GEOSTATISTICAL MODEL FOR CONCENTRATIONS OR FLOW RATES IN STREAMS: SOME RESULTS

EDWIGE POLUS-LEFEBVRE[1], CHANTAL DE FOUQUET[1], CAROLINE BERNARD-MICHEL[2], NICOLAS FLIPO[1] and MICHEL POULIN[1]

[1] Centre de Géosciences, MINES ParisTech, Fontainebleau, France

[2] now at MISTIS, INRIA Rhône-Alpes, Saint-Ismiers Cedex, France

ABSTRACT

Estimating concentrations or flow rates along a stream network requires specific Random Functions (RF) models. We propose a construction by "streams", which combines one-dimensional-RF defined on each path between sources and outlet. The model properties are examined, namely the consistency conditions at the confluences for different variables. In practice, the data are spatially too scarce for a precise inference of the covariances. To verify if a phenomenological model can be used to guide the geostatistical modelling, measurements are compared to the output of the ProSe model. The similarity of results is convincing for discharge, and acceptable for the nitrates concentration.

INTRODUCTION

Estimating concentrations or pollutant loads along a stream network makes specific Random Functions models necessary. Indeed, the usual geostatistical models were developed for Euclidean spaces and are not valid anymore for the graph topology: the Bochner and Schoenberg theorems that give the spectral characterization of the covariance or variogram, explicitly refer to the Euclidean distance (Chilès and Delfiner, 1999 for example). Using the curvilinear distance along a stream network, authors (Ver Hoef *et al.*, 2006) give an example of negative eigenvalues of the covariance matrix, with the spherical model. The variances calculated with this "model" can thus be negative too.

We first recall (de Fouquet & Bernard-Michel, 2006) the principles of a RF model defined on directed trees, constructed by combining 1D RF. Then the critical inference question is examined: on the treated cases, the data are spatially too sparse for a precise inference of the covariance. Can a phenomenological model then be used as a mock-up to guide the geostatistical modelling? To answer this question, the experimental data are compared to the ProSe model outputs, with a detailed study of time and space variograms.

RANDOM FUNCTION DEFINED ON A TREE

The few models found in the literature are briefly recalled, and we present the construction principle of a wide class of RF models.

Bibliographic Elements

To model the fluvisol width in a part of the Hérault hydrographic network (Monestiez *et al.*, 2005) or the drain ditches of the Roujan Basin (Bailly *et al.*, 2006), the authors construct a RF on a tree from the outlet to the sources. Given all the values downstream of a confluence, a hypothesis of conditional independence between points located on different rivers upstream of this confluence is made. On each river, all one-dimensional covariance models are admissible; the covariance between points from different rivers is not stationary.

To estimate the heavy metals concentration along a stream network, Ver Hoef *et al.* (2006) use a construction in the opposite direction, from the sources to the outlet. Hypothesizing the independence between rivers upstream of their confluence, they adapt the classical moving average method, distributing the "kernel" on the half line among the rivers upstream of the confluences. Using this model, others (Cressie *et al.*, 2006) combine Euclidean and curvilinear distances.

In the same way, Bruno *et al.* (2001) developed a "contribution model" of the metals regional distribution from measurements made in sediments along the stream. The value at a point is the sum of the value immediately upstream and of a local contribution. The points along the stream are expressed by curvilinear distance, while the local contribution is expressed with the Euclidean distance.

Stationarity

In the following, a covariance is said to be "stationary" when it only depends on the curvilinear distance between two points, and non-stationary when it depends on the two points separately. If the covariance depends on the river orientation, the global covariance (on the whole graph) is considered as non-stationary.

Combination of One-Dimensional RF

Construction Principle (Bernard-Michel, 2006)

Let us suppose two affluents, with respective discharges d_1, d_2 and concentrations c_1, c_2 immediately upstream of their confluence. Immediately downstream of the confluence, the discharge is $d = d_1 + d_2$ and the concentration $c = \frac{d_1}{d_1 + d_2} c_1 + \frac{d_2}{d_1 + d_2} c_2$, i.e. a linear combination of the variables defined on the affluents. Similar equations are obtained for specific discharge

(discharge/watershed area) or residues. At the confluence, all these variables show a discontinuity.

Let us prolong the affluents as distinct "water paths", whose gathering forms the "rivers". For each path, from one source to the outlet, we define a discharge and a concentration, which are functions of the curvilinear abscissa counted from the outlet. Downstream of the successive confluences from the sources, the paths discharges cumulated themselves to form the global discharge, and the river concentration is given by a combination of the different paths concentrations, in accordance with their relative discharge. This procedure gives a general method to construct RF on a tree, by combining 1D processes.

Definitions and Notations

The usual geographic terminology is used together with the one relative to graphs. Some notations are taken from Ver Hoef *et al.* (2006). The hydrographic network is represented by a tree whose vertex (or nodes) are sources, confluences or the supposed unique outlet. A "river" is a path from a source to the outlet; each river corresponds to a unique source, and vice versa. Two points (or edges) are stream-connected if they belong to at least one common river. Two unconnected points (or edges) belong to different rivers, upstream of their confluence.

The rivers are indexed in capital letters and the edges with small letters (Figure 1). The edge immediately downstream of a source is indexed by the source. Each river has the index of its upstream edge. There are two ways to spot any point of the river network s_i: its curvilinear distance s positively counted from the outlet (where $s = 0$), or the number i of the edge it belongs to. The upstream vertex u_i belongs to the edge i when the downstream vertex is supposed to be the upstream part of the next edge (going with the current) and does not belong to edge i.

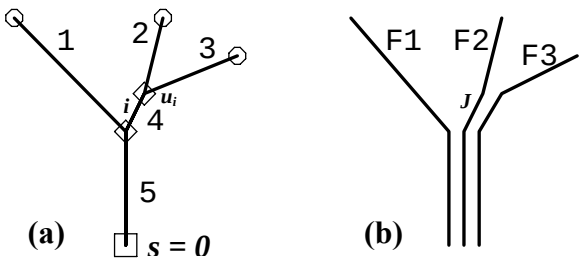


Figure 1: Tree description. (**a**) definition of edges and (**b**) rivers.

The indexes of the edges belonging to the river J upstream of the edge i, this one being excluded, is noted B_{iJ}. V_i refers to the rivers set going through the edge i. The confluence abscissa of the rivers I and J is noted u_{IJ} while u_{ij} refers to the confluence abscissa of the rivers going through the edges i and j. The curvilinear distance along the tree is:

- $d\left(s_i, t_j\right) = \left|s - t\right|$ on the whole river;
- $d\left(s_i, t_j\right) = \left(s_i - u_{ij}\right) + \left(t_j - u_{ij}\right)$ between unconnected points.

The length of the river J is the curvilinear abscissa u_J of its source. The confluences can be denoted by their downstream edge.

The "paths" are in bijection with the rivers. The RF or "component" Y_J is defined on the F_J "path"(Figure 1b), which is a segment whose length is u_J. The RF Z represents the discharge or a concentration along the river network and can be defined on the tree indexed by its edges. The covariance of Z is written as a function of *(s - t)* when it only depends on the curvilinear distance between the points, or as a function of s_i and s_j when it also depends on the edges.

Combination of Stationary RF

Let us consider a tree with N sources and "components" Y_J, $1 \le J \le N$, centered and with any 1D covariance $C_J(h)$. At each confluence, let's attribute to each upstream edge k a weight w_k (Figure 2). On the tree, the RF is defined as:

$$Z\left(s_i\right) = \sum_{J \in V_i} \left(\prod_{k \in B_{iJ}} w_k \right) Y_J(s) \qquad (1)$$

In this linear combination, the *coefficients* of the Y_J components of the paths going trough the edge i are equal to the product of all *weights* of the edges strictly upstream of i (from the source u_J). The sources are treated like confluences with only one upstream edge, whose coefficient is one. On the edge $i = I$ immediately downstream of a source, $Z(s_i)=Y_I(s)$.

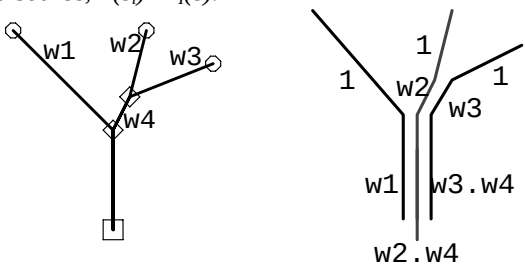


Figure 2: Combining rivers. On the left, weights assigned to the edges and on the right, resulting coefficients along the rivers.

Simple Case of Independent Components

Let's first assume the mutual independency of the Y_J components, all with the same covariance $C(h)$. In the combination (1), unconnected points don't have any common component. Given the spatial independency of the Y_J the covariance between these points is null.

For two connected points located on different edges, with *i* upstream of *j*, only the common paths components, those going through the edge *i*, have a contribution in the covariance. The Z covariance between these points is, for $s > t$ and $V_i \subset V_j$:

$$C_Z(s_i, t_i) = C_1(s-t) \sum_{j \in V_i} \left(\prod_{k \in B_{iJ}} w_k \right) \left(\prod_{l \in B_{jJ}} w_l \right) \quad (2)$$

Along the whole river, the Z covariance is proportional to $C_1(s-t)$ up to a factor varying according to the edges *i* and *j*. The weights of all the edges located between sources and each point contribute to the covariance (2). On any edge *i* Z is a linear combination of the same Y_J components, $J \in V_i$, and its covariance is:

$$C_Z(s_i, t_i) = C_1(s-t) \sum_{J \in V_i} \prod_{k \in B_{iJ}} w_k^2 \quad (3)$$

In particular, $Var\, Z(s_i) = C_1(0) \sum_{J \in V_i} \prod_{k \in B_{iJ}} w_k^2$

On an edge, the Z variance is constant and its covariance is stationary. At the confluences Z is discontinuous in squared mean, meaning that its covariance is discontinuous. Z is non-stationary on a river and thus on the tree: its variance generally changes at each confluence. For a defined curvilinear distance, the covariance depends on common sources of the considered points, intermediate confluences, and edges weight.

This model, constructed from the sources to the outlet, makes the description of the concentrations along a hydrographic network possible. When they are known (for instance calculated from the drained watershed area), the relative discharges play a part via the confluences weights.
For the discharges, the mass conservation condition is respected if each edge weight is 1. The RF Z is constructed by summation of its components Y_J:

$$Z(s_i) = \sum_{J \in V_i} Y_J(s) \quad (4)$$

When variance $C_1(0)$ is identical for all the paths, the Z variance is at each point proportional to the paths number: constant by edge, it increases from the sources to the outlet.

Let's now suppose that at each confluence the squared weights sum is equal to 1. For a *n* edges confluence:

$$\sum_{j=1}^{n} w_j^2 = 1 \quad (5)$$

The demonstrations given by Ver Hoef *et al.* (2006) remain valid for any covariance C_1. By recurrence on the successive confluences from the sources, we show that on each edge the covariance (3) is equal to $C_1(s-t)$. The Z variance $C_1(0)$ is then constant on the tree. For two connected points s_i and t_j separated by at least one confluence, the only weights contributing to the covariance are the ones of the confluences located between s_i and t_j:

$$C_Z\left(s_i, t_j\right) = C_1(s-t)\prod_{k\in B_{ij}} w_k^2 \tag{6}$$

The Ver Hoef model corresponds to a particular case, where covariance C_1 is the autoconvolution of a kernel f defined on the half line. The Y_J components are constructed by convolution of a random orthogonal measurement with kernel f. In the combination by "paths", any covariance C_1 can be used. In case of convolution kernel f can be symmetric for instance.

More generally let's introduce a weighting function $a_J(s)$ by path. The linear combination:

$$Z\left(s_i\right) = \sum_{J\in V_i} a_J\left(s_i\right) Y_J\left(s\right) \tag{7}$$

defines a RF with non-stationary variance and covariance, given respectively by:

$Var\, Z\left(s_i\right) = C_1(0)\sum_{J\in V_i}\left(a_J\left(s_i\right)\right)^2$, and

$$C_Z\left(s_i, t_j\right) = C_1(s-t)\sum_{K\in V_i\cap V_j} a_K\left(s_i\right) a_K\left(t_j\right) \tag{8}$$

The initial model corresponds to a constant by edge function, with:

$a_J\left(s_i\right) = \prod_{k\in B_{iJ}} w_k$.

A weighting function constant by edge gives a Z covariance stationary by edge. For discharges, $a_J(s)$ is constant and equal to 1 along each river ((4) and (7)).

This model is easily extended to correlated components Y_J or to different covariances C_J (de Fouquet and Bernard-Michel, 2006). Modifying the operator acting on the Y_J gives other classes of RF models on tree: the linear combination (7) can be replaced by an average of any order, or by a product, the minimum or the maximum. The previous models can be extended to other graph types by combining components defined on the paths connecting two "end" nodes.

EXPERIMENTAL RESULTS

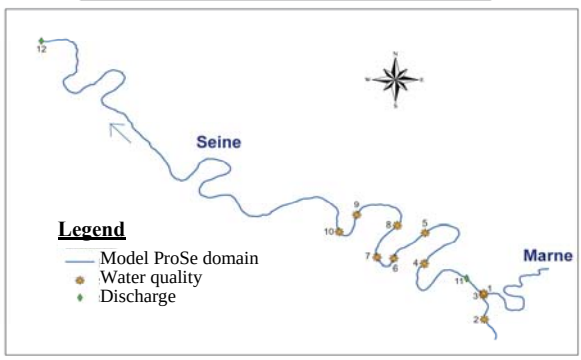
In practice, inferring 1D covariances C_J becomes problematic because of the scarcity of data points. The French RNB (National Basin Network) has been providing monthly measurements for more than fifteen years, but with rarely more than one measurement site per edge (Bernard-Michel, 2007).

A solution to this problem could be to use a deterministic model as a mock-up provided it represents pretty well the reality. The flexibility of the time and space output resolution of the model should enable us to make a detailed variographic analysis and fitting. Preliminary results of the comparison between measurements and a deterministic model are now presented.

The Deterministic Model ProSe

The ProSe model (Even et al., 1998, 2004, 2007; Flipo et al., 2004) is composed of three modules: hydrodynamic, transport and biogeochemical. The conceptual scheme is based on a macroscale simulation of the micro-organisms dynamics that govern the transformation of many components (organic matter, nutrients, oxygen). Only two compartments, water column and sediments, are simulated here because biogeochemical reactions due to periphyton are less important in large rivers (Flipo et al. 2004). In the following, the output of the model is called "ProSe values". The simulation was run for year 2003, for the rivers Marne and Seine, upstream of the greater Paris to the estuary (200km). The model gives three outputs per hour, namely at ten measurement sites (Figure 3).
Note useful hereafter: Veolia Water provided daily integrated data, obtained from hourly sampling, which were used for boundary conditions. More details about these and the model parameterization can be found in Poulin (2006).

The discharge is very well reproduced by the ProSe model. The Nash criterion, often used to qualify the simulated discharge in hydrology and representing the variance part explained by the model, is superior to 0.97. This model can thus be used to infer the variographic model, or even directly to estimate the discharge between measurement sites. This allows us to go on with the variographic analysis of the specific discharge (discharge/watershed area) initiated by Bernard-Michel (2006).



1	Choisy
2	Ivry
3	Alfortville
4	Suresnes
5	Argenteuil
6	Chatou
7	Bougival
8	Sartrouville
9	Conflans
10	Poissy
11	Austerlitz
12	Poses

Figure 3: Modeled area. Sampling sites of water quality (dots) and discharge (diamonds).

Nitrates Measurements and "ProSe values"

The Interdistrict Federation for Sewage of Greater Paris (SIAAP) provides weekly measurements at nine sites over the Seine river (including Choisy) and one site over the Marne river (Figure 3). The nitrates concentrations are chosen for this exploratory study because of their environmental impact. Their "slowly changing" behaviour is known. Since the "ProSe values" are only available for the year 2003, we first verified on the experimental data (from 2001 to 2003) that the nitrates in 2003 didn't show a particular behaviour.
For these "instantaneous" data, the measurement day is known but not the sampling hour. We thus arbitrarily choose to take the "ProSe values" at noon for

the comparison. It should only introduce a weak deviation, given the high continuity at the origin of the time variogram (Figure 4(a)). At Choisy, the comparison of the weekly measurements and the "ProSe values" at "likely hours of measurement", from 8.00 to 12.00 a.m. and from 1.30 to 5.00 p.m., shows that the daily variability of the ProSe nitrates values stay low compared to the gap between this model and the measurements.

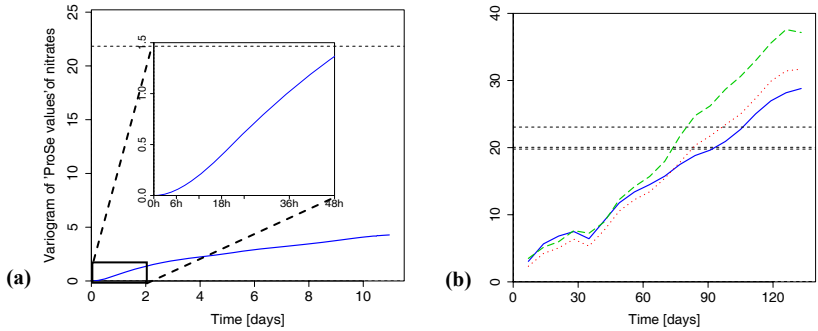


Figure 4: Nitrates time variogram at Chosiy: **(a)** Short time step according to "ProSe values", **(b)** measurements (line) and "ProSe values" (dashed) simple and cross (dotted) variograms.

The superposition of measurements and "Prose values" times series (Figure 5(a)) shows that if the deterministic model correctly reproduces the amplitude of the seasonal variations, with *minima* during the summer and *maxima* during the winter, ProSe tends to overestimate the concentrations. This correlation is confirmed by the scatter diagram between each measurement and the "ProSe values" at "likely hours of measurement" on the same day (Figure 5(b)).

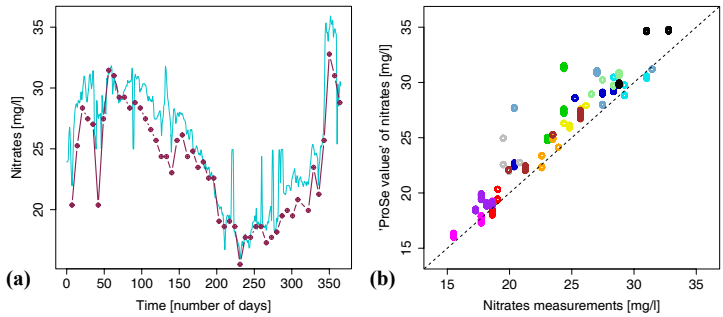


Figure 5: **(a) N**itrates chronicles at Choisy: "ProSe values" (line) and measurements (points). **(b)** Correlation between nitrates measurements and "ProSe values" at "likely hours of measurement".

The model's overestimation can be explained, at least partly, by the different measurement procedures used by Veolia Water (automatic hourly sampling) and SIAAP (manual weekly sampling). Indeed the comparison between the two data sets in Choisy – the only location where both measurements are available – showed greater values for Veolia Water data, used as boundary conditions.

Time Variability

Simple and cross time variograms show that ProSe reproduces the temporal variability of nitrate concentrations, even if a greater variability can be noticed on the cross variogram after 60 days (Figure 4(b)). The cross variogram is very similar to the simple ones, showing a high correlation between the experimental and ProSe time series. The ProSe model can thus be used to study the nitrates temporal evolution in more details.

Figure 6 presents the simple and cross temporal variograms between different measurement sites located on a same river (left) and on different rivers (centre and right). The best correlations are observed for Choisy and Ivry located on the same river, and for Choisy and Alfortville respectively located on the rivers Seine and Marne (Figure 6). The correlation is weaker for Ivry (Seine) and Alfortville (Marne) although they are the closest sites.

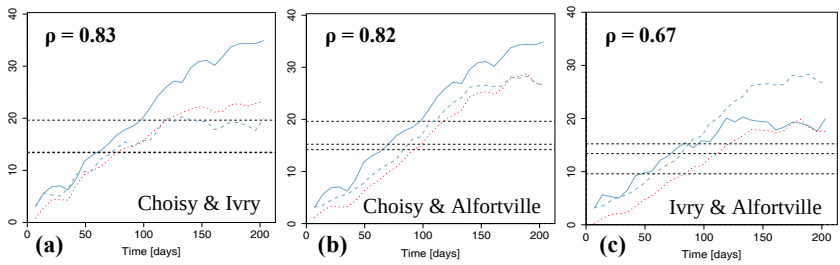


Figure 6: Simple and cross temporal variograms of nitrates measurements at Choisy and Ivry (Seine), and Alfortville (Marne). **(a)** Between sites located on the same segment; **(b)** and **(c)** between sites located on different rivers upstream of their confluence. The correlation coefficient is reported.

Spatial Variability along Edges

The spatial variation of nitrate concentrations seems more problematic. Figure 7 shows the mean *instantaneous* variogram between sampling sites along the stream. Couples of sites located on both sides of a singularity (confluence, treatment plants…) are not taken into account.

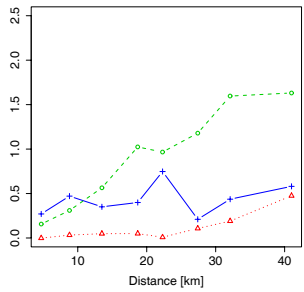


Figure 7: Mean spatial variograms of instantaneous nitrates in 2003 on a river segment. Measurements (line), "ProSe values" (dashed) and cross (dotted) variograms.

While the variogram of "ProSe values" is increasing the spatial structure of measurements seems less clear. The cross variogram is low, indicating a poor spatial correlation between measurements and ProSe concentrations. This is due

to the different sampling techniques between boundary conditions and SIAAP data. Nevertheless the values of spatial variograms are very low compared to time variograms, with a ratio greater than ten. Thus the influence of the sampling techniques on spatial experimental variogram seems to be important.

CONCLUSIONS

Discharges and nitrate concentrations simulated by ProSe are in agreement with the data. A comparison between Veolia Water and SIAAP data could help understanding the model overestimation. However, a detailed variographic analysis on the whole modeled network is therefore possible using the ProSe model as an approximate mock-up to choose consistent classes of RF models. RF models along hydrographic networks become useful in the scope of implementing the Water European Framework directive, to construct estimators for concentrations or loads on "water masses". In addition integrating the phenomenological model will improve the consistency and the precision of the geostatistical model. On another hand, geostatistics could help improving the efficiency of the Prose model by providing co-kriging boundary conditions.

ACKNOWLEDGEMENTS

The authors are very grateful to the SIAAP members for providing the data.

REFERENCES

Bailly, J.-S., Monestiez, P. and Lagacherie, P. (2006) *Exploring Spatial Variability along Drainage Networks with Geostatistics*, Mathematical Geology, 38(5), in press.

Bernard-Michel, C. (2006) *Indicateurs Géostatistiques de la Pollution dans les Cours d'Eau*, Thèse de Doctorat, Ecole des Mines de Paris.

Bruno, R., Palumbo, V. and Bonduà, S. (2001) *Identification of Regional Variability Component by Geostatistical Analysis of Stream Sediments*, Geostatistics for Environmental Applications, III, pp. 113-123.

Chilès, J.-P. and Delfiner, P. (1999) *Geostatistics: Modeling Spatial Uncertainty*, Wiley, New York.

Cressie, N., Frey, J., Harch, B. and Smith, M. (2006) *Spatial Prediction on a River Network*, Journal of Agricultural Biological Environmental Statistics, in press.

Even, S., Poulin, M., Garnier, J., Billen, G., Servais, P., Chesterikoff, A. and Coste, M. (1998) *River Ecosystem Modelling. Applications of the PROSE Model to the Seine river (France)*, Hydrobiologia, 373/374, p. 27–45.

Even, S., Mouchel, J-M., Servais, P., Flipo, N., Poulin, M., Blanc, S., Chabanel, M. and Paffoni, C. (2007) *Modeling the Impacts of Combined Sewer Overflows on the River Seine Water Quality*. Sci Total Environ ;375:140–51. doi:10.1016/j.scitotenv.2006.12.007.

Flipo, N., Even, S., Poulin, M., Tusseau-Vuillemin, M.-H., Améziane, T. and Dauta, A. (2004) *Biogeochemical Modelling at the River Scale: Plankton and Periphyton Dynamics—Grand Morin Case Study, France*, Ecological Modelling, 176, pp. 333–47.

Fouquet C. de, Bernard-Michel C. (2006) *Modèles Géostatistiques de Concentrations ou de Débits le Long des Cours d'Eau.* C.R. Géosciences 338(2006) 307-318.

Monestiez, P., Bailly, J.-S., Lagacherie, P. and Voltz, M. (2005) *Geostatistical Modelling of Spatial Processes On Directed Trees: Application to Fluvisol Extent*, Geoderma, 128(3-4), pp.179-191.

Poulin, M. (2006) *Réalisation de Simulations PROSE Année de Référence 2003*, Technical report Centre de Geosciences – Ecole des Mines de Paris.

Ver Hoef, J.-M., Peterson, E. and Theobald, D. (2006) *Spatial Statistical Models that Use Flow and Stream Distance*, Environmental Ecological Statistics, in press.