MOX-Report No. 32/2022

An object-oriented approach to the analysis of spatial complex data over stream-network domains

Barbi, C.; Menafoglio, A; Secchi, P.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

An object-oriented approach to the analysis of spatial complex data over stream-network domains

Chiara Barbi[1*] , Alessandra Menafoglio[1] and Piercesare Secchi[1]

[1]MOX – Department of Mathematics, Politecnico di Milano, Milano, Italy
[*]chiara.barbi@polimi.it

Abstract

We address the problem of spatial prediction for Hilbert data, when their spatial domain of observation is a river network. The reticular nature of the domain requires to use geostatistical methods based on the concept of Stream Distance, which captures the spatial connectivity of the points in the river induced by the network branching. Within the framework of Object Oriented Spatial Statistics (O2S2), where the data are considered as points of an appropriate (functional) embedding space, we develop a class of functional moving average models based on the Stream Distance. Both the geometry of the data and that of the spatial domain are thus taken into account. A consistent definition of covariance structure is developed, and associated estimators are studied. Through the analysis of the summer water temperature profiles in the Middle Fork River (Idaho, USA), our methodology proved to be effective, both in terms of covariance structure characterization and forecasting performance.

Keywords: Geostatistics, Functional Data Analysis, Stream Distance, Kriging

1 Introduction

The need to analyse and extract useful information from extremely complex and varied data has certainly been a central challenge for the statistical community in recent years. The statistical methods formulated for scalar data are not usable in those –increasingly frequent– contexts in which the data are featured by a high complexity (such as curves, surfaces or images). For this reason, Functional Data Analysis (FDA, Ramsay and Silverman (2005)) and Object Oriented Data Analysis (OODA, Marron and Alonso (2014)) have attracted great interest among researchers and extensive effort has been made in developing functional versions for a wide range of classical statistical methods. Whenever data are georeferenced, however, the complexity of the data is compounded by the need to take into account the dependence between observations induced by their spatial proximity. A relatively large body of literature has recently focused on developing methods

of spatial statistics for general types of data objects, including functional data, distributions and data belonging to Riemannian manifolds. These efforts lie within the domain of Object Oriented Spatial Statistics (O2S2, Menafoglio and Secchi (2017)), a recent system of ideas for the analysis of spatial complex data, founded on a strong geometrical approach to the data analysis. The methods developed so far allow one to model the dependence among data, perform dimensionality reduction, as well as perform prediction at unsampled locations within the domain (Horváth and Kokoszka, 2012; Menafoglio and Secchi, 2017; Mateu and Giraldo, 2021). However, all these methods are focused on Euclidean spatial domains, or on mildly non-Euclidean spatial regions that, locally, admit a Euclidean representation (see Menafoglio et al. (2018, 2021)). As a matter of fact, vast areas of geosciences study random processes which naturally develop over non-Euclidean settings, where the closeness between data locations is naturally expressed through the shortest path (i.e., the geodesic) induced by the physics of the phenomenon. For instance, when studying aquatic variables in a stream network system, the proximity among sites is better represented by the *water distance* which separates the locations, rather than by the Euclidean shortest path, which does not account for the topology and connectivity of the network.

Although relevant for an increasing number of industrial and environmental applications (see, e.g., Menafoglio and Secchi (2019)), working with non-Euclidean spatial domains poses challenges, because the usual parametric families (e.g., spherical, Matérn) for the covariance among observations may be no longer positive semi-definite under a non-Euclidean metric (Curriero, 2006). Nonetheless, in a few cases, it is possible to derive ad-hoc parametric families, which are well-suited to the topology of the domain under study. This is the case of the models for stream networks proposed and extensively studied by Ver Hoef et al. (2006); Peterson et al. (2007); Ver Hoef and Peterson (2010); Peterson and Ver Hoef (2010); Cressie et al. (2006). These models are built upon a moving average construction of Yaglom (1987), and precisely account for the dependence among observations induced by their water distance (named *stream distance*). This approach yields valid covariance models and proper estimation procedures for spatial data, which can be used whenever their domain of reference can be represented as a binary tree – the water flowing from its root to its leafs.

Although these innovative models exhibit an incredible potential, their range of action is still limited to scalar data. As a matter of fact, while sensors typically record relevant variables continuously along time, previous works need to compress this rich set of information into scalar summaries (e.g., the monthly average temperature, the average weekly dissolved oxygen), inevitably leading to a loss of information. The aim of this work is to overcome these limitations, extending the theory of Ver Hoef et al. (2006); Cressie et al. (2006) to general object data, provided that these can be embedded in a (separable) Hilbert space. This setting includes, e.g., the case of functional data (which are typically embedded in the space L^2 of square-integrable functions) as well as that of distributional data (for which the embedding in a Bayes Hilbert space can be used, Van Den Boogaart

et al. (2014)). To the authors' knowledge, the only existing work enabling the analysis of functional data over a stream network is that by Haggarty et al. (2014). Motivated by the clustering analysis of temporal profiles of nitrate concentrations along the River Tweed (Scotland), these authors propose to model the spatial covariance among observations through the valid models of Ver Hoef et al. (2006), by grounding on integral summaries of the functional data. However, even though the framework of Haggarty et al. (2014) uses typical concepts of FDA, it does not provide a characterization of the infinite-dimensional random field generating the data (being the covariance actually based on scalar summaries), and thus only allows for unsupervised (explorative) analyses. As a key element of innovation with respect to existing literature, we here provide a direct construction of a functional moving average process distributed over a stream network, that creates a solid foundation upon which developing a strategy for variographic analysis and estimation of the spatial covariance structure, which can ultimately be used for the scope of spatial prediction.

The remaining part of this work is organized as follows. Section 2 presents a review of the models of Ver Hoef et al. (2006), which is instrumental to the extension of the construction to Hilbert data presented in Section 3. Section 4 proposes estimators for the spatial dependence of the field under stationary and non-stationary assumptions, and presents the associated Kriging predictors. Section 5 discusses two illustrative simulated examples, while Section 6 reports a summary of the supporting simulation studies – included in the Supplementary Material. Finally, Section 7 discusses the application of the proposed methods to a case study dealing with temperature profiles along the Middle Fork river (Idaho, USA).

2 Stream network models for scalar observations

In this Section, a brief review of the models proposed by Ver Hoef et al. (2006) for scalar data distibuited over a stream network is given. The reader is referred to Ver Hoef et al. (2006), Peterson et al. (2007),Ver Hoef and Peterson (2010), Peterson and Ver Hoef (2010) for further details. The stream networks considered in this work are topologically modelled as dendritic networks made up of a finite number of stream segments indexed by $i = 1, 2, \ldots$. To each segment, which can be represented as a line, is associated a unique direction, that is the direction of the water flow. Having assumed the network to be dendritic, there will always be a single most-downstream point, to which from now on we will refer to as *the outlet*. It is therefore possible to define the "upstream distance" for each point in a network as the length of the path (on the network) that connects the point with the outlet.

Let the whole set of stream segment indices be denoted as I. The most downstream location in the $i-th$ segment is denoted as l_i, whereas the most upstream location is u_i. The index set of stream segments upstream of a point s_i belonging to $i \in I$, will be $U_{s_i} \subseteq I$; segment i is excluded from U_{s_i}. Analogously, $D_{s_i} \subseteq I$ is the index

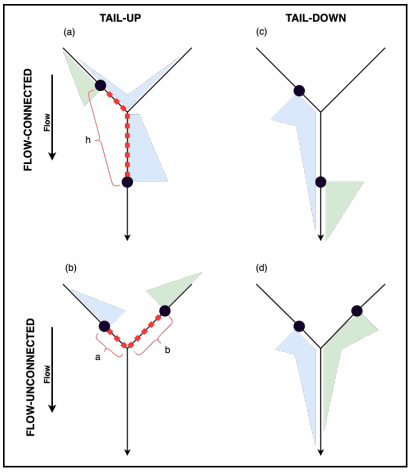

Figure 1. Representation of flow-connected (a,c) and flow-unconnected (b,d) locations, and moving average functions for tail-up (a,b) and tail-down (c,d) models. Locations on the stream network are indicated as black circles; stream-distances between locations are indicated as dashed red lines in (a) and (b). Modified from Peterson and Ver Hoef (2010).

set of all stream segments downstream of s_i, including the segment i containing s_i. Using these definitions, we can say that two locations, s_i and s_j, on a stream network are "flow-connected" (FC) if $D_{s_i} \cap D_{s_j} = D_{s_i}$ or $D_{s_i} \cap D_{s_j} = D_{s_j}$. In other words, water must flow from one site to another in order for the pair to be considered flow connected (see Figure 1a and 1c). In the following, we denote by B_{s_i,s_j} the set of stream segments between two locations s_i, s_j, including the segment for the upstream location but excluding the segment for the downstream location. The same definition holds if we want to identify the segments between location s_i and segment j, for which we will use the notation $B_{s_i,[j]}$.

Given the notation introduced above, it is possible to define the stream distance as the shortest distance between two locations, with the constrain that all displacements are taken along the network.

$$d(s_i, s_j) = \begin{cases} |s_i - s_j| & \text{if } s_i \text{ and } s_j \text{ are flow-connected,} \\ (s_i - u) + (s_j - u) & \text{otherwise.} \end{cases} \quad (1)$$

Here u is the nearest junction downstream which is common to both flow-unconnected locations. Consider now two flow-unconnected locations. Conventionally, we will use a to indicate the shortest distance to u while b indicates the largest one. We use h for the distance between two FC locations (see Figure 1b).

We are now able to enter the core of the models proposed by Ver Hoef et al. (2006). To build the random process $\{Z(s), s \in D\}$ on the stream network domain D, these authors generalize the moving-average construction of Yaglom (1987),

originally designed on $\mathbb{R}^1$, to the topology of D. Yaglom (1987) defines the element $Z(s)$ of a random process on $\mathbb{R}^1$ as

$$Z(s) = \int_{-\infty}^{+\infty} g(x - s|\boldsymbol{\theta}) dW(x) \tag{2}$$

where $W(x)$ is a white noise process and $g(x|\boldsymbol{\theta})$ is called the moving-average (MA) function, which is defined on $\mathbb{R}^1$ and assumed to be squared integrable. To account for the topology of the domain, Ver Hoef et al. (2006) use the same construction, but compute the integral in (2) piece-wise, summing up the contribution from each segment of the network associated with non-null values of the MA function $g(x|\boldsymbol{\theta})$. The key idea is that the overlap between the MA function of one random variable and that of another give rise to a partial correlation between these two variables. Notice that the moving average function could go in both directions, up and down the stream with respect to flow, and this choice will discriminate whether the final model will be a *tail-up* or *tail-down*, respectively.

Moreover, recall that, when $W(x)$ is a Brownian motion, by Ito Isometry it follows that

$$\mathbb{E}\left[\left(\int_{-\infty}^{+\infty} g(x - s|\boldsymbol{\theta}) dW(x)\right)^2\right] = \int_{-\infty}^{+\infty} \left(g(x - s|\boldsymbol{\theta})\right)^2 dx.$$

Hence, from the moving average construction (2) it is possible to obtain the autocovariance between two elements of the field $Z(s)$ and $Z(s + h)$ as

$$C_t(h|\boldsymbol{\theta}) = \begin{cases} \int_{-\infty}^{+\infty} (g(x|\boldsymbol{\theta}))^2 dx + \eta & h = 0 \\ \int_{-\infty}^{+\infty} g(x|\boldsymbol{\theta}) g(x - h|\boldsymbol{\theta}) dx & h > 0, \end{cases} \tag{3}$$

where η is the nugget effect. From this construction, when $D \subset \mathbb{R}^1$, several classes of models can be obtained (e.g., spherical, exponential, Mariah; see Yaglom (1987)). Analogously, parametric classes are obtained by Ver Hoef et al. (2006) by computing the integrals in (3) piece-wise along the stream network. The expressions of the MA model for $Z(s)$ and the corresponding autocovariance functions are recalled hereafter for tail-up and tail-down models respectively.

Tail-up Models In the tail-up models, the support of the moving average functions is not null only moving in the upstream direction (Figure 1a and 1b). Obviously, if two locations are not flow connected, the corresponding tail up moving average will never overlap (Figure 1b), hence null covariance is associated to two flow unconnected random variables. The way that $g(x|\boldsymbol{\theta})$ gets split as we go upstream plays a crucial role to ensure the stationarity of the spatial process. Segment weights ω_k are used to proportionally split the function between upstream segments when the MA function reaches a confluence in the network, eventually obtaining the following expression for the element $Z(s_i)$

$$Z(s_i) = \int_{s_i}^{u_i} g(x - s_i|\boldsymbol{\theta}) dW(x) + \sum_{j \in U_{s_i}} \left(\prod_{k \in B_{s_i,[j]}} \sqrt{\omega_k} \right) \int_{l_j}^{u_j} g(x - s_i|\boldsymbol{\theta}) dW(x). \tag{4}$$

In (4), for each segment i in the network, the weights associated to the two segments j and k in which segment i splits are such that $0 \leq \omega_j, \omega_k \leq 1$ and $\omega_j + \omega_k = 1$. Note that the weights ω_k may be chosen as to reflect specific hydrological characteristics of each segment, such as discharge, watershed area or flow volume (Ver Hoef et al., 2006).

The covariance between two random elements $Z(s_i)$, $Z(s_j)$ defined by (4) is then given by

$$C(s_i, s_j|\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } s_i \text{ and } s_j \text{ are not flow connected} \\ C_t(0|\boldsymbol{\theta}) & \text{if } s_i = s_j \\ \pi_{i,j} C_t(h|\boldsymbol{\theta}) & \text{otherwise.} \end{cases} \tag{5}$$

where $\pi_{i,j} = \prod_{k \in B_{s_i,s_j}} \sqrt{\omega_k}$, h is the stream distance between the two flow connected locations on the stream network, and the (unweighted) covariance functions $C_t(0|\boldsymbol{\theta})$ are obtained by using moving average functions in one dimension without any branching given in (3). Therefore $C_t(0|\boldsymbol{\theta})$ may share the same expression of the valid covariance models in $\mathbb{R}^1$ derived by Yaglom (1987). To conclude, autocorrelation in the tail-up models depends on flow-connected hydrologic distance, on the number of confluences found in the path between two sites and finally on the weights assigned to each segment. Imposing zero autocorrelation when sites are not flow connected makes these models particularly appropriate when the variable of interest is dominated by flow (e.g. organisms or materials that move passively downstream like pollutants, waterborne chemicals and so on).

Tail-down models In contrast to the tail-up models, tail-down (TD) models arise when the MA function is non-zero only downstream of a location. This means that the "tail" of the moving average functions points in the flow direction (Figure 1c and 1d). Therefore, the tail-down random variable has the following expression:

$$Z(s) = \int_{-\infty}^{s} g(s - x|\boldsymbol{\theta}) dW(x). \tag{6}$$

As shown in Figure 1c and 1d and by direct computations, autocorrelation in tail-down models is allowed both for flow-connected and flow-unconnected locations. Moreover, since the MA functions do not split at the junctions, introducing a weighting procedure is not needed anymore. As before, more overlap in the MA function implies more autocorrelation. Some examples of the tail-down covariance structures are given in Table 1. Due to their characteristic of allowing correlation for both connected and not connected pairs of sites, tail-down models are particularly indicated for modeling variables, such as fish or aquatic insects, that can move both upstream and downstream.

Table 1. Covariograms and Semivariograms for tail-up and tail-down models. $\theta_r, \theta_v \in \mathbb{R}^+$ are respectively the range and the sill parameters. Recall that b denotes the longest of the distances to the common downstream junction, and a denotes the shortest one; h is the total stream distance (see Figure 1). The notation FC and FU is used to denote respectively that s_i and s_j are flow-connected or flow-unconnected.

Name	Covariogram	Semivariogram
Tail-up Linear with Sill	$\mathrm{C}(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} \theta_v & \text{if } s_i = s_j \\ \pi_{i,j}\theta_v\left(1-\frac{h}{\theta_r}\right)\mathbb{1}_{(\frac{h}{\theta_r}\leq 1)} & \text{if FC} \\ 0 & \text{if FU.} \end{cases}$	$\gamma(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } s_i = s_j \\ \theta_v - \pi_{i,j}\theta_v\left(1-\frac{h}{\theta_r}\right) & \text{if FC and } h \leq \theta_r \\ \theta_v & \text{if } h > \theta_r \text{ or if FU.} \end{cases}$
Tail-up Spherical	$\mathrm{C}(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} \theta_v & \text{if } s_i = s_j \\ \pi_{i,j}\theta_v\left(1-\frac{3}{2}\frac{h}{\theta_r}+\frac{1}{2}\frac{h^3}{(\theta_r)^3}\right)\mathbb{1}_{(\frac{h}{\theta_r}\leq 1)} & \text{if FC} \\ 0 & \text{if FU.} \end{cases}$	$\gamma(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } h = 0 \\ \theta_v - \pi_{i,j}\theta_v\left(1-\frac{3}{2}\frac{h}{\theta_r}+\frac{1}{2}\frac{h^3}{(\theta_r)^3}\right) & \text{if FC and } h \leq \theta_r \\ \theta_v & \text{if } h > \theta_r \text{ or if FU.} \end{cases}$
Tail-up Exponential	$\mathrm{C}(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} \theta_v & \text{if } s_i = s_j \\ \pi_{i,j}\theta_v exp(-h/\theta_r) & \text{if FC} \\ 0 & \text{if FU.} \end{cases}$	$\gamma(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } h = 0 \\ \theta_v - \pi_{i,j}\theta_v(exp(-h/\theta_r)) & \text{if FC} \\ \theta_v & \text{if FU.} \end{cases}$
Tail-up Mariah	$\mathrm{C}(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} \theta_v & \text{if } s_i = s_j \\ \pi_{i,j}\theta_v\left(\frac{log(h/\theta_r+1)}{h/\theta_r}\right)\mathbb{1}_{(h>0)} & \text{if FC} \\ 0 & \text{if FU.} \end{cases}$	$\gamma(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } h = 0 \\ \theta_v - \pi_{i,j}\theta_v\left(\frac{log(h/\theta_r+1)}{h/\theta_r}\right) & \text{if } s_i \text{ and } s_j \text{ are FC} \\ \theta_v & \text{if } s_i \text{ and } s_j \text{ are FU.} \end{cases}$
Tail-down Linear with Sill	$\mathrm{C}_d(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} \theta_v\left(1-\frac{h}{\theta_r}\right)\mathbb{1}_{(\frac{h}{\theta_r}\leq 1)} & \text{if FC} \\ \theta_v\left(1-\frac{b}{\theta_r}\right)\mathbb{1}_{(\frac{b}{\theta_r}\leq 1)} & \text{if FU.} \end{cases}$	$\gamma(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } h = 0 \\ \theta_v\frac{h}{\theta_r} & \text{if FC and } h \leq \theta_r \\ \theta_v\frac{b}{\theta_r} & \text{if } b < \theta_r \text{ and if FU} \\ \theta_v & \text{otherwise.} \end{cases}$
Tail-down Spherical	$\mathrm{C}_d(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} \theta_v\left(1-\frac{3}{2}\frac{h}{\theta_r}+\frac{1}{2}\frac{h^3}{(\theta_r)^3}\right)\mathbb{1}_{(\frac{h}{\theta_r}\leq 1)} & \text{if FC} \\ \theta_v\left(1-\frac{3}{2}\frac{a}{\theta_r}+\frac{1}{2}\frac{b}{\theta_r}\right)\left(1-\frac{b}{\theta_r}\right)^2\mathbb{1}_{(\frac{b}{\theta_r}\leq 1)} & \text{if FU} \end{cases}$	$\gamma(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } h = 0 \\ \theta_v\left(\frac{3}{2}\frac{h}{\theta_r}-\frac{1}{2}\frac{h^3}{(\theta_r)^3}\right) & \text{if FC and } h \leq \theta_r \\ \theta_v - \theta_v\left(1-\frac{3}{2}\frac{a}{\theta_r}+\frac{1}{2}\frac{b}{\theta_r}\right)\left(1-\frac{b}{\theta_r}\right)^2 & \text{if FU and } b < \theta_r \\ \theta_v & \text{otherwise.} \end{cases}$
Tail-down Exponential	$\mathrm{C}_d(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} \theta_v exp(-h/\theta_r) & \text{if FC} \\ \theta_v exp(-(a+b)/\theta_r) & \text{if FU} \end{cases}$	$\gamma(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } h = 0 \\ \theta_v(1-exp(-h/\theta_r)) & \text{if FC} \\ \theta_v(1-exp(-(a+b)/\theta_r)) & \text{if FU.} \end{cases}$
Tail-down Mariah	$\mathrm{C}_d(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} \theta_v\left(\frac{log(h/\theta_r+1)}{h/\theta_r}\right) & \text{if FC, } h > 0 \\ \theta_v & \text{if FC, } h = 0 \\ \theta_v\left(\frac{log(a/\theta_r+1)-log(b/\theta_r+1)}{(a-b)/\theta_r}\right) & \text{if FU, } a \neq b \\ \theta_v\left(\frac{1}{a/\theta_r+1}\right) & \text{if FU, } a = b \end{cases}$	$\gamma(s_i,s_j\|\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } h = 0 \\ \theta_v\left(1-\frac{log(h/\theta_r+1)}{h/\theta_r}\right) & \text{if FC, } h > 0 \\ \theta_v\left\langle 1-\frac{log(a/\theta_r+1)-log(b/\theta_r+1)}{(a-b)/\theta_r}\right\rangle & \text{if FU, } a \neq b \\ \theta_v\left\langle 1-\frac{1}{a/\theta_r+1}\right\rangle & \text{if FU, } a = b. \end{cases}$

3 Functional random fields over stream network domains

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and H a separable Hilbert space, equipped with operations $(+, \cdot)$ and inner product $\langle \cdot, \cdot \rangle$, inducing the norm $\|\cdot\|$. Following Menafoglio et al. (2013), we consider the case of real-valued functional observations and assume that each element of H is a function $\mathcal{X} : \tau \to \mathbb{R}$, τ being a compact subset of $\mathbb{R}$. Denote by D the spatial domain, and let

$$\{\mathcal{X}_s, s \in D \subseteq \mathbb{R}^d\} \tag{7}$$

be a functional random field valued in H. The theory of random processes on Hilbert spaces is well established when D is a subset of $\mathbb{R}^d$ (see, e.g., Bosq, 2000); in the following, we elaborate on the case of D being a stream network domain, defining the field (7) by direct construction. In this work, we will always assume the square-integrability of the process, i.e., that each element $\mathcal{X}_s$, $s \in D$, of the random field is such $\mathbb{E}[\|\mathcal{X}_s\|^2] < \infty$; we denote this as $\mathcal{X}_s \in L^2(\Omega; H)$. As in the usual setting of geostatistics, we consider that a partial observation of the field is available at given (non-random) spatial locations $s_1, ..., s_n$ in D, and denote the functional dataset as $\mathcal{X}_{s_1}, ..., \mathcal{X}_{s_n}$.

Following Bosq (2000), for any s_1, s_2 in D, we define the cross-covariance operator between the elements $\mathcal{X}_{s_1}$ and $\mathcal{X}_{s_2}$ of (7) as the operator $\mathcal{C}_{s_1,s_2} : H \to H$ acting on the (non-random) element $x \in H$ as

$$\mathcal{C}_{s_1,s_2} x = \mathbb{E}[\langle \mathcal{X}_{s_1} - m_{s_1}, x \rangle (\mathcal{X}_{s_2} - m_{s_2})]$$

with m_{s_1} (m_{s_2}) the mean of the process in s_1 (s_2). The family of cross-covariance operators $\{\mathcal{C}_{s_1,s_2}, s_1, s_2 \in D\}$ fully defines the second-order properties of the field (Bosq (2000), Kokoszka and Horváth (2012)). A *(global)* measure of dependence for the process (7) is instead provided by the so-called *trace-covariogram* (Giraldo, 2009; Menafoglio et al., 2013). This is defined as the (real-valued) function $C : D \times D \to \mathbb{R}$:

$$C(s_1, s_2) = \mathbb{E}\big[\langle \mathcal{X}_{s_1} - m_{s_1}, \mathcal{X}_{s_2} - m_{s_2} \rangle\big]. \tag{8}$$

Note that $C(s_1, s_2)$ defines a scalar product on $L^2(\Omega; H)$ and it is positive definite. Moreover, $C(s_1, s_2)$ coincides with the trace of the cross-covariance operator $\mathcal{C}_{s_1,s_2}$ (Menafoglio et al. (2013)).

Recall also that the field (7) is *second-order stationary* if (i) the mean is spatially constant ($\mathbb{E}[\mathcal{X}_s] = m$ for all $s \in D$), and (ii) the family of cross-covariance operators is stationary, i.e., if there exist a family of operators $\{\mathcal{C}_h, h \in \mathbb{R}^d\}$ such that $\mathcal{C}_{s_1,s_2} = \mathcal{C}_h$ for all s_1, s_2 satisfying $s_1 - s_2 = h$. The assumption of *global* second-order stationarity requires, instead of condition (ii), that (ii') the trace-covariogram is stationary, i.e., that there exist a function $\tilde{C}$ such that $\tilde{C}(h) = C(s_1, s_2)$ for all s_1, s_2 satisfying $s_1 - s_2 = h$.

Hörmann and Kokoszka (2011) show that every functional random process (7) with constant mean can be expressed through the following basis expansion

$$\mathcal{X}_s = m + \sum_{k\geq 1} \xi_k(s) e_k. \tag{9}$$

Here $\{e_k, k \geq 1\}$ is an orthonormal basis of H and the random coefficients $\xi_k(s) = \langle \mathcal{X}_s - m, e_k \rangle$ are the projections of the functional random variable $\mathcal{X}_s$ on the orthonormal basis. These coefficients determine both the stationarity and the covariance structure of the functional process. To ease the notation, we herafter assume the process to be zero mean.

3.1 Functional moving-average models on the real line

We now use the direct construction (9) to show the existence of a functional version of the MA random variables defined in (2). We first set $D = \mathbb{R}^1$, and consider N independent, zero mean, second-order stationary and isotropic scalar random fields, $\{\xi_k(s), s \in D\}$ for $k = 1, .., N$. We further assume that each scalar random field is defined through a MA model

$$\xi_k(s) = \int_{-\infty}^{+\infty} g^{(k)}(x - s|\boldsymbol{\theta}) dW_k(x), \tag{10}$$

In (10), each $g^{(k)}(x - s|\boldsymbol{\theta})$ needs to be square integrable for the stochastic integral to be well defined, i.e., $\int_{-\infty}^{+\infty} |g^{(k)}(x - s|\boldsymbol{\theta})|^2 dx < +\infty$.

Let us now focus on a truncated version of (9), obtained as

$$\mathcal{X}_s^{(N)} = \sum_{k=1}^{N} \xi_k(s) e_k, \tag{11}$$

where $\{e_k, k \geq 1\}$ is an orthonormal basis of H. In this case, each $\mathcal{X}_s$ is valued in $H^{(N)}$, where $H^{(N)} = span\{e_1, ..., e_N\}$ is the finite-dimensional Hilbert space generated by the N orthonormal vectors $e_1, ..., e_N$. Moreover, $\mathcal{X}_s$ is square-integrable (i.e., $\mathcal{X}_s \in L^2(\Omega; H^{(N)})$) as

$$\begin{aligned} \mathbb{E}\left[\left\|\mathcal{X}_s^{(N)}\right\|^2\right] &= \mathbb{E}\left[\sum_{k=1}^{N}\left(\int_{-\infty}^{+\infty} g^{(k)}(x - s|\boldsymbol{\theta}) dW_k(x)\right)^2\right] \\ &= \sum_{k=1}^{N} \int_{-\infty}^{+\infty} \left(g^{(k)}(x - s|\boldsymbol{\theta})\right)^2 dx < +\infty, \end{aligned} \tag{12}$$

thanks to the Ito isometry and to the fact that each g is deterministic and square integrable. Hence, the variable $\mathcal{X}_s^{(N)}$ has finite second moment, which guarantees the existence of the family of cross-covariance operators for the process $\{\mathcal{X}_s^{(N)}, s \in D\}$. It is worth highlighting that the boundedness of the last sum in (12) is due

to the finiteness of the orthonormal basis being considered. Letting $N \to +\infty$, the square-integrability of $\mathcal{X}_s$ is only obtained if the sequence $\{\xi_k(s)\}_{k\geq 1}$ belongs to $l^2(\Omega;\mathbb{R})$ (i.e., if $\sum_{k\geq 1}\mathbb{E}[\xi_k(s)^2] < \infty$). This can be guaranteed including additional assumptions on each moving average function $g(x|\boldsymbol{\theta})$ (see Appendix A). In this case, the MA random field (9) has a well-defined family of cross-covariance operators.

Note that the covariance functions $C_k(s_1, s_2) = \mathbb{E}[\xi_k(s_2)\xi_k(s_2)]$ of the scalar random fields appearing in (9) completely characterize the family of cross-covariance operators of the field $\{\mathcal{X}_s, s \in D\}$, thus also its trace-covariogram (see, e.g., Hörmann and Kokoszka, 2011; Menafoglio et al., 2013). In particular, the trace-covariogram of the process (9), obtained from the moving average construction, is

$$C(s_i, s_j) = \sum_{k=1}^{N} \mathbb{E}\Big[\xi_k(s_i)\xi_k(s_j)\Big] = \sum_{k=1}^{N} C_t^{(k)}(h|\boldsymbol{\theta}).$$

where $C_t^{(k)}(h|\boldsymbol{\theta})$ is the autocovariance function of the k-th scalar random field $\{\xi_k(s), s \in D\}$ (defined as in (3)). Since the family of valid covariograms is a convex cone and each $C_t^{(k)}(h|\boldsymbol{\theta})$ is a valid covariogram for the k-th random field, $C(s_i, s_j)$ is clearly a valid covariance function. On the other hand, any of the valid covariance models available in D can be used to provide a valid covariance model for the field $\{\mathcal{X}_s, s \in D\}$.

3.2 Functional tail-up and tail-down models

The approach just introduced can be extended to a stream network domain D. Indeed, the previous arguments still hold true if we assume that each scalar random field is represented as a tail-up model, i.e., as (see also eq. (4))

$$\begin{aligned}\xi_k(s) = &\int_s^{u_i} g^{(k)}(x - s|\boldsymbol{\theta})dW(x) + \\ &+ \sum_{j\in U_s}\left(\prod_{n\in B_{s,[j]}}\sqrt{\omega_n}\right)\int_{l_j}^{u_j} g^{(k)}(x - s|\boldsymbol{\theta})dW(x).\end{aligned} \tag{13}$$

In this case, when the functional random process is built by direct construction as in (9), the covariance function associated with each random field ξ_k is a combination of the covariance functions of the scalar field defined in (5). The trace-covariogram of the functional process is then easily obtained by linearity as

$$C(s_i, s_j) = \begin{cases} 0 & \text{if } s_i \text{ and } s_j \text{ are not flow connected} \\ \sum_{k\geq 1} C_t^{(k)}(0|\boldsymbol{\theta}) & \text{if } s_i = s_j \\ \pi_{i,j}\left(\sum_{k\geq 1} C_t^{(k)}(h|\boldsymbol{\theta})\right) & \text{otherwise.} \end{cases} \tag{14}$$

In (14), we adopt the same weighting structure as the scalar case, because the weights $\pi_{i,j}$ are related to the geometry of the stream-network domain rather than that of the data objects.

Concerning the tail-down models, the procedure is even simpler, since in this case the weights are not needed. Each scalar random field is obtained as (see (6))

$$\xi_k(s) = \int_{-\infty}^{s} g^{(k)}(s - x|\boldsymbol{\theta})dW(x) \tag{15}$$

where $g^{(k)}(s - x|\boldsymbol{\theta})$ is a unilateral tail-down function with nonzero values only on the negative (i.e., downstream) side of s as in the tail-down model introduced in Section 2. Similarly as in the tail-up case, the trace-covariogram function for the functional tail-down process is straightforwardly obtained as

$$C(s_i, s_j) = \sum_{k\geq 1} C_d^{(k)}(s_i, s_j|\boldsymbol{\theta}), \tag{16}$$

where $C_d^{(k)}(s_i, s_j|\boldsymbol{\theta})$ is the covariance function associated to the k-th scalar tail-down random field, whose expression may be of the kind presented in Table 1. In the light of expressions (14) and (16), one may wonder whether (and under which conditions) the families $C_t^{(k)}(h|\boldsymbol{\theta})$ and $C_d^{(k)}(s_i, s_j|\boldsymbol{\theta})$ are closed under conic combinations, i.e., if (and when) linear combinations, with positive weights, of valid covariance functions in the same parametric family still belong to the same family. If this was the case, the trace-covariogram of the functional process built in (9) would belong to the same family as those of the 1D processes $\{\xi_k(s), s \in D\}, k = 1, \ldots, N$. Concerning $C_t^{(k)}(h|\boldsymbol{\theta})$, it is well-known from scalar geostatistics that finite conic combinations of valid models are closed if and only if they belong to the same family and share the same range parameter. The same applies to $C_d^{(k)}(s_i, s_j|\boldsymbol{\theta})$, as can be straightforwardly derived from the expressions in Table 1. As such, if the scalar fields $\{\xi_k(s), s \in D\}$, $k = 1, ..., N$, share the same valid model and the same range parameter, the trace-covariogram of the functional process (9) will belong to the same family and share the same range as the scalar fields, but will have a sill equal to the sum of the sills of the scalar fields.

4 Model estimation and spatial prediction

The scope of this Section is to propose and discuss methods to estimate the models introduced in the previous section, both under stationary and non-stationary conditions. In the stationary case, model estimation typically reduces to estimating the covariance structure of the field. In the non-stationary case, a drift term generally needs to be estimated as well.

4.1 Estimation of the spatial covariance under stationarity

In this work, we will estimate the parameters of the covariance models proposed in Section 3 by estimating the trace-semivariogram of the field, which is defined, under global second-order stationarity as

$$\gamma(s_1, s_2) = \frac{1}{2}\mathbb{E}[\|\mathcal{X}_{s_1} - \mathcal{X}_{s_2}\|^2]$$

and is related with the trace-covariogram through the well-known relation $\gamma(s_1, s_2) = C(s_1, s_1) - C(s_1, s_2)$ (see, e.g., Menafoglio et al., 2013).

As in scalar geostatistics, estimation of the trace-semivariogram can be performed by first determining an empirical estimator and then fitting a valid model. As discussed in Section 3, all the valid models in use in the scalar case can be adopted in the functional case too; for convenience, the semivariogram models derived by Ver Hoef et al. (2006) are reported in Table 1. The semivariograms in Table 1 are defined piecewise, depending on the connectedness of the pair being considered. From now on, the portion of a semivariogram associated to flow-connected (flow-unconnected) locations will be denoted as "the flow-connected (flow-unconnected) portion of the semivariogram".

4.1.1 Empirical Semivariograms for Stream Networks

From Section 2 and the expressions in Table 1, it should be clear that, for both tail-up and tail-down models, the covariance structure among observations does not depend only on the stream distance but also on other characteristics such as flow connectedness, weights attributed to the stream segments, and/or distances to a common junction. This dependence – which motivates the use of the notation $\gamma(s_i, s_j|\boldsymbol{\theta})$ instead of $\gamma(s_i - s_j|\boldsymbol{\theta})$ – highlights the inadequacy in this context of the empirical semivariogram proposed in the Euclidean setting. In the functional case, the empirical estimator of the trace-semivariogram from data $\mathcal{X}_{s_1}, ..., \mathcal{X}_{s_n}$ observed over a Euclidean domain would read (see, e.g. Giraldo, 2009; Menafoglio et al., 2013)

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(s_i,s_j)\in N(h)} \left\|\mathcal{X}_{s_i} - \mathcal{X}_{s_j}\right\|^2 \tag{17}$$

where $N(h) = \{(s_i, s_j) : s_i - s_j \approx h\}$ and $|N(h)|$ is its cardinality.

In the scalar setting, Zimmerman and Ver Hoef (2017) propose and discuss modifications of the (scalar) empirical estimator to deal with stream networks and stream distances. They thus derive the *flow-unconnected stream-distance* (FUSD) semivariogram and the *flow-connected stream-distance* (FCSD) semivariogram, able to deal both with the peculiar topology of a stream network and with the stream distance. Following the approach of Zimmerman and Ver Hoef (2017), we here study functional counterparts of these estimators, eventually aiming to fit the parameters of a valid model to the most appropriate empirical estimator.

Flow-Unconnected Stream-Distance (FUSD) Trace-Semivariogram The FUSD empirical trace-semivariogram is computed only from those site-pairs that are flow-unconnected and, for such pairs, it is a function of the stream distance only. The FUSD trace-semivariogram is thus defined as

$$\hat{\gamma}_{FUSD}(h_k) = \frac{1}{2|N(\mathcal{U}_k)|} \sum_{(s_i,s_j)\in N(\mathcal{U}_k)} \left\|\mathcal{X}_{s_i} - \mathcal{X}_{s_j}\right\|^2, \qquad k = 1, ..., K_{\mathcal{U}}, \tag{18}$$

where $N(\mathcal{U}_k) = \{(s_i, s_j) : d(s_i, s_j) \approx h_k, U_{s_i} \cap U_{s_j} = \varnothing\}$ is the set of flow-unconnected pairs separated by a stream-distance approximately equal to h_k, and $|N(\mathcal{U}_k)|$ is its cardinality. Note that, if $\{\mathcal{X}_s, s \in D\}$ follows a pure tail-up model, the flow-unconnected portion of its semivariogram is constant and corresponds to the sill, as the variables associated to flow-unconnetted pairs are uncorrelated (see Table 1). In this case, $\hat{\gamma}_{FUSD}$ is an unbiased estimator for the flow-unconnected portion of the trace-semivariogram and an estimate of the sill is obtained as

$$\bar{\gamma}_{FUSD} = \frac{\sum_{k=1}^{K_{\mathcal{U}}} |N(\mathcal{U}_k)| \hat{\gamma}_{FUSD}(h_k)}{\sum_{k=1}^{K_{\mathcal{U}}} |N(\mathcal{U}_k)|}, \tag{19}$$

with $K_{\mathcal{U}}$ the number of bins in which the set of stream distances is partitioned. An alternative way of estimating the sill, related with the FCSD trace-semivariogram, is discussed in the following. On the other hand, if $\{\mathcal{X}_s, s \in D\}$ follows a pure tail-down model, the flow-unconnected portion of its semivariogram in general does not depend on the total stream distance (i.e., $h = a+b$, see Fig. 1) but on the two stream distances from sites within a site-pair to their common junction (i.e., a and b, see Fig. 1). Therefore, in this case, the FUSD empirical semivariogram is not enough to characterize the spatial dependence of flow-unconnected sites. It is worth noticing that an exception occurs if the tail-down component has an exponential semivariogram; indeed, in this case the flow-unconnected part of the semivariogram is a function of the total stream distance only; consequently, $\hat{\gamma}_{FUSD}(\cdot)$ remains unbiased for it.

Flow-Connected Stream-Distance (FCSD) Trace-Semivariogram The FCSD trace-semivariogram differs from the FUSD trace-semivariogram by being computed from site-pairs that are flow-connected rather than flow-unconnected. Thus, it is defined as

$$\hat{\gamma}_{FCSD}(h_k) = \frac{1}{2|N(\mathcal{C}_k)|} \sum_{(s_i,s_j)\in N(\mathcal{C}_k)} \left\|\mathcal{X}_{s_i} - \mathcal{X}_{s_j}\right\|^2, \qquad k = 1, ..., K_{\mathcal{C}}, \tag{20}$$

where $N(\mathcal{C}_k) = \{(s_i, s_j) : d(s_i, s_j) \approx h_k, U_{s_i} \cap U_{s_j} \neq \varnothing\}$, is the set of flow-connected pairs separated by a stream distance approximately h_k, and $|N(\mathcal{C}_k)|$ is its cardinality.

Note that now, if $\{\mathcal{X}_s, s \in D\}$ follows a pure tail-down model, the flow connected portion of its semivariogram is a function of the stream distance h

between locations. In this case, the well-known valid models in use in scalar geostatistics can be used for parametric modelling (see Table 1). Moreover, similarly as for the scalar case (Zimmerman and Ver Hoef, 2017), $\hat{\gamma}_{FCSD}(\cdot)$ is an unbiased estimator for the flow-connected portion of the trace-semivariogram of the process.

On the other hand, if $\{\mathcal{X}_s, s \in D\}$ is a pure tail-up process, $\hat{\gamma}_{FCSD}(\cdot)$ is not fully appropriate to estimate its covariance structure, because in those cases the flow-connected portion of the trace-semivariogram is a function of the stream distance and of the spatial weights (see Table 1). In Subsec. 4.1.2 we will further expand on this point.

4.1.2 Parameters estimation

In this Section we outline the steps to estimate a parametric model for the trace-semivariogram whenever the underlying process that generated the data is assumed to follow either a *pure tail-up* or a *pure tail-down* model. We will thus focus on the characterization of the covariance structure only in case of pure models (tail-up or tail-down). However, note that, in general, mixtures of tail-up and tail-down models may arise and these are, in principle, more flexible to describe the spatial dependence. The extension of the proposed procedure to the case of mixture models will be discussed in Section 8.

The identification of the underlying process – based on empirical trace-semivariograms – is a rather hard task. For this purpose, we propose a slightly simplified version of the strategy of Zimmerman and Ver Hoef (2017), made of two steps. The FUSD trace-semivariogram is examined first. If it appears to be relatively flat, we adopt a pure tail-up model. Otherwise, a tail-down model is assumed.

We recall that, in the scalar setting, if a pure tail-up model cannot be assumed, the strategy of Zimmerman and Ver Hoef (2017) would advocate a further inspection to discriminate whether the model should be a pure tail-down or a mixture of tail-up and tail-down. In this framework, Liu (2019) proposed non-parametric tests for pure tail-down and pure tail-up dependence on stream networks. The extension of this type of tests to the case of functional data is left for future research. Nevertheless, the problem of mixed models identification was not addressed by Liu (2019) either. With this premise in mind, we propose the following procedure to estimate a valid trace-semivariogram model $\gamma(\cdot, \cdot|\boldsymbol{\theta})$, and the associated trace-covariogram $C(\cdot, \cdot|\boldsymbol{\theta})$.

(i) Estimate the empirical FCSD trace-semivariogram $\hat{\gamma}_{FCSD}(h_k)$, $k = 1, ..., K_{\mathcal{C}}$. from the observations $x_{s_1}, ..., x_{s_n}$ using (20).

(ii) Estimate the empirical FUSD trace-semivariogram $\hat{\gamma}_{FUSD}(h_k)$, $k = 1, ..., K_{\mathcal{U}}$. from the observations $x_{s_1}, ..., x_{s_n}$ using (18).

a. If $\hat{\gamma}_{FUSD}(h_k)$, $k = 1, ..., K_{\mathcal{U}}$, is compatible with a pure nugget model, assume the process to be pure tail-up.

b. If $\hat{\gamma}_{FUSD}(h_k)$, $k = 1, ..., K_{\mathcal{U}}$, is not compatible with a pure nugget model, assume the process to be pure tail-down.

(iii) Fit a valid model $\gamma(\cdot, \cdot|\boldsymbol{\theta})$ to the empirical FCSD trace-semivariogram $\hat{\gamma}_{FCSD}(h_k)$ and get $\hat{\boldsymbol{\theta}}$.

(iv) Obtain the trace-covariogram as $C(\cdot, \cdot|\hat{\boldsymbol{\theta}})$ plugging $\hat{\boldsymbol{\theta}}$ in the stream-network trace-covariograms expressions in Table 1.

Whenever the underlying process can be assumed to follow a pure tail-down model (see point (ii) above) the outlined procedure can be applied without hindrance. This follows from the fact that the flow-connected portion of the tail-down trace-semivariograms in Table 1 has exactly the same expression as the classical models. In case of a pure tail-up model, instead, the approach presents limitations, because the FCSD trace-semivariogram does not account for the spatial weights $\pi_{i,j}$ in (14) (see also Sect. 4.1.1). Note that step (iii). consists of fitting a standard valid model, and neglects the weights $\pi_{i,j}$ within the variograms of Table 1. These classical geostatistical models are recovered when including weights $\pi_{i,j} = 1$ in the expressions of the theoretical semivariogram in Table 1; in the following the semivariogram associated with weights $\pi_{i,j} = 1$ will be denoted as *unweighted flow-connected semivariogram*. In fact, the empirical trace-semivariogram (20) is used in step (iii). precisely as an estimator of the unweighted flow-connected semivariogram. Nevertheless, the FCSD semivariogram is a biased estimator for the unweighted flow-connected semivariogram (see Appendix B). As shown in the simulation study presented in the Supplementary Material (Barbi et al. (2022)) this bias may adversely affect the analyst's ability to correctly determine the range of spatial dependence among flow connected sites (i.e., the range estimates tend to be negatively biased). A similar problem is discussed by Zimmerman and Ver Hoef (2017), who eventually proposed an adjusted empirical estimator (FCWA), which accounts for the weights and is unbiased for the unweighted flow-connected semivariogram. A modification of the empirical trace-semivariogram (20) that follows the same line of Zimmerman and Ver Hoef (2017) (named FCWA2), is developed and studied via simulation in the Supplementary Material (Barbi et al. (2022)). These developments are not reported in the outlined procedure because, despite their unbiasedness, the adjusted estimators proved to be characterized by extremely high variance, hindering their use in practice (see Sect. 6 for a summary of the simulation results). The same simulations show that the range underestimation does not heavily affect the Kriging performances, thus suggesting that the use of FCSD trace-semivariogram should be anyway preferred to its adjusted (unbiased) version.

4.2 Estimating the spatial dependence in the non-stationary case

The methods developed so far assume second-order stationarity. If stationarity does not hold, we propose to use the non-stationary model of Menafoglio et al. (2013), which decouples the elements of the random field $\{\mathcal{X}_s : s \in D\}$ in a non-stationary mean term m_s (the drift) and a zero-mean stationary residual component δ_s, i.e.,

$$\mathcal{X}_s = m_s + \delta_s. \tag{21}$$

In the following, we use for the stochastic residual δ_s the models built in Section 3, and denote by $C(s_1, s_2|\boldsymbol{\theta})$, $\gamma(s_1, s_2|\boldsymbol{\theta})$ the trace-covariogram and trace-semivariogram of δ_s, respectively. A model for the drift term is needed to allow for the estimate of $C(\cdot,\cdot|\boldsymbol{\theta})$, $\gamma(\cdot,\cdot|\boldsymbol{\theta})$, as these are assessed from the (estimated) residuals. We consider a linear model, i.e.,

$$\mathcal{X}_s = \sum_{l=0}^{L} a_l f_l(s) + \delta_s \quad s \in D, \tag{22}$$

where $f_0(\boldsymbol{s}) = 1$ for all $s \in D$, $f_l(\cdot), l = 1, \ldots, L$, are known functions of the spatial variable $s \in D$ and $a_l(\cdot) \in H, l = 0, \ldots, L$, are functional coefficients independent of the spatial location.

Estimation of the linear model (22) can follow the very same lines as in the case of a Euclidean spatial domain, broadly discussed by Menafoglio et al. (2013). These authors propose a generalized least-squares (GLS) estimator for the coefficients a_l, based on the covariance matrix of the residuals $\delta_{s_1}, ..., \delta_{s_n}$. The very same procedure can be used in our setting, provided that the covariance matrix Σ is interpreted in terms of the stream-network trace-covariogram $C(\cdot,\cdot|\boldsymbol{\theta})$, i.e., $\Sigma_{i,j} = C(s_i, s_j|\boldsymbol{\theta})$. For brevity of exposition, the iterative algorithm for estimating the model parameters in the non-stationary case is deferred to Appendix C.

4.3 Kriging Prediction

Having estimated the model, spatial prediction at a target site s_0 in D can be performed by using the theory of object-oriented kriging presented in Menafoglio et al. (2013) (see also Menafoglio and Secchi (2017) for a recent review). In this setting, the kriging predictor is defined via a linear combination of the data that have been observed:

$$\mathcal{X}^*_{s_0} = \sum_{i=1}^{n} \lambda^*_i \mathcal{X}_{s_i}. \tag{23}$$

Here the weights $\lambda^*_1, ...\lambda^*_n \in \mathbb{R}$ are found as to minimize the global variance of the prediction error under the unbiasedness constraint, i.e.,

$$(\lambda^*_1, \ldots \lambda^*_n) = \underset{\substack{\lambda_1,\ldots\lambda_n\in\mathbb{R}:\\ \mathcal{X}^{\boldsymbol{\lambda}}_{s_0}=\sum_{i=1}^n \lambda_i \mathcal{X}_{s_i}}}{\operatorname{argmin}} \quad Var(\mathcal{X}^{\boldsymbol{\lambda}}_{s_0} - \mathcal{X}_{s_0}) \quad \text{subject to} \quad \mathbb{E}[\mathcal{X}^{\boldsymbol{\lambda}}_{s_0}] = m_{s_0}. \tag{24}$$

Using the general model (22) – which clearly reduces to the stationary case when $L = 0$ – the global optimum of problem (24) is obtained by solving the following linear system

$$\begin{pmatrix} C(s_1,s_1|\boldsymbol{\theta}) & \dots & C(s_1,s_n|\boldsymbol{\theta}) & 1 & f_1(s_1) & \dots & f_l(s_1) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ C(s_n,s_1|\boldsymbol{\theta}) & \dots & C(s_n,s_n|\boldsymbol{\theta}) & 1 & f_1(s_n) & \dots & f_l(s_n) \\ 1 & \dots & 1 & 1 & 0 & \dots & 0 \\ f_1(s_1) & \dots & f_1(s_n) & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_l(s_1) & \dots & f_l(s_n) & 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu_0 \\ \mu_1 \\ \vdots \\ \mu_L \end{pmatrix} = \begin{pmatrix} C(s_0,s_1|\boldsymbol{\theta}) \\ \vdots \\ C(s_0,s_n|\boldsymbol{\theta}) \\ 1 \\ f_1(s_0) \\ \vdots \\ f_L(s_0) \end{pmatrix} \tag{25}$$

where $\mu_0, \dots, \mu_L$ are Lagrange multipliers. Provided that the covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ is positive definite and that the design matrix of the linear model (22) is of full rank, the linear system admits a unique solution. It is worth highlighting that, even if the drift coefficients are not directly included in the Kriging system, their estimation is necessary in order to assess the trace-covariogram of the residual process $\{\delta_s, s \in D\}$, as discussed in Section 4.2.

5 Two Simulated Examples

The procedure outlined in Section 4.1.2 is here be applied to two simulated examples, one for the tail-up case and one for the tail-down case. In both examples, we consider the stream network domain D represented in Figure 2a, characterized by 250 segments and $n = 200$ observation points; this was generated using the `SNN` package (Ver Hoef et al. (2014)) in `R` (R Core Team (2020)). Zero mean functional random processes are simulated by exploiting the construction

$$\mathcal{X}_s = \sum_{k=1}^{N} \xi_k(s) e_k, \tag{26}$$

which is analogous to (9), with $m = 0$. Here, $\{e_k, k \geq 1\}$ denotes the Fourier orthonormal basis of $H = L^2([0,1])$, and N is set to $N = 7$. Parameters for the generation of the scalar fields $\{\xi_k(s), s \in D\}$, $k = 1, ..., N$, are specific of the examples, and are detailed below. The fields $\{\xi_k(s), s \in D\}$, $\{\xi_j(s), s \in D\}$ are assumed to be independent for $j \neq k$; each $\{\xi_k(s), s \in D\}$ is finally assumed to be Gaussian.

5.1 Estimating the trace-covariogram in a pure tail-down model

In this example, for each field $\{\xi_k(s), s \in D\}$ appearing (26), a tail-down exponential model is used with sill $\theta_v^{(k)}$, range $\theta_r^{(k)}$ and nugget $\eta^{(k)}$ parameters set to $(\theta_v^{(k)}, \theta_r^{(k)}, \eta^{(k)}) = (5, 6.5, 0)$, respectively. Therefore the theoretical model for the functional process (26) is a tail-down exponential model with parameters

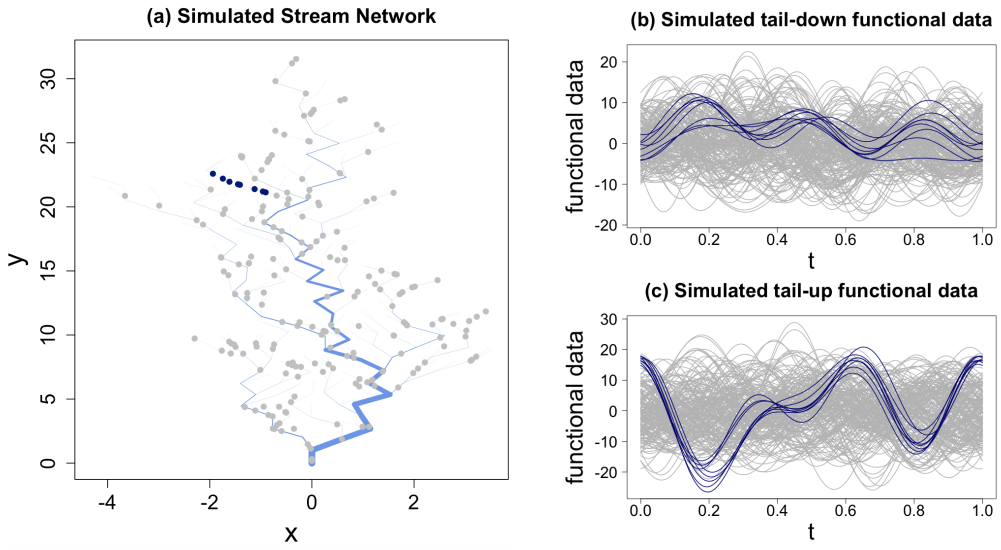

Figure 2. (a) Simulated stream network. Points indicate the locations of observed data. Blue lines indicate the stream network; their thickness is proportional to the stream segments orders. (b) Simulated functional data for the tail-down model. (c) Simulated functional data for the tail-up model.

$(\theta_v, \theta_r, \eta) = (35, 6.5, 0)$ (see the final remarks in Section 3.2). The functional dataset in Figure 2b was obtained by combining the realizations of the $N = 7$ scalar random fields sampled at the $n = 200$ locations displayed in Figure 2a.

Figure 3a displays the FCSD and FUSD empirical estimators (see eq. (20) and (18)), which share – as expected – the same non-trivial structure of spatial dependence. Both trace-semivariograms were obtained considering 15 lags and a maximum distance equal to half the maximum distance in the stream network. Clearly, the number of flow-unconnected pairs (represented through the dimensions of the circles in Fig. 3a) is much larger than the connected ones, as evidenced by the larger size of the yellow circles compared to the blue ones.

Although, in general, the flow-connected portion of the trace-semivariogram (blue circles) is the only one that should be considered to retrieve the parameter estimates (see the strategy outlined in Section 4.1.2), we may here consider also the FUSD for the purpose. Indeed, recall that, for the special case of an exponential tail-down model, the FUSD empirical estimator is unbiased for the flow-unconnected portion of the trace-semivariogram, as, in this case, the latter depends only on the stream distance $h = a + b$ (see Section 4.1.1 and Table 1). Fitting the trace-semivariogram parameters separately to the FCSD and the FUSD yields the results reported in Table 2. The slight overestimation of both the sill and the range obtained by fitting the FUSD semivariogram may be due to the variability of the FUSD estimator. The FCSD estimates seem to be more accurate, instead, being very close to the reference values ($\theta_v = 35$ and $\theta_r = 6.5$).

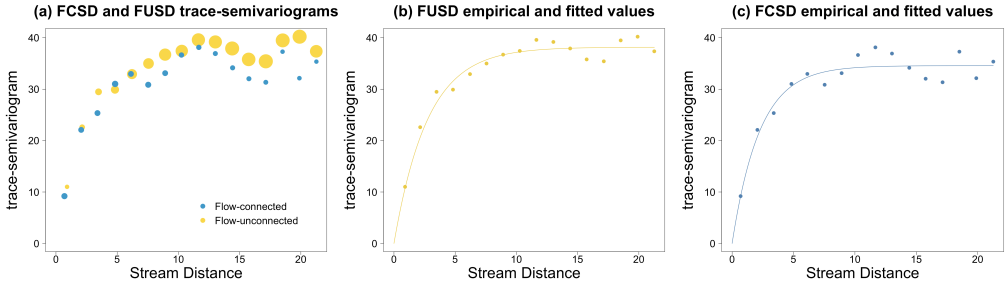

Figure 3. (a) Empirical FCSD trace semivariogram (blue) and FUSD trace semivariogram (yellow). The dot's sizes are proportional to the number of pairs for each binned distance class. (b) Empirical and Fitted FCSD trace-semi variograms. (c) (c) Empirical and Fitted FUSD trace-semi variograms.

Table 2. Estimated parameters of the Tail-Down covariance structure. Generating parameters were set to $\theta_v = 35$, and $\theta_r = 6.5$.

	$\hat{\theta}_v$	$\hat{\theta}_r$
FCSD	34.57	6.73
FUSD	38.16	8.26

5.2 Estimating the trace-covariogram in a pure tail-up model

We here consider a functional tail-up model, built as in (26), and assuming for each field $\{\xi_k(s), s \in D\}$ a tail-up spherical model with sill $\theta_v^{(k)}$, range $\theta_r^{(k)}$ and nugget $\eta^{(k)}$ parameters set to $(\theta_v^{(k)}, \theta_r^{(k)}, \eta^{(k)}) = (10, 8.5, 0)$, respectively. Therefore, the resulting functional process (26) is again a tail-up spherical model, but characterized by the parameters $(\theta_v, \theta_r, \eta) = (70, 8.5, 0)$. The functional dataset in Figure 2 (bottom right panel) was obtained by combining the $N = 7$ independent realizations of the scalar random fields at the $n = 200$ sampling locations in D, with the first $N = 7$ elements of the Fourier basis $\{e_k, k = 1, ..., 7\}$.

The empirical FUSD trace-semivariogram (18) and the empirical FCSD trace-semivariogram (20) are displayed in Figure 4a. They were computed considering 15 lags and a maximum distance equal to half the maximum distance in the network. The FUSD semivariogram appears to be flat, as expected in a pure tail-up model. On the contrary, the flow connected pairs are featured by a non-trivial spatial dependence. In particular, the FCSD semivariogram exhibit a clear downward concavity near the origin, settling towards a sill not far from the value of the FUSD semivariogram. Indeed, recall that the FUSD semivariogram the can be used to unbiasedly estimate the variogram sill in a pure tail-up model using expression (19). Here, as well as in the case study presented in Section 7, to retrieve estimated parameters, we fit a spherical model to the FCSD, as shown in Figure 4b. This leads to the following parameters estimates: $\hat{\theta}_v = 67.87, \hat{\theta}_r = 5.53$. Note that the estimated sill is close to the reference value $\theta_v = 70$; however the

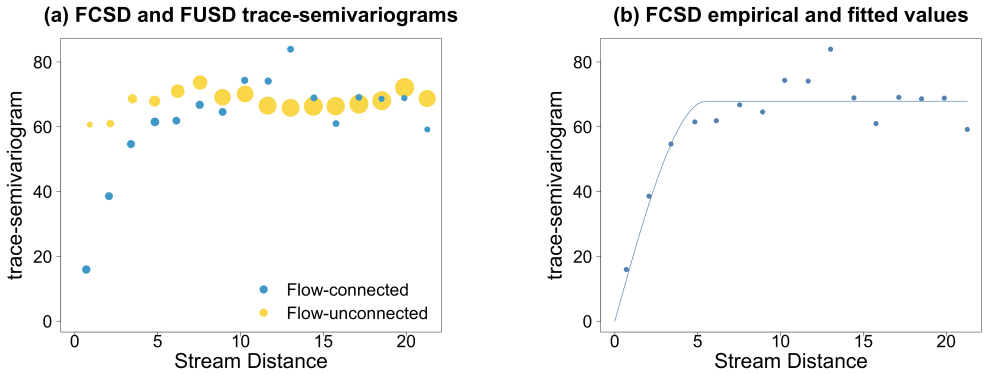

Figure 4. (a) Empirical FCSD trace semivariogram (blue) and FUSD trace semivariogram (yellow). The dots sizes are proportional to the number of pairs for each class of distances. (b) Empirical FCSD and fitted trace-semivariograms.

range is underestimated, the reference value being $\theta_r = 8.5$. This tendency is confirmed by the results of the simulation study described in Section 6, and it is due to the fact that FCSD trace-semivariogram neglects the weights π_{ij} appearing in (14), but the simulation process clearly accounts for them. Finally, the sill estimated from the average of the FUSD trace-semivariogram (see eq. (19)) is $\hat{\theta}_v = 68.35$, again rather close to the reference value.

6 Summary of the supporting simulation studies

In this Section, we briefly describe the simulation study reported in the Supplementary Material. This material introduces and tests two alternative estimators to FCSD, that, unlike FCSD, account for the weights π_{ij} appearing in (14). The first estimator is named Flow-Connected Weigth-Adjusted (FCWA) and is fully analogous to the one proposed by Zimmerman and Ver Hoef (2017) for the scalar case; the second estimator, named FCWA2, improve on FCWA by trying to reduce its variability. The simulation study compares the validity of these three estimators when used to estimate the unweigthed flow connected trace-semivariogram in a pure tail-up model and the corresponding covariance parameters. It also assesses the impact of the estimators on the kriging performances.

In all the tested scenarios, the spatial domain and sampled locations were fixed as in Section 5 and the functional random field $\{\mathcal{X}_s, s \in D\}$ was built using the construction (26), with $N = 7$. Here, the elements $\{e_k, k = 1, 2, ..., 7\}$ represent again the orthonormal basis of $H = L^2([0, 1])$ generated by the first $N = 7$ Fourier basis functions, and $\{\xi_k(s), s \in D\}$ are second-order stationary Gaussian random fields, with parameters set analogously as in Section 5. A Monte Carlo analysis was run by simulating $B = 500$ independent realizations of the functional fields, keeping as fixed the parameters of the scalar fields $\{\xi_k(s), s \in D\}$.

Comparison between FCSD, FCWA and FCWA2 For each of the $B = 500$ simulations, the three empirical trace-semivariograms FCSD, FCWA and FCWA2 were computed and compared with the corresponding theoretical values of the unweighted semivariogram $\gamma_{uw}(h)$ at the selected distances. The main results of this simulation follows.

- The FCSD trace-semivariogram is positively biased, in particular at the first lags, where it happened to almost double the theoretical value. This causes underestimations of the range when the FCSD semivariogram is used to fit a parametric model.
- The FCWA trace-semivariogram is featured by an extremely high variance and it is highly sensitive to the presence of outliers among the weights.
- The FCWA2 trace-semivariogram shows slightly lower variance and higher robustness w.r.t. the presence of outliers in the weights. FCWA2 seems to provide better results than FCWA1 at small distances.

Parameter estimation For the $B = 500$ simulated datasets, each empirical estimator (FCSD, FCWA, FCWA2) was used to estimate the sill and the range. The sill was also estimated via FUSD, applying equation (19).

- The fitting procedure implemented within the `R` package `gstat` (Pebesma, 2004) converged 497/500 times for the FCSD fitting, 317/500 for the FCWA and 356/500 for FCWA2; this evidences an instability in FCWA and FCWA2 due to their high variance.
- The FUSD estimate for the sill outperforms the others in terms of Root Mean Square error (RMSE).
- A non-negligible underestimation of the range is observed when fitting the FCSD semivariogram (both with respect to the mean and the median).
- FCWA and FCWA2 are more accurate on average than FCSD (although slightly overestimating the theoretical range), but they show higher RMSE due to their extremely high variability.
- Setting the sill to the FUSD estimate during the fitting procedure allows one to slightly reduce the variance in the range estimate of both FCWA and FCWA2.

Sensitivity of Kriging to range underestimation For each simulated dataset, a leave-one-out cross validation analysis was run to evaluate the performance of the kriging predictor obtained by plugging-in the parameters estimated via FCSD, FCWA and FCWA2 in the trace-covariogram. For each dataset and for each method (FCSD, FCWA1 and FCWA2) the prediction error was computed.

- Simulations show a notable presence of rather high outliers in terms of errors, when the empirical semivariogram is fitted via FCWA and FCWA2. These high errors are due to the very high variability of the estimators, which happen to be associated with fitted parameters extremely far from the theoretical ones. In these extreme cases, the kriging prediction are

compromised.

- Using FCSD, the distribution of the errors shows a slightly higher median than for FCWA and FCWA2, but proves more stable without outliers. The range underestimation does not affect significantly the Kriging performance.

As a result of the simulation study, we conclude that the extremely high variability of the weights-adjusted empirical semivariograms FCWA and FCWA2 makes them unusable from a practical point of view. The mere fact that in a large number of cases the variogram fit does not reach convergence leads us to consider these estimators too unstable for real applications. Furthermore, their use does not seem to be encouraged by better predictive performance either. These results are in full agreement with the conclusions of Zimmerman and Ver Hoef (2017) for the scalar case. The use of the FCSD is thus recommended, despite its bias.

7 A case study: Analysis of Middle Fork River Water Temperatures

Middle Fork River and Dataset The data analysed in this Section consist of the maximum daily water temperatures recorded between 15 July 2005 and 31 August 2005 at different locations of the Middle Fork river in Idaho, USA. The data, which can be found and downloaded at `https://www.fs.fed.us/rm/boise/AWAE/projects/SSN_STARS/software_data.html`, has been pre-processed and created as part of an NCEAS Workshop in April 2011 (National Center for Ecological Analysis and Synthesis). The Middle Fork River is a 104-mile-long (167 km) river in central Idaho. Its elevation ranges from 919 meters above sea level (at its mouth) to 2.100 meters. In Figure 5a the Middle Fork river is depicted together with the $N = 157$ observation sites. The daily maximum water temperature (in $^{\circ}C$) have been recorded at each of the 157 locations for 47 days in the aforementioned summer period (15 July 2005 - 31 August 2005). We embedd the data in the Hilbert Space $H = L^2$ of the square integrable functions endowed with the usual scalar product. The raw data (Figure 5b) were smoothed via spline smoothing with a roughness penality (Figure 5c). The number of basis functions ($nb = 49$) and the smoothing parameter ($\lambda = 5$) were chosen through a non-parametric leave-one-out cross validation approach to avoid overfitting. The watershed area accumulated downstream (km^2), was used to compute the weights for the tail up models, as a proxy variable for flow volume (see Ver Hoef and Peterson (2010)). Two covariates, namely the elevation of the upper stream segment node on which a temperature sensor was located (m) and the upstream distance between the stream outlet and the site (m), were used to model the drift term.

Geostatistical analysis The stationarity of the random field is evaluated from the empirical trace-semivariograms, computed by considering 13 distance classes with bins of equal size up to a maximum distance of 63.11 km (Figure 6). Visual

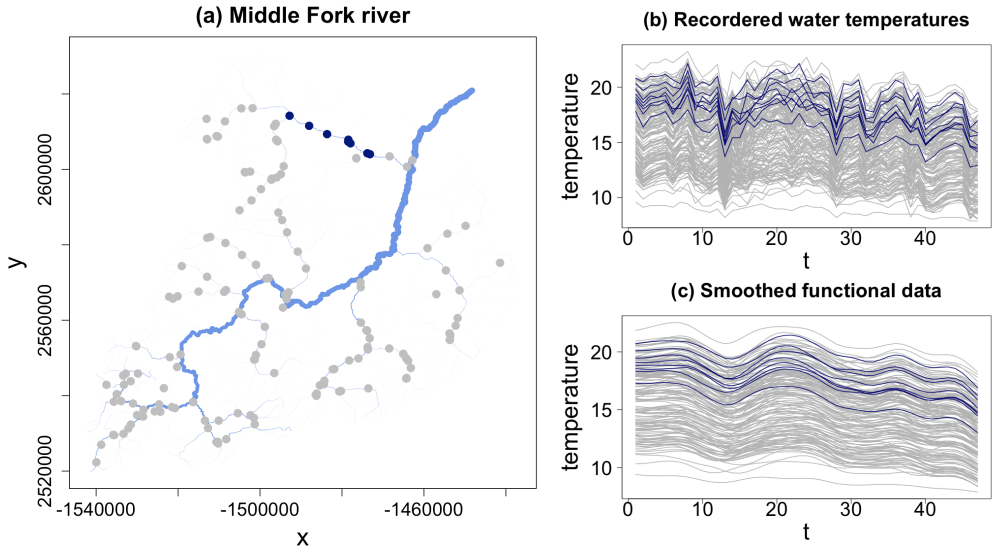

Figure 5. (a)Middle Fork river. Points indicate the locations of observed data. Blue lines indicate the stream network and their thickness is proportional to the stream segments orders. (b) Water temperatures at the 157 locations of the Middle Fork River from 15 July 2005 to 31 August 2005. (c) Smoothed functional data.

inspection of the trace-semivariograms suggest that a non-stationarity assumption is appropriate for the random field, since the FCSD trace-semivariogram (blue dots in Figure 6) seems to increase without bound, beside crossing the flow unconnected trace semivariogram (yellow dots in Figure 6), indicating a trend contamination aligned with flow. This behaviour (crossing components and unbounded growth) is indeed evidence of an unmodeled drift in upstream distance (Zimmerman and Ver Hoef (2017)). Following the approach devised in Subection 4.2, a drift term is thus included in the model.

We here consider as covariate for the drift term the variables $\{x, y\}$ ={elevation, distance upstream}, which are appropriate to describe a drift term aligned with flow (see Zimmerman and Ver Hoef (2017) for the scalar case). For the selection of the functional form for the drift, we follow the approach of Menafoglio et al. (2013), who suggest to consider polynomial forms for the drift term, and select the optimal one through cross-validation. Here, each candidate model is evaluated in terms of kriging performances, quantified through the sum of squared errors

$$SSE_i = \|\mathcal{X}_{s_i} - \mathcal{X}^*_{s_i}\|^2, \quad i = 1, ..., n, \tag{27}$$

where $\mathcal{X}^*_{s_i}$ stands for the kriging prediction of $\mathcal{X}_{s_i}$ when this is left out of the sample. At this stage, a simplified version of the Universal Kriging predictor with a covariance structure of pure nugget is employed, thus providing the prediction which would have been obtained via FDA linear models (indeed in this case the UK predictor reduces to the drift estimate). We thus considered as candidate

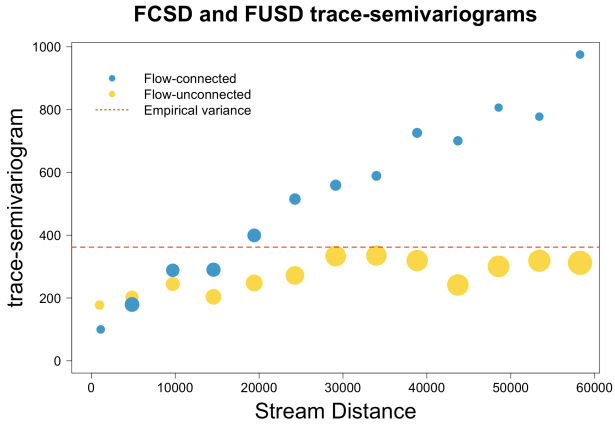

Figure 6. Trace-semivariograms for the Middle Fork temperatures. Empirical estimates of the flow-connected (blue symbols) and flow-unconnected (yellow symbols) trace-semivariograms; the empirical variance is reported as a dashed red line.

models the 31 polynomials of order lower than 2 (excluding the case of the sole intercept as drift). For each candidate drift, the empirical trace-semivariograms of the residuals are computed to verify that the optimal drift model selected according to the introduced criterion gives rise to a stationary residual. The best model is the following:

$$m(s,t) = a_0(t) + a_1(t)x + a_2(t)y + a_3(t)x^2 + a_4(t)y^2 + a_5(t)xy, \quad t \in \tau = [1, 47]. \tag{28}$$

Figure 7a reports the empirical FCSD and FUSD trace-semivariogram of the residuals of model (28), when these are estimated via OLS. The shape of the flow connected trace-semivariogram (in blue) is not that of a pure nugget, suggesting that the residuals are spatially correlated. On the other hand, the flow unconnected empirical semivariogram (in yellow) is compatible with a pure nugget model, suggesting the use of a tail-up model for the field (see Section 4.1.2). In the following, we shall thus consider an exponential tail-up model with sill θ_v, range θ_r and nugget η (see Table 1). Analogous results were obtained, in a scalar case, by Liu (2019), who used non-parametric testing for selecting the model for the average water temperature over part of the domain of our study.

Having chosen the drift and the covariance models, the model parameters are estimated by means of the generalized least square criterion outlined in Section 4.2 and Appendix C. Figure 7b displays the FCSD empirical trace-semivariogram together with the fitted variogram model, characterized by estimated parameters: $(\hat{\theta}_v, \hat{\theta}_r, \hat{\eta}) = (68.83, 25885.44, 93.59)$. Note that, as broadly discussed in Section 4.1.2, interpretation of $\hat{\theta}_r$ requires particular care, as it could be affected by a negative bias. Kriging is eventually performed at a grid of new locations along the stream network, by plugging-in the estimated parameters $\hat{\boldsymbol{\theta}} = (\hat{\theta}_v, \hat{\theta}_r, \hat{\eta})'$ in the linear system (25). To evaluate the performance of the Kriging predictor, a leave-one-out cross validation (LOOCV) procedure is applied, considering as

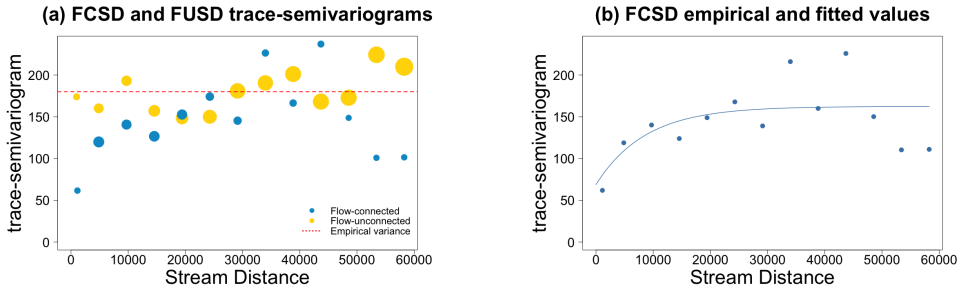

Figure 7. (a) Empirical trace semivariogram of the residuals obtained with an OLS estimate of the drift. (b) Empirical and fitted trace-semivariograms of the residuals obtained with WLS.

measure of discrepancy between the true value and the predicted one the SSE defined in (27) and the relative SSE performance index, given by

$$SSE_i^{(rel.)} = \frac{SSE_i}{\|\mathcal{X}_{s_i}\|^2}. \tag{29}$$

The statistics shown in Table 3 prove the satisfactory forecasting performance of the method. Figure 8a displays with colors the SSE_i for each location on the river; a part for a few locations associated with a high estimation error ($SSE > 700$, red dots), which possibly mark influential/outlying data, the kriging predictor works properly.

Figure 9 shows the original data (Fig. 9a) together with the UK estimates (Fig. 9b) and the corresponding kriging residuals (Fig. 9c). Note that the significant reduction of the total SSE, $SSE = \sum_{i=1}^{n} SSE_i$, attained with a tail-up covariance structure ($SSE = 20864.89$) as opposed to a pure nugget ($SSE = 29485.93$), confirms the ability of the former to capture in a greater extent the stochastic variability of the residual process. Finally, Figure 10 provides a representation of the observed (Fig. 10a) and predicted (via LOOCV, Fig. 10b) average temperatures (the average being taken over the summer period) together with the corresponding marginal distributions (Fig. 10c). Cross-validation results exhibit a narrower range of values than the data and this is a sign of the Kriging smoothing effect. Comparison between Fig. 10a and b confirms the validity of the proposed method, which is able to reproduce the main spatial patterns in the data.

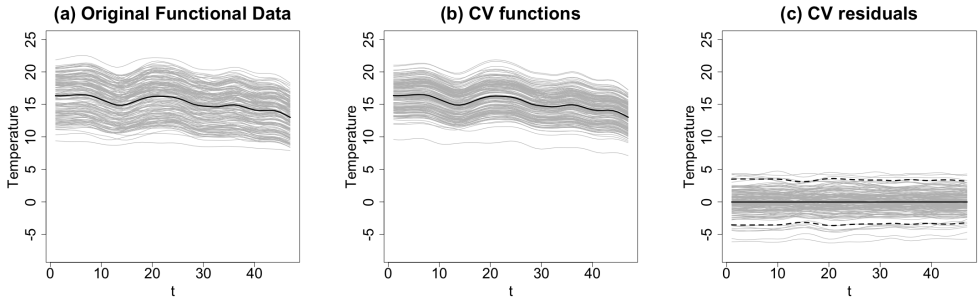

Figure 9. Cross-validation analysis. (a) Original data (in grey) and relative sample mean (in black). (b) Data predicted via Universal Kriging (in grey) and their mean (black). (c) Difference between original and predicted data (in grey), their mean $\hat{m}_r$ (in black) and the (approximate) point-wise confidence band $m_r(z) + 2\hat{\sigma}_r(z)$(dashed black), where $\hat{\sigma}_r(z)$ is the (point-wise) standard deviation estimated from the cross-validation residual.

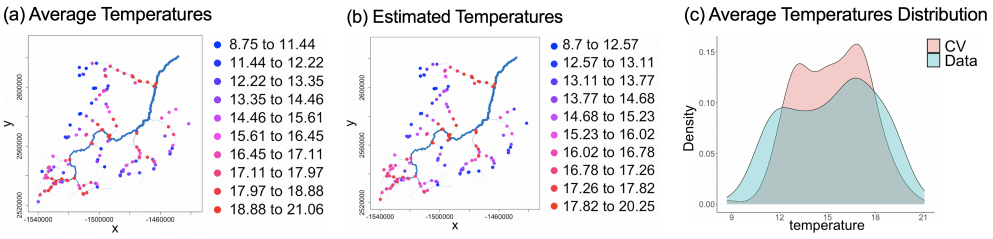

Figure 10. (a): Observed average temperatures over the summer period for the locations on the Middle Fork River. (b): Estimated average temperatures via leave-one-out Universal Kriging for the Middle Fork data. (c): Distributions of the observed average temperatures together with the cross-validation average temperatures.

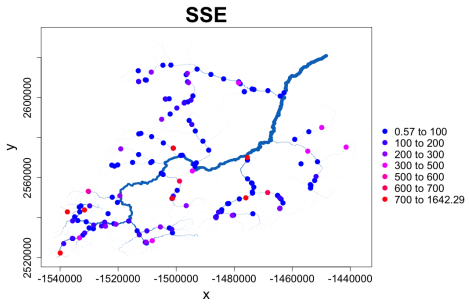

Figure 8. SSE leave-one-out error for each location on the Middle Fork River

	SSE	$SSE^{(rel)}$
Min	0.574	$3.44 \cdot 10^{-5}$
Median	43.346	$4.19 \cdot 10^{-3}$
Mean	132.897	$1.6 \cdot 10^{-2}$
Sum	20864.89	2.52

Table 3. Summary indices of the distribution of SSE and $SSE^{(rel)}$

8 Conclusion

In this work new geostatistical methods for complex data distributed over a stream network have been proposed. First of all, the theoretical construction presented in Section 3 allows one to develop a strategy for variographic analysis and estimation of the spatial covariance structure, which proved to be effective in terms of prediction performance for both tail-up and tail-down models. When tested on real data, the methodology achieved a good prediction performance. However, further research should be carried out for tail-up models, aiming at a better empirical estimator of the semivariogram that takes into account the weights characterising the river topology. This could potentially allow one to achieve better estimates for the range of correlation of the field. In addition, extension to mixed models should certainly be the object of future works. To this end, an adaptation of the procedure proposed in this work is envisioned as a nested structure for the fitting of the FCSD trace-semivariogram. Further research may also be devoted to relax the hypothesis of homogeneous covariance structure when dealing with large stream networks. In fact, allowing varying dependence structures on different sub-networks could lead to interesting developments in the direction of modeling strong spatial non-stationarities. Concerning the topological description of the stream network, one should note that the current work, as well as the literature focused on the scalar case, only enables one to analyse data over one-dimensional stream segments. However, allowing the representation of stream segments to be also equipped with information about their depth and thickness, might notably enrich the geostatistical analysis, especially in cases where the stream network includes large sub-streams and lakes. Finally, the application of the developed models to contexts other than those of a river network is possible, and would be extremely topical in contexts like electricity grids, traffic and transportation systems, and road networks. Indeed, whenever it is advisable to define the stream distance with respect to the topology of a network featured by the presence of a flow rather than based on a Euclidean distance, the proposed approach should be considered. Here, extensions of the considered class of models will deserve further research to allow for the analysis of data distributed over non-binary trees and general networks for which valid covariance models are yet to be studied.

Appendix

A Infinite Dimensional Functional Process

We here discuss the conditions that allows one to consider an infinite-dimensional construction for the functional process, i.e., to represent the element $\mathcal{X}_s$ as the limit as $N \to \infty$ of (11) and obtain a well-defined global covariance function.

Therefore, we formally define the functional random field as

$$\mathcal{X}_s = \sum_{k=1}^{\infty} \xi_k(s) e_k, \tag{30}$$

where $\{e_k\}$ is an orthonormal basis of H and $\{\xi_k(s), s \in D\}$, $k \geq 1$, denote independent, zero-mean, second-order stationary and isotropic scalar random fields, defined through the moving average construction described in Section 2. A minimal assumption for the *existence* of the cross covariance operators C_{s_i,s_j} is that the random variables $\mathcal{X}_s$ have finite second moments, i.e., $\mathbb{E}[\|\mathcal{X}_s^2\|] < \infty$, for all s in D. For the functional random process (30) this is equivalent to the following condition

$$\sum_{k=1}^{\infty} \mathbb{E}[\xi_k(s)^2] = \sum_{k=1}^{\infty} \int_{-\infty}^{+\infty} \left(g^{(k)}(x - s|\boldsymbol{\theta})\right)^2 dx < \infty, \quad \text{for all } s \in D. \tag{31}$$

In this context, the dependence of $g(\cdot|\boldsymbol{\theta})$ on the parameters $\boldsymbol{\theta}$ plays an important role since one should impose conditions on $\boldsymbol{\theta}$ to ensure the finiteness of the series. It is worth highlighting that the autocovariance function of each scalar random field $\{\xi_k(s), s \in D\}$ admits a compact form

$$\mathbb{E}[\xi_k(s)\xi_k(s+h)] = C^{(k)}(h|\boldsymbol{\theta}) = \theta_v^{(k)} \rho_k(h/\theta_r^{(k)}), \tag{32}$$

where $\rho_k(\cdot)$ is a positive correlation function that depends on the type of moving average function of the k-th random field. Plugging-in (32) in (31), we get

$$\sum_{k=1}^{\infty} \mathbb{E}[\xi_k(s)^2] = \sum_{k=1}^{\infty} \int_{-\infty}^{+\infty} \left(g^{(k)}(x - s|\boldsymbol{\theta})\right)^2 dx = \sum_{k=1}^{\infty} C^{(k)}(0|\boldsymbol{\theta}) = \sum_{k=1}^{\infty} \theta_v^{(k)}$$

Therefore $\mathcal{X}_s$ belongs to $L^2(\Omega; H)$ provided that we assume the summability of the series of $\theta_v^{(k)}$, i.e.,

$$\sum_{k=1}^{\infty} \theta_v^{(k)} < \infty. \tag{33}$$

Under condition (33), one can prove by direct computations that $\theta_v^{(k)}$ are the eigenvalues of the covariance operator $\mathcal{C}_{s,s}$. Moreover, under the square integrability assumption, the cross-covariance operator $\mathcal{C}_{s_i,s_j}$ exists and it is a symmetric trace-class Hilbert-Schmidt operator (Bosq, 2000). Finally, its trace is well defined by $\sum_{k=1}^{\infty} \langle C_{s_i,s_j} e_k, e_k \rangle$, as the series converges absolutely for every orthonormal basis in H and the sum does not depend on the choice of the basis (Zhu, 2007). The identity

$$C(s_i, s_j) = \sum_{k=1}^{\infty} \langle C_{s_i,s_j} e_k, e_k \rangle$$

can be proved as in Menafoglio et al. (2013), by exploiting the Parseval identity and the Lebesgue's dominated convergence theorem for series. Note that to apply the latter theorem, the requirement $\mathbb{E}[\|\mathcal{X}_s\|^2] < \infty$ is crucial.

B Bias of the FCSD empirical estimator

We here discuss on the bias of the FCSD empirical estimator proposed in Subsection 4.1.1 when $\{\mathcal{X}_s, s \in D\}$ is represented by a tail-up model. Recall that the expression of the trace-semivariogram expression for tail-up models is

$$\gamma(s_i, s_j) = \begin{cases} 0 & \text{if } s_i = s_j \text{ (i.e if } h = 0) \\ \theta_v & \text{if } s_i \text{ and } s_j \text{ are not flow connected} \\ \theta_v - \pi_{i,j} C_t(h|\boldsymbol{\theta}) & \text{otherwise.} \end{cases} \tag{34}$$

Here, $C_t(h|\boldsymbol{\theta})$ is the trace-covariogram for the moving average functional process on the real line, which is related to the *unweighted flow-connected trace-semivariogram* $\gamma(h)$ by the relation

$$C_t(h|\boldsymbol{\theta}) = C_t(0) - \gamma(h) = \theta_v - \gamma(h). \tag{35}$$

Plugging-in (35) in (34) we get

$$\gamma(s_i, s_j) = \begin{cases} 0 & \text{if } s_i = s_j \text{ (i.e if } h = 0) \\ \theta_v & \text{if } s_i \text{ and } s_j \text{ are not flow connected} \\ \theta_v - \pi_{i,j}\big(\theta_v - \gamma(h)\big) & \text{otherwise.} \end{cases} \tag{36}$$

From this expression we can easily see that, in the absence of weights (i.e., setting $\pi_{i,j} = 1$) $\gamma(s_i, s_j)$ effectively would correspond to $\gamma(h)$, which is targetted by FCSD. Concerning the flow-connected portion of expression (36), i.e., focusing on locations flow-connected s_i, s_j, one has

$$\frac{1}{2}\operatorname{Var}_H(\mathcal{X}_{s_i} - \mathcal{X}_{s_j})_H = \theta_v - \pi_{i,j}\theta_v + \pi_{i,j}\gamma(h), \tag{37}$$

that, rearranging the terms, reads

$$\gamma(h) = \theta_v + \frac{1}{2\pi_{i,j}}\operatorname{Var}_H(\mathcal{X}_{s_i} - \mathcal{X}_{s_j}) - \frac{1}{\pi_{i,j}}\theta_v.$$

Given that the FCSD empirical estimator (20) is an unbiased estimator for $\frac{1}{2}\operatorname{Var}_H(\mathcal{X}_{s_i} - \mathcal{X}_{s_j})$, it straightforwardly follow that it is biased for the unweighted flow-connected semivariogram $\gamma(h)$, unless $\pi_{i,j} = 1$ for all i, j. Unbiased estimators named FCWA and FCWA2 are derived and studies in the Supplementary Material, adjusting for the bias of the FCSD according to eq. (37).

C Drift estimation

We here briefly recall the procedure which can be used to estimate the linear model in (22). The model for the vector of observations $\boldsymbol{\mathcal{X}} = (\mathcal{X}_{s_1}, ..., \mathcal{X}_{s_n})^T$ can be expressed as

$$\boldsymbol{\mathcal{X}} = \mathbb{F}a + \boldsymbol{\delta}, \tag{38}$$

where $\boldsymbol{a} = (a_0, \ldots, a_L)^T$ is the vector of (functional) coefficients, $\boldsymbol{\delta} = (\delta_{s_1}, \ldots, \delta_{s_n})^T$ is the random vector of spatially-correlated residuals and $\mathbb{F} \in \mathbb{R}^{n \times (L+1)}$ is the design matrix, i.e., $\mathbb{F}_{i,l} = (f_l(s_i))$. Menafoglio et al. (2013) propose to estimate the functional coefficients $\boldsymbol{a}$ given $\boldsymbol{\mathcal{X}}$ based on a generalized least square criterion (GLS) with weighting matrix Σ^{-1}, i.e., the inverse of the $n \times n$ covariance matrix Σ of $\boldsymbol{\mathcal{X}}$. Since $\hat{\boldsymbol{a}}^{GLS}$ depends itself on Σ, which is usually unknown, the following iterative algorithm, can be used for its actual computation.

Algorithm 1 (Menafoglio et al. (2013))**.** Given a realization $\boldsymbol{x} = (x_{s_1}, \ldots, x_{s_n})$ of $\boldsymbol{\mathcal{X}}$, represented as in (22):

1. Estimate the drift vector $\boldsymbol{m}$ through the OLS method and set $\hat{\boldsymbol{m}} = \hat{\boldsymbol{m}}^{OLS}$, with $\hat{\boldsymbol{m}}^{OLS} = \mathbb{F}(\mathbb{F}^T\mathbb{F})^{-1}\mathbb{F}^T\boldsymbol{x}$.
2. Compute the residual estimate $\hat{\boldsymbol{\delta}} = (\hat{\delta}_{\boldsymbol{s}_1}, \ldots, \hat{\delta}_{\boldsymbol{s}_n})$ by difference: $\hat{\delta} = \boldsymbol{x} - \hat{m}$.
3. Estimate the trace-semivariogram $\gamma(\cdot, \cdot)$ of the residual process $\{\boldsymbol{\delta}_s, s \in D\}$ from $\hat{\delta}$ first with the FCSD empirical estimator (20) and then fitting to this a valid model $\gamma(\cdot; \boldsymbol{\theta})$, obtaining $\hat{\boldsymbol{\theta}}$. Plug-in $\hat{\boldsymbol{\theta}}$ in the stream-network trace-covariogram expression of Σ (see Table 1, $\Sigma_{i,j} = C(s_i, s_j|\boldsymbol{\theta})$) yielding $\hat{\Sigma}$ (with $\hat{\Sigma}_{i,j} = C(s_i, s_j|\hat{\boldsymbol{\theta}})$).
4. Estimate the drift vector $\boldsymbol{m}$ with $\hat{\boldsymbol{m}}^{GLS}$, obtained from x using: $\hat{\boldsymbol{m}}^{GLS} = \mathbb{F}(\mathbb{F}^T\Sigma^{-1}\mathbb{F})^{-1}\mathbb{F}^T\Sigma^{-1}\boldsymbol{x}$.
5. Repeat 2.-4. until convergence.

References

Barbi, C., Menafoglio, A., and Secchi, P. (2022), "Supplementary material for: An object-oriented approach to the analysis of spatial complex data over stream-network domains," *MOX report*, .

Bosq, D. (2000), *Linear Processes in Function Spaces* New York, NY: Springer-Verlag.

Cressie, N., Frey, J., Harch, B., and Smith, M. (2006), "Spatial Prediction on a River Network," *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 127–150.

Curriero, F. (2006), "On the Use of Non-Euclidean Distance Measures in Geostatistics," *Mathematical Geology*, 38, 907–926.

Giraldo, R. (2009), Geostatistical analysis of functional data, PhD thesis, Universitat Politecnica da Catalunya, Barcelona.

Haggarty, R., Miller, C., and Scott, E. (2014), "Spatially Weighted Functional Clustering of River Network Data," *Journal of the Royal Statistical Society*, 64, 491–506.

Horváth, L., and Kokoszka, P. (2012), *Inference for Functional Data with Applications*, Springer Series in Statistics, New York, NY: Springer.

Hörmann, S., and Kokoszka, P. (2011), "Consistency of the Mean and the Principal Components of Spatially Distributed Functional Data," *Bernoulli*, 19, 1535–1558.

Kokoszka, P., and Horváth, L. (2012), *Inference for Functional Data with Applications,* New York, NY: Springer-Verlag.

Liu, Z. V. (2019), Testing covariance structure of spatial stream network data using Torgegram components and subsampling., PhD thesis, University of Iowa.

Marron, J. S., and Alonso, A. M. (2014), "Overview of Object Oriented Data Analysis," *Biometrical Journal*, 56(5), 732–753.

Mateu, J., and Giraldo, R. (2021), *Geostatistical Functional Data Analysis* Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd.

Menafoglio, A., Gaetani, G., and Secchi, P. (2018), "Random domain decompositions for object-oriented Kriging over complex domains," *Stochastic Environmental Research and Risk Assessment*, 32, 3421–3437.

Menafoglio, A., Pigoli, D., and Secchi, P. (2021), "Kriging Riemannian Data via Random Domain Decompositions," *Journal of Computational and Graphical Statistics*, 30(3), 709–727.

Menafoglio, A., and Secchi, P. (2017), "Statistical analysis of complex and spatially dependent data: A review of Object Oriented Spatial Statistics," *European Journal of Operational Research*, 258, 401–410.

Menafoglio, A., and Secchi, P. (2019), "O2S2: A new venue for computational geostatistics," *Applied Computing and Geosciences*, 2, 100007.

Menafoglio, A., Secchi, P., and Dalla Rosa, M. (2013), "A Universal Kriging Predictor for Spatially Dependent Functional Data of a Hilbert Space," *Electronic Journal of Statistics*, 7, 2209 – 2240.

Pebesma, E. J. (2004), "Multivariable geostatistics in S: the gstat package," *Computers and Geosciences*, 30, 683–691.

Peterson, E., Theobald, D., and Ver Hoef, J. (2007), "Geostatistical modelling on stream networks: Developing valid covariance matrices based on hydrologic distance and stream flow," *Freshwater Biology*, 52, 267 – 279.

Peterson, E., and Ver Hoef, J. (2010), "A mixed-model moving-average approach to geostatistical modeling in stream networks," *Ecology*, 91, 644–51.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis*, Springer Series in Statistics New York, NY: Springer.

Van Den Boogaart, K. G., Egozcue, J., and Pawlowsky-Glahn, V. (2014), "Bayes Hilbert Spaces," *Australian & New Zealand Journal of Statistics*, 56, 171–194.

Ver Hoef, J., and Peterson, E. (2010), "A Moving Average Approach for Spatial Statistical Models of Stream Networks," *Journal of the American Statistical Association*, 105, 6–18.

Ver Hoef, J., Peterson, E. E., Clifford, D., and Shah, R. (2014), "SSN: An R Package for Spatial Statistical Modeling on Stream Networks," *Journal of Statistical Software*, 56(3), 1–45.

Ver Hoef, J., Peterson, E., and Theobald, D. (2006), "Spatial statistical models that use flow and stream distance," *Environmental and Ecological Statistics*, 13, 449–464.

Yaglom, A. M. (1987), *Correlation Theory of Stationary and Related Random Functions*, Vol. 1 of *Springer Series in Statistics,* New York: Springer-Verlag.

Zhu, K. (2007), *Operator Theory in Function Spaces (Second ed.)* American Mathematical Society.

Zimmerman, D. L., and Ver Hoef, J. M. (2017), "The Torgegram For Fluvial Variography," *Journal of Computational and Graphical Statistics*, 13, 253–264.

Supplementary material for: An object-oriented approach to the analysis of spatial complex data over stream-network domains

Chiara Barbi[1*] , Alessandra Menafoglio[1] and Piercesare Secchi[1]

[1]MOX – Department of Mathematics, Politecnico di Milano, Milano, Italy
[*] chiara.barbi@polimi.it

1 Derivation of weight adjusted empirical estimators

Aiming to derive unbiased empirical estimators for the unweighted flow-connected trace-semivariogram for tail-up models, let us consider the relation for two flow-connected locations s_i, s_j separated by the stream distance h_k, derived in Appendix B:

$$\gamma(h_k) = \theta_v + \frac{1}{2\pi_{i,j}} \mathrm{Var}_H(\mathcal{X}_{s_i} - \mathcal{X}_{s_j}) - \frac{1}{\pi_{i,j}}\theta_v \tag{1}$$

Summing over all the couples of flow connected locations characterized by a distance h_k on both sides of (1) and rearranging the terms yields

$$\gamma(h_k) = \theta_v + \frac{1}{2}\mathrm{Var}_H\left(\frac{1}{\sqrt{\pi_{i,j}}}(\mathcal{X}(s_i) - \mathcal{X}(s_j))\right) - \frac{\theta_v}{|N(\mathcal{C}_k)|}\sum_{(s_i,s_j)\in N(\mathcal{C}_k)}\frac{1}{\pi_{i,j}},$$

where $N(\mathcal{C}_k) = \{(s_i, s_j) : d(s_i, s_j) = h_k, U_{s_i} \cap U_{s_j} \neq \varnothing\}$, is the set of flow-connected pairs separated by a stream distance h_k, and $|N(\mathcal{C}_k)|$ is its cardinality. To obtain an estimator $\hat{\gamma}(h_k)$ for $\gamma(h_k)$ recall that the flow-unconnected portion of the trace-semivariogram of a tail-up model is a constant function corresponding to the partial sill θ_ν. Therefore, $\hat{\gamma}_{FUSD}$ is unbiased for this flow-unconnected portion of the semivariogram and an estimate for the partial sill θ_v is given by equation (19) of the main manuscript. Moreover, an empirical estimator for $\mathrm{Var}_H\left(\frac{1}{\sqrt{\pi_{i,j}}}(\mathcal{X}(s_i) - \mathcal{X}(s_j))\right)$ is given by

$$\frac{1}{|N(\mathcal{C}_k)|}\sum_{(s_i,s_j)\in N(\mathcal{C}_k)}\frac{\|\mathcal{X}(s_i) - \mathcal{X}(s_j)\|^2}{\pi_{i,j}},$$

leading to

$$\hat{\gamma}_{FCWA}(h_k) = \bar{\gamma}_{FUSD} - \frac{1}{2|N(\mathcal{C}_k)|} \sum_{(s_i,s_j)\in N(\mathcal{C}_k)} \frac{2\bar{\gamma}_{FUSD} - \|\mathcal{X}(s_i) - \mathcal{X}(s_j)\|^2}{\pi_{i,j}}. \quad (2)$$

Equation (2) defines a weight-adjusted flow-connected semivariogram (hereafter named FCWA), which is analogous to the definition given in Zimmerman and Ver Hoef (2017) for scalar data. In practice, a tolerance is used in defining $N(\mathcal{C}_k)$, by considering pairs of locations whose distance is approximately h_k.
The estimator $\hat{\gamma}_{FCWA}(h_k)$ is unbiased for the unweighted flow-connected portion of the semivariogram of a pure tail-up model but, as demonstrated in our Montecarlo simulation presented in Section 2, it is characterised by having a very high variability. This large variance is due to the division by $\pi_{i,j}$ when computing $\hat{\gamma}_{FCWA}(h_k)$ in stream segments associated with small weights. In an attempt to reduce this variability, the following modified estimator will also be studied

$$\hat{\gamma}_{FCWA2}(h_k) = \bar{\gamma}_{FUSD} - \frac{1}{\hat{\pi}}\bar{\gamma}_{FUSD} + \frac{1}{2|N(\mathcal{C}_k)|\hat{\pi}} \sum_{(s_i,s_j)\in N(\mathcal{C}_k)} \|\mathcal{X}(s_i) - \mathcal{X}(s_j)\|^2. \quad (3)$$

In (3) $\hat{\pi}$ is the average of all the weights of the couples considered in $N(\mathcal{C}_k)$ and $\bar{\gamma}_{FUSD}$ is defined as before. This expression can be obtained straightforwardly from equation (1). The estimator in (3) is, as FCWA, unbiased for the unweighted flow-connected portion of the semivariogram of a pure tail-up model. Furthermore, we hope that averaging the weights will mitigate the effect of small outliers among the weights π_{ij} associated to the couples $(s_i, s_j) \in N(\mathcal{C}_k)$.

2 A comparison between FCSD, FCWA and FCWA2

A simulation analysis is presented to compare the validity of the three empirical estimators (FCSD, FCWA and FCWA2) when used to estimate the unweigthed flow-connected semivariogram in a pure tail-up model and the corresponding covariance parameters.
In all the tested scenarios, the stream network with 250 segments and 200 observation sites employed in Section 5 of Barbi et al. (2022) has been used. B=500 independent realizations of the following functional field have been simulated:

$$\mathcal{X}_s = \sum_{k=1}^{N} \xi_k(s) e_k \quad (4)$$

where $N = 7$, the elements $\{e_k, k = 1, 2, ..., 7\}$ represent the orthonormal basis of $H = L^2([0, 1])$ generated by the first $N = 7$ Fourier basis functions and $\{\xi_k(s), s \in D\}$ are independent second-order stationary Gaussian random fields, characterized by a tail-up Spherical models with sill, range an nugget equal to

$(\theta_v^{(k)}, \theta_r^{(k)}, \eta^{(k)}) = (10, 8.5, 0)$. Therefore the theoretical model for the functional process (4) is a tail-up Spherical model with parameters $(\theta_v, \theta_r, \eta) = (70, 8.5, 0)$. For each of the B=500 simulations, the three empirical semivariograms FCSD, FCWA and FCWA2 were computed by considering 15 lags and a maximum distance equal to half the maximum distance in the network.
Table 1 displays the value of the theoretical unweighted semivariogram $\gamma(h)$ (i.e., Spherical with $\theta_v = 70$ and $\theta_r = 8.5$) together with the average of the 500 esitimated FCSD, FCWA and FCWA2 trace-semivariograms at selected distances and their standard deviation. Notice that, to make this comparision reliable, $\gamma(h)$ at each distance is not the pointwise value of the theoretical semivariogram. Instead, at each lag, it is the average of the theoretical values evaluated at the same classes of distances used to obtain the three empirical semivariograms. Comparison of these three values at each distance h indicates that the FCSD semivariogram is positively biased, in particular at the first lags, where it almost doubles the theoretical value. This appears to be the reason why we experienced a rather underestimated range by fitting the theoretical parametric model to the FCSD trace-semivariogram. On the other hand, the average FCWA and FCWA2 semivariograms reach the sill more slowly and as expected, they are both unbiased estimators of the unweighted flow connected semivariogram. Nevertheless their standard deviation is unacceptably high, especially as the distance increases. It is worth noticing, however, that there is a slight improvement in terms of variability of FCWA2 with respect to FCWA, at least in the first lags.

The FCWA and FCWA2 semivariogram's variability can be unacceptably high as exemplified in Figure 1. These plots represents all the 500 estimated trace-semivariogram in the three cases (FCSD, FCWA and FCWA2) together with their average (red line) and the theoretical model (black line). The FCWA and FCWA2 variograms are characterized by a much larger variation than the FCSD trace-semivariogram. As the distance increases, in many of the 500 iterations, both the FCWA and the FCWA2 semivariograms assume unnaturally high values (in absolute value). The high variance is due to the division by small values of $\pi_{i,j}$, $\hat{\pi}$ when computing $\hat{\gamma}_{FCWA}(h_k)$, $\hat{\gamma}_{FCWA2}(h_k)$ respectively. The slight improvement in term of variances of the FCWA2 is easily explainable considering the expressions of the FCWA in (2) and FCWA2 in (3). In the former, the weight $\pi_{i,j}$ of every couple of locations (s_i, s_j) considered in the bin whose representative distance is h_k appears at the denominator. Hence we are considering $|N(\mathcal{C}_k)|$ different weights. If one of these weights were to be excessively small, then the second term in expression (2) would explode in absolute value. In other words, FCWA is highly sensitive to to the existence of outliers among the weights. In fact, estimator FCWA2, taking the mean of all the $|N(\mathcal{C}_k)|$ weights, partially mitigate this drawback. Indeed, FCWA2 seems to provide satisfactory results at least for small distances. However, as the distance increases, also FCWA2 presents very large values, as the higher the distance between two locations (s_i, s_j), the smaller will be the associated weight $\pi_{i,j}$ (see Section 2 in Barbi et al. (2022) for

Table 1. Comparison between FCSD, FCWA and FCWA2 as estimators of the unweighted semivariogram $\gamma_{uw}(h)$ at 15 lags.

$\boldsymbol{Distances(h)}$	**0.74**	**2.23**	**3.66**	**5.11**	**6.61**	**8.05**	**9.46**	**10.97**
$\boldsymbol{\gamma(h)}$	9.04	26.79	42.31	55.37	64.96	69.47	70.00	70.00
$\mathbf{E}[\hat{\gamma}_{\boldsymbol{FCSD}}(\boldsymbol{h})]$	18.97	45.80	59.23	65.44	68.38	69.94	69.80	69.63
$\boldsymbol{\sigma}[\hat{\gamma}_{\boldsymbol{FCSD}}(\boldsymbol{h})]$	1.66	4.30	6.15	7.22	7.97	8.51	9.67	10.22
$\mathbf{E}[\hat{\gamma}_{\boldsymbol{FCWA}}(\boldsymbol{h})]$	8.89	26.45	41.42	55.59	64.27	70.13	69.97	67.74
$\boldsymbol{\sigma}[\hat{\gamma}_{\boldsymbol{FCWA}}(\boldsymbol{h})]$	4.68	13.40	23.19	30.11	31.34	49.22	60.52	54.33
$\boldsymbol{E}[\hat{\gamma}_{\boldsymbol{FCWA2}}(\boldsymbol{h})]$	8.52	26.04	41.80	55.02	64.42	69.35	68.80	68.04
$\boldsymbol{\sigma}[\hat{\gamma}_{\boldsymbol{FCWA2}}(\boldsymbol{h})]$	1.85	6.41	13.35	19.53	22.93	33.81	37.58	39.99

$\boldsymbol{Distances(h)}$	**12.37**	**13.94**	**15.30**	**16.80**	**18.23**	**19.73**	**21.18**
$\boldsymbol{\gamma_{uw}(h)}$	70.00	70.00	70.00	70.00	70.00	70.00	70.00
$\mathbf{E}[\hat{\gamma}_{\boldsymbol{FCSD}}(\boldsymbol{h})]$	69.58	69.74	70.06	70.20	70.09	70.05	69.98
$\boldsymbol{\sigma}[\hat{\gamma}_{\boldsymbol{FCSD}}(\boldsymbol{h})]$	11.06	11.38	11.76	12.97	13.03	12.51	13.62
$\mathbf{E}[\hat{\gamma}_{\boldsymbol{FCWA}}(\boldsymbol{h})]$	67.30	68.00	69.96	71.41	70.30	69.87	68.95
$\boldsymbol{\sigma}[\hat{\gamma}_{\boldsymbol{FCWA}}(\boldsymbol{h})]$	61.34	63.83	78.50	80.78	83.37	91.18	95.16
$\boldsymbol{E}[\hat{\gamma}_{\boldsymbol{FCWA2}}(\boldsymbol{h})]$	67.76	68.29	69.77	70.56	70.00	69.69	69.30
$\boldsymbol{\sigma}[\hat{\gamma}_{\boldsymbol{FCWA2}}(\boldsymbol{h})]$	44.96	50.60	67.39	66.21	68.61	73.84	80.52

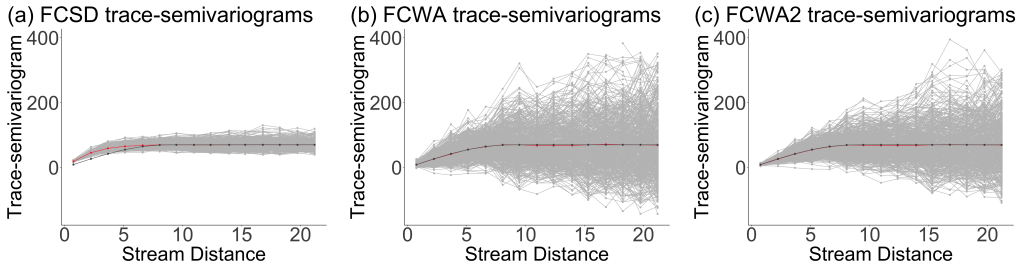

Figure 1. Plot of the (a) FCSD, (b) FCWA, (c) FCWA2 trace-semivariograms for each of the 500 simulations together with the average (red line) and the theoretical unweighted trace-semivariogram (black line).

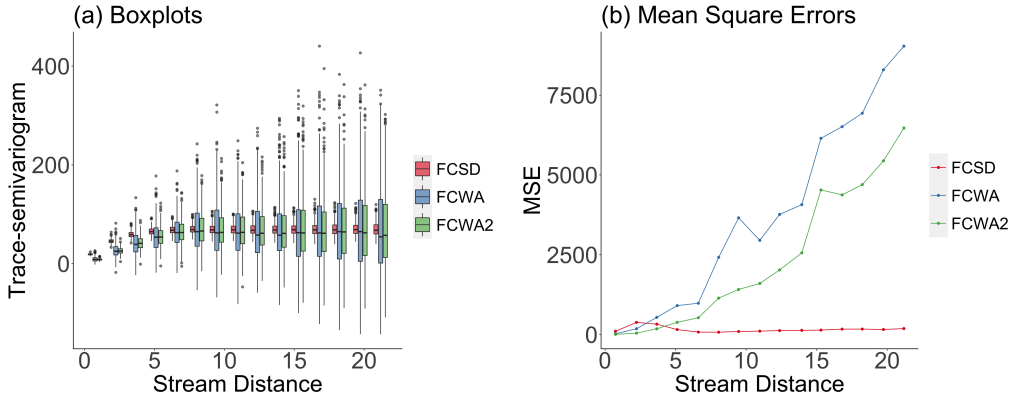

Figure 2. (a) Boxplots for the FCSD trace-semivariograms, FCWA and FCWA2 trace-semivariograms. (b) Mean Square Errors of the three estimators (FCSD, FCWA and FCWA2) for the unweighted trace-semivariograms.

weights definition). Further research is needed on alternative estimators of the unweighted flow-connected portion of the semivariogram of a pure tail-up model that balance bias and variance better than $\hat{\gamma}_{FCWA}(h_k)$ and $\hat{\gamma}_{FCWA2}(h_k)$ do.

The observations made are further confirmed by looking at Figure 2a, displaying the boxplots of the semivariograms at each lag. Here, the boxplots highlight the excessive variability of both FCWA and FCWA2 trace-semivariograms. Figure 2b shows the Mean Square Error of each estimator at each distance. The MSE associated to FCWA and FCWA2 is increasing for increasing values of the stream distance, consistent with the results available in the literature for the scalar case.

We now dive a little deeper in the results of this simulation in terms of parameters estimates. For the B=500 simulated datasets, each empirical estimator (FCSD, FCWA, FCWA2) is exploited to estimate the sill and the range (using the `R` package `gstat`, (Pebesma (2004)). Table 2 displays the relevant statistics for this simulation. To automatize the procedure, the choice of the initial paramters in the `fit.variogram` function, has been done as follows:

1. The initial partial sill was set as the median of the last four estimates of the empirical semivariogram;

2. The initial range was set as the minimum distance in which the empirical semivariogram reach the 95% of the sill.

In this simulation the nugget has not been fitted. Moreover, despite this convention, the fitting procedure implemented in `fit.variogram` did not always reach convergence: it converged 497/500 times for the FCSD fitting, 317/500 for the FCWA and 356/500 for FCWA2. The cases in which convergence was not reached has been removed from the analysis. In Table 2, FUSD refers to the sill estimate obtained by $\bar{\gamma}_{FUSD}$. Concerning this last estimate, it is clear that it outperforms the others in terms of Root Mean Square error (RMSE). Regarding the range estimates, we can appreciate the strong underestimation obtained by fitting the FCSD semivariogram (both with respect to the mean and the median). On the other hand, if FCWA and FCWA2 are more accurate on average (they sligthly overestimate the theoretical range), they have higher RMSE due to their incredible high variability. Figure 3a and Figure 3b present the distributions of these estimates. All the considerations that have been made appear even more clearly by looking at these graphs.

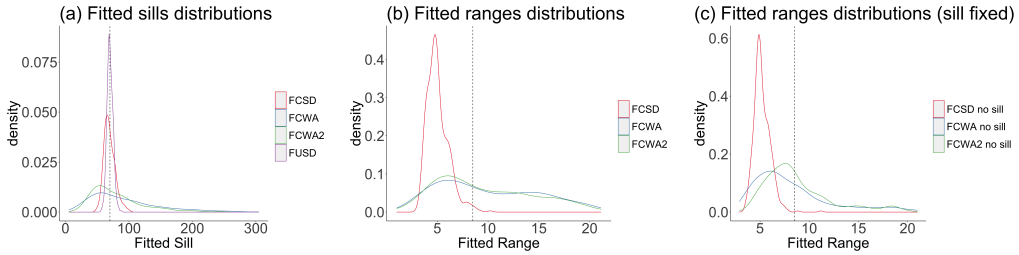

Figure 3. Distributions of the fitted ranges (a) and sills (b) obtained via FCSD, FCWA and FCWA2. (c) Distributions of the fitted ranges obtained by fitting FCSD, FCWA and FCWA2 empirical trace-semivairograms and fixing the sill to the FUSD estimate.

Table 2. Simulation results in terms of sill (θ_v) and range (θ_r) estimation.

	Theoretical			$\hat{\theta}_v$			$\hat{\theta}_r$		
Mod	θ_v	θ_r	Empirical	Median	Mean	RMSE	Median	Mean	RMSE
Sph	70	8.5	FCSD	68.05	69.35	9	4.87	5.04	3.62
			FCWA	79.14	91.28	54.62	8.89	9.98	5.17
			FCWA2	68.97	79.39	40.47	8.45	9.64	4.87
			FUSD	69.69	70.11	4.63			

Finally, encouraged by the accuracy in the sill estimation (obtained via FUSD), we exploited it in an attempt to improve the range estimate. Indeed, the theoretical models have also been fitted to the trace-semivariograms fixing the sill to the FUSD estimate. The results, reported in Table 3, are sligtly better in terms of RMSE, as far as FCWA and FCWA2 are concerned. Indeed, also by looking at Figure 3c - which displays the distributions of the so obtained estimates - it is clear that the parameters estimated via FCWA and FCWA2 exhibit a lower variance than before.

Table 3. Simulation results for the empirical trace-semivariograms in terms of range estimates.

	Theoretical			$\hat{\theta}_r$		
Mod	θ_v	θ_r	Empirical	Median	Mean	RMSE
Sph	70	8.5	FCSD	5.04	5.2	3.4
			FCWA	7.07	8.18	3.19
			FCWA2	7.9	8.81	3.47

3 Sensitivity of Kriging to range underestimation

Beyond variographic estimation and variance characterisation, our interest lies in kriging prediction. Thus, in this Section, the impact that a rough estimate of the range would entail is investigated. In particular, for each simulation, a leave-one-out cross validation approach (LOOCV) has been employed to evaluate the performance of the kriging predictor obtained by means of the parameters estimated in that specific iteration (via FCSD, FCWA and FCWA2). The procedure is summarized as follows.

For each simulation $b = 1, \dots, 500$:

1. Fit the empirical semivariogram (FCSD, FCWA or FCWA2) to the valid model and obtain the parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\theta}_v, \hat{\theta}_r)$. (provided the `fit.variogram` function reaches convergence).
2. for each location s_i, $i = 1, \dots, 200$:
   (a) remove the data point $\mathcal{X}_{s_i}$ from the b-th simulated dataset.
   (b) Retrieve the covariance structure by plugging the estimate $\hat{\boldsymbol{\theta}}$ in the theoretical parametric model. Hence, predict the removed functional data point $\mathcal{X}_{s_i}$ via Kriging interpolation of the remaining data. Let $\mathcal{X}^*_{s_i}$ be the estimated data point.
   (c) compute the error $\|\mathcal{X}_{s_i} - \mathcal{X}^*_{s_i}\|^2$

3. Compute the performance index associated to the $b-th$ simulation.

$$SSE^b = \sum_{i=1}^{200} \|\mathcal{X}_{s_i} - \mathcal{X}^*_{s_i}\|^2$$

Figure 4 displays the violin plot of the distribution of the SSE^b for the three different parameter estimation methods.

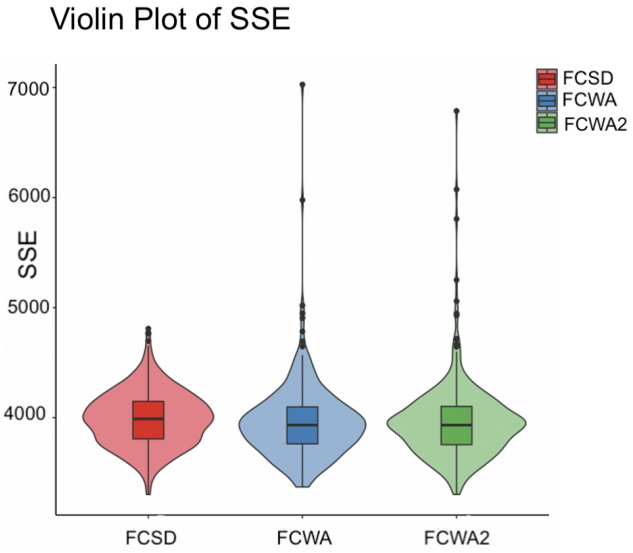

Figure 4. Violin plots (and boxplots) of the sum of square errors $SSE^b, b = 1 \dots 500$

The first thing we notice from Figure 4 is the presence of rather high outliers when the empirical semivariogram was fitted via FCWA and FCWA2. Indeed, since these estimators are featured by a strong variability, it is possible to come up with highly unlikely parameters estimates (either too high or too low). Consequently, in these extreme cases, the kriging prediction is compromised. On the other hand, FCSD shows a slightly higher median. However, at least in this specific case, the range underestimation does not seem to heavily affect the Kriging performance. Some research has been done in literature to investigate Kriging robustness. Nevertheless greater attention has been posed on the misspecification of the variogram family rather than bad parameters estimates. What emerges from the classical texts (Cressie (1993), Chilès and Delfiner (2012)) is that the range does affect the Kriging predictor, but certainly to a lesser extent than, for example, the behaviour of the semivariogram near the origin (the nugget effect).

Beyond these general considerations, the conclusion we draw from this simulation study is that the excessive variability of the weights adjusted empirical

semivariograms FCWA and FCWA2 makes them unusable from a practical point of view. The mere fact that, in a large number of cases, the fit does not reach convergence leads us to consider these estimators too unstable for real applications. Furthermore, their use does not seem to be encouraged by better predictive performance either. Therefore, pending on the considerations in Zimmerman and Ver Hoef (2017) and on the results of this analysis, we encourage the use of $\hat{\gamma}_{FCSD}(\cdot)$ without modification to characterize flow-connected dependence, despite its bias. Nevertheless it is not unusual for a theoretically biased estimator - such as FCSD - to perform better in practice with respect to an unbiased one.

References

Barbi, C., Menafoglio, A., and Secchi, P. (2022), "An object-oriented approach to the analysis of spatial complex data over stream-network domains," *MOX report*, .

Chilès, J., and Delfiner, P. (2012), *Geostatistics: Modeling Spatial Uncertainty, Second Edition*, Wiley Series in Probability and Statistics John Wiley and Sons, Inc.

Cressie, N. A. C. (1993), *Correlation Theory of Stationary and Related Random Functions, Vol. I*, Wiley Series in Probability and Statistics John Wiley and Sons, Inc.

Pebesma, E. J. (2004), "Multivariable geostatistics in S: the gstat package," *Computers and Geosciences*, 30, 683–691.

Zimmerman, D. L., and Ver Hoef, J. M. (2017), "The Torgegram For Fluvial Variography," *Journal of Computational and Graphical Statistics*, 13, 253–264.

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

31/2022 Bortolotti, T; Peli, R.; Lanzano, G; Sgobba, S.; Menafoglio, A
Weighted functional data analysis for the calibration of ground motion models in Italy

30/2022 Bonetti S.; Botti M.; Antonietti P.F.
Discontinuous Galerkin approximation of the fully-coupled thermo-poroelastic problem

29/2022 Fumagalli, I.; Polidori, R.; Renzi, F.; Fusini, L.; Quarteroni, A.; Pontone, G.; Vergara, C.
Fluid-structure interaction analysis of transcatheter aortic valve implantation

28/2022 Ciarletta, P.; Pozzi, G.; Riccobelli, D.
The Föppl–von Kármán equations of elastic plates with initial stress

26/2022 Orlando, G.
A filtering monotonization approach for DG discretizations of hyperbolic problems

25/2022 Cavinato, L; Gozzi, N.; Sollini, M; Kirienko, M; Carlo-Stella, C; Rusconi, C; Chiti, A; Ieva, F.
Perspective transfer model building via imaging-based rules extraction from retrospective cancer subtyping in Hodgkin Lymphoma

27/2022 Lazzari J., Asnaghi R., Clementi L., Santambrogio M. D.
Math Skills: a New Look from Functional Data Analysis

24/2022 Cappozzo, A.; McCrory, C.; Robinson, O.; Freni Sterrantino, A.; Sacerdote, C.; Krogh, V.; Par
A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events

23/2022 Masci, C.; Ieva, F.; Paganoni, A.M.
A multinomial mixed-effects model with discrete random effects for modelling dependence across response categories

22/2022 Regazzoni, F.; Pagani, S.; Quarteroni, A.
Universal Solution Manifold Networks (USM-Nets): non-intrusive mesh-free surrogate models for problems in variable domains