$$

For ease of reference $J^T J$ was renamed $F_S$. To summarize, Equation (5) defines the optimal spatial and temporal distribution of samples under the following conditions:

(a) The model is accurate (model structure and prior parameter values are reasonable);
(b) The measurement errors are independent of each other and of the time and location they were sampled; their variance is a constant. If this assumption does not hold but the measurement covariance matrix is known a priori (e.g. proportional to the measured quantity), the more general Equation (4) can be used.
(c) The model parameters will be estimated by minimizing a weighted least squares cost function, using a weighting proportional to the inverse of the measurement covariance matrix $cov_e$.
(d) If the model is nonlinear $F_S$ is exactly proportional to $cov_p^{-1}$ only for infinite sample size, otherwise it is still a measure of the maximal achievable parameter precision.

Under the above assumptions, applying the sampling scheme satisfying Equation (5) will provide the most "informative" set of measurements, in the sense that they will allow to estimate the unknown model parameter(s) with a lowest uncertainty (determinant of the covariance matrix). This way, eventually, these samples will permit to perform the most precise model simulations.

### 2.2. Design of an optimal sampling scheme in practice

The practical difficulty remains in "trying" all combinations of $(R,T)$. Due to the high combinatorial complexity of this optimisation problem, a theoretical solution is only known for very special cases, and even solving the problem numerically is very difficult (Fedorov and Hackl, 1997). Conventional gradient-based optimisation techniques often fail because of model nonlinearities and nonconvexity (McPhee and Yeh, 2006; Catania and Paladino, 2009). Therefore, to start, a reasonable subset of all locations and times must be chosen. This is partly dictated by practical constraints known beforehand, e.g. some areas of the domain are inaccessible or there is a limitation on the number of samples per unit of time that can be processed due to experimental or storage constraints. The remaining possibilities should then be tested in an efficient way, to keep the number of combinations feasible. It is important here to note that the Jacobian matrix (and thus the model output) does not have to be computed again for every experimental setup. Instead, the model is run $2n_p + 1$ times ($n_p$ = number of model parameters) with slightly different parameter values and the outputs are stored for all locations and times decided in advance. In this study we chose to store the outputs at all nodes of mesh and every 30 min. By subtracting the different model runs, a finite difference approximation of the sensitivities is obtained for all those locations and times. These elements form what we call the "meta-Jacobian". Building this meta-Jacobian does not cost (significantly) more time than building a smaller Jacobian matrix, it only requires some memory space. Then the actual Jacobians associated with a given experimental setup (sampling location + times) can be formed by selecting elements from the meta-Jacobian. This procedure is in fact quite time effective because model runs are very expensive and only a very small number of them are needed – these can be performed in advanced and one meta-Jacobian can thus be used for all experimental design analyses. Furthermore, as a global (although discretised) search is performed, there is more certainty of having found the global optimum, at least within the discretisation precision.

The final question is how to perform this "subsampling" of meta-Jacobian elements. To guarantee that the "globally" optimal solution will be found, all combinations (already reduced by considering practical constraints) must be considered. But in practice a sequential procedure, fixing one sample (time and location) per step, seems the most feasible.

Depending on the specific application the search for the maximal information may be done differently. In the examples shown below, it was decided to split the determination of timing $(T)$ and positioning $(R)$, instead of trying to optimise everything at once. In some examples, the timing of the samples was fixed beforehand. In real applications this can correspond e.g. to the case where a sampling "protocol" must be followed consisting of a fixed number of samples taken at predefined time instants. With the timing being fixed, the actual optimisation using the above criterion now only concerns the positioning (and number) of these sampling "campaigns", which is much more feasible from the combinatorial point of view. More details on the procedure followed are given in the respective Results sections.

For real applications, it is only reasonable to admit that practical constraints of the experiment and on the field will highly influence the actual possible sampling schemes. It is best to include all constraints from the start, but this is not always feasible. For instance, it would be a huge amount of work to classify the whole domain in "accessible" and "non-accessible for sampling". Therefore, the reasonable strategy seems to perform a first optimisation (taking into account all constraints available but knowing that some solutions may still be impossible). If, by confronting the proposed samplings to the real world, some of them appear to be impossible, one can further optimise the sampling setup, by taking into account the newly "discovered" constraints. This could consist of repeating the full optimisation but with the new constraints included (possibly in an iterative way) or, alternatively, of only a local optimisation (which is satisfactory assuming that the new constraints do not greatly change the results).

A final note on prior data: if older measurements are available, these can be included in the analysis. The Jacobian matrix can be extended with fixed rows representing these measurements; in this way the prior data (or actually their locations and timings) will contribute to $F_S$ and can also influence the optimal design for the future sampling. This could be worthwhile e.g. in cases where older data are only available for a part of the domain. This knowledge is then included, such that the outcome of the analysis will probably (but dependent on the model) indicate that new samples are to be taken preferentially in the domain unsampled so far. If the uncertainties associated with these prior data are known and different from the

(expected) uncertainties that will be associated with the future measurements, the weighting matrix can be constructed and Equation (4) can be used.

## 3. Application to the Scheldt Estuary

To investigate its potential for the problem of sampling *E. coli* in the Scheldt Estuary, the results for a number of simplified examples are shown in the next section. They are simplified in the following aspects:

(1) The modelled processes contain some simplifications (see Section 3.2), although much attention has been paid to an accurate representation of the tide as this is expected to be a key player in the Scheldt theatre.
(2) Only the *E. coli* specific parameters are taken into the analysis. In theory, also hydrodynamic parameters could be included as some of them may be badly known too.
(3) Not all point sources of *E. coli* really present are considered. Instead we are interested in studying the influence of point source locations and magnitude to the optimal sampling scheme.
(4) Specific constraints like accessibility or cost are not taken into account.

Therefore, this study is meant to identify the general trends influencing the information that can be retrieved from different sampling designs, in order to explore the potential of this kind of analysis for future (more realistic) applications. Before discussing the results, some information on the study area, the model and the variable of interest (*E. coli* concentration) is given.

### 3.1. The Scheldt Estuary

The Scheldt River flows from northwestern France, through northern Belgium, ending in the North Sea in the southwestern part of the Netherlands (Fig. 1). In this study we will concentrate on the estuarine part going from Antwerpen (B) to the mouth joining the North Sea. The river and its tributaries drain a densely populated area, where both active industries and intensive agriculture and animal farming have developed. As a consequence, the Scheldt Estuary receives extremely polluted water, although recently a certain improvement has been noted compared to 1970s (Meire et al. (2005) and references therein), especially since 2007 when the big Brussels' waste water treatment plant started operating (Schoonjans, 2007). An important dynamical feature in the studied area is the tide, its major components being semi-diurnal (lunar tide M2 and solar tide S2). When referring to the tidal cycle in this study, the M2 period of 12 h25' is meant, as this component has by far the largest amplitude (four times the second most important, S2, amplitude). The Scheldt Estuary is considered a macrotidal system, with its large tidal ranges (mean neap and spring ranges are 2.7 and 4.5 m, respectively) and huge water volumes transferred during the tide (approximately 200 times more water entering the estuary during flood than the average freshwater discharge during one tidal cycle (Vanderborght et al., 2007)). As a consequence of this relatively small river discharge, the transit time through the estuary is estimated to be 1–3 months (Soetaert and Herman, 1996). Another consequence is that the water column is generally well mixed.

### 3.2. Hydrodynamical – reactive tracer model

A model to describe the dynamics of the variable of interest is necessary in this methodology. In fact, it is not directly the model output that is of use, but its derivative with respect to the parameters, to form the Jacobian matrix. In this study this derivative is approximated by finite difference between model runs (see Section 4.1). The technical details on the model are summarised in this section.

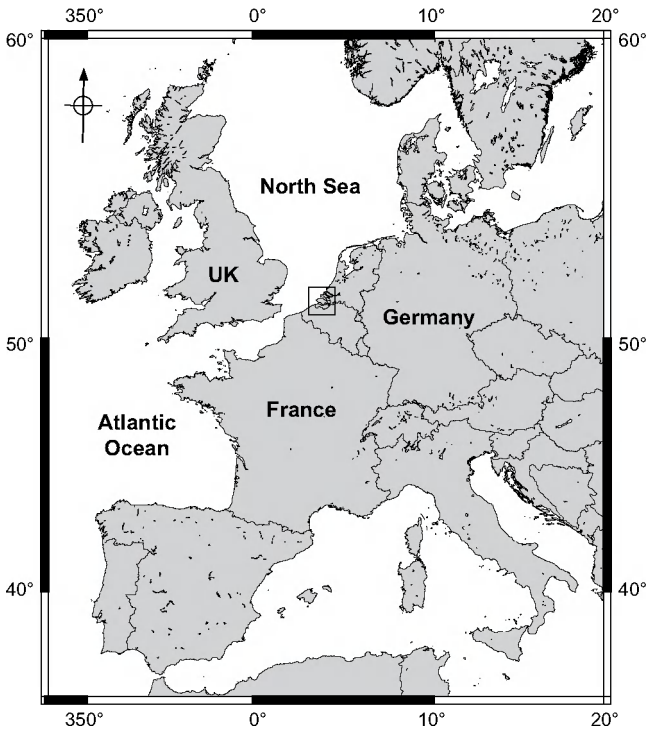


**Fig. 1.** Map indicating the location of the Scheldt Estuary.

The *E. coli* dynamics are modelled using a hydrodynamical model coupled to a reactive tracer module, forming a new application of the Second-generation Louvain-la-Neuve Ice-ocean Model, abbreviated as SLIM (http://www.climate.be/SLIM/). The hydrodynamic part of the model solves the (depth-averaged) shallow water (thus depth-averaged) equations using the finite element method (Lambrechts et al., 2008a; Comblen et al., 2008) with linear discontinuous elements ($P_1^{DG}$) for all variables. The finite element method allows the use of an unstructured mesh which has the advantage that the grid size can be adapted in time and space according to the need of detail. The mesh used in this study was constructed using Gmsh (Gmsh, 2008; Lambrechts et al., 2008b) and is shown in Fig. 2. The whole continental shelf has been included in the domain, such that the tide can be neatly imposed at the boundary, i.e. at the shelf break (more details below). However, the mesh is refined closer to the estuary and to the coastlines. This way, a finer resolution (of about 200 m) is achieved in the areas of most interest, simultaneously with a reasonable total number of grid cells (about 10 000).

*E. coli* is modelled as a reactive tracer in two dimensions, according to (e.g. Breton and Salomon, 1995; Padilla et al., 1997; Naithani et al., 2007):

$$\frac{\partial HC}{\partial t} + \nabla \cdot (H\underline{u}C) = \nabla \cdot (KH\underline{\nabla} C) + H(P - D), \tag{6}$$

where $C(r, t)$ represents the concentration of *E. coli* being dependent on location and time, $H(r, t)$ and $\underline{u}(r, t)$ are the water height and depth-averaged velocity, respectively, which are computed by the hydrodynamical part of the model. $K$ stands for the horizontal eddy viscosity, which is further described by a Smagorinsky's parameterization (Smagorinsky, 1963; Lambrechts et al., 2008a,b). This incorporates unresolved turbulent features and boundary layers along the coastlines, by making $K$ dependent on the local

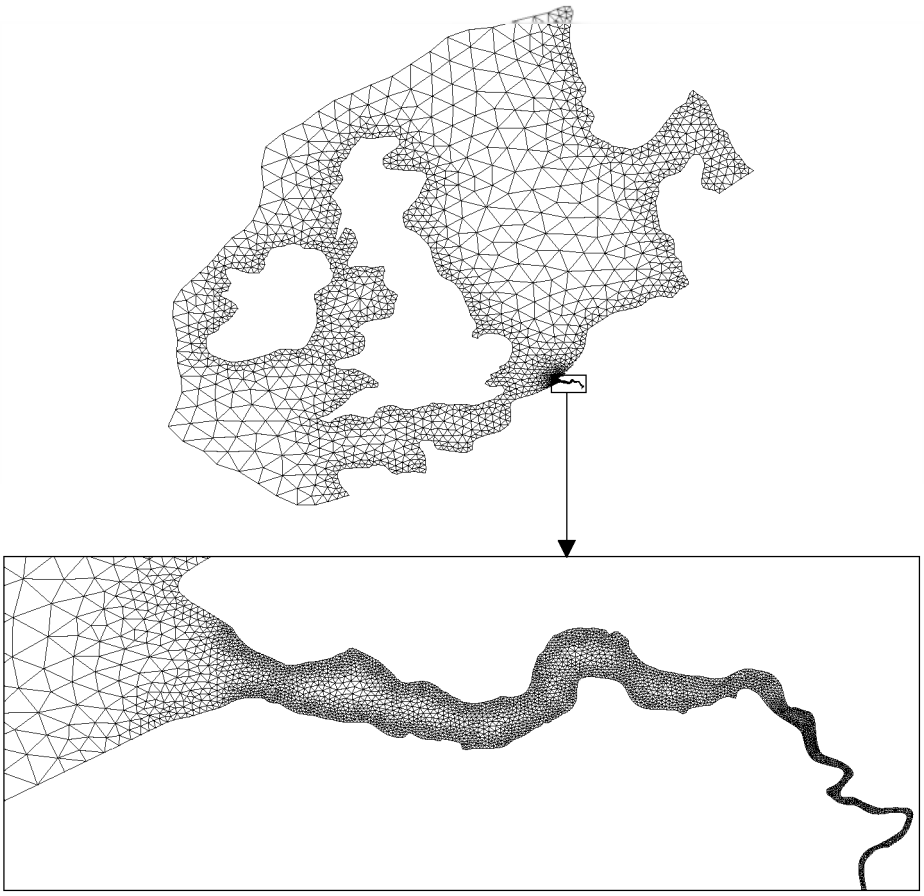

**Fig. 2.** Model domain with mesh used for this study.

mesh size and flow structure. A reactive tracer is transported by advection and diffusion (the second and third terms of the equation, respectively), according to the hydrodynamics, and at the same time it is subject to some specific dynamics, represented by the production ($P$) and destruction ($D$) terms. In the case of *E. coli*, these specific dynamics were assumed to be relatively simple. Being outside their natural habitat, the faecal bacteria's dynamics are assumed to be limited to disappearance processes. More specifically, they are assumed to disappear due to mortality and sedimentation, both according to a first order relation (e.g. Steets and Holden, 2003):

$$D = k_{mort}C + \frac{v_{sed}}{H}C, \tag{7}$$

with $k_{mort}$ the mortality rate constant, $v_{sed}$ is the sedimentation rate. This way, the disappearance is modelled as a first order decay (Pichot and Barbette, 1978; Kay et al., 2005; Servais et al., 2007b), but with a first refinement considering a decay constant varying with water height. Further refinements can be made, e.g. taking into account the variation of $k_{mort}$ with temperature, salinity, turbidity, flow, etc., but for this to be useful the exact dependencies should be well quantified. No processes "produce" *E. coli*, except for their injection into the domain by the considered point sources.

The finite element method is also used to model the advection–diffusion part of the tracer equation. But for the reactive part, we only solve an ordinary differential equation (ODE) on each degree of freedom of the concentration field. It is therefore just a pointwise equation acting as a source/sink term on the advection–diffusion equation.

### 3.3. Model inputs

The influence of the initial conditions becomes negligible after some time because of the frictional and viscous dissipation for the hydrodynamical equations, because of the diffusion and destruction terms in the tracer equation. The actual time needed before the initial conditions' influence becomes negligible is relatively short thanks to the open boundary conditions. So, any initial condition can be used.

Coasts are considered impermeable and frictionless. The hydrodynamical model is forced by the tide at the frictionless open boundary. We consider that the concentration of *E. coli* is zero outside the domain, such that the tracer can only leave the domain through the open boundary. *E. coli* is primarily injected into the domain by waste water treatment plants (WWTPs) acting as point sources (Garcia-Armisen and Servais, 2007). In the examples shown below it is not intended to take into account all known point sources and no upstream pollution is considered to enter the domain. Instead, the performance of the method is investigated for some simplified situations. This allows to better interpret the results and make some general statements on the sampling strategy.

It would be closest to reality to model the point sources as Dirac delta functions based in the discharge points. Their implementation in finite element models is not trivial and several previous studies devoted some efforts in giving a theoretical and numerical framework for treating this kind of point sources (Alvarez-Vazquez et al., 2002a,b; Scott, 1973). However, in our particular case (using discontinuous shape functions and Eulerian advection schemes without limiting technology), the implementation of a Dirac input would lead to large and unrealistic inter-element jumps and grid-scale oscillations.

Therefore, i.e. to guarantee numerical robustness, the point sources are modelled by Gaussian-shaped inputs into the model domain (Kärnä et al., submitted for publication), i.e. the concentration injected at each time step is not concentrated in one point but spreads over some surface of the domain, centred around the actual source point. However, to be realistic, the standard deviation of the Gaussian-shaped input is chosen sufficiently small compared to the grid size (more on this in Section 4.1). Fig. 3 shows a snapshot of the *E. coli* concentrations in the estuary during one of the simulations discussed below, with 8 point sources continuously injecting bacteria in the domain. Note that the values are not intended to be realistic, and indeed probably are far from it, because arbitrary source fluxes were used (see Section 4.2.3.).

### 3.4. Model parameters

The tunable model parameters are given in Table 1. As already mentioned in Introduction of this section, only the *E. coli* specific parameters are considered in this study. Note that value ranges are available for these parameters, from previous studies (e.g. Steets and Holden, 2003; Garcia-Armisen et al., 2006; Servais et al., 2007a). A representative average of these values is used as "typical" and is used in the simulations.

### 3.5. E. coli measurements

The sampling design procedure described above will be applied to make statements towards an optimal sampling design for *E. coli* concentration in the Scheldt Estuary. Therefore, it seems relevant to briefly present some information on *E. coli* measurements: the experimental procedure and a discussion of existing datasets.

The methods traditionally used for the enumeration of *E. coli* in water are plate count methods with different specific media and incubation conditions (Rompré et al., 2002). Today, numerous chromogenic and fluorogenic agar media are available on the market to enumerate *E. coli* (Manafi, 2000); they allow an easy detection of the β-D-glucuronidase, an enzyme specific to *E. coli*. Practically, immediately after returning to the lab, the sample is filtrated through a 0.45 μm-pore-size sterile membrane. The filter is then incubated on a selective agar medium for 24–36 h at temperature in the range 36–44 °C; incubation duration and temperature are depending on the selective agar medium used. After incubation, *E. coli* colonies (detected by colour or fluorescence) are enumerated and the data are expressed in *E. coli* number per 100 ml of water sample.

In order to avoid fluctuations in *E. coli* numbers between sample collection and analysis when using this type of microbiological method, samples must be proceeded within 12 h and kept at 4 °C between collection and analysis. Accordingly, in order to assure prompt analysis, the number of samples that can be collected during a one-day sampling campaign and analysed by a single analyst is limited to 20–30. Furthermore, if these samples are to be distributed over different locations, logistic considerations restrict the number of locations that can be visited during one day to only a few. These constrains should be taken into account when setting up the optimal sampling distribution.

Presently, no useful dataset concerning *E. coli* concentrations is available for the Scheldt Estuary. The few existing datasets all exhibit the disadvantages of low sampling frequency and very poor spatial coverage. Indeed, water quality standards in terms of fecal bacteria concentrations are routinely controlled only in bathing sites during the bathing period (June to September) and there are no controlled bathing sites in the study area. In addition, even when fecal pollution is monitored, it is done with a typical sampling frequency of one per several weeks, which is evidently too sparse to make any statements about the specific dynamics of the system under study, i.e. primarily the tide (main cycle period is ±12 h) for the hydrodynamics and the semi-exponential disappearance of the faecal bacteria themselves (typical decay time is ±21 h). Therefore, a new sampling scheme, which will enable to capture these frequencies, is of real interest. In the next section, the potential of the proposed methodology to answer this question will be investigated.

## 4. Results and discussion

### 4.1. Computation and validation of numerical Jacobian

The Jacobian matrix, necessary to compute the $F_S$ (Equation (5)), is approximated by first order finite differences between two model runs. For this, the model outputs at discrete positions and times are used. More precisely, the Jacobian is computed at all corners of the triangular grid cells, i.e. at the nodes, and every 30 min.

The finite difference approach was validated for a simplified case where the analytical expression of the Jacobian matrix with

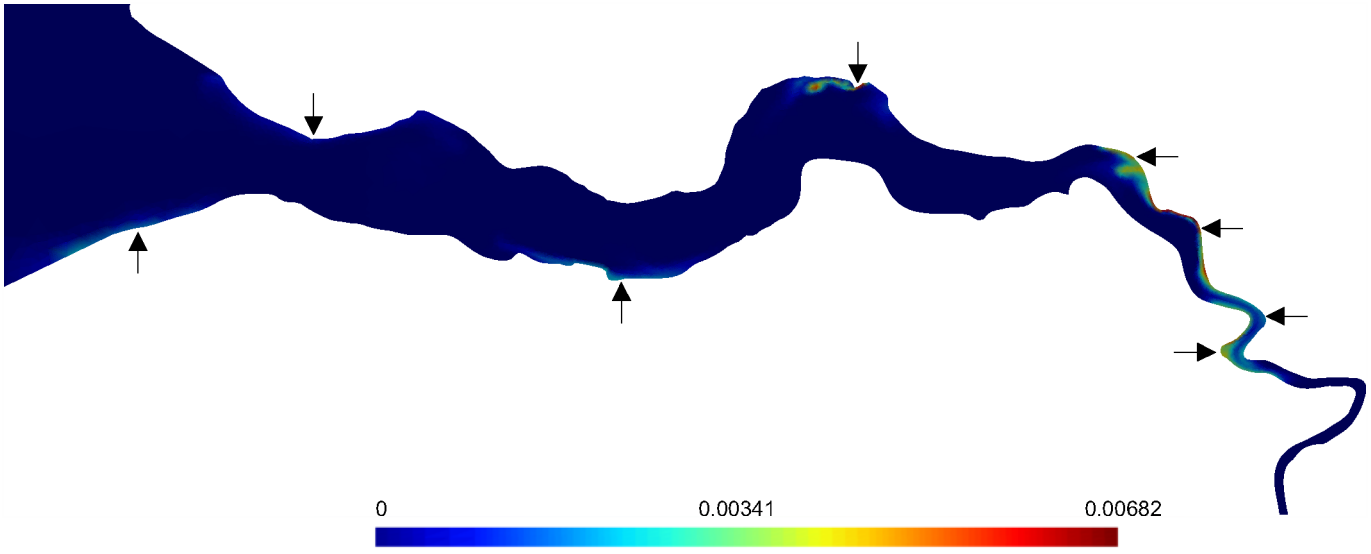


**Fig. 3.** Snapshot of simulation with 8 point sources (indicated by the arrows) injecting *E. coli* into the estuary (in units per $m^3$).

**Table 1**
Overview of model parameters under study.

| Symbol | Description | Typical value |
|---|---|---|
| $k_{mort}$ | Mortality rate constant of *E. coli* | $1.25 \times 10^{-5}\ s^{-1}$ |
| $v_{sed}$ | Sedimentation rate of *E. coli* | $5.55 \times 10^{-6}\ m\,s^{-1}$ |

respect to $k_{mort}$ could be derived. This theoretical test showed that there are two sources of error: the discretisation error introducing numerical diffusion and the finite size of the Gaussian-shaped input. In the Scheldt application both are minimized by using small triangles and a standard deviation for the Gaussian which is small compared to the grid size. This should remove the latter error but the numerical diffusion is probably not completely negligible because the grid size is still larger than the physical diffusion length scale $K/\|\underline{u}\|$ (of the order of 0.1–1 m in the Scheldt Estuary).

### 4.2. One sampling location with fixed sampling times

In this section, some examples are shown where a single sampling location is to be selected. For reasons of simplicity, the sampling timing is fixed beforehand. How this timing was chosen is explained in the next paragraph; subsequently two case studies are discussed, with respectively one and eight point sources.

#### 4.2.1. Fixing the sampling times

As the tide is a major process in the domain, the samples at any location were fixed to approximately cover one cycle of the major tidal component, M2 (12 h25′), such that the start time of sampling (relative to the tidal phase) is arbitrary. In addition, the experimental restrictions were taken into account (cf. Section 3.5), resulting in the following sampling timing: once every half an hour a sample is taken, during 12.5 h, resulting in 25 samples. With this setup the sampling covers one tidal cycle, and can be performed by one person, without having to store the samples too long before returning to the lab. Including travel times and preprocessing in the lab, this still makes a working day of more than 17 h.

#### 4.2.2. Optimal sampling location; one source

As a first example, some situations with one point source of *E. coli* are shown, in order to illustrate the importance of the local topography and related hydrodynamics. The point sources always discharge *E. coli* at a constant rate, but as only one source is considered the actual magnitude is of no importance. As it is expected that *E. coli* is primarily injected into the domain from the exterior (WWTPs, canal discharges, locks of the harbours, etc), it is most realistic to place sources along the coasts of the domain. Note that nothing in the model or experimental design procedure prohibits placing sources inside the domain, only in this case study it is believed that the relevant point sources discharge along (or very close to) the coastlines. The timing of the samples was fixed as said above. Then the single optimal location is sought where these 25 samples should be taken. This is done by computing the $F_S$ at every node of the mesh. If necessary, another spatial discretisation is possible, as with the finite element approach the model output can be computed at any position. The optimal sampling position is then the one maximizing $F_S$ (Equation (5)).

Fig. 4 shows the spatial distribution of $F_S$ (only with respect to $k_{mort}$, i.e. $v_{sed}$ is fixed) for a source close to Terneuzen, in the broader part of the estuary (see arrow in Fig. 4). According to the $F_S$ the

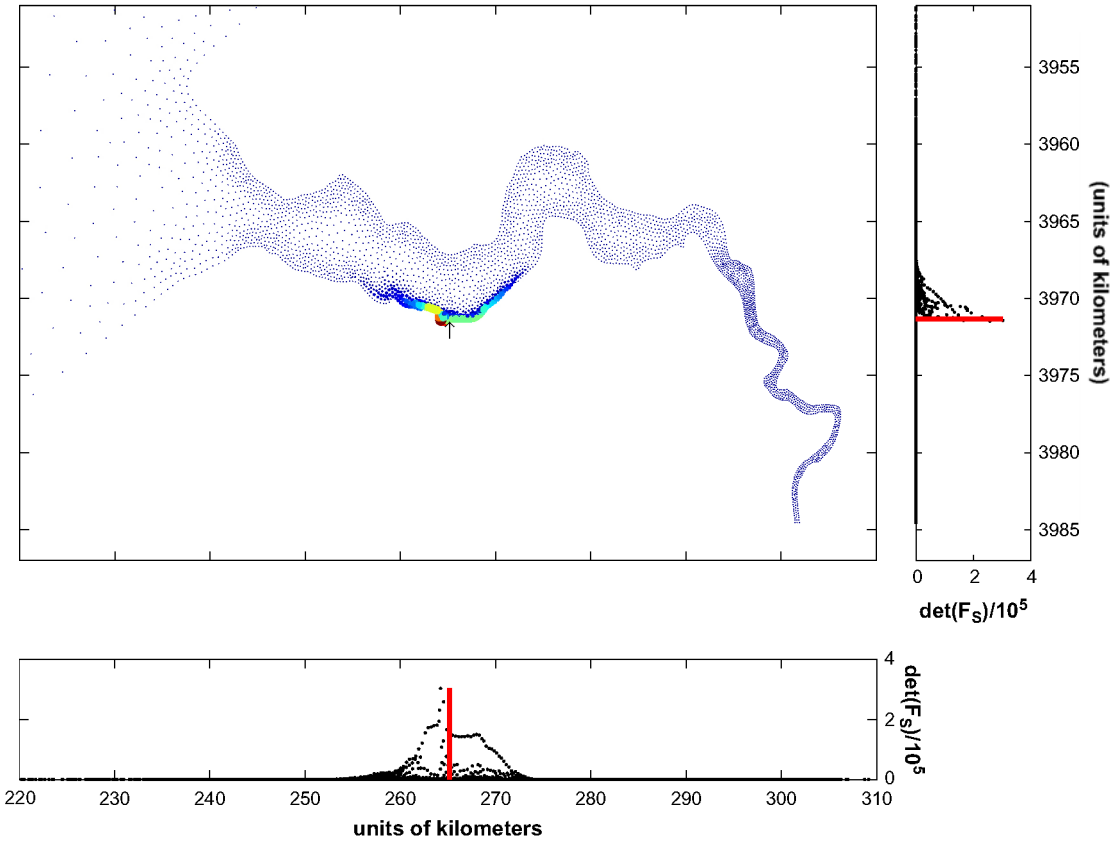


**Fig. 4.** One source at Terneuzen: spatial distribution of scaled Fisher information matrix ($F_S$), only considering $k_{mort}$, i.e. the information content of a sampling campaign of 12.5 h (1 sample / 30 min) for all considered locations. Central figure: dots are sized and colored proportionally to the value of the $F_S$, arrow represents the point source. This figure serves to spatially localise the information maxima. Side figures: $F_S$ as a function of $x$ and $y$ coordinates (i.e. projections), red lines indicate source location. These figures facilitate a more quantitative interpretation of the results. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

optimal sampling location is slightly downstream of the source, not symmetric and steeply peaked. This shape is thought to be attributable to the particular tidal dynamics. Indeed, when the simulations are performed with a simple sinusoidal tidal forcing at the mouth of the estuary, $F_S$ as a function of the $x$ coordinate has an almost Gaussian shape (results not shown).

The outcome is not only dependent on the hydrodynamics, but also on the model parameters that are considered in the analysis. In Fig. 5, the results are plotted for the same situation as above but considering the two model parameters $k_{mort}$ and $v_{sed}$. As now two parameters are considered, the determinant of $F_S$ was computed to get a scalar information criterion at every position. The spatial distribution of the information looks somewhat different although the peak location is the same. Note however that the information is not optimal at the source itself. To put numbers on it, sampling at the source point will deliver parameter variances respectively 38 and 47% (for $k_{mort}$ and $v_{sed}$) higher than those found by sampling at the optimal location.

The above examples show how an "informative" area can be visualised, with a maximum indicating the optimal sampling location. Other features of the information spread (secondary peaks or flat maximum) can be useful knowledge for planning a sampling campaign in real applications. Depending on the local topography and hydrodynamics, the results can differ. In some cases, the optimal sampling location may lie upstream of the point source, or the analysis may indicate that there are several locations delivering measurements with equivalently high information.

#### 4.2.3. Optimal sampling location; eight sources

This section will show that when more than one point source is present and only one location can be sampled, it is of even more interest to perform an information or experimental design analysis. Indeed, as long as we know there is only one source, it is quite sure that the optimal sampling location will be close to that single source. In the more realistic situation with several sources, determining the optimal sampling location is not so obvious anymore.

In Fig. 6 the results are shown considering eight point sources, eight being a more realistic number. In Fig. 3 a snapshot was shown of the simulated *E. coli* concentrations in the estuary. They are positioned along the estuary at known potential inputs of pollution, like canal discharge points and harbour locks (potentially important because WWTPs discharge in the harbour). However, their importance in terms of discharge in *E. coli* is not well known. Therefore, we assumed that they are all equal and assigned arbitrary values to the source fluxes ($1\ s^{-1}$). $F_S$ is computed with respect to both parameters $k_{mort}$ and $v_{sed}$. It is clear that not all sampling locations deliver the same information, although the sources discharge bacteria at the same rate. Instead it appears that sampling close to the most seaward sources delivers negligible information compared to the vicinity of the sources in the narrower part.

### 4.3. Influence of the sampling timing

So far, we fixed the sampling timing beforehand and only optimised the location where the samples should be taken. In this section, we will investigate how changing the timing of the sampling influences the results (while still optimising the location too). First, the question is investigated whether the timing of a sampling campaign relative to the tidal cycle is relevant. To address this question, the following simulation test was performed. A fixed sampling "protocol" was considered, consisting of 7 samples, taken 1 per half an hour, covering a sampling period of 3 h.

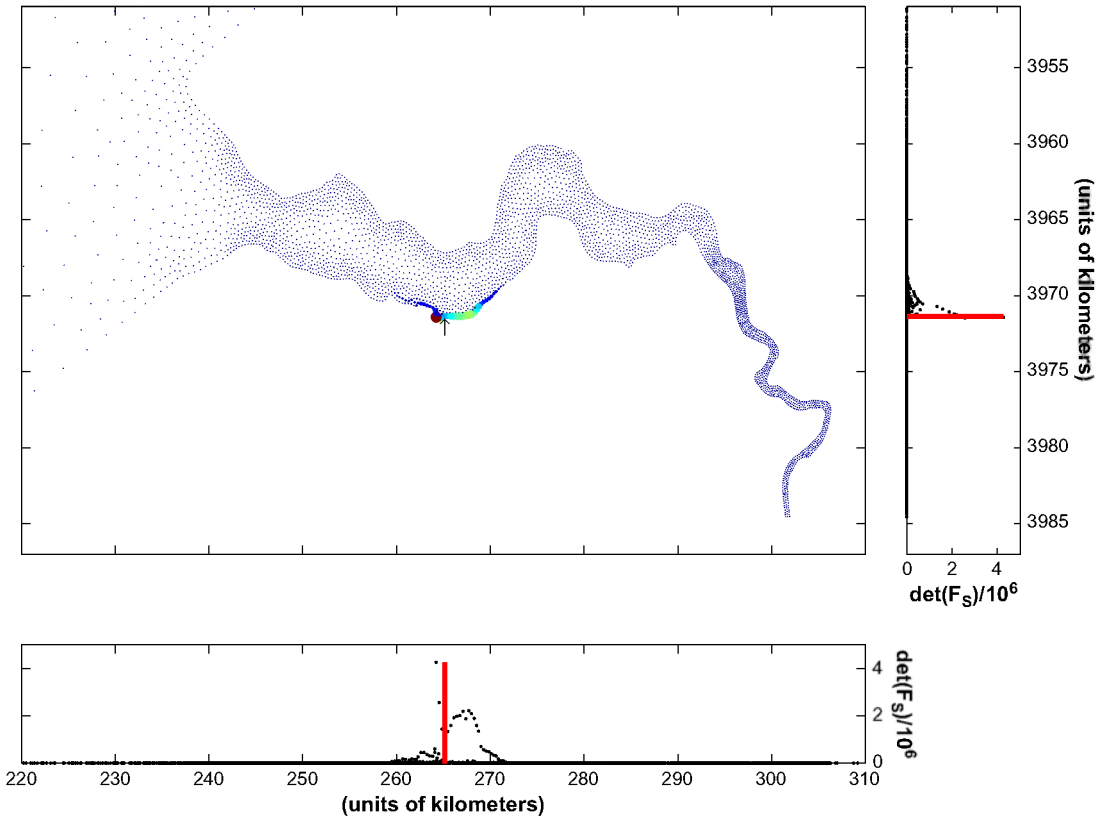


**Fig. 5.** One source at Terneuzen: spatial distribution of det($F_S$), considering both $k_{mort}$ and $v_{sed}$. (cf. Fig. 4) Central figure: dots are sized and colored proportionally to the value of the det($F_S$), arrow represents the point source. Side figures: det($F_S$) as a function of $x$ and $y$ coordinates, red lines indicate source location. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

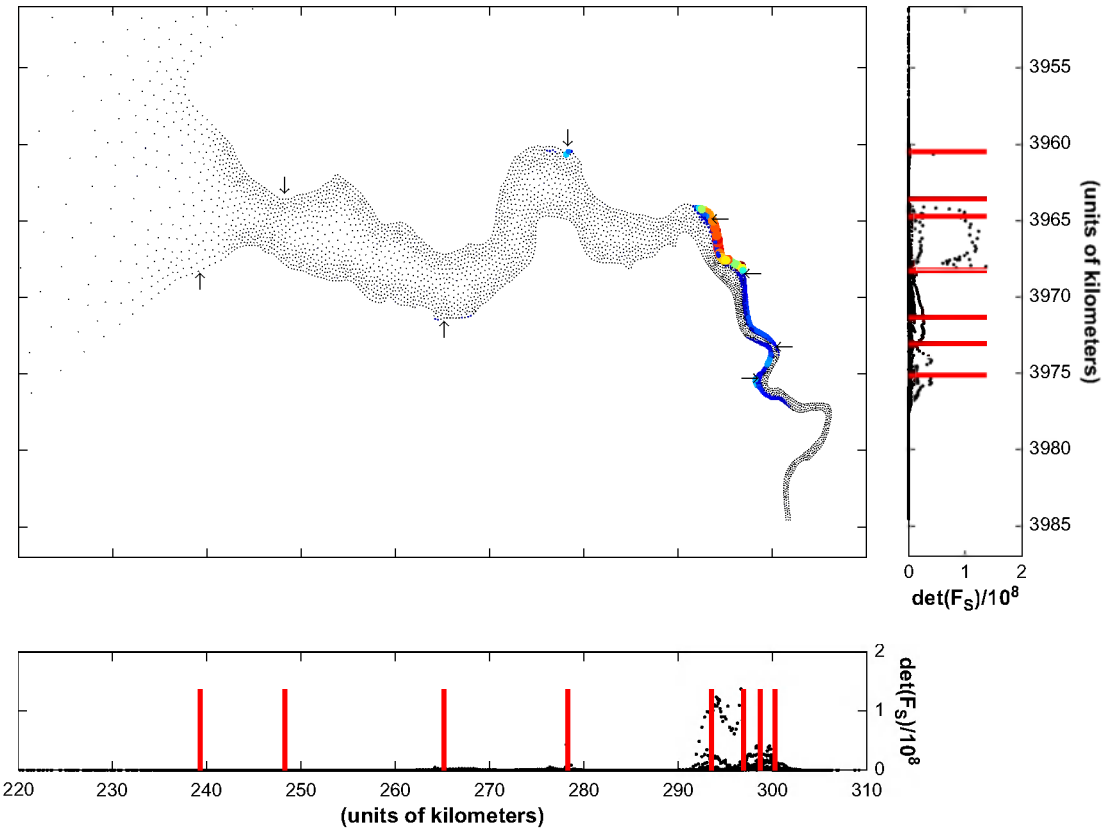


**Fig. 6.** Eight sources: spatial distribution of det($F_S$), considering both parameters $k_{mort}$ and $v_{sed}$. Central figure: dots are sized and colored proportionally to the value of the det($F_S$), arrows represent the point sources. Side figures: det($F_S$) as a function of $x$ and $y$ coordinates, red lines indicate source positions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

This sampling period is short relative to the main tidal period of approximately 12 h and 25 min; therefore we suspect that the information that can be retrieved from one such sampling campaign will vary with the start time of the campaign. In other words, we wonder whether there is a preference to start the sampling campaign at high tide, or low tide or any other time in the tidal cycle.

Fig. 7 shows the results for the maximum information that can be retrieved by the 3 h campaign depending on its starting time. In this first example, $F_S$ is computed only with respect to $k_{mort}$. To identify any relation with the tide, the water height is plotted on the same time axis, as it is modelled at the source position (the single source at Terneuzen is considered here and as we know the maximum information is close to the source, this location seems representative). There is a clear periodicity in the information evolution, closely matching the main tidal periodicity. In fact both series are almost in anti-phase: the maximum information that can be retrieved from the sampling campaign is highest for those

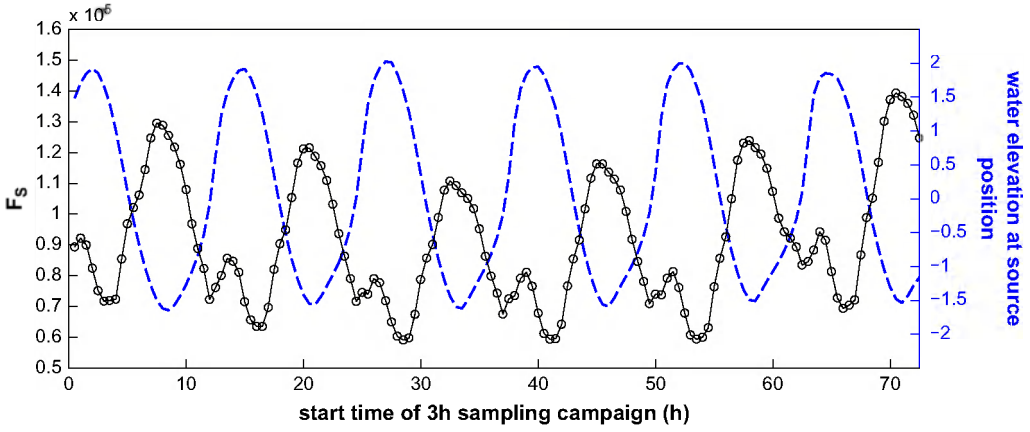


**Fig. 7.** Relation between maximum information (black line with circles) retrieved from a 3 h sampling campaign (1 sample every half an hour) and starting time of the campaign. The simulation considers a single source at Terneuzen (Section 4.2.2) and only parameter $k_{mort}$ for the computation of $F_S$. In the same plot water elevation (blue dotted line) at the starting time (different $y$ axis) at the source position is shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

campaigns that started just before low tide. A second, smaller information peak is visible just before high tide. These phases in the tide are both associated with low water velocities, and the low tide is, obviously, characterised by low water level. These features may increase the sample information because they tend to induce an accumulation of tracer concentration. High concentrations will probably be associated with high (in absolute values) sensitivities to changes in model parameters, and this is exactly what defines the information potential of the samples.

Note that this exercise in fact consists of the optimisation of both sampling location and start time. Indeed, for every sampling campaign (starting at a different time) the maximal information shown in Fig. 7 corresponds to a different optimal sampling location, namely that sampling location delivering the most informative measurements for that campaign. As can be expected, this location depends on the phase of the tide as well. Roughly said, for campaigns during rising tide, the optimal sampling location is upstream of the source; during falling tide it is downstream.

The results for only $v_{sed}$ considered, are quite similar (not shown). Note that this means that when both parameters are taken into account, and trace($F_S$) is used as the information criterion, instead of det($F_S$), also similar results are obtained.

Conversely, when repeating the exercise with both parameters and det($F_S$) as information criterion, a different picture is found (Fig. 8). Indeed, the information, now expressed by det($F_S$) still exhibits a periodicity strongly related to the tide, but the peaks are shifted. There is still a gain in information when sampling around low tide, but the highest information is achieved 2–3 h before high tide. Sampling a little bit later and during the biggest part of falling tide results in much lower information. This is not straightforward to explain in terms of hydrodynamics anymore, and is probably due to the fact that to maximize det($F_S$) the off-diagonal terms, related to the interaction between the two parameters, play a role, which is not the case if only one parameter is considered or the trace criterion. In fact, a sampling may be optimal to estimate the two parameters individually, or to minimise their individual variances, but not optimal to reduce their covariances – which is probably the case here. This observation confirms that the optimal sampling setup depends on which parameters are to be estimated with the eventual measurements. More generally, it illustrates the expectation that there is no single best sampling, the optimal sampling depends on what you want to do with it.

Looking at the achievable parameter precision quantitatively, the difference between two sampling campaigns starting at different times can be huge. For instance, when comparing the two campaigns indicated by arrows in Fig. 8: a campaign starting at time 53.5 h (just after high tide) will result in a parameter variance nine ($k_{mort}$) to eleven ($v_{sed}$) times the variance found when the sampling started at time 49.5 h (just before high tide). This means that to achieve comparable precisions ten times more samples would be needed in the first case than in the second case.

Besides the start time of a campaign, the sampling frequency may also be a factor influencing the retrievable information. The sampling frequency may be changed by varying the number of samples or the total campaign period. The first case is irrelevant as more samples will always deliver more information. But the effect of the second case is not so obvious a priori. To investigate how changing the sampling frequency (by changing the campaign period and keeping seven samples) affects the maximal information that can be retrieved from the sampling campaign, the preceding exercise was extended to take into account varying sampling frequencies in addition to sampling location and campaign starting time. Several cases were considered:

(a) the information is optimised with respect to location and starting time, and the variation of this maximal information as a function of sampling frequency (or actually campaign period) is plotted;
(b) the information is only optimised for location, i.e. for a campaign fixed to start at $t = 0$;
(c) the information is only optimised for start timing, i.e. for a campaign fixed at the source position.

The results are summarised in Fig. 9, showing that apparently the maximal information as a function of sampling frequency does not exhibit any systematic pattern. That is to say, the optimal sampling frequency could not be related to any other feature, like the tidal periodicity. The highest information is found for the shortest campaign but it is not clear whether this is a systematic result or found by coincidence. Yet, the parameter variances can be improved by choosing the best sampling frequency, so it is certainly not irrelevant to consider sampling frequency as an optimisation variable.

### 4.4. Selecting several sampling locations

So far, we have considered problems where only one sampling location is to be selected. In this section, the selection of several sampling locations was investigated. A sequential procedure is proposed, such that the information criterion is optimised for one location at a time. With a fixed sampling timing, the first step is then identical to the selection of a single sampling location (Section 4.2). In the next step, this location is fixed, and the additional information that can be retrieved from a second simultaneous sampling at another location is maximised.

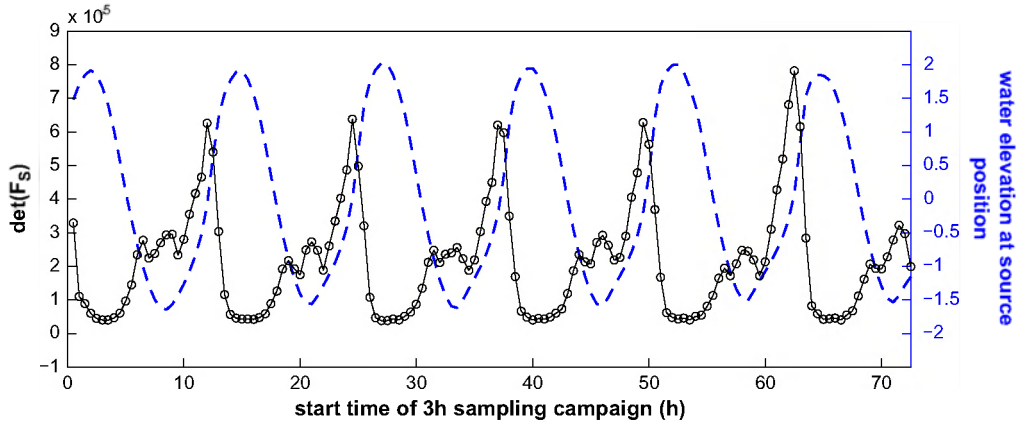


**Fig. 8.** Similar to Fig. 7 but with both parameters $k_{mort}$ and $v_{sed}$ considered.

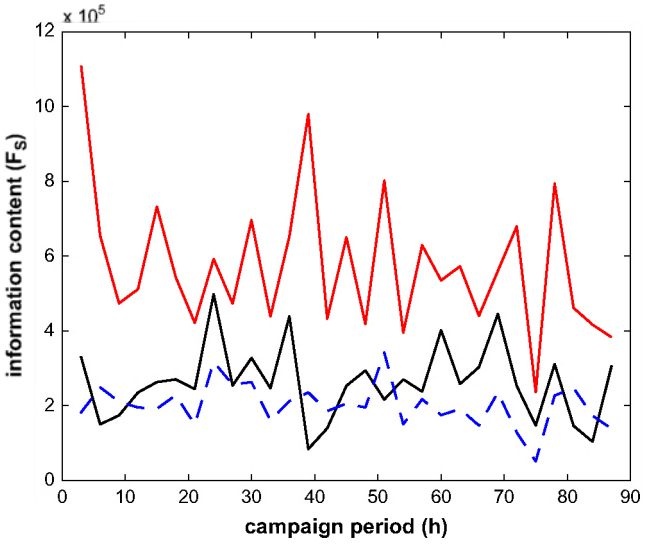


**Fig. 9.** Evolution of (maximal) information content ($F_S$) of a 7 samples campaign as a function of the campaign period. $F_S$ is computed considering both parameters $k_{mort}$ and $v_{sed}$. Thick red line: information at optimal sampling location and starting time. Thin black line: information at optimal sampling location but for fixed starting time at $t=0$. Blue dashed line: information for optimal starting time and at fixed location (source position). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

First, note that if only one model parameter is considered in the analysis, there is only one single optimal sampling location, no matter how many can be selected. Indeed, the criterion which is maximised is $J(R,T)^T J(R,T)$, where the $T$ stands for the fixed sampling times and $R$ defines the different sampling locations. As $J(R,T) = [J(r_1,T);\ J(r_2,T);\ \ldots]$ it is easily seen that the information criterion will be maximal if all the one-location criteria are maximal and this will obviously happen for one and the same location:

$$J(R,T)^T J(R,T) = \sum_{i=1}^{n_l} J(r_i,T)^T J(r_i,T) \leq n_l \max_r \left( J(r,T)^T J(r,T) \right). \quad (8)$$

If more than one parameter is included, whether this property (the information from several sampling locations is the sum of the "informations" from the single locations) still holds depends on the criterion used. If trace($F_S$) is used, the property is still valid, but for det($F_S$) it is not so obvious due to the mixed off-diagonal terms. In other words, if the information is expressed by trace($F_S$), only one single sampling location is optimal, no matter how many simultaneous replicates can be taken. So, no additional information is to be gained by performing two simultaneous sampling campaigns at two different locations. For $D$-optimal design, theoretically different locations may appear to be optimal. But based on some tests, if different locations come out of the analysis at all, they are very close to each other. Consequently, if a fixed sampling timing is applied, not much information is to be gained by repeating this sampling at a different location, compared to simply performing a replicate experiment at the same location.

## 5. Conclusion and perspectives

The potential of the information criterion related to the Fisher information matrix, expressing the magnitude of the covariance matrix of the model parameters, has been investigated to derive optimal sampling schemes for the calibration of a simplified model for *E. coli* concentration in the Scheldt Estuary. From the results, the method appears useful. Accurately timing and placing samplings have been shown to significantly reduce the parameter variances – indeed up to a factor of ten in the investigated examples – which will in turn influence the reliability of the calibrated model.

Considering the method, we would like to make two remarks with outlooks towards future developments. First, the experimental design procedure used here, based on the minimizing the parameter covariance matrix, delivers a design which is only locally optimal, because it depends on the values of the model parameters used to compute the matrix. To cope with this dependence, and at the same time include the fact that the model parameter values are not well known but fall within a well known range, a robust experimental design method could be applied. Such an approach computes the optimal experiments for the parameters varying within predefined ranges, and selects the ultimate optimal experiment based on some robustness rule. For instance, the most representative experiment may be selected, by taking the "centroid" experiment or the experimental conditions found by cluster analysis (Dror and Steinberg, 2006); an even more robust approach is to select the experiment performs best for the worst case parameter values (minimax approach, Rojas et al., 2007; Sun, 2007; Sun and Yeh, 2007). Such an approach may be interesting for environmental applications, as many model parameters are intrinsically not fixed but can vary according to varying environmental conditions, even within one system.

A second remark refers to the lack of "real" optimisation algorithm in this study. Rather, a "grid search" optimisation was performed, i.e. the optimisation variables (time and location) were only considered at discrete values, but they were considered at *all* these discrete values. Admittedly, this procedure will only give the optimal setup within the precision of the discretisation, so it has to be chosen to be relevant with respect to the crucial length and time scales. But the advantages are that within the search points the global optimum is found, and the "optimisation" is relatively fast: only $2n_p+1$ model runs are needed, to compute the meta-Jacobian by finite difference of model outputs, and the model outputs can be stored at all times and locations a priori decided. On the other hand, if more variables have to be optimised simultaneously, "trying" all combinations becomes impossible because of the exponentially increasing combinatorial complexity. Therefore, it would be interesting to find the optimal experimental variables using an efficient and global optimisation algorithm but which still requires "samples" from the meta-Jacobian.

For real applications, and the Scheldt case in particular, some recommendations may be formulated. First, for reliable results the model should include realistic and quantified point sources. Furthermore it is recommended to adopt a pragmatic attitude towards the optimal sampling scheme. The experimental design outcome should be regarded as a guideline and in terms of general trends, rather than as an inescapable fact. This is especially so because in real applications the *actually* optimal sampling design is a compromise between maximised information and practical constraints. If these constraints are known beforehand (e.g. sampling off the banks is more expensive because a boat is needed), they may be included in the information criterion to form a "cost function" expressing the overall "value for money" associated with a sample. Some previous studies already considered a maximum budget that should not be exceeded as a constraint in their experimental design analysis (Wagner, 1995; Sciortino et al., 2002; Catania and Paladino, 2009), but the experimental cost was simply set proportional to the number of samples and thus did not depend on the other variables of the sample. However, in general not all practical constraints are known in advance. Summarizing, the experimental design procedure can be used to find the optimal sampling conditions, taking into account all quantifiable information influencing the quality and cost of the measurements. Once this "theoretical" optimum has been found, it can further be

confronted with practical constraints, to enable the planning of an effective sampling campaign.

For *E. coli* in the Scheldt the current problem is the lack of prior data. Since the model is then the only information at hand, the proposed experimental design seems useful. In a future situation where data will be available, an iterative procedure can be applied: using the data the model parameters (and may be even the model structure) can be improved and this new model can be used together with the old data to determine where next samples should be taken. This way, at each stage all available information is used to decide on the next step. As such, this approach to experimental design illustrates (again) that models and observations should not be regarded separately but instead that they are the two indispensable pieces at our disposal to resolve the puzzle of reality.

## Acknowledgements

The authors wish to specially thank the Vlaamse Milieumaatschappij (Yves Vanderstraeten) and the Waterbouwkundig Laboratorium (Marc Wouters) for providing data. Anouk de Brauwere is postdoctoral researcher of the Research Foundation Flanders (FWO) and is currently working at the Université Catholique de Louvain as a postdoctoral intercommunity collaborator of the Francqui Foundation. Jonathan Lambrechts is supported by the Fonds pour la formation a la Recherche dans l'Industrie et dans l'Agriculture (FRIA). Eric Deleersnijder is a Research associate with the Belgian National Fund for Scientific Research (FRS-FNRS). The research was conducted within the frameworks of research project GOA22/DSWER4 and the Methusalem Fund (METH 1), funded by the Flemish Government; as well as the Interuniversity Attraction Poles TIMOTHY (IAP VI.13) and DYSCO (IAP VI.4), funded by the Belgian Science Policy (BELSPO). SLIM is developed under the auspices of the programme ARC 04/09-316 (Communauté française de Belgique).

## References

Alvarez-Vázquez, L.J., Martínez, A., Rodríguez, C., Vázquez-Méndez, M.E., 2002a. Mathematical analysis of the optimal location of wastewater outfalls. The IMA Journal of Applied Mathematics 67, 23–39.

Alvarez-Vázquez, L.J., Martínez, A., Rodríguez, C., Vázquez-Méndez, M.E., 2002b. Numerical optimization for the location of wastewater outfalls. Computational Optimization and Applications 22, 399–417.

Alvarez-Vázquez, L.J., Martínez, A., Vázquez-Méndez, M.E., Vilar, M.A., 2006. Optimal location of sampling points for river pollution control. Mathematics and Computers in Simulation 71, 149–160.

Breton, M., Salomon, J.C., 1995. A 2D long term advection–dispersion model for the Channel and Southern North Sea. Part A: validation through comparison with artificial radionuclides. Journal of Marine Systems 6, 495–513.

Catania, F., Paladino, O., 2009. Optimal sampling for the estimation of dispersion parameters in soil columns using an iterative genetic algorithm. Environmental Modelling and Software 24, 115–123.

Comblen, R., Legrand, S., Deleersnijder, E., Legat, V., 2008. A finite element method for solving the shallow water equations on the sphere. Ocean Modelling, doi:10.1016/j.ocemod.2008.05.004.

Dixon, W., Chiswell, B., 1996. Review of aquatic monitoring program design. Water Research 30, 1935–1948.

Dror, H.A., Steinberg, D.M., 2006. Robust experimental design for multivariate generalized linear models. Technometrics 48 (4), 520–529, doi:10.1198/004017006000000318.

Edberg, S.C., Rice, E.W., Karlin, R.J., Allen, M.J., 2000. *Escherichia coli*: the best biological drinking water indicator for public health protection. Journal of Applied Microbiology 88, 106S–116S.

Fedorov, V.V., 1972. Theory of Optimal Experiments. Academic Press, New York, 292 pp.

Fedorov, V.V., Hackl, P., 1997. Model-Oriented Design of Experiments. Lecture Notes in Statistics. Springer, New York.

Fewtrell, L., Bartram, J., 2001. Water quality: guidelines, standards and health. In: World Health Organization Water Series. IWA Publishing, London (UK).

Garcia-Armisen, T., Thouvenin, B., Servais, P., 2006. Modelling faecal coliforms dynamics in the Seine Estuary, France. Water Science and Technology 54 (3), 177–184.

Garcia-Armisen, T., Servais, P., 2007. Respective contributions of point and non point sources of *E. coli* and enterococci in a large urbanized watershed (the Seine river, France). Journal of Environmental Management 82 (4), 512–518.

Gmsh, 2008. Gmsh: a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities, by Christophe Geuzaine and Jean-François Remacle, Version 2.1.0., http://www.geuz.org/gmsh/.

Jacquez, J.A., 1998. Design of experiments. Journal of the Franklin Institute 335B, 259–279.

Kärnä, T., Deleersnijder, E., de Brauwere, A., Simple test cases for validating a finite element unstructured grid fecal bacteria transport model, Applied Mathematical Modelling, submitted for publication.

Kay, D., Stapleton, C.M., Wyer, M.D., McDonald, A.T., Crowther, J., Paul, N., Jones, K., Francis, C., Watkins, J., Wilkinson, J., Humphrey, N., Lin, B., Yang, L., Falconer, R.A., Gardner, S., 2005. Decay of intestinal enterococci concentrations in high-energy estuarine and coastal waters: towards real-time $T_{90}$ values for modelling faecal indicators in recreational waters. Water Research 39, 655–667.

Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11 (1), 137–148.

Knopman, D.S., Voss, C.I., 1989. Multiobjective sampling design for parameter estimation and model discrimination in groundwater solute transport. Water Resources Research 25 (10), 2245–2258.

Lambrechts, J., Hanert, E., Deleersnijder, E., Bernard, P.-E., Legat, V., Remacle, J.-F., Wolanski, E., 2008a. A multi-scale model of the hydrodynamics of the whole Great Barrier Reef. Estuarine, Coastal and Shelf Science 79, 143–151.

Lambrechts, J., Comblen R., Legat V., Geuzaine C., Remacle J.-F., 2008b. Multiscale mesh generation on the sphere, Ocean Dynamics 58 (5-6), 461–473, doi 10.1007/s10236-008-0148-3.

Lo, S.L., Kuo, J.T., Wang, S.M., 1996. Water quality monitoring network design of Keelung river, northern Taiwan. Water Science and Technology 34 (12), 49–57.

Manafi, M., 2000. New development in chromogenic and fluorogenic culture media. International Journal of Food Microbiology 60, 205–218.

Marengo, E., Todeschini, R., 1992. A new algorithm for optimal, distance-based experimental design. Chemometrics and Intelligent Laboratory Systems 16, 37–44.

McPhee, J., Yeh, W.W.-G., 2006. Experimental design for groundwater modeling and management. Water Resources Research 42, W02408, doi:10.1029/2005wr003997.

Meire, P., Ysebaert, T., Van Damme, S., Van den Bergh, E., Maris, T., Struyf, E., 2005. The Scheldt Estuary: a description of a changing ecosystem. Hydrobiologia 540, 1–11.

Naithani, J., Darchambeau, F., Deleersnijder, E., Descy, J.-P., Wolanski, E., 2007. Study of the nutrient and plankton dynamics in Lake Tanganyika using a reduced-gravity model. Ecological Modelling 200, 225–233.

Padilla, F., Secretan, Y., Leclerc, M., 1997. On open boundaries in the finite element approximation of two-dimensional advection–diffusion flows. International Journal For Numerical Methods In Engineering 40, 2493–2516.

Pichot, G., Barbette, J., 1978. Estimation des taux moyens de disparition des bactéries fécales dans les eaux côtières belges de la mer du Nord. Revue Internationale. d'Océanographie Médicale LI–LII, 115–126.

Pukelsheim, F., 2006. Optimal Design of Experiments. SIAM, Philadelphia, USA, 454 pp.

Rensfeld, A., Mousavi, S., Mossberg, M., Söderström, T., 2008. Optimal sensor locations for nonparametric identification of viscoelastic materials. Automatica 44, 28–38.

Rojas, C.R., Welsh, J.S., Goodwin, G.C., Feuer, A., 2007. Robust optimal experiment design for system identification. Automatica 43 (6), 993–1008.

Rompré, A., Servais, P., Baudart, J., De Roubin, M.R., Laurent, P., 2002. Methods of detection and enumeration of coliforms in drinking water: a review. Journal of Microbiological Methods 49, 31–54.

Sanders, T.G., Adrian, D.D., 1978. Sampling frequency for river quality monitoring. Water Resources Research 14, 569–576.

Sanders, T.G., 1982. Representative sampling location criterion for rivers. Water South Africa 8 (4), 169–172.

Schoonjans, J., 2007. Stikstof in een antropogeen vervuilde rivier (de Zenne): Distributie, speciatie en biodegradeerbaarheid van de organische fractie, M.Sc. Thesis, Vrije Universiteit Brussel, Brussels, Belgium.

Sciortino, A., Harmon, T.C., Yeh, W.W.-G., 2002. Experimental design and model parameter estimation for locating a dissolving dense nonaqueous phase liquid pool in groundwater. Water Resources Research 38 (5), 1057, doi:10.1029/2000wr000134.

Scott, R., 1973. Finite element convergence for singular data. Numerical Mathematics 21, 317–327.

Servais, P., Billen, G., Goncalves, A., Garcia-Armisen, T., 2007a. Modelling microbial water quality in the Seine river drainage network: past, present and future situations. Hydrology and Earth Systems Sciences 11, 1581–1592.

Servais, P., Garcia-Armisen, T., George, I., Billen, G., 2007b. Fecal bacteria in the rivers of the Seine drainage network (France): sources, fate and modelling. Science of the Total Environment 375 (1–3), 152–167.

Sharp, W.E., 1971. A topologically optimum water-sampling plan for rivers and streams. Water Resources Research 7, 1641–1646.

Smagorinsky, J., 1963. General Circulation experiments with the primitive equations. Monthly Weather Review 91 (3), 99–164.

Soetaert, K., Herman, P., 1996. Estimating estuarine residence times in the Westerschelde Estuary (S.W. Netherlands) using a box model with fixed dispersion coefficients. Hydrobiologia 311, 215–224.

Steets, B.M., Holden, P.A., 2003. A mechanistic model of runoff-associated fecal coliform fate and transport through a coastal lagoon. Water Research 37, 589–608.

Steinberg, D.M., Hunter, W.G., 1984. Experimental design: review and comment. Technometrics 26 (2), 71–97.
Sun, A.Y., 2007. A robust geostatistical approach to contaminant source identification. Water Resources Research 43, W02418, doi:10.1029/2006WR005106.
Sun, N.-Z., Yeh, W.W.-G., 2007. Development of objective-oriented groundwater models: 2. Robust experimental design. Water Resources Research 43, W02421, doi:10.1029/2006WR004888.
Vandenberghe, V., van Griensven, A., Bauwens, W., 2002. Detection of the most optimal measuring points for water quality variables: application to the river water quality model of the river Dender in ESWAT. Water Science and Technology 46, 1–7.
Vanderborght, J.P., Folmer, I.M., Aguilera, D.R., Uhrenholdt, T., Regnier, P., 2007. Reactive-transport modelling of C, N and $O_2$ in a river–estuarine–coastal zone system: application to the Scheldt Estuary. Marine Chemistry 106, 92–110.
Vanrolleghem, P., Coen, F., 1995. Optimal design of in-sensor-experiments for on-line modelling of nitrogen removal processes. Water Science and Technology 31 (2), 149–160.
Wagner, B.J., 1995. Sampling design methods for groundwater modeling under uncertainty. Water Resources Research 31 (10), 2581–2591.
Whitfield, P.H., 1988. Goals and data collection designs for water quality monitoring. Water Resources Bulletin 24, 775–780.