MOX-Report No. 40/2026

A Convolution Process for Sea Surface Temperature Hot-Spot Identification in the Mediterranean Sea

Marchesin, L.; Menafoglio, A.; Secchi, P.

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

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

A Convolution Process for Sea Surface Temperature Hot-Spot Identification in the Mediterranean Sea

Leonardo Marchesin[1], Alessandra Menafoglio[1] and Piercesare Secchi[1]

Abstract

Sea surface temperature (SST) is a fundamental determinant of global climate dynamics and economic activity. Reliable projections of future SST patterns depend critically on a rigorous characterization of the underlying spatial random field. In this study, we introduce a novel convolution-based covariance framework tailored to geostatistical domains constrained by physical barriers and influenced by vector-driven flows. By discretizing the continuous marine domain into a directed linear network that preserves the orientation of ocean currents, we construct a moving-average stochastic process whose dynamic is encoded via a Markovian transition-probability matrix on the network's vertices. The induced covariance structure emerges as a weighted combination of a spatial kernel and flow-dependent weights, giving rise to a complex estimation problem. To stabilize inference, we propose a penalized estimator that regularizes covariance parameters while enforcing consistency with known hydrodynamic properties. We then embed this covariance model into a Monte Carlo simulation framework to refine RCP-based SST projections and to identify thermal "hot spots" of heightened ecological risk. Our approach delivers a statistically principled framework that prevents physical inconsistencies – such as correlations across land barriers – providing a robust basis for quantifying uncertainty in future SST forecasts and for guiding targeted environmental assessments.

Keywords: Network-based spatial processes; Flow-informed covariance modeling; Moving-average processes; Sea surface temperature; Extreme events; Hot-spot identification.

1 Introduction

Sea surface temperature (SST) is a vital component in sustaining life and global prosperity. Extreme values of SST can lead to the loss of part of the biodiversity in the sea (Smith et al., 2023; Dayan et al., 2023). In particular, the Mediterranean sea is a hot-spot for climate change (Giorgi, 2006): despite its limited extent it hosts a large part of the World's marine wildlife and plants. Accurately assessing the risks associated with rising water temperatures requires moving beyond point predictions, which fail to capture the full spectrum of variability necessary for such a complex and sensitive issue. Consequently, robust statistical methods are essential for evaluating the risk of exceeding critical thresholds and conducting insightful analyses (Hazra and Huser, 2021; Bolin and Lindgren, 2015; French and Sain, 2013).

[1]MOX - Department of Mathematics, Politecnico di Milano, Milan, Italy.

Future SST projections are commonly studied using Representative Concentration Pathway (RCP) scenarios, which simulate future greenhouse gas concentrations and their impacts on global systems. These scenarios are implemented within complex climate models using ensemble methods that aim to capture the intricate interactions between greenhouse gas concentrations and Earth's features (Taylor et al., 2012). However, such projections are limited to spatial point estimates of the mean SST, restricting their ability to provide comprehensive statistical insights into potential risks and uncertainties. This work introduces a new framework for spatial statistics which allows for physics-based predictive distribution of future SST, enabling assessment and uncertainty quantification of the risks associated with rising temperatures.

Spatial statistics concerns the analysis of spatially indexed data and the dependence relationships arising therein. Traditionally, dependence among observations has been characterized via covariance functions, with most methodologies formulated under the assumption of a Euclidean distance metric (Cressie, 1993). However, when one substitutes a non-Euclidean distance into these classical parametric covariance models, the resulting covariance function may fail to satisfy the requirements of positive definiteness, producing relations that are invalid for statistical inference (Curriero, 2006).

There are many scenarios where an Euclidean framework cannot accurately represent the domain. Spatial connectivity, driven by underlying physical phenomena, may create stronger relationships between particular pairs of locations while weakening others. Additionally, discontinuities (for instance, a gap within the spatial domain) may render some connections infeasible. In such cases, the traditional Euclidean metric proves inadequate for capturing the complexities of the domain. Therefore, alternative metrics need to be employed to more effectively and comprehensively describe the relationships between locations. Water resources, in particular, represent a typical scenario where the challenges mentioned above emerge.

Figure 1 illustrates the extent of the area studied in this work, where a velocity field shapes the structure of a spatial domain. The figure displays the point estimates projections of the sea-surface temperatures in the northern Thyrrenian Sea in August 2050, with arrows representing the prevailing marine currents. These currents impose directional connectivity across the domain, and any spatial modeling framework that aims to accurately capture relationships between locations must incorporate these flow-driven constraints. Accounting for the physical structure induced by the velocity field is essential for a more faithful representation of spatial dependence.

This work introduces a novel framework for developing spatial statistical models that explicitly incorporate an underlying velocity field governing the domain. The objective is to capture and represent domain connectivity through a covariance structure informed by the flow dynamics, thereby enabling spatial statistical analysis, uncertainty quantification, and simulation studies that reflect the physical interactions implied by geographic configuration.

The first step in the analysis involves defining a suitable representation of the spatial domain that accounts for the directionality and connectivity induced by the current field. To this end, the linear network representation is adopted, as it provides a natural and effective structure for modeling connected domains. This representation captures the directional and relational characteristics of the system, with the flow dynamics encoded directly through the network's edges, preserving the physical coherence of the spatial dependencies.

Recent advancements in spatial statistics have rigorously formalized the theory of random fields on metric graphs (Anderes et al., 2020; Bolin et al., 2024). These foundational works provide valid covariance models based on the geodesic metric, effectively overcoming the limitations of

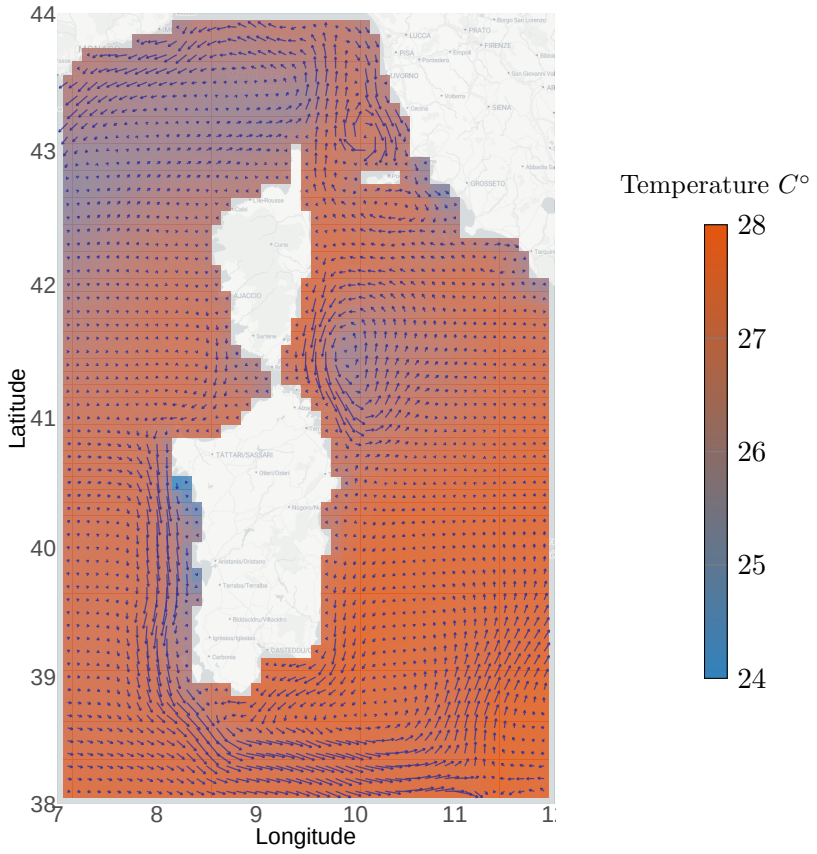


Figure 1: The spatial domain of the northern Tyrrhenian Sea is depicted with the point estimates projections of mean sea-surface temperatures in August 2050, provided by the RCP 4.5 emissions pathway. The arrows represent the velocity fields. Sardinia, Corsica, and the Italian peninsula coastline are clearly recognizable.

Euclidean distances in non-convex domains. However, standard metric graph models are typically symmetrical, assuming that dependence relies solely on the distance along the graph, regardless of direction. While this assumption holds for many diffusive processes, it does not account for the strong advection characteristic of sea currents.

To model spatial dependence over complex, irregular and physics-driven domains, one widely used framework is based on stochastic partial differential equations (SPDEs) (Bolin and Lindgren, 2011; Lindgren et al., 2011; Clarotto et al., 2024). These methods construct Gaussian random fields with Matérn-type covariance structures by solving SPDEs, providing a principled bridge between continuous-domain models and sparse representations on triangulated meshes. This framework has proven effective in many applications, particularly due to its computational efficiency and scalability. However, classical SPDE approaches typically rely on symmetric differential operators, which naturally yield spatially symmetric covariance structures. This assumption presents challenges when modeling processes on directed networks, where edge asymmetry breaks operator self-adjointness. Moreover, extending SPDEs to accommodate nonstationarity, anisotropy, or complex boundary behavior often requires careful design of the differential operator and mesh structure. While there are emerging efforts to relax these assumptions, a general and practical framework for covariance modeling on directed networks remains underdeveloped.

Another closely related methodology is spatial smoothing with differential regularization proposed by Clemente et al. (2026); Tomasetto et al. (2024). This method applies spatial smoothing to the available data, considering a differential penalization that incorporates possible prior knowledge about the phenomenon, expressed through a partial differential equation (Ramsay, 2002; Azzimonti et al., 2015). However, unlike geostatistical methods, spatial regularization does not provide a spatial covariance function, but rather captures the spatial variability through a deterministic function, estimated through smoothing. Having an (estimated) covariance model at our disposal is crucial not only for enabling spatial prediction but also for simulating multiple realistic scenarios—an essential requirement in studies focused on uncertainty quantification and scenario analysis. This ability to generate diverse and plausible scenarios captures the inherent variability in the data and gives geostatistical methods a significant advantage over purely smoothing-based approaches.

In this work, we propose a novel class of valid covariance models for directed linear networks, based on a new construction of a moving-average-type convolution process. Our approach is inspired by the moving average framework developed for stream network systems (Ver Hoef et al., 2006; Peterson et al., 2007; Cressie et al., 2006), where the process value at a given vertex is defined as a local average of random noise weighted by a spatial kernel. In the proposed framework the flow dynamics governing the moving average construction are modeled stochastically, through a Markov chain defined on the network. This stochastic encoding of the transport mechanism allows the model to naturally propagate the information about the velocity field into the covariance structure.

The remainder of this work is organized as follows. Section 2 introduces the northern Tyrrhenian Sea domain of analysis and its approximation as a linear network. Section 3 defines the proposed convolution-based process and presents the functional form of the resulting covariance model. Section 4 outlines the covariance estimation procedure. Section 5 summarizes the results of the simulation study. Section 6 applies the methodology to the spatial field defined by the sea surface temperature of the northern Tyrrhenian Sea, estimates and analyzes its spatial dependence structure and uses it to simulate multiple realizations of the sea surface temperature, leveraged for conducting an extreme event analysis. Proofs of all theoretical results, along with a simulation study and detailed algorithmic specifications, are provided in the Appendix.

2 Representing a physics-driven spatial domain

2.1 SST and current data

The Copernicus Climate Change Service (C3S, Copernicus Climate Change Service, 2020) provides Representative Concentration Pathway (RCP) water temperature projections up to 2099, evaluated using the European Regional Seas Ecosystem Model (ERSEM, Butenschön et al., 2016). These projections offer high-resolution data on various environmental variables, including sea surface temperature (SST) and sea currents, across the entire northern hemisphere.

Representative Concentration Pathways (RCPs) are named according to the range of radiative forcing values they are expected to achieve by the year 2100 (van Vuuren et al., 2011). In particular, we consider the RCP 4.5. It is described as an intermediate scenario. It is more optimistic than the current situation, as it assumes a moderate reduction of emissions.

For our analysis, we employ monthly mean SST for August, which represents the period of maximum thermal stress on marine ecosystems, both for the sea currents and the SST. We focus

on the northern Mediterranean basin (specifically, the northern Tyrrhenian Sea), as this region may provide thermal refuge for marine organisms when other areas experience extreme warming during peak summer conditions. Understanding projected temperature changes in this region is essential for assessing habitat persistence under future climate scenarios (Heino et al., 2009).

Figure 1 illustrates the spatial extent of the study area. It presents projected August sea temperature and velocity fields for the mid-twenty-first century (2050), serving as representative snapshots of future warming under the RCP 4.5 emissions pathway. The available temperature projections are point estimates, which, while informative, are insufficient on their own for deriving statistically robust conclusions about critical regions. To enhance inferential reliability, a detailed analysis of the associated spatial dependence – and therefore, uncertainty – is essential.

Specifically, let $\{Y_s,\ s \in D\}$ be a Gaussian random field representing sea surface temperature (SST) over the analysis domain D (Tandeo et al., 2011). We observe this field at locations $\{s_1, \ldots, s_n\}$, yielding the partial realization $\{Y_{s_1}, \ldots, Y_{s_n}\}$. The field's mean surface is prescribed by the Copernicus SST projections. The primary objective of this work is to characterize the covariance structure of SST. To this end, we compute yearly empirical residuals by comparing Copernicus projections with observed SST data. Indeed, the projection period begins in 2006 and overlaps with available satellite data that can be found in the Copernicus Marine Environment Monitoring Service (CMEMS E.U. Copernicus Marine Service Information, 2020); the observations are provided until 2022. For this study, the selected CMEMS data set describes temperature and currents field over the Mediterranean Sea, and it is available on a regular grid. These data can be used to compute empirical residuals between RCPs and observations, which are then integrated with observed current-velocity measurements to inform the estimation of a physics-informed spatial covariance model.

Since the data are provided on a regular grid, it is natural to represent D as a linear network, where each grid node is a vertex and the current field induces a directed edge structure between adjacent neighbors. This construction requires no boundary conditions, which are instead needed by PDE-based approaches and are particularly difficult to specify reliably over the irregular coastal geometries of the Mediterranean basin.

2.2 Linear network domain

A directed linear network in $\mathbb{R}^2$ is defined as a pair $(\mathcal{L}, V)$ where $\mathcal{L} = \cup_i l_i$ is the finite set of all directed edges that make up the network and $V \subset \mathbb{R}^2$ is the finite set of vertices. Each edge is defined as $l = l_{[a,b]} = \{u \in \mathbb{R}^2 : u = ta + (1-t)b;\ 0 \leq t \leq 1;\ a, b \in V\}$. The intersection of any two edges is either empty or a vertex. Let $\mathcal{N} = V \cup \mathcal{L}$ to be the whole domain of points in $\mathbb{R}^2$ obtained as the union of all vertices and points on the edges of the linear network $(\mathcal{L}, V)$. Hence, any $u \in \mathcal{N}$ either lies on some unique edge $l_{[a,b]}$, or is a vertex.

In fact, a linear network may more generally be defined in any metric space, with vertices embedded in that space and edges representing connections between them. For what follows, a fundamental attribute of each edge is its length as measured by the chosen metric. The specific metric is problem-dependent – any appropriate distance function may be employed, and multiple metrics can even coexist within the same ambient space. In this work, we adopt the edge length – $length(l_{[a,b]}) = \|a - b\|_2$, $\|\cdot\|_2$ denoting the Euclidean distance – as the distance metric between the vertices of an edge.

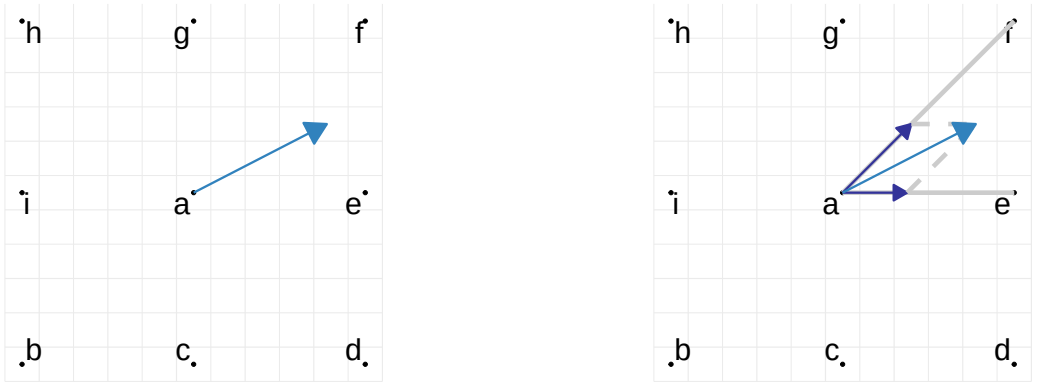
A path $p = p_{(v,x)}$ between two vertices v and x in V is an ordered finite sequence of connected edges in $\mathcal{L}$ such that $p_{(v,x)} = \{l_i = l_{[a_i,b_i]} \in \mathcal{L},\, i = 1, ..., n : a_1 = v;\, b_n = x; b_i = a_{i+1}\}$. The length of the path is the sum of the lengths of all edges of the path, i.e., $length(p_{(v,x)}) = |p_{(v,x)}| = \sum_{i=1}^{n} \|a_i - b_i\|_2$. In the following, we denote by $\mathcal{P}(v, x)$ the set of all possible paths from vertex v to x. Note that in a directed network the set of paths in different directions can be different – i.e., the paths from v to x might be different than those from x to v. If there are no paths connecting v and x, then $\mathcal{P}(v, x)$ is the empty set. Finally, observe that cyclic paths may occur. For instance, the set $\mathcal{P}(v, x)$ contains as distinct elements both the path $p_{(v,x)}$ and the path $p_{(v,x)} \cup p_{(x,x)}$.

More in general, the natural definition of the sub-path $p_{(u,x)}$ between a point in the network $u \in \mathcal{N}$ and a vertex $x \in V$ goes as follows. Let $u \in l_{[v,b_1]}$. If $b_1 = x$, then $p_{(u,x)} = [u, x]$ is the sub-edge and its length is $|p_{(u,x)}| = lenght([u, x]) = \|u - x\|_2$. If $b_1 \neq x$, let $p_{(v,x)} = \left\{l_{[v,b_1]}, l_2 \ldots, l_n\right\}$ and set $p_{(u,x)} = \{[u, b_1], l_2, \ldots, l_n\}$. Denoting $\{l_2, \ldots, l_n\}$ as $p_{(b_1,x)}$, we say that $p_{(b_1,x)} \subseteq p_{(u,x)} \subseteq p_{(v,x)}$. Finally, we define the length of the sub-path $p_{(u,x)}$ as $|p_{(u,x)}| = length([u, b_1]) + |p_{(b_1,x)}|$. Note that, although individual edge lengths coincide with Euclidean distances, the distance between a point $u \in \mathcal{N}$ and a vertex $x \in V$ constrained to the network topology may differ substantially from the direct Euclidean distance of the two points in $\mathbb{R}^2$.

To construct a linear network $(\mathcal{L}, V)$ approximating the water domain D represented in Figure 1, we outline a procedure for defining its vertices and edges, which leverages the regularity of the grid $\mathcal{G}$ on which both current and temperature data are collected. The vertices V correspond to the nodes $\{s_1, ..., s_n\}$ of $\mathcal{G}$ within the water domain D. Note that only those nodes for which we observe a temperature value are included, thus explicitly excluding non-water regions of the domain. The edges in $\mathcal{L}$ are designed to uphold the directionality dictated by the velocity field, ensuring that the network faithfully represents the natural connectivity of locations linked by water flow. Specifically, on a regular grid each node has eight surrounding neighbors, as illustrated in Figure 2a. An edge connects a vertex $a \in V$ to one of its eight adjacent nodes on the grid. Drawing inspiration from river topography (Tarboton, 1997; Tesfa et al., 2011), the non-null velocity vector $\mathbf{v}_a$ observed at the vertex a naturally suggests two likely flow directions – those most closely aligned with its orientation. The two directed edges are defined as the vectors along these two directions, whose sum – according to the parallelogram rule – equals $\mathbf{v}_a$. As illustrated in Figure 2b, these connect a to the two most adjacent nodes, e and f in the figure. We include the edge $l_{[a,e]}$ in $\mathcal{L}$ as an edge of the network if and only if $e \in V$, and the same goes for $l_{[a,f]}$. If e or f do not belong to V, the vertex a is called an *outlet*. If the vertex a has no incoming edges, it is called a *source*. Note that a vertex can simultaneously be a source and a outlet. Repeating this procedure for all vertices in V, a complete directed linear network $(\mathcal{L}, \mathcal{V})$ is constructed, as represented in Figure 3 where we use different colors to identify sources and outlets. The network corresponds to current data in Figure 1.

In this work, velocity and temperature data are co-located on the same regular rectangular grid. We exploit this co-location in constructing the network representation. More generally, when observations are not co-located or are irregularly spaced, one could define a sufficiently fine grid such that the observations lie on it. Extension to such non-co-located settings will be explored in future work.

The connectivity patterns induced by the sea currents drive the network construction. We focus on the direction of the velocity field in each point, hence overlooking the magnitude of the velocity of these vectors. If this additional information were to be incorporated, we could account for both the Euclidean distance between vertices and the velocity of the water in that direction by setting the edge length as proportional to the time required for the water flow to travel the edge



(a) 8 locations around the point of the grid.

(b) Definition of the edges.

Figure 2: Construction of the edges.

– namely, proportional to the ratio between the Euclidean distance of the edge and the magnitude of the corresponding water velocity vector. Note that this change would not affect the following construction.

We do not impose restrictive boundary conditions on the perimeter vertices; in particular, we allow water to both exit and enter the domain. This choice reflects the physical structure of the system: the study area, as illustrated in Figure 1, is inside of a bigger system, comprised of the surrounding seas and the open ocean. Consequently, the domain D should be regarded as an open system, with water entering from external sources and necessarily flowing outward beyond its boundaries without any specific constraint. Thus the network has some specific points, the *sources*, from which water come in from outside, and some points from which the water flows outside D, the *outlets*.

It should be noted that, although the topology of the linear network has a clear influence on the subsequent covariance structure, the mathematical construction of the following sections does not rely on the specific procedure adopted here to encode through a graph the physics-driven knowledge of the case study under consideration. Indeed, any alternative procedure capable of producing a meaningful linear directed graph could be employed.

2.3 A Markov chain representation

Crucial to the following construction is the definition of how the physical phenomenon underlying the spatial system influences the dynamics within the domain supporting the random process. In this work, the physics of water within the linear network is modeled as a Markov chain $(V \cup \{S\}, \pi)$, whose state space is the union of the set V of vertices of the network with an external absorbing state S, called the *sink*. At each vertex, multiple potential paths forward may exist; each edge of the network is thus assigned a transition probability by the transition matrix π. The Markovian framework provides a simple yet powerful approach for modeling network dynamics, preserving the system's complexity and generality. Moreover, this approach allows an intuitive understanding of the network's behavior.

To construct the Markov chain, we start by defining its transition probability matrix π. For every vertex $a \in V$, the velocity $\mathbf{v}_a$ is decomposed along the two directions identified by the rule

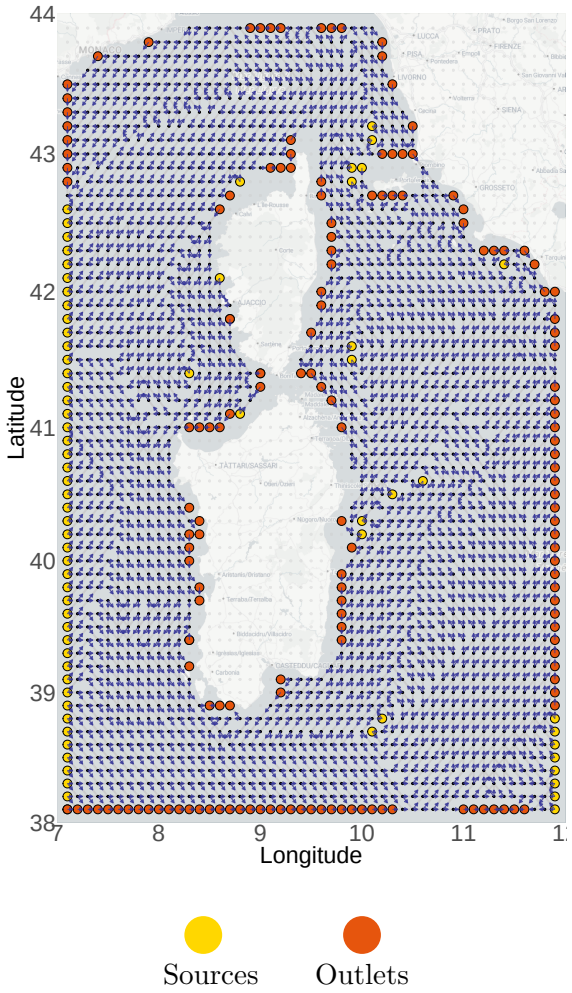


Figure 3: The linear network in the whole domain of analysis. In blue the directed edges composing the network are represented. Grey points unconnected with the others are non-water locations, explicitly excluded from the domain of analysis. The sources and the outlet points are highlighted.

illustrated in Figure 2. The magnitudes of the resulting two components, normalized by their sum, are collected in a weighted adjacency matrix M. Details on the computation of M are reported in the Appendix. If a component identifies an edge $l_{[a,e]} \in \mathcal{L}$ of the network, the transition probability $\pi_{[a,e]}$ is defined as its normalized magnitude. If a component points to a node of the grid $\mathcal{G}$ that does not belong to D, – i.e. a is an outlet – then its normalized magnitude is added to $\pi_{[a,S]}$. Finally, we set $\pi_{[S,S]} = 1$. All remaining transition probabilities are set to zero. Note that each row of π sums to one. This construction captures the directionality of the underlying velocity field in a manner consistent with the network's topology. Let $\{X_n\}_{n\geq 0}$ be a Markov chain on $V \bigcup \{S\}$ with transition matrix π.

As detailed in Section 2.2, water enters the system from external sources and eventually flows out. This implies that all vertices of V are transient states of $\{X_n\}_{n\geq 0}$, with S as the unique absorbing state.

We aim at defining the random path connecting two vertices $v, x \in V$ in the network. Fix $v \in V$

and condition on $X_0 = v$. Then $\{X_n \mid X_0 = v\}_{n\geq 0}$ is the random trajectory of the Markov chain starting from v. Given the starting point v, let $L_x = \sup\{n \geq 0 : X_n = x\}$ be the last hitting time of x. Since x is transient, the chain visits x at most finitely many times, with probability one. Now set $T(v, x)$ to be the random path that records the trajectory from v up to the last visit to x. Formally, if $L_x > 0$,

$$T(v, x) = \big(l_{[X_0=v,X_1]}, \ldots, l_{[X_{L_x-1}, X_{L_x}]}\big),$$

which is an element of $\mathcal{P}(v, x)$, equipped with the discrete σ-algebra; naturally, $T(v, x) = \emptyset$, if $L_x = 0$. Note that $T(v, x)$ may trace a cyclic path, consistently with the definition of $\mathcal{P}(v, x)$.

For $p_{(v,x)} \in \mathcal{P}(v, x)$,

$$\mathbb{P}\big(T(v, x) = p_{(v,x)}\big) = \left(\prod_{l_{[a,b]} \in p_{(v,x)}} \pi_{[a,b]}\right) U(x), \tag{1}$$

where the product aggregates the transition probabilities $\pi_{[a,b]}$ along the edges composing the path $p_{(v,x)}$ (counted with multiplicity for cyclic paths), and $U(x)$ denotes the probability of never returning to x after arrival. Under transience, $U(x) > 0$ for all $x \in V$. If x is an *acyclic* vertex – meaning no path connects x to itself – then the return probability is zero, implying $U(x) = 1$.

We define a path-dependent distance, denoted by $dist_{T(v,x)}(\cdot, \cdot)$, by evaluating the length along a specific realization of the random path connecting two vertices. Specifically, given two vertices $v, x \in V$, consider the realization $T(v, x) = p_{(v,x)} = \{l_1, \ldots, l_n\}$ and, for $u \in l_1$, the sub-path $p_{(u,x)} \subseteq p_{(v,x)}$. Then, set $dist_{T(v,x)}(u, x) = |p_{(u,x)}|$. We set the distance to infinity in case the vertices are not connected according to $T(v, x)$ – i.e. $dist_{T(v,x)}(\cdot, x) = +\infty$ if $T(v, x) = \emptyset$.

Finally, for ease of exposition, let us assume that the collection of Markov chains starting from different vertices $v \in V$ are independent. In fact, this condition does not impose practical restrictions on the resulting covariance model, as further remarked in Section 3.

3 A convolution process on linear networks

3.1 Moving average construction

The primary objective of this section is to derive a novel, valid covariance structure over the topologically complex domain of a linear network. It is important to note that our primary interest lies not in modeling the underlying generative process itself, but in obtaining a flexible and mathematically valid spatial dependence structure. To achieve this, we introduce a moving average convolution process building upon the stream network strategies of Ver Hoef et al. (2006) and Ver Hoef and Peterson (2010). We utilize this process to deduce a closed-form, positive-definite covariance function.

Choosing a linear network $\mathcal{N}$ as the domain for this convolution process offers the key analytical advantage of evaluating spatial contributions via line integrals. By constraining the velocity field to the edges and vertices, the formulation naturally admits an integral representation along the network structure. While this moving average construction is inspired by the dendritic stream network models of Ver Hoef and Peterson (2010), we integrate an additional source of randomness. This extension is necessary to accommodate the greater topological complexity—such as cycles

and multidirectional connectivity—inherent to general ocean current networks compared to strictly directed streams.

In the following construction, we denote by W a Wiener process independent of the Markov process $\{X_n\}_{n\geq 0}$, introduced in the previous section, and by $\int_A [\cdot]\, W(du)$ the stochastic integral over $A \subset \mathbb{R}$. We let g be a square-integrable function, which will play the role of moving average function. Finally, for $v, x \in V$, we denote by X_1 the (random) right vertex of the first edge in the path $T(v, x)$.

Given a family of positive parameters β_p indexed by paths between vertices in V, we now define a process $\{Z_x, x \in V\}$ on the network $\mathcal{N}$. For $x \in V$, set

$$Z_x = \sum_{v \in V} \int_{L_{[v,X_1]}} \frac{g\left(dist_{T(v,x)}\left(u, x\right)\right)}{\sqrt{\beta_{T(v,x)}}} W(du). \tag{2}$$

In (2), the integral is constrained to the (random) segment $l_{[v,X_1]}$, ensuring that each vertex $v \in V$ contributes exactly once to the process, and only if it is connected to x via the path $T(v, x)$. Moreover, the random paths $T(v, x)$'s influence both the interval of integration and the distance between points in the network. Square-integrability of g guarantees the well-definition of (2). The role of $\beta_{T(v,x)}$ will become clearer in the following subsections, where it plays a key part in ensuring a relaxed condition for stationarity.

For a more thorough interpretation of the process, the terminology of upstream and downstream points will be borrowed from works on stream networks, together with the specific type of images depicting the effects of the moving averages functions (Ver Hoef et al., 2006; Peterson et al., 2007; Ver Hoef and Peterson, 2010). A single realization of the process Z_x can be viewed as representing an individual unit within a larger flow. The random quantity modeled by the process corresponds to this unit, which evolves according to the specific realizations of the random variables T's. When the moments are evaluated, "global" measures for the whole flow are deducted.

For instance, refer now to Figure 6. The dependency is represented in the picture by the colored areas, picturing the moving average kernels. A split v^* in the linear network generates a dependence between the vertex v before the split and the (possibly more than two) vertices after it, say x_1 and x_2 – where the terms *before* (resp. *after*) and *upstream* (resp. *downstream*) are given according to the direction of the flow. The flow passing through $l_{[v,v^*]}$ influences the two downstream vertices x_1 and x_2 differently. According to the realizations of the random variables $T(v, \cdot)$'s, at each split, each individual unit of the flow in the network follows one and only one direction. In Figure 6a, the unit selects path toward the vertex x_1, hence leading to the realizations $T_{(v,x_1)} = \{l_{[v,v^*]}, l_{[v^*x_1]}\}$, and $T_{(v,x_2)} = \emptyset$. Thus, the random value of the process in x_1, i.e. Z_{x_1}, depends on the Wiener process defined over $l_{[v,v^*]}$, while Z_{x_2} does not. Conversely, in Figure 6b a different realization is pictured, and Z_{x_2} depends on the Wiener process of $l_{[v,v^*]}$, while Z_{x_1} does not. When two points are not connected – i.e., $T(v, x) = \emptyset$ – their distance is set to infinite, which removes any contribution from the upstream vertex to the random value of the process at the downstream vertex.

The following propositions provides the expressions of the moments of the process $\{Z_x, x \in V\}$ defined in Equation (2). Proofs are reported in the Appendix.

Proposition 1. *The process $\{Z_x, x \in V\}$ is zero-mean: for $x \in V$, $\mathbb{E}[Z_x] = 0$.*

We evaluate the moments of the process in Equation (2) by first exploiting the stochastic integral, and secondly evaluating the distribution of the $T(v, \cdot)$.

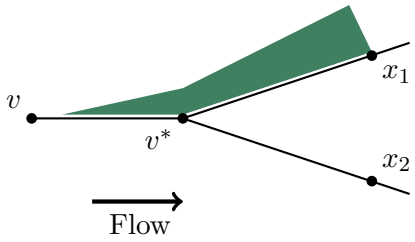


Figure 4: (a)

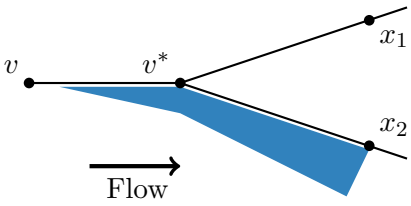


Figure 5: (b)

Figure 6: A split v^* in the network. The two areas green (up) and blue (down) represent the moving average kernels of the two random values defined in the two downstream vertices (x_1 and x_2 respectively). At the split, the Markov chain will select either one of the two vertices downstream. In (a), $T_{(v,x_1)}(\omega) = p_{(v,x_1)}$ where $p_{(v,x_1)}$ is the path connecting v and x_1, while $T_{(v,x_2)}(\omega) = \emptyset$. Conversely, in (b) the opposite happens: $T_{(v,x_1)}(\omega) = \emptyset$; $T_{(v,x_2)}(\omega) = p_{(v,x_2)}$.

Proposition 2. *For x, y in V,*

$$Cov(Z_x, Z_y) = \sum_{v \in V} \mathbb{E}\left[\int_{L_{[v,X_1]}} \frac{g\left(dist_{T(v,x)}(u,x)\right) g\left(dist_{T(v,y)}(u,y)\right)}{\sqrt{\beta_{T(v,x)}\beta_{T(v,y)}}} du\right]. \tag{3}$$

In particular,

$$Var(Z_x) = Cov(Z_x, Z_x) = \sum_{v \in V} \mathbb{E}\left[\int_{L_{[v,X_1]}} \frac{g^2\left(dist_{T(v,x)}(u,x)\right)}{\beta_{T(v,x)}} du\right]. \tag{4}$$

Note that the variance of the process in Equation (4) is finite. The sum is on a finite number of vertices, and each term is given by finite integrals – because of the square-integrability of g.

Importantly, the assumption of path independence does not limit the flexibility of our model, which aims at specifying a valid second-order covariance structure on linear networks. In the convolution framework, integrating against the Wiener process effectively neutralizes any joint dependence between any two paths $T(v_1, x)$ and $T(v_2, x)$ during the evaluation of the second moment (see the proof of Proposition 2). The resulting spatial covariance depends exclusively on the marginal path probabilities.

Given the non-Euclidean topology of the domain and the structure of the covariance in (3), enforcing classical second-order stationarity is generally infeasible. Instead, inspired by approaches used in stream network models (Ver Hoef et al., 2006), we adopt a relaxed notion of stationarity based on spatial homogeneity of the first two moments–specifically, constant mean and constant variance.

3.2 Weighted covariance model

We enforce stationary variances following the approach of (Ver Hoef et al., 2006; Ver Hoef and Peterson, 2010). The parameters β_p's are used to this end. Indeed, for $v, x \in V$, we specify $\beta_{p_{(v,x)}}$

as depending on the total probability mass in the end points of the edges composing the path, and on the probability of not returning to the endpoint:

$$\beta_{p_{(v,x)}} = \left(\prod_{l_{[a,b]} \in p_{(v,x)}} \Big[\sum_k \pi_{[k,b]} \Big] \right) U(x). \tag{5}$$

Note that, for any edge $l_{[a,b]}$ belonging to a path with a non zero probability of realization, the factor $\sum_k \pi_{[k,b]} > 0$. Moreover, as observed in Section 2.3, $U(x) > 0$ for each $x \in V$.

The following results provide the expressions for the variance and covariance of the random process defined in (2), under the specification of the normalization constants $\beta_{p(v,x)}$ set in (5).

Proposition 3. *Let $(\mathcal{L}, V)$ be a linear network, $\{Z_x, x \in V\}$ be the process defined in Equation (2) and the normalization constants β's be defined as in Equation (5). Then for every $x \in V$ the marginal variance is constant and equals*

$$Var(Z_x) \;=\; \int_0^{+\infty} g^2(r)\, dr. \tag{6}$$

Note the the variance in Equation (6) is finite thanks to the square-integrability of the moving average function g, as noted in Section 3.1. More generally, constant marginal variance holds for any choice of the constants β's inducing a *unit-influx* normalization such that $\sum_k \pi_{[k,x]}/\beta_{p(k,x)} = 1$, analogously to the river network literature (Ver Hoef et al., 2006; Cressie et al., 2006).

For $x, y \in V$, we denote by $U(x, y)$ the probability of not returning to either x or y, when the chain starts from x. In addition, let $\tilde{\mathcal{P}}(x, y)$ denote the set of paths from x to y that do not contain cycles at the initial vertex x. Notably, for an acyclic point x^*, $\tilde{\mathcal{P}}(x^*, y) = \mathcal{P}(x^*, y)$.

Proposition 4. *Let $(\mathcal{L}, V)$ be a linear network, $\{Z_x, x \in V\}$ be the process defined in Equation (2) and the normalization constants β's be defined as in Equation (5). Let $x, y \in V$, $x \neq y$.*

(i) If $\tilde{\mathcal{P}}(x, y) \bigcup \tilde{\mathcal{P}}(y, x) \neq \emptyset$,

$$Cov(Z_x, Z_y) = \sum_{p \in \tilde{\mathcal{P}}(x,y) \bigcup \tilde{\mathcal{P}}(y,x)} w_p C(|p|),$$

where, for any two vertices $v_1, v_2 \in V$,

$$w_{p_{(v_1,v_2)}} = \left(\prod_{l_{[a,b]} \in p_{(v_1,v_2)}} \frac{\pi_{[a,b]}}{\sqrt{\sum_k \pi_{[k,b]}}} \right) \frac{U(v_2, v_1)}{\sqrt{U(v_1)\, U(v_2)}}, \tag{7}$$

and $C(\cdot)$ is the spatial covariance function defined, for $h \geq 0$, as

$$C(h) = \int_0^{+\infty} g(r) g(r + h) dr.$$

(ii) If $\tilde{\mathcal{P}}(x, y) \bigcup \tilde{\mathcal{P}}(y, x) = \emptyset$,

$$Cov(Z_x, Z_y) = 0.$$

The covariance structure is highly sensitive to the choice of the moving average function. In particular, specific forms of g result in classical parametric expressions for the spatial covariance function, analogous to those introduced in Ver Hoef et al. (2006); Peterson and Ver Hoef (2010); Barbi et al. (2023). Table 1 illustrates three common choices used as spatial covariance function $C(\cdot)$, with their corresponding moving average kernels g. These representations of C depend on two parameters which have a similar role as those used in classical geostatistics (Cressie, 1993): the range $\theta_r > 0$, which scales the distance between vertices, and the sill $\theta_s > 0$, which sets the overall variance at any given vertex. Indeed, note that by setting $\int_0^{+\infty} g(r)^2 dr = \theta_s$ the unique parameter controlling the amount of variability in the covariance structure is the sill parameter $C(0) = \theta_s$.

With this notation, for any two vertices x and y, the covariance model of the process defined in Equation (2), with the family of parameter β's defined as in Equation (5) can be represented as

$$Cov(Z_x, Z_y) = \begin{cases} C(0) = \theta_s & \text{if } x \equiv y, \\ 0 & \text{if } x \neq y;\ \tilde{\mathcal{P}}(x,y) \bigcup \tilde{\mathcal{P}}(y,x) = \emptyset, \\ \sum_{p \in \tilde{\mathcal{P}}(x,y) \bigcup \tilde{\mathcal{P}}(y,x)} w_p C(|p|) & \text{if } x \neq y;\ \tilde{\mathcal{P}}(x,y) \bigcup \tilde{\mathcal{P}}(y,x) \neq \emptyset. \end{cases} \tag{8}$$

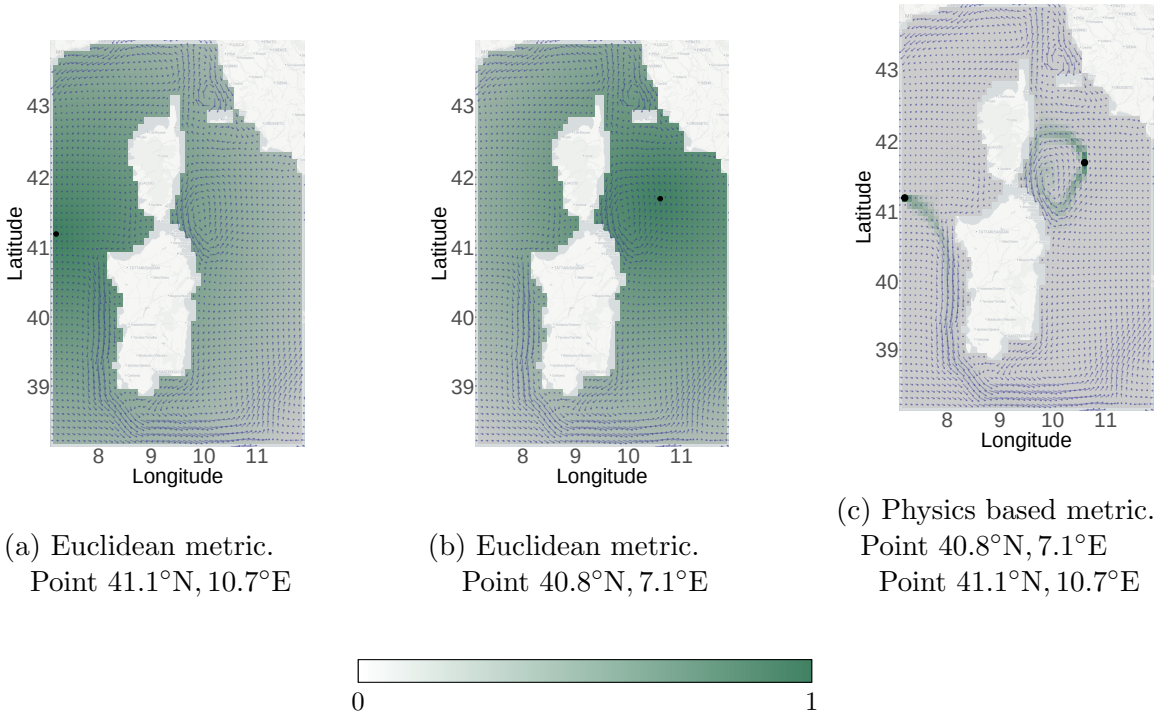
The formulation in Equation (8) recovers the weighting structure typical of stream network models (Ver Hoef et al., 2006; Ver Hoef and Peterson, 2010). In particular, the model is consistent with the stream network construction when the domain is restricted accordingly, as shown in the Appendix. We note that we here assume the range θ_r and the sill θ_s to be spatially homogeneous. The extension to spatially varying parameters is left to future research.

Note that the covariance between pairs of vertices depends solely on the set of possible paths connecting them within the linear network. In particular, it is determined by the path weights and their associated distances. This construction enables the direct encoding of classical geostatistical effects – such as scale and marginal variance – onto the network domain itself, without relying on Euclidean embeddings or latent spatial representations. As a result, the model naturally adapts to the topology of the network, preserving interpretability while remaining intrinsic to the domain.

To illustrate the distinction between our network-based approach and conventional Euclidean modeling, Figure 7 presents covariance heatmaps for both frameworks. Both use an exponential covariance with $\theta_s = 1$ and range set to ten times the maximum distance in each metric. Reference points are located on the western (40.8°N, 7.1°E) and eastern (41.1°N, 10.7°E) coasts of Sardinia and Corsica. The Euclidean model assumes isotropy, with covariance depending solely on straight-line distance. This produces unrealistic dependencies: points on opposite sides of Sardinia exhibit strong covariance despite having no hydrological connection due to the island barrier. In contrast, the network-based model respects directional constraints imposed by sea currents, yielding anisotropic covariance structures that reflect true marine connectivity. The resulting patterns

Table 1: Traditional moving average and spatial covariance function.

	Moving average function	**Spatial covariance function**
Linear with sill	$g(r) = \sqrt{\theta_s}\, \mathbb{I}_{\{0 \leq r/\theta_r \leq 1\}}$	$C(h) = \theta_s\,(1 - h/\theta_r)\mathbb{I}_{\{0 \leq h/\theta_r \leq 1\}}$
Spherical	$g(r) = \sqrt{3\,\theta_s}(1 - r/\theta_r)\mathbb{I}_{\{0 \leq r/\theta_r \leq 1\}}$	$C(h) = \theta_s(1 - \frac{3}{2}h/\theta_r + \frac{1}{2}h^3/\theta_r)\mathbb{I}_{\{0 \leq h/\theta_r \leq 1\}}$
Exponential	$g(r) = \sqrt{2\,\theta_s}e^{-r/\theta_r}\mathbb{I}_{\{0 \leq r/\theta_r\}}$	$C(h) = \theta_s e^{-h/\theta_r}$



(a) Euclidean metric. Point 41.1°N, 10.7°E

(b) Euclidean metric. Point 40.8°N, 7.1°E

(c) Physics based metric. Point 40.8°N, 7.1°E Point 41.1°N, 10.7°E

Figure 7: Spatial heatmaps for different covariance models. The color gradient indicates the covariance of each location with respect to the reference point (black dot). In the proposed framework, the effects of both influence points are superimposed in a single map, whereas the Euclidean framework requires two separate maps to avoid visual interference between the individual influence points.

are shaped by flow dynamics, capturing spatial dependencies consistent with underlying physical transport processes.

4 Estimation of the covariance structure

When data become available, the primary goal of a geostatistical analysis is to estimate the covariance structure of the underlying random field. However, the specific form introduced in Equation (8) presents challenges that make classical estimation procedures difficult to apply directly. In particular, the summation over paths complicates the use of standard empirical covariances based on point pairs, as it introduces ambiguity in attributing observed covariance to specific distances.

Weighted covariance structures arise in stream network geostatistics (Ver Hoef and Peterson, 2010), where efforts to develop unbiased estimators incorporating these weights have faced significant difficulties. Such estimators tend to exhibit high variability and have proven unreliable in practice, leading the literature to favor biased but more stable unweighted alternatives (Zimmerman and Ver Hoef, 2016). In this work, we construct a novel penalized estimator that explicitly accounts for both the path summation in the covariance model and the associated weights, using penalization to ensure stability and reliability.

We develop an estimation algorithm assuming the stationary variance condition introduced in Section 3.2. This assumption is implemented by defining the family of parameters β's as in

Equation (5). The weights w_p are used in the sense of their definition given in Equation (7).

First notice that for a zero-mean stationary process $\{Z_x, x \in V\}$ whose covariance function has a finite sill θ_s,

$$Cov(Z_x, Z_y) = \theta_s - \frac{Var(Z_x - Z_y)}{2} \tag{9}$$

for all $x, y \in V$.

We begin by estimating the sill parameter, θ_s, which represents the variance of the process. Consider pairs of vertices that are unconnected. For such pairs, the covariance is zero and the variance of their difference reduces to $2\,\theta_s$. Define $\mathcal{H}_\infty = \{(x, y) \in V \times V : \tilde{\mathcal{P}}(x, y) \bigcup \tilde{\mathcal{P}}(y, x) = \emptyset\}$ as the set of all such pairs. This yields the unbiased estimator

$$\hat{\theta}_s = \frac{1}{2|\mathcal{H}_\infty|} \sum_{(x,y)\in\mathcal{H}_\infty} (Z_x - Z_y)^2. \tag{10}$$

Let now $\boldsymbol{\gamma} \in \mathbb{R}^n$ denote the vector of empirical semi-variances computed from all distinct pairs of vertices, where n is the number of pairs in the linear network. That is, the i-th component of $\boldsymbol{\gamma}$ is

$$\gamma_i = \frac{(Z_{x_i} - Z_{y_i})^2}{2}$$

if (x_i, y_i) is the i-th pair among the distinct pairs of vertices of the linear network. Because of Equation (9), we consider the method-of-moments system of estimating equations

$$\hat{\theta}_s - \gamma_i = \sum_{p\in\tilde{\mathcal{P}}(x_i,y_i)\cup\tilde{\mathcal{P}}(y_i,x_i)} w_p C(|p|). \tag{11}$$

where each equation is indexed by a pair of distinct vertices (x_i, y_i).

Following standard geostatistical practice (Cressie, 1993), we now reduce the dimensionality of the system (11) by introducing distance classes $\mathcal{H}_1, \ldots, \mathcal{H}_l$ that partition all path lengths into l disjoint bins. We approximate the covariance function C via l representative values $C(h_j)$ for $j = 1, \ldots, l$, where h_j denotes the mean path distance within bin $\mathcal{H}_j$. Let $\boldsymbol{C} = (C(h_1) \ldots C(h_l))^\top$ be the vector of unknown covariance values. Then, the system of equations (11) is approximated by the linear system:

$$\hat{\theta}_s \mathbf{1} - \boldsymbol{\gamma} = W\,\boldsymbol{C}, \tag{12}$$

where $\mathbf{1} \in \mathbb{R}^n$ and the weight matrix $W \in \mathbb{R}^{n\times l}$ aggregates the path weights. Specifically, the entry $W_{[i,j]}$ corresponds to the i-th vertex pair and the j-th distance lag:

$$W_{[i,j]} = \sum_{\substack{p\in\tilde{\mathcal{P}}(x_i,y_i)\cup\tilde{\mathcal{P}}(y_i,x_i)\\ |p|\in\mathcal{H}_j}} w_p. \tag{13}$$

Thus, $W_{[i,j]}$ represents the total weight of all paths connecting the i-th pair that fall within the j-th distance class.

Inverting Equation (12) poses challenges analogous to those encountered in stream network geostatistics (Zimmerman and Ver Hoef, 2016). The weight matrix W is frequently ill-conditioned,

rendering ordinary least-squares estimates unstable. To enforce stability, we adopt a penalized least-squares approach. The estimator is obtained by minimizing the regularized objective function:

$$\mathcal{L}(\boldsymbol{C}) = \frac{1}{2}\|W\boldsymbol{C} - (\hat{\theta}_s\mathbf{1} - \boldsymbol{\gamma})\|^2 + \frac{1}{2}\lambda\|\boldsymbol{C}\|^2 \tag{14}$$

where $\lambda > 0$ is a regularization parameter. The ridge-type penalty is a standard choice in inverse problems; it stabilizes the solution by reducing the variance of the estimator.

Proposition 5. *The penalized estimator for the covariance function of the process is*

$$\hat{\boldsymbol{C}} = \left(W^TW + \lambda I\right)^{-1} W^T\left(\hat{\theta}_s\mathbf{1} - \boldsymbol{\gamma}\right). \tag{15}$$

Moreover, $\lambda \geq \frac{\|W^T(\hat{\boldsymbol{\theta}}_s - \boldsymbol{V})\|_\infty}{\hat{\theta}_s} - \min_i \delta_i$, *where* $\delta_i = \left|\left(W^TW\right)_{(i,i)}\right| - \sum_{j\neq i}\left|\left(W^TW\right)_{(i,j)}\right|$, *guarantees that* $\|\hat{\boldsymbol{C}}\|_\infty \leq \hat{\theta}_s$.

The regularization parameter λ governs the bias-variance trade-off: larger values enforce stability at the expense of bias, while smaller values approach the non regularized solution, exhibiting higher variance. In our framework, we select the minimal λ that satisfies the admissibility condition established in Proposition 5.

Once the discretized covariance estimates $\hat{\boldsymbol{C}}$ are obtained, we fit a parametric model (e.g., see Table 1) to capture the continuous spatial dependence. Specifically, we estimate the range parameter θ_r by minimizing the distance between the theoretical model and the non-parametric estimates. This two-stage approach results in a fully specified covariance structure for the spatial process.

We remark that a variogram estimator could theoretically be derived via the identity

$$\hat{\gamma}(h) = \hat{\theta}_s - \hat{C}(h).$$

5 Summary of the supporting simulation studies

We briefly describe the behavior of the proposed covariance structure in a simulation setting; a detailed account is provided in the Appendix.

We consider multiple realizations from a Gaussian field whose covariance structure is specified through a network-based exponential model as built in Section 3. We assess the ability of our approach to estimate this covariance structure from the simulated data by comparing our framework with the classical Euclidean approach. The comparison is conducted across five values of the range parameter, while keeping the sill parameter fixed. The domain is a less refined version of the network in Figure 3, constructed from real data following the procedure in Section 2.2. The covariance matrix between vertices is then built in a train-test setting.

The first step concerns estimation of the covariance parameters θ_s and θ_r. The sill is consistently recovered, while the range parameter is systematically underestimated. This behavior is consistent with findings from analogous studies in the stream-network setting (Zimmerman and Ver Hoef, 2016; Barbi et al., 2023). Nevertheless, the Euclidean framework fails to retain meaningful spatial dependence, producing semi-variogram estimates close to a flat line, whereas our estimator recovers

a meaningful spatial structure even under range underestimation. The result of this comparison is confirmed quantitatively via mean squared estimation error.

We then assess covariance reconstruction, measuring how closely the matrix built from estimated parameters approximates the true one. Our framework clearly outperforms the Euclidean baseline on both considered metrics.

Finally, we evaluate out-of-sample prediction via ordinary kriging. Our approach reconstructs the spatial field more accurately, while the Euclidean framework fails to capture the dependence induced by the velocity field.

The simulation results confirm the superior performance of our construction and estimation procedures compared with Euclidean benchmarks. They highlight how physics-driven analyses allow for a more faithful representation of the spatial dependence structure, which in turn has a significant impact on the predicted patterns.

6 Case Study: quantifying uncertainty for RCP scenarios

6.1 Estimation of the residual's covariance structure

To generate simulations that capture realistic spatial patterns, we first characterize the residual covariance structure of the water temperature projections (Section 2.1). We define residuals as the point-wise difference between RCP projections and satellite observations over the overlapping period 2006–2022. These residuals serve as the basis for selecting and estimating the parametric covariance model. Under a time-invariant covariance assumption, residuals and covariance structure are evaluated and estimated independently for each year.

Data alignment is achieved by assigning each satellite observation to its nearest RCP grid point. Given the fine resolution of the RCP grid, hereafter we neglect the spatial discretization error resulting from this assignment. We first estimate the bias occurring in the projection. Then, using the velocity field derived from the satellite data (E.U. Copernicus Marine Service Information (2020)), we build, for each year, the linear network according to the procedure described in Section 2.2. Based on the unbiased residual observations at the network vertices, we estimate the residuals' covariance structure following the estimation procedure presented in Section 4, separately for each year from 2006 to 2022.

Figure 8 displays the annual empirical covariance functions using functional box-plots (Sun and Genton, 2011). The estimates are computed over $l = 15$ distance bins. The visualization highlights the functional depth of the curves: the dark shaded area represents the 50% central region (the "bag"), while the lighter area delineates the 90% envelope. The black solid line indicates the point-wise mean, which is used for parameter estimation. We selected the exponential kernel as it best describes the observed spatial decay among the candidates in Table 1.

Leveraging the estimated residual covariance structure, we simulate realizations of Sea Surface Temperature (SST) within the Gaussian framework established in Section 2.1. Formally, we treat the SST field $\{Y_s, s \in D\}$ as a Gaussian Random Field observed in $\{s_1, \dots, s_n\} \in D$, hence $\{Y_{s_1}, \dots, Y_{s_n}\} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$. The mean is estimated by $\hat{\boldsymbol{\mu}}$, the RCP projections corrected by the bias estimated in the observational period 2006-2022. The covariance is estimated by $\hat{\Sigma}$ which follows the construction detailed in Section 3, leveraging the latter estimated parameters. The network

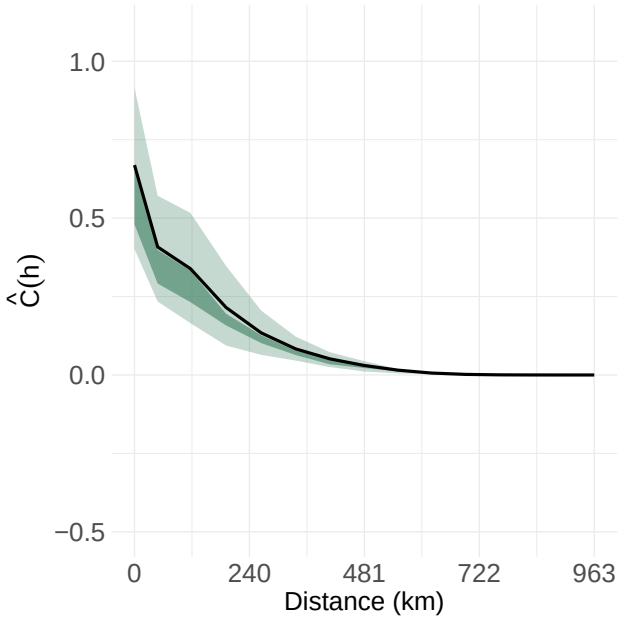


Figure 8: Empirical covariance functions for the period 2006–2022 and point-wise mean. The estimated parameters from the point-wise mean are $\hat{\theta}_s = 0.66$ and $\hat{\theta}_r = 154$ km.

used to evaluate $\hat{\Sigma}$ is the one depicted in Figure 3, describing the physical trend in the prediction year of 2050.

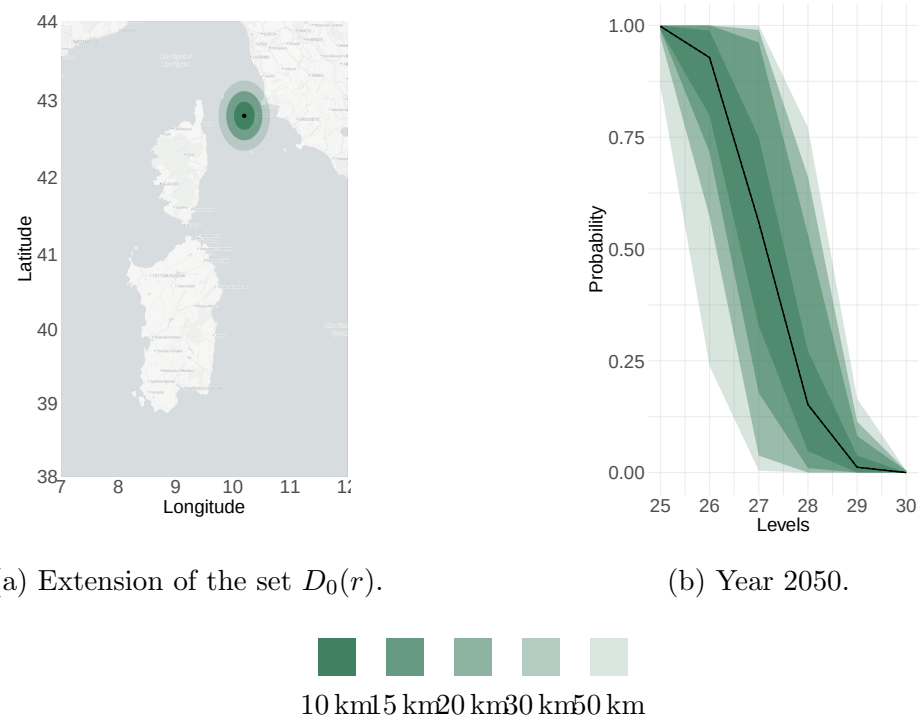
To perform the extreme event analysis, we generate $M = 500$ Monte Carlo realizations of the SST field $\{Y_{s_1}, \ldots, Y_{s_n}\}$, enabling a probabilistic assessment of future climate scenarios and the identification of high-risk regions.

6.2 Extreme event analysis

As a first analysis, we evaluate joint exceedance probabilities (Hazra and Huser, 2021). We investigate joint exceedance probabilities for the Elba Island region using RCP 4.5 SST projections through August 2050. Specifically, for a given temperature t, we estimate the probabilities of two distinct spatial events within a Euclidean neighborhood $D_0(r)$ of radius r centered on the island: the union event, $\cup_{s \in D_0(r)}\{Y_s > t\}$, which happens when in at least one location in $D_0(r)$ the SST exceeds the threshold t; and the intersection event, $\cap_{s \in D_0(r)}\{Y_s > t\}$, which happens when the temperature in all locations in $D_0(r)$ exceed t. We examine radii $r \in \{0, 10, 15, 20, 30, 50\}$ km. Note that for $r = 0$, the two probabilities collapse to the marginal exceedance probability.

A critical distinction underpins our analysis: while the region of interest $D_0(r)$ is defined geometrically using Euclidean distance, the probabilistic assessment relies on the flow-directed network structure. The Euclidean metric serves only to identify the monitoring area; but the joint probabilities are governed by the anisotropic influence of ocean currents. By incorporating directionality, our framework captures flow-driven connectivity that a purely Euclidean dependence model would miss, leading to a more realistic representation of spatial risk.

Figure 9 displays these probabilities for varying radii. The shaded area between the lower curve (union event) and the upper curve (intersection event) corresponds to the probability of partial



(a) Extension of the set $D_0(r)$.

(b) Year 2050.

Figure 9: Joint exceedance probabilities for neighborhoods $D_0(r)$ centered on the Elba Island for varying radii r (in km). For each radius, the shaded band is bounded below by the probability that *all* vertices within $D_0(r)$ exceed the threshold, and bounded above by the probability that *at least one* vertex exceeds it.

exceedance—where extreme temperatures occur within the region but do not saturate it. The opacity of these bands increases with the radius r, visually encoding the effect of spatial extent. For the year 2050, the probability of exceedance remains notably high for thresholds up to $27°C$, signaling a significant shift in the thermal baseline. Even at the extreme threshold of $29°C$, the probability drops substantially but remains non-negligible.

We further characterize extreme behavior by identifying regions where the random field exceeds a threshold t with a specified probability $1-\alpha$ (Bolin and Lindgren, 2015; Hazra and Huser, 2021). Formally, we study the random excursion set defined as $E_{t^+} = \{s \in D : Y_s > t\}$. Following the methodology of French and Sain (2013) and Hazra and Huser (2021), we construct two specific credible regions: the outer excursion set S_{t^+}, which contains the random set E_{t^+} with high probability ($\mathbb{P}(E_{t^+} \subseteq S_{t^+}) = 1-\alpha$); and the inner excursion set $S^C_{t^-}$, which is contained within E_{t^+} with high probability ($\mathbb{P}(S^C_{t^-} \subseteq E_{t^+}) = 1-\alpha$). In our analysis, we set the credibility level at $1-\alpha = 0.95$. Figure 10 displays these regions for the year 2050 and the threshold temperatures $t \in \{25°C, 27°C, 30°C\}$, overlaid with the velocity field.

The analysis reveals distinct risk dynamics as the threshold increases. At $t = 25°C$, the inner set $S^C_{t^-}$ already encompasses specific areas in the southeastern domain—particularly east of Sardinia—indicating a high-confidence of exceedance. Conversely, the outer set S_{t^+} covers the entire domain, implying that no location can be confidently excluded from the risk of surpassing $25°C$. At the more extreme threshold of $30°C$, the risk becomes highly localized. The outer set identifies potential critical areas in southern and eastern Sardinia, as well as the Tyrrhenian sector between

Sardinia and mainland Italy (within the Italian EEZ).

These results provide a robust quantitative basis for risk assessment. The inner set $S_{t^-}^C$ demarcates zones of imminent hazard. Meanwhile, the outer set S_{t^+} defines the maximum potential extent of the event, guiding the spatial allocation of monitoring resources to ensure that no exceedance goes undetected.

The bottom row compares our results with the classical Euclidean framework. Specifically, we apply the same yearly analysis on the empirical residuals from 2006-2022. The covariance parameters (sill and range) are estimated by fitting a parametric model to the point-wise mean of the empirical variograms. The two frameworks yield notably different hot-spot estimates, particularly at the $25°C$ threshold. By incorporating velocity field data, the proposed framework narrows the extent of the high-risk hot-spot in southeastern Sardinia—a direct result of the pronounced currents in that area. Additionally, at the $30°$ threshold, the spatial pattern of the low-risk hot-spot along the eastern coast of Corsica differs visibly between the two frameworks.

7 Conclusions and discussion

This work establishes a rigorous mathematical framework for modeling spatial processes governed by intrinsic directionality. By constructing valid covariance models on directed linear networks, we address the fundamental challenge of capturing flow-driven dependence. Central to our methodology is the integration of a Markovian dynamic with a specialized weighting procedure, which ensures both the positive definiteness of the covariance function and the stationarity of the process. Furthermore, the development of a penalized estimator tailored to this weighted setting resolves the instabilities inherent in network-based estimation.

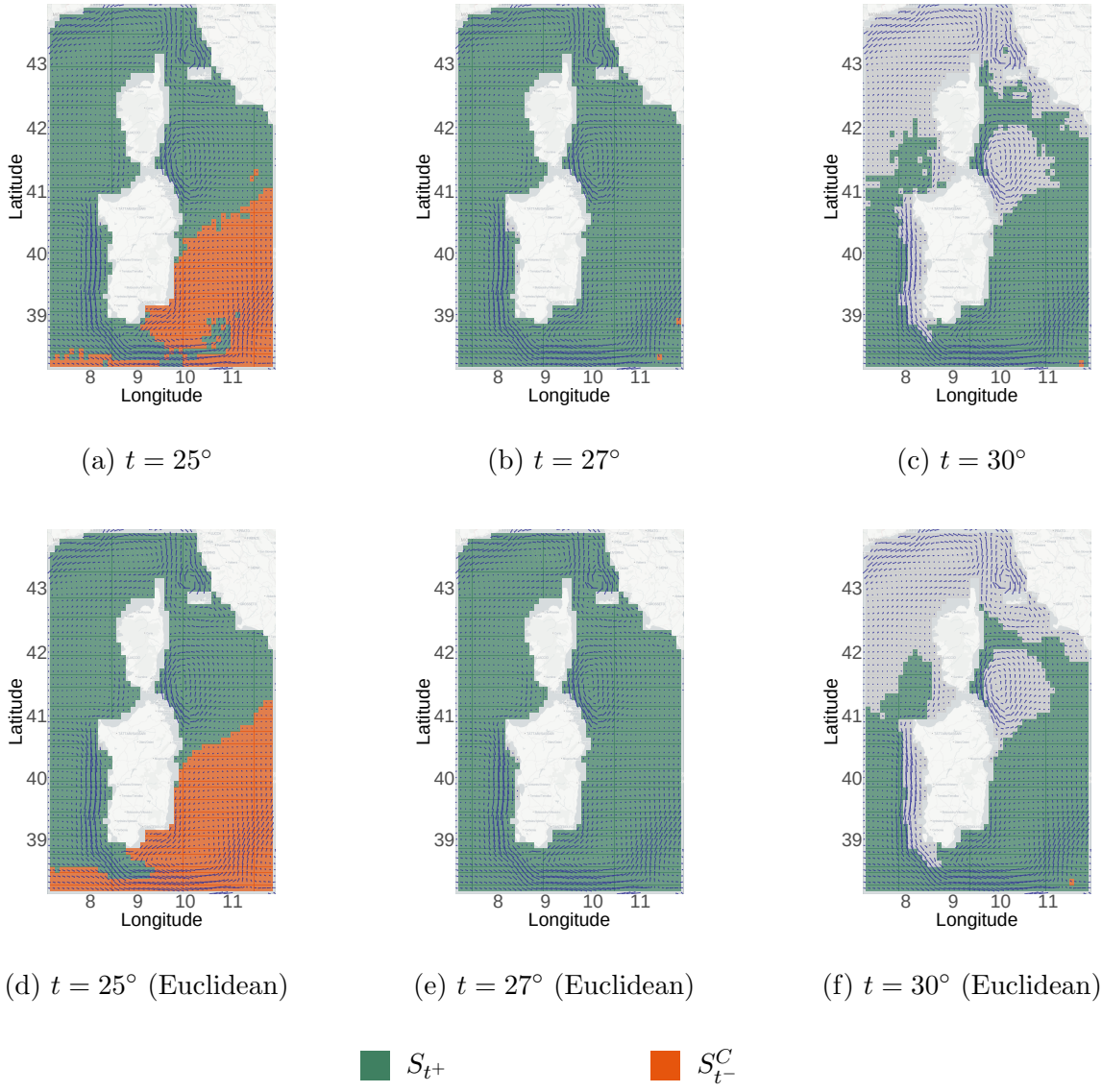
The practical power of this framework is demonstrated through the analysis of projected water temperatures using the ERSEM model. By encoding ocean currents into the network topology, we derived a flow-informed covariance structure. This approach enabled a physics-consistent simulation of temperature fields, providing a probabilistic assessment of risk areas in the Mediterranean.

Beyond marine applications, the proposed framework offers a general solution for systems where topology and transport shape spatial dependence, such as river networks, urban traffic flows, and wind corridors. Its adaptability makes it a versatile tool for diverse domains requiring non-Euclidean geostatistical modeling.

Looking forward, several research avenues emerge to extend the flexibility and scope of the proposed framework.

A primary theoretical objective is to relax the transience assumption to accommodate self-contained (recurrent) networks. Generalizing the moving average rationale to closed systems—where mass is preserved rather than dissipated—would significantly broaden the model's applicability. Crucially, this extension would lay the groundwork for bridging our graph-based approach with state-of-the-art techniques based on Stochastic Partial Differential Equations (SPDEs) (Bolin and Lindgren, 2011) and spatial regression on manifolds (Azzimonti et al., 2015), effectively linking discrete network processing with continuous domain geostatistics.

In parallel, enhancing the physical fidelity of the model is a key priority. This includes incorporating a temporal dimension to model spatiotemporal evolution and exploring non-Markovian dynamics. Moving beyond the Markovian dynamic would allow for the representation of more com-



(a) $t = 25°$ (b) $t = 27°$ (c) $t = 30°$

(d) $t = 25°$ (Euclidean) (e) $t = 27°$ (Euclidean) (f) $t = 30°$ (Euclidean)

Figure 10: Hotspot estimates for RCP 4.5 – Year 2050. Top row: New framework. Bottom row: Euclidean Framework.

plex physical phenomena, such as inertia or momentum in flow-driven systems, thereby offering a more realistic approximation of environmental processes.

Finally, the framework is well-positioned to integrate with Object-Oriented Spatial Statistics (Menafoglio and Secchi, 2017). Extending the domain from point-referenced observations to complex spatial objects—such as trajectories, curves, or sub-regions moving along the network—would open new frontiers for ecological and environmental studies involving structured data.

Acknowledgments

L. Marchesin gratefully acknowledges the financial support of his PhD fellowship from Polis-Lombardia. P. Secchi acknowledges the PRIN2022 project CoEnv - Complex Environmental Data and modelling (CUP2022E3RY23) founded by the European Union - NextGenerationEU program and by the Italian Ministry for University and Research (MUR). All authors acknowledge the support of MUR, grant Dipartimento di Eccellenza 2023–2027. The simulations discussed in this work were, in part, performed on the HPC Cluster of the Department of Mathematics of Politecnico di Milano which was funded by MUR grant Dipartimento di Eccellenza 2023-2027.

References

Anderes, E., J. Møller, and J. G. Rasmussen (2020, aug). Isotropic covariance functions on graphs and their edges. *The Annals of Statistics 48*(4), 2478–2503.

Azzimonti, L., L. M. Sangalli, P. Secchi, M. Domanin, and F. Nobile (2015). Blood flow velocity field estimation via spatial regression with pde penalization. *Journal of the American Statistical Association 110*(511), 1057–1071.

Barbi, C., A. Menafoglio, and P. Secchi (2023). An object-oriented approach to the analysis of spatial complex data over stream-network domains. *Spatial Statistics 58*, 100784.

Bolin, D. and F. Lindgren (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *The Annals of Applied Statistics 5*(1), 523–550.

Bolin, D. and F. Lindgren (2015, jan). Excursion and contour uncertainty regions for latent gaussian models. *Journal of the Royal Statistical Society Series B: Statistical Methodology 77*(1), 85–106.

Bolin, D., A. B. Simas, and J. Wallin (2024, may). Gaussian whittle–matérn fields on metric graphs. *Bernoulli 30*(2), 1611–1639.

Butenschön, M., J. Clark, J. N. Aldridge, J. I. Allen, Y. Artioli, J. Blackford, J. Bruggeman, P. Cazenave, S. Ciavatta, S. Kay, G. Lessin, S. van Leeuwen, J. van der Molen, L. de Mora, L. Polimene, S. Sailley, N. Stephens, and R. Torres (2016). Ersem 15.06: a generic model for marine biogeochemistry and the ecosystem dynamics of the lower trophic levels. *Geoscientific Model Development 9*(4), 1293–1339.

Clarotto, L., D. Allard, T. Romary, and N. Desassis (2024, aug). The spde approach for spatio-temporal datasets with advection and diffusion. *Spatial Statistics 62*, 100847.

Clemente, A., E. Arnone, J. Mateu, and L. M. Sangalli (2026, apr). Nonparametric estimators over metric graphs. *Biometrika*, asag029.

Copernicus Climate Change Service, C. (2020). Marine biogeochemistry data for the northwest european shelf and mediterranean sea from 2006 up to 2100 derived from climate projections. Copernicus Climate Change Service (C3S) Climate Data Store (CDS).

Cressie, N. (1993). Statistics for spatial data.

Cressie, N., J. Frey, B. Harch, and M. Smith (2006). Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics 11*(2), 127.

Curriero, F. C. (2006). On the use of non-euclidean distance measures in geostatistics. *Mathematical Geology 38*(8), 907–926.

Dayan, H., R. McAdam, M. Juza, S. Masina, and S. Speich (2023). Marine heat waves in the mediterranean sea: An assessment from the surface to the subsurface to meet national needs. *Frontiers in Marine Science 10*.

E.U. Copernicus Marine Service Information, C. (2020). Multi observation global ocean 3d temperature salinity height geostrophic current and mld. Marine Data Store (MDS).

French, J. P. and S. R. Sain (2013, sep). Spatio-temporal exceedance locations and confidence regions. *The Annals of Applied Statistics 7*(3), 1421–1449.

Giorgi, F. (2006). Climate change hot-spots. *Geophysical Research Letters 33*(8).

Hazra, A. and R. Huser (2021, jun). Estimating high-resolution red sea surface temperature hotspots, using a low-rank semiparametric spatial model. *The Annals of Applied Statistics 15*(2), 572–596.

Heino, J., R. Virkkala, and H. Toivonen (2009). Climate change and freshwater biodiversity: detected patterns, future trends and adaptations in northern regions. *Biological Reviews 84*(1), 39–54.

Lindgren, F., H. Rue, and J. Lindström (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(4), 423–498.

Menafoglio, A. and P. Secchi (2017). Statistical analysis of complex and spatially dependent data: A review of object oriented spatial statistics. *European Journal of Operational Research 258*(2), 401–410.

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

Peterson, E. E. and J. M. Ver Hoef (2010). A mixed-model moving-average approach to geostatistical modeling in stream networks. *Ecology 91*(3), 644–651.

Ramsay, T. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society Series B: Statistical Methodology 64*(2), 307–319.

Smith, K. E., M. T. Burrows, A. J. Hobday, N. G. King, P. J. Moore, A. S. Gupta, M. S. Thomsen, T. Wernberg, and D. A. Smale (2023). Biological impacts of marine heatwaves. *Annual Review of Marine Science 15*, 119–145.

Sun, Y. and M. G. Genton (2011, jan). Functional boxplots. *Journal of Computational and Graphical Statistics 20*(2), 316–334.

Tandeo, P., P. Ailliot, and E. Autret (2011, aug). Linear gaussian state-space model with irregular sampling: application to sea surface temperature. *Stochastic Environmental Research and Risk Assessment 25*(6), 793–804.

Tarboton, D. G. (1997). A new method for the determination of flow directions and upslope areas in grid digital elevation models. *Water Resources Research 33*(2), 309–319.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012, apr). An overview of cmip5 and the experiment design. *Bulletin of the American Meteorological Society*.

Tesfa, T. K., D. G. Tarboton, D. W. Watson, K. A. T. Schreuders, M. E. Baker, and R. M. Wallace (2011). Extraction of hydrological proximity measures from dems using parallel processing. *Environmental Modelling & Software 26*, 1696–1709.

Tomasetto, M., E. Arnone, and L. M. Sangalli (2024). Modeling anisotropy and non-stationarity through physics-informed spatial regression. *Environmetrics 35*(8), e2889.

van Vuuren, D. P., J. Edmonds, M. Kainuma, K. Riahi, A. Thomson, K. Hibbard, G. C. Hurtt, T. Kram, V. Krey, J.-F. Lamarque, T. Masui, M. Meinshausen, N. Nakicenovic, S. J. Smith, and S. K. Rose (2011, aug). The representative concentration pathways: an overview. *Climatic Change 109*(1), 5.

Ver Hoef, J. and E. Peterson (2010). A moving average approach for spatial statistical models of stream networks. *Journal of the American Statistical Association 105*, 6–18.

Ver Hoef, J. M., E. Peterson, and D. Theobald (2006). Spatial statistical models that use flow and stream distance. *United States Department of Commerce: Staff Publications*.

Zimmerman, D. and J. Ver Hoef (2016). The torgegram for fluvial variography: Characterizing spatial dependence on stream networks. *Journal of Computational and Graphical Statistics 26*.

A Proofs of the propositions

Proposition 1. *The process* $\{Z_x, x \in V\}$ *is zero-mean: for* $x \in V$, $\mathbb{E}[Z_x] = 0$.

Proof. By the tower property and independence between the Wiener process W and the random variables T, we have

$$\begin{aligned}
\mathbb{E}[Z_x] &= \mathbb{E}\left[\sum_{v\in V}\int_{L_{[v,X_1]}} \frac{g\big(dist_{T(v,x)}(u,x)\big)}{\sqrt{\beta_{T(v,x)}}}\, W(du)\right] \\
&= \sum_{v\in V}\mathbb{E}\left[\mathbb{E}\left[\int_{L_{[v,X_1]}} \frac{g\big(dist_{T(v,x)}(u,x)\big)}{\sqrt{\beta_{T(v,x)}}}\, W(du) \;\middle|\; T(v,x)\right]\right] \\
&= \sum_{v\in V}\mathbb{E}\left[\int_{L_{[v,X_1]}} \frac{g\big(dist_{T(v,x)}(u,x)\big)}{\sqrt{\beta_{T(v,x)}}}\, \mathbb{E}\left[W(du) \;\middle|\; T(v,x)\right]\right] \\
&= \sum_{v\in V}\mathbb{E}\left[\int_{L_{[v,X_1]}} \frac{g\big(dist_{T(v,x)}(u,x)\big)}{\sqrt{\beta_{T(v,x)}}}\, \mathbb{E}\left[W(du)\right]\right] \;=\; 0,
\end{aligned}$$

since the Wiener process has zero mean.

□

Proposition 2. *For* $x, y \in V$,

$$Cov(Z_x, Z_y) = \sum_{v\in V}\mathbb{E}\left[\int_{L_{[v,X_1]}} \frac{g\left(dist_{T(v,x)}(u,x)\right) g\left(dist_{T(v,y)}(u,y)\right)}{\sqrt{\beta_{T(v,x)}\,\beta_{T(v,y)}}}\, du\right]. \tag{A1}$$

In particular,

$$Var(Z_x) = Cov(Z_x, Z_x) = \sum_{v\in V}\mathbb{E}\left[\int_{L_{[v,X_1]}} \frac{g\left(dist_{T(v,x)}(u,x)\right)^2}{\beta_{T(v,x)}}\, du\right]. \tag{A2}$$

Proof. We condition on the joint realization of all random paths $\{T(v,x),\, T(v',y)\}_{v,v'\in V}$, which are independent of W. By the tower property,

$$\mathbb{E}[Z_x Z_y] = \mathbb{E}\big[\,\mathbb{E}\big[Z_x Z_y \;\big|\; \{T(v,x),\, T(v',y)\}_{v,v'\in V}\big]\big].$$

Expanding the product $Z_x Z_y$ yields a double sum over $v, v' \in V$:

$$\begin{aligned}
\mathbb{E}\big[Z_x Z_y \;\big|\; \{T(v,x),\, T(v',y)\}_{v,v'\in V}\big] = \sum_{v\in V}\sum_{v'\in V}\mathbb{E}\Bigg[&\int_{L_{[v,X_1]}} \frac{g\Big(dist_{T_{(v,x)}(u,x)}\Big)}{\sqrt{\beta_{T_{(v,x)}}}} W(du) \\
&\int_{L_{[v',X_1']}} \frac{g\Big(dist_{T_{(v',y)}}(u',y)\Big)}{\sqrt{\beta_{T_{(v',y)}}}} W(du') \;\Bigg|\; \{T(v,x),\, T(v',y)\}_{v,v'\in V}\Bigg].
\end{aligned}$$

We now apply the white-noise isometry. Recall that for two square-integrable deterministic functions f_1, f_2 and two edges l_1, l_2,

$$\mathbb{E}\Big[\int_{l_1} f_1(u)\, W(du) \int_{l_2} f_2(u)\, W(du)\Big] = \begin{cases} \int_l f_1(u)\, f_2(u)\, du & \text{if } l_1 = l_2 =: l, \\ 0 & \text{if } l_1 \cap l_2 = \emptyset. \end{cases}$$

Since $L_{[v,X_1]}$ and $L_{[v',X_1']}$ are the first edges traversed by the Markov chains started at v and v' respectively, if $v \neq v'$ they are distinct edges of the network. Therefore, all cross terms with $v \neq v'$ vanish, and the double sum reduces to

$$\mathbb{E}\big[Z_x Z_y \big| \{T(v,x), T(v',y)\}_{v,v' \in V}\big] = \sum_{v \in V} \int_{L_{[v,X_1]}} \frac{g\big(dist_{T(v,x)}(u,x)\big)\; g\big(dist_{T(v,y)}(u,y)\big)}{\sqrt{\beta_{T(v,x)}\, \beta_{T(v,y)}}} du$$

Taking the outer expectation with respect to the T's yields (3). Setting $y = x$ gives (4). □

Proposition 3. *Let $(\mathcal{L}, V)$ be a linear network, $\{Z_x, x \in V\}$ be the process defined in Equation (2) of the main text, and the normalization constants β's be defined as in Equation (5) of the main text. Then for every $x \in V$ the marginal variance is constant and equals*

$$Var(Z_x) \;=\; \int_0^{+\infty} g^2(r)\, dr. \tag{A3}$$

Proof. Consistent with the framework established by Ver Hoef et al. (2006); Ver Hoef and Peterson (2010), we address the unbounded domain of the moving average by introducing a virtual upstream node acting as a generic boundary condition. This ensures total probability mass preservation for the source nodes. Pushing the upstream boundary to infinity ensures that the influence of any specific external condition vanishes, effectively decoupling the internal covariance structure from unobserved upstream inputs.

The first edge of $T(v,x)$ is indicated with $L(v, X_1)$; if $T(v,x) = p(v,x)$, its first edge is therefore $l(v, x_1)$. By Proposition 2,

$$\begin{aligned} Var(Z_x) &= \sum_{v \in V} \mathbb{E}\left[\int_{L_{[v,X_1]}} \frac{g\left(dist_{T(v,x)}(u,x)\right)^2}{\beta_{T(v,x)}}\, du\right] \\ &= \sum_{v \in V} \sum_{p(v,x) \in \mathcal{P}(v,x)} \mathbb{P}\left(T(v,x) = p(v,x)\right)\, \frac{1}{\beta_{p(v,x)}} \int_{l_{[v,x_1]}} g\left(|p(u,x)|\right)^2\, du. \end{aligned} \tag{A4}$$

Along the edge $l_{[v,x_1]}$, the map $u \mapsto r := |p(u,x)|$ is the arc-length parameter, so $dr = du$ while the integration range is $[\,|p(x_1,x)|,\, |p(v,x)|\,]$. Swapping summation and integral, one obtains:

$$Var(Z_x) = \int_0^{\infty} g(r)^2\, S_x(r)\, dr,$$

where the auxiliary function $S_x(r)$ is defined as:

$$S_x(r) = \sum_{v \in V} \sum_{p(v,x) \in \mathcal{P}(v,x)} \frac{\mathbb{P}(T(v,x) = p(v,x))}{\beta_{p(v,x)}}\, \mathbb{I}_{\{|p(x_1,x)| \leq r < |p(v,x)|\}}. \tag{$\star$}$$

We claim that $S_x(r) \equiv 1$ for all $r \in \mathbb{R}$. Note that $S_x(r)$ is a piecewise constant function by construction; discontinuities are possible only at values of r corresponding to the length of a path ending in x. We demonstrate that the function is constant by analyzing the conservation of probability mass across a generic branching point, as illustrated in Figure A1.

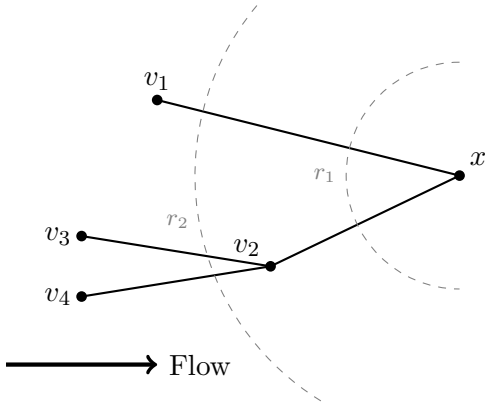


Figure A1: Schematic representation of the network for the proof of mass conservation.

Consider the first radius $r_1 \leq |l_{[v_2,x]}|$ shown in the figure. The only paths for which the indicator function is non-zero are $p(v_1, x)$ and $p(v_2, x)$. In this scenario, the normalization constants β's are given, up to the multiplying probability $U(x)$, by

$$\beta_{p(v_1,x)} = \beta_{p(v_2,x)} \propto \pi_{[v_1,x]} + \pi_{[v_2,x]} \coloneqq A.$$

Thus, the non-zero probabilities and normalization constants are:

- $\mathbb{P}(p(v_1, x)) = \mathbb{P}(T(v_1, x) = p(v_1, x)) = U(x) \cdot \pi_{[v_1,x]}$,
- $\mathbb{P}(p(v_2, x)) = \mathbb{P}(T(v_1, x) = p(v_2, x)) = U(x) \cdot \pi_{[v_2,x]}$,
- $\beta_{p(v_1,x)} = \beta_{p(v_2,x)} = U(x) \cdot A$.

The expression for $S_x(r_1)$ becomes:

$$S_x(r_1) = \frac{\pi_{[v_1,x]}}{A} + \frac{\pi_{[v_2,x]}}{A} = \frac{\pi_{[v_1,x]} + \pi_{[v_2,x]}}{\pi_{[v_1,x]} + \pi_{[v_2,x]}} = 1.$$

We now show that $S_x(r)$ remains equal to 1 when moving to a radius $r_2 > |l_{[v_2,x]}|$, hence crossing the junction v_2. The path $p(v_2, x)$ is no longer included in the sum, but is replaced by the upstream paths originating from v_3 and v_4. The paths contributing to $S_x(r_2)$ are $p(v_1, x)$, $p(v_3, x)$, $p(v_4, x)$. Let $B = (\pi_{[v_3,v_2]} + \pi_{[v_4,v_2]})$. The relevant probabilities and normalization constants β's are:

- $\mathbb{P}(p(v_3, x)) = \mathbb{P}(T(v_3, x) = p(v_3, x)) = U(x) \cdot \pi_{[v_3,v_2]}\, \pi_{[v_2,x]}$,
- $\mathbb{P}(p(v_4, x)) = \mathbb{P}(T(v_4, x) = p(v_4, x)) = U(x) \cdot \pi_{[v_4,v_2]}\, \pi_{[v_2,x]}$,
- $\beta_{p(v_3,x)} = \beta_{p(v_4,x)} = U(x) \cdot A \cdot B$.

Substituting these into the expression for $S_x(r_2)$, one obtains:

$$\begin{aligned} S_x(r_2) &= \frac{\mathbb{P}(p(v_1,x))}{\beta_{p(v_1,x)}} + \frac{\mathbb{P}(p(v_3,x))}{\beta_{p(v_3,x)}} + \frac{\mathbb{P}(p(v_4,x))}{\beta_{p(v_4,x)}} \\ &= \frac{\pi_{[v_1,x]}}{A} + \frac{\pi_{[v_3,v_2]}\,\pi_{[v_2,x]}}{A \cdot B} + \frac{\pi_{[v_4,v_2]}\,\pi_{[v_2,x]}}{A \cdot B} \\ &= \frac{\pi_{[v_1,x]}}{A} + \frac{\pi_{[v_2,x]}}{A}\left(\frac{\pi_{[v_3,v_2]} + \pi_{[v_4,v_2]}}{B}\right) \\ &= \frac{\pi_{[v_1,x]}}{A} + \frac{\pi_{[v_2,x]}}{A} \\ &= 1. \end{aligned}$$

This conservation of mass applies to any bifurcation. Hence, by induction on r, S_x is a constant function equal to 1, completing the proof. □

Proposition 4. *Let $(\mathcal{L}, V)$ be a linear network, $\{Z_x, x \in V\}$ be the process defined in Equation (2) of the main text, and the normalization constants β's be defined as in Equation (5) of the main text. Let $x, y \in V$, $x \neq y$.*

(i) If $\tilde{\mathcal{P}}(x,y) \bigcup \tilde{\mathcal{P}}(y,x) \neq \emptyset$,

$$Cov(Z_x, Z_y) = \sum_{p \in \tilde{\mathcal{P}}(x,y) \bigcup \tilde{\mathcal{P}}(y,x)} w_p C(|p|),$$

where, for any two vertices $v_1, v_2 \in V$,

$$w_{p(v_1,v_2)} = \left(\prod_{l_{[a,b]} \in p(v_1,v_2)} \frac{\pi_{[a,b]}}{\sqrt{\sum_k \pi_{[k,b]}}} \right) \frac{U(v_2,v_1)}{\sqrt{U(v_1)\,U(v_2)}}, \tag{A5}$$

and $C(\cdot)$ is the spatial covariance function defined, for $h \geq 0$, as

$$C(h) = \int_0^{+\infty} g(r) g(r+h) dr.$$

(ii) If $\tilde{\mathcal{P}}(x,y) \bigcup \tilde{\mathcal{P}}(y,x) = \emptyset$,

$$Cov(Z_x, Z_y) = 0.$$

Proof. As before, let $L(v, X_1)$ be the first edge of $T(v,x)$, and $l(v,x_1)$ be the first edge of the realization $T(v,x) = p(v,x)$. Note that two simultaneous realizations $T(v,x) = p(v,x)$ and $T(v,y) = p(v,y)$ must share the first edge $l(v,x_1)$.

By Proposition 2,

$$\begin{aligned} &Cov(Z_x, Z_y) \\ &\quad = \sum_{v \in V} \sum_{p(v,x) \in \mathcal{P}(v,x)} \sum_{p(v,y) \in \mathcal{P}(v,y)} \frac{\mathbb{P}(T(v,x) = p(v,x), T(v,y) = p(v,y))}{\sqrt{\beta_{p(v,x)}\,\beta_{p(v,y)}}} \\ &\quad \int_{l_{[v,x_1]}} g(|p(u,x)|)\, g(|p(u,y)|) du. \end{aligned}$$

To simplify notation, for any path p, let $\Pi(p) = \prod_{l \in p} \pi_l$ and $B(p) = \prod_{l_{[a,b]} \in p} \sum_k \pi_{[k,b]}$, so that $\mathbb{P}(p) = \Pi(p)\, U(x)$ and $\beta_p = B(p)\, U(x)$, where x is the terminal node of p.

We now turn to (i). Given the initial state v of the Markov chain, L_x and L_y are both finite with probability one, because x and y are transient. Moreover they are different, since $x \neq y$. In the representation of $Cov(Z_x, Z_y)$, we partition each contribution to the external sum according to the order of the last visits: $I_v(x,y)$ will denote the contribution from realizations such that $L_x < L_y$, and $I_v(y,x)$ the contribution from realizations such that $L_x > L_y$.

Let's focus on $I_v(x,y)$. Since $L_x < L_y$, the trajectory $T(v,y)$ can be decomposed as the concatenation of $T(v,x)$ and $\tilde{T}(x,y) = (l_{[X_{L_x}, X_{L_x+1}]}, \dots, l_{[X_{L_y-1}, X_{L_y}]})$, that is the trajectory of the chain from the last visit to x to the last visit to y. Note that $\tilde{T}(x,y)$ cannot revisit x; hence $\tilde{T}(x,y) \in \tilde{\mathcal{P}}(x,y)$. Moreover:

$$\begin{aligned}
&\mathbb{P}(T(v,x) = p(v,x), T(v,y) = p(v,y)) \\
&= \mathbb{P}\Big(T(v,x) = p(v,x),\ \tilde{T}(x,y) = p(x,y)\Big) = \Pi(p(v,x)) \cdot \Pi(p(x,y)) \cdot U(y,x),
\end{aligned}$$

where $U(y,x)$ is the probability of never returning to either x or y after reaching y, as defined in the main text. The path $p(x,y) \in \tilde{\mathcal{P}}(x,y)$ is the path realizing $\tilde{T}(x,y)$, consistent with $p(v,x)$ and $p(v,y)$. Conversely, any $p(x,y) \in \tilde{\mathcal{P}}(x,y)$, and any $p(v,x) \in \mathcal{P}(v,x)$ identify a $p(v,y) \in \mathcal{P}(v,y)$.

Now note that $B(p(v,y)) = B(p(v,x))\, B(p(x,y))$, and therefore

$$\beta_{p(v,x)}\, \beta_{p(v,y)} = B(p(v,x))\, U(x) \cdot B(p(v,y))\, U(y) = B(p(v,x))^2\, B(p(x,y))\, U(x)\, U(y).$$

We now turn to the integral over $l_{[v,x_1]}$ appearing in the term $I_v(x,y)$. Setting $h_p = |p(x,y)|$ and $r = |p(v,x)|$, we have $|p(v,y)| = r + h_p$.

Summing all up, we obtain

$$\begin{aligned}
I_v(x,y) = \sum_{p(v,x) \in \mathcal{P}(v,x)} \sum_{p(x,y) \in \tilde{\mathcal{P}}(x,y)} & \frac{\Pi(p(v,x))\, \Pi(p(x,y))\, U(y,x)}{B(p(v,x))\, \sqrt{U(x)\, U(y)\, B(p(x,y))}} \\
& \int_{|p(x_1,x)|}^{|p(v,x)|} g(r)\, g(r + h_p)\, dr.
\end{aligned}$$

Rearranging, we isolate the dependence on $p(x,y)$ to factor out the weight $w_{p(x,y)}$ as defined in (7),

$$
\begin{aligned}
&I_v(x,y)\\
&\qquad = \sum_{p(x,y)\in\tilde{\mathcal{P}}(x,y)} \frac{\Pi(p(x,y))\,U(y,x)}{\sqrt{U(x)\,U(y)\,B(p(x,y))}}\\
&\qquad \int_0^\infty \sum_{p(v,x)\in\mathcal{P}(v,x)} \frac{\Pi(p(v,x))}{B(p(v,x))}\mathbb{I}_{\{|p(x_1,x)|\le r<|p(v,x)|\}} g(r)\,g(r+h_p)\,dr\\
&\qquad = \sum_{p(x,y)\in\tilde{\mathcal{P}}(x,y)} w_{p(x,y)}\\
&\qquad \int_0^\infty \sum_{p(v,x)\in\mathcal{P}(x,y)} \frac{\Pi(p(v,x))}{B(p(v,x))}\mathbb{I}_{\{|p(x_1,x)|\le r<|p(v,x)|\}} g(r)\,g(r+h_p)\,dr.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\sum_{v\in V} I_v(x,y)\\
&\qquad = \sum_{p(x,y)\in\tilde{\mathcal{P}}(x,y)} w_{p(x,y)}\\
&\qquad \int_0^\infty \sum_{v\in V}\sum_{p(v,x)\in\mathcal{P}(x,y)} \frac{\Pi(p(v,x))}{B(p(v,x))}\mathbb{I}_{\{|p(x_1,x)|\le r<|p(v,x)|\}} g(r)\,g(r+h_p)\,dr\\
&\qquad = \sum_{p(x,y)\in\tilde{\mathcal{P}}(x,y)} w_{p(x,y)} \int_0^{+\infty} S_x(r)\,g(r)\,g(r+h_p)\,dr,
\end{aligned}
$$

where, as in the proof of Proposition 3,

$$
S_x(r) = \sum_{v\in V}\sum_{p(v,x)} \frac{\Pi(p(v,x))}{B(p(v,x))}\,\mathbb{I}_{\{|p(b_1,x)|\le r<|p(v,x)|\}}.
$$

However, within the proof of Proposition 3, we already proved that $S_x(r)\equiv 1$ for all $r\ge 0$, so the integral reduces to $C(h_p)=\int_0^{+\infty} g(r)\,g(r+h_p)\,dr$.

The same argument applies symmetrically to $I_v(y,x)$, with the last sum running over $\tilde{\mathcal{P}}(y,x)$. Summing $\sum_{v\in V} I_v(x,y)$ to $\sum_{v\in V} I_v(y,x)$ completes the proof of part (i).

To prove (ii), note that if $\tilde{\mathcal{P}}(x,y)\bigcup\tilde{\mathcal{P}}(y,x)=\emptyset$, then x and y are not connected on the network $\mathcal{L}$. Hence $Cov(Z_x,Z_y)=0$.

□

B Data Management and Algorithm

B.1 Data Preprocessing and Network Construction

As detailed in Section 2 of the main text, the analysis relies on sea surface temperature and velocity fields from the CMEMS and C3S databases. The spatial domain is restricted to the bounding

box 38°N–44°N and 7°E–12°E. To ensure physical consistency when computing the Euclidean distance matrix and paths' lengths for the covariance models, the original geographic coordinates (WGS84) are projected onto the ETRS89-extended / LAEA Europe coordinate reference system (EPSG:3035). A land-mask is applied by pruning any edges connected to vertices with missing oceanographic data, strictly confining the network to valid water regions.

The network topology and transition probabilities are constructed by decomposing the observed velocity field onto the regular grid. For a given vertex a with a valid velocity vector $\boldsymbol{v}_a$, we identify the two valid neighbors, say e and f, whose connecting unit vectors $\boldsymbol{d}_{[a,e]}$ and $\boldsymbol{d}_{[a,f]}$ most closely align with $\boldsymbol{v}_a$ in terms of angle, following the procedure outlined in Sectin 2.2 of the main text.

Let v_a^n and v_a^e denote the northward and eastward components of $\boldsymbol{v}_a$. Similarly, let $d^n_{[a,e]}$ and $d^e_{[a,e]}$ be the corresponding components of the unit vector $\boldsymbol{d}_{[a,e]}$. We define a weighted adjacency matrix $\tilde{M}$ containing the flow magnitudes along the directed edges. In fact, $\tilde{M}_{[a,e]}$ and $\tilde{M}_{[a,f]}$, are obtained by solving the exact linear system:

$$\begin{bmatrix} d^n_{[a,e]} & d^n_{[a,f]} \\ d^e_{[a,e]} & d^e_{[a,f]} \end{bmatrix} \begin{bmatrix} \tilde{M}_{[a,e]} \\ \tilde{M}_{[a,f]} \end{bmatrix} = \begin{bmatrix} v_a^n \\ v_a^e \end{bmatrix}. \tag{A6}$$

This decomposition guarantees that $\tilde{M}_{[a,e]}\boldsymbol{d}_{[a,e]} + \tilde{M}_{[a,f]}\boldsymbol{d}_{[a,f]} = \boldsymbol{v}_a$. The operation is repeated for all nodes of the grid $\mathcal{G}$. The magnitudes of the two components in $\tilde{M}$, scaled by their sum, populate the weighted adjacency matrix M. Indeed, M collects the inputs to build the transition matrix π, as detailed in Section 2.3 of the main text.

Specific software execution steps, including raw variable extraction and data formatting pipelines, are provided in the repository's README.

B.2 Computation of the Non-Return Probabilities

The covariance derivations in the main text (Section 3) require the non-return probability $U(x, y)$, defined as the probability that the Markov chain, starting at node $x \in V$, never returns to the set $A = \{x, y\} \subset V$.

Let $H^A = \inf\{n \geq 0 : X_n \in A\}$ denote the first hitting time of the set A. By conditioning on the first step of the chain $(X_1 = x_1)$, $U(x, y)$ is expressed as:

$$U(x, y) = \sum_{x_1 \notin A} \pi_{[x,x_1]} \left(1 - \mathbb{P}(H^A < \infty | X_0 = x_1)\right). \tag{A7}$$

Thus, the problem reduces to computing the hitting probabilities $\mathbb{P}(H^A < \infty | X_0 = x_1)$ for all neighbors x_1 of x different from y.

Directly simulating or enumerating paths to solve for hitting probabilities is computationally prohibitive. Instead, we leverage the matrix $\mathbf{G}$, whose generic entry $G_{[v_1,v_2]}$ is indexed by $v_1, v_2 \in V$ and is equal to the expected number of visits to v_2 when $X_0 = v_1$:

$$\mathbf{G} = \mathbb{E}\left[\sum_{n=0}^{\infty} \mathbb{I}_{\{X_n = v_2\}}\right] = (\mathbf{I} - \pi_V)^{-1}, \tag{A8}$$

where π_V is the sub matrix of the transition matrix π restricted to the vertices in V, i.e. excluding the absorbing state S. Since all states in V are transient, the spectral radius of π_V is less than 1, and thus$(\mathbf{I} - \pi_V)$ is invertible.

Applying the Strong Markov Property to the hitting time H^A, the expected number of visits to a target $v \in A$ from any starting node x_1 can be decomposed by conditioning on the specific node $k \in A$ where the chain first hits the set:

$$G_{[x_1,v]} = \sum_{k \in A} \mathbb{P}(X_{H^A} = k | X_0 = x_1)\, G_{[k,v]}.$$

Indeed, since $A = \{x, y\}$, we can express this relationship in matrix form for the unknowns $\mathbb{P}(X_{H^A} = x|X_0 = x_1)$ and $\mathbb{P}(X_{H^A} = y|X_0 = x_1)$:

$$\begin{bmatrix} G_{[x_1,x]} & G_{[x_1,y]} \end{bmatrix} = \begin{bmatrix} \mathbb{P}(X_{H^A} = x|X_0 = x_1) & \mathbb{P}(X_{H^A} = y|X_0 = x_1) \end{bmatrix} \begin{bmatrix} G_{[x,x]} & G_{[x,y]} \\ G_{[y,x]} & G_{[y,y]} \end{bmatrix}.$$

Since $\mathbb{P}(H^A < \infty|X_0 = x_1) = \mathbb{P}(X_{H^A} = x|X_0 = x_1) + \mathbb{P}(X_{H^A} = y|X_0 = x_1)$, we obtain:

$$\mathbb{P}(H^A < \infty|X_0 = x_1) = \begin{bmatrix} G_{[x_1,x]} & G_{[x_1,y]} \end{bmatrix} \begin{bmatrix} G_{[x,x]} & G_{[x,y]} \\ G_{[y,x]} & G_{[y,y]} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{A9}$$

To evaluate $U(x, y)$ efficiently across the domain, we implement the following procedure using sparse linear algebra:

1. Extract the 2×2 submatrix $\mathbf{G}_{AA} = \begin{bmatrix} G_{[x,x]} & G_{[x,y]} \\ G_{[y,x]} & G_{[y,y]} \end{bmatrix}$ and compute the coefficient vector
$$\boldsymbol{b} = \mathbf{G}_{AA}^{-1}\mathbf{1}$$
2. For any downstream neighbor $x_1 \notin A$ connected to x, compute $h^A_{x_1} = [G_{[x_1,x]}\ G_{[x_1,y]}]\boldsymbol{b}$.
3. Evaluate the final probability using the transition probabilities from x:
$$U(x, y) = \sum_{x_1 \notin A} \pi_{[x,x_1]}(1 - h^A_{x_1}).$$

Note that, once $\mathbf{G}$ has been evaluated, the previous procedure is totally parallelizable, making the algorithm efficient and scalable.

We also note that, for all $x \in V$, the probability of not returning to $U(x)$ is simply the reciprocal of $G_{[x,x]}$.

B.3 Covariance Matrix Evaluation via the Exponential Kernel

A significant computational advantage arises when adopting the Exponential covariance function, $C(h) = \theta_s \exp(-h/\theta_r)$. The exponential function possesses the semigroup property:

$$\exp\left(-\frac{h_1 + h_2}{\theta_r}\right) = \exp\left(-\frac{h_1}{\theta_r}\right) \cdot \exp\left(-\frac{h_2}{\theta_r}\right).$$

This property implies that the covariance contribution of any path factors into the product of the contributions of its individual edges. Consequently, the summation over all possible paths (including cyclic ones) can be computed exactly via matrix inversion, utilizing the Von Neumann series expansion. This completely bypasses the need for explicit path enumeration or heuristic pruning.

Let $\mathbf{P}$ be the matrix encoding the transition probabilities between vertices in V, weighted by the normalization coefficients; that is the generic element of $\mathbf{P}$ is

$$p[v_1, v_2] = \pi_{[v_1,v_2]} / \sqrt{\sum_{k \in V} \pi_{[k,v_2]}},$$

for $v_1, v_2 \in V$. Let $\mathbf{D}$ be the matrix of Euclidean distances between connected vertices. We construct a decayed transition matrix $\mathbf{R}$ via the Hadamard (element-wise) product:

$$R_{[v_1,v_2]} = p[v_1, v_2] \cdot \exp\left(-\frac{D_{[v_1,v_2]}}{\theta_r}\right).$$

The entry $R_{[v_1,v_2]}$ represents the transition weight from v_1 to v_2 discounted by the spatial correlation decay associated with that single step.

The total covariance accumulated along all paths composed of k steps is given by $\mathbf{R}^k$. Summing over all possible path lengths $k \in \{0, 1, \ldots, \infty\}$ yields the geometric series:

$$\sum_{k=0}^{\infty} \mathbf{R}^k = (\mathbf{I} - \mathbf{R})^{-1}. \tag{A10}$$

Because all vertices $v \in V$ are transient and the spatial decay strictly bounds the entries of $\mathbf{R}$ below the corresponding ones of the transition matrix π, the spectral radius of $\mathbf{R}$ is strictly less than 1, guaranteeing the convergence of the series.

The raw inverse matrix accumulates the effect of self-loops (recirculation starting and ending at the origin node). To derive the correct covariance constrained to acyclic paths $\tilde{\mathcal{P}}(x, y) \bigcup \tilde{\mathcal{P}}(x, y)$ as defined in the main text, the exact final covariance matrix is computed through the following algebraic operations:

1. **Neumann Inversion:** Compute the raw accumulation matrix $\mathbf{S} = (\mathbf{I} - \mathbf{R})^{-1}$. The diagonal entry $S_{[v,v]}$ represents the infinite sum of all cyclic paths starting and ending at $v \in V$.

2. **Source Cycle Elimination:** Normalize the matrix to remove the inflation caused by self-loops at the origin. We divide each row by its corresponding diagonal entry to yield the normalized matrix $\mathbf{S}^*$, where $S^*_{[v_1,v_2]} = S_{[v_1,v_2]} / S_{[v_1,v_1]}$, for $v_1, v_2 \in V$.

3. **Non-return probabilities weighting:** Apply the matrix of pairwise non-return probabilities $\mathbf{U}$ and scale by the marginal non-return probabilities. For $x, y \in V$, let

   $$U_{[x,x]} = U(x),\ U_{[x,y]} = U(y, x)$$

   and define $\mathbf{U}$ to be the matrix whose generic entry, indexed by the vertices in V, is $U_{[x,y]}$. Then set

   $$\tilde{\mathbf{\Sigma}} = \mathbf{S}^* \mathbf{U}.$$

4. **Symmetrization and Scaling:** The final spatial covariance matrix is symmetrized and scaled by the marginal variance θ_s, with the diagonal explicitly reset to θ_s to ensure exact variance matching:

$$\mathbf{\Sigma}_{final} = \theta_s \cdot (\tilde{\mathbf{\Sigma}} + \tilde{\mathbf{\Sigma}}^T), \quad \text{diag}(\mathbf{\Sigma}_{final}) = \theta_s.$$

The matrix $\mathbf{\Sigma}_{final}$ collects the covariances $Cov(Z_x, Z_y)$ of the process $\{Z_x, x \in V\}$, when the assumptions of Propostion 4 are satisfied and the covariance function is Exponential.

C Simulation study

C.1 Simulation setting

The objective of this simulation study is to evaluate the differences between the proposed framework and the conventional Euclidean approach. Specifically, we examine the ability of the new framework to reconstruct the spatial domain after generating synthetic data from its covariance structure.

We use a real velocity field as the basis for our simulations, specifically the data collected in 2022. Once the velocity field and the spatial points at which water temperature was observed were obtained, we constructed a network (using the procedure outlined in the main text) and performed all simulations on this network. Figure A2 illustrates the network used for the simulation study. It is coarser than the network employed for the case study illustrated in the main text. This choice stems from the computational complexity of running simulations on a highly detailed domain.

For the simulation, $M = 100$ replicates of the random field $\boldsymbol{Y} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ were generated, with Σ following the covariance structure defined by the proposed framework. The covariance function is set to be Exponential: the sill parameter θ_s was fixed equal to 1 throughout the study to facilitate interpretability in terms of scale. For the range parameter θ_r, we tested the newly defined process under five different range values (50, 87, 125, 162, 200 km in network distance) to assess how this parameter affects the behavior of the process.

Once the synthetic data were simulated, each iteration involved splitting the field into training and test sets, with the test set comprising one-fifth of the observations. This partitioning was randomized across all iterations to ensure robustness.

To quantify the predictive gain yielded by the network-based topology, we benchmarked the proposed method against a standard geostatistical alternative. For each replicate, we fitted a stationary, isotropic Gaussian process based on standard Euclidean distances. This model represents the baseline geostatistical approach. Consequently, it explicitly ignores the complex non-convexity of the domain (e.g., correlations crossing landmasses) and treats the space as homogeneous, disregarding the directional transport induced by the velocity field.

C.2 Covariance Parameters Estimation

We first assess the estimators' ability to recover the parameters governing the random field. The estimation procedure follows the two-step algorithm detailed in Section 4 of the main text.

Figure A3 displays the kernel density estimates of the sill parameter ($\hat{\theta}_s$) obtained over the $M = 100$ replicates. Both the proposed physics-informed framework and the classical Euclidean

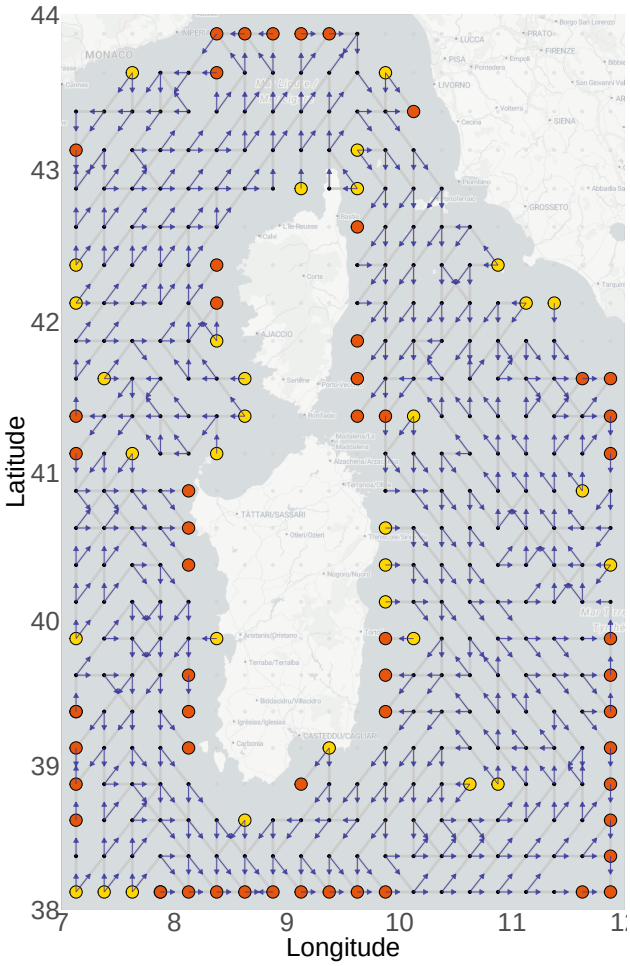


Figure A2: Linear network used in the simulation.

approach yield unbiased distributions centered around the true value ($\theta_s = 1$). This result confirms that the total process variability is correctly identified by both methods, as it is primarily driven by the marginal variance of the data rather than the spatial dependence structure.

A sharp contrast emerges when examining the spatial dependence structure. Figure A4 presents the empirical covariance functions estimated under the Euclidean framework. Crucially, the Euclidean model fails to detect any significant spatial correlation, collapsing to a pure nugget effect at a significantly small range in the vast majority of replicates. This failure stems from a topological mismatch: pairs of points that are spatially proximal in Euclidean space (e.g., separated by a landmass) may be distant in the hydrographic network. The Euclidean metric "short-circuits" these connections, interpreting the high dissimilarity between these points as noise at short lags, thereby masking the true underlying signal.

Conversely, the proposed framework (Figure A5) successfully reconstructs the decaying profile of the covariance function, confirming its ability to capture the complex connectivity encoded in the velocity field. However, we observe a systematic negative bias in the estimation of the range parameter $\hat{\theta}_r$. This underestimation is a known phenomenon in network geostatistics (Zimmerman

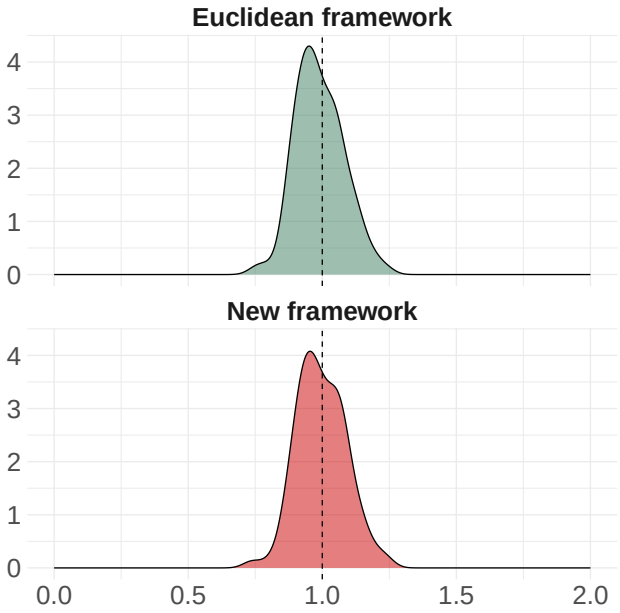
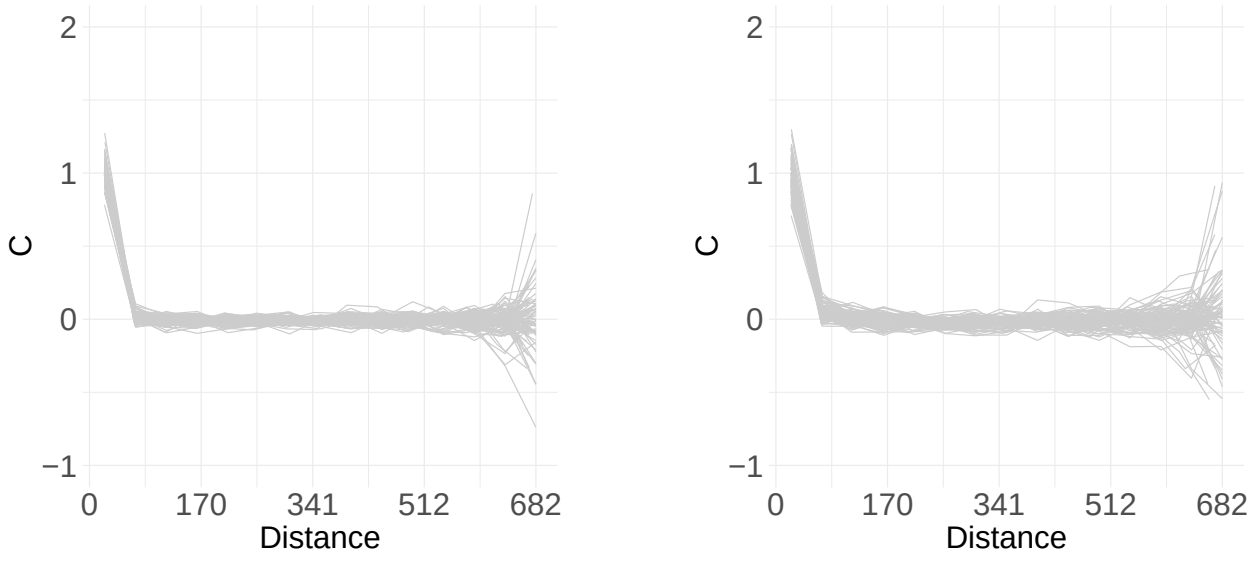


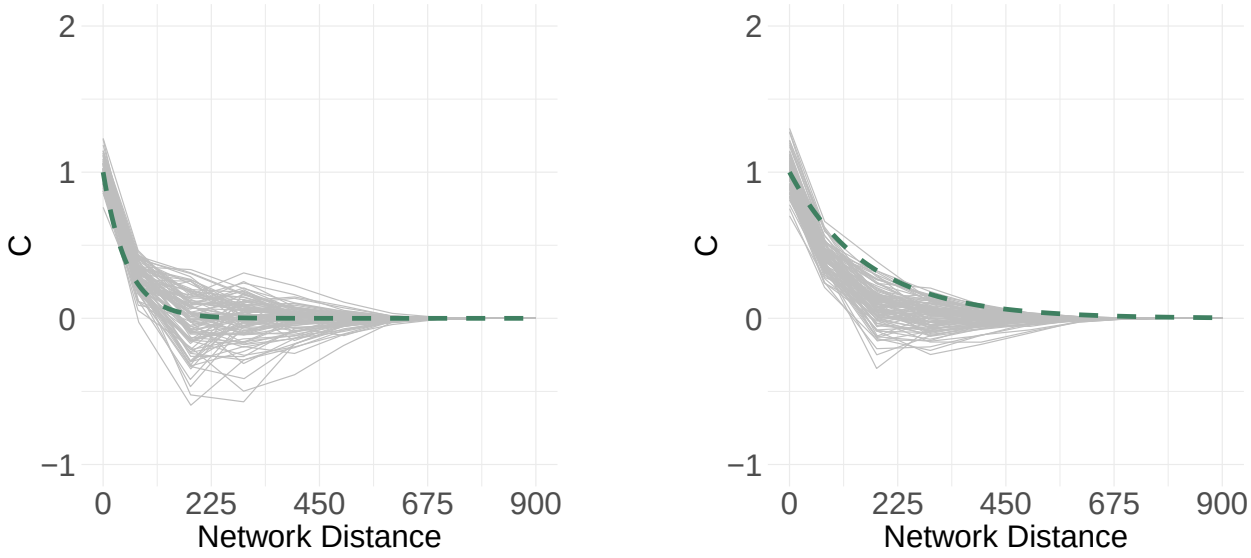
Figure A3: Empirical density of sill parameter estimates ($\hat{\theta}_s$) over 100 replicates. The vertical line indicates the true value.



(a) True Range = 50 km

(b) True Range = 162 km

Figure A4: Empirical covariance functions estimates (gray lines) under the Euclidean framework. The lack of structure indicates a failure to capture the process dependence.



(a) True Range = 50 km

(b) True Range = 162 km

Figure A5: Empirical covariance function estimates under the proposed framework (gray lines) compared to the true generating function (green dashed line).

and Ver Hoef, 2016), often attributed to the confounding between network topology and metric distance. In our framework, this bias is further exacerbated by the regularization term λ in the penalized least squares objective. Despite this bias in the magnitude of θ_r, the framework correctly identifies the *existence* and *shape* of the spatial decay, which—as shown in the following section—is sufficient to outperform the Euclidean benchmark in predictive tasks.

To quantify the visual findings, we evaluated the Mean Squared Error (MSE) associated with the recovery of the exponential covariance structure. For the proposed framework, the MSE was computed as the average squared difference between the estimated covariance function $\hat{C}(h)$ and the true generating function $C(h)$.

Defining a comparable metric for the Euclidean framework is non-trivial, as the Euclidean distance h_{eucl} does not translate directly into network distance. However, since the Euclidean model estimates a pure nugget effect, we compare the true covariance function against the following surrogate:

$$\hat{C}_{eucl}(h) = \begin{cases} \hat{\theta}_{s,eucl} & \text{if } h = 0, \\ 0 & \text{if } h > 0. \end{cases}$$

This function mimics the behavior of the Euclidean framework, under which no spatial correlation is detected. We then compute the MSE by comparing the true generating function $C(\cdot)$ against the two estimated covariance functions. Figure A6 reports the distribution of the MSE values across the replicates.

To further dissect the source of the range underestimation observed in Figure A5, we performed a diagnostic experiment by fixing the sill parameter to its true value ($\hat{\theta}_s \equiv \theta_s$) during the optimization. This allows us to isolate the estimation of the range θ_r from potential identifiability issues between

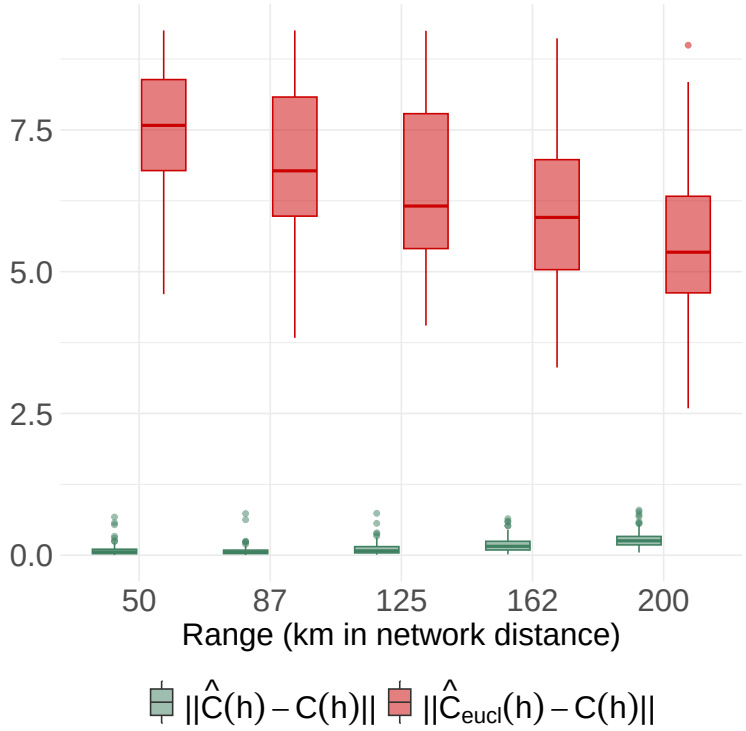


Figure A6: Distribution of Mean Squared Error (MSE) in covariance function estimation. The proposed framework, left/green, is compared with the euclidean benchmark, right/red.

the sill and the range. Figure A7 presents the ensemble of estimated covariance functions under this constrained setting. Even with the sill correctly specified, a negative bias in the covariance magnitude persists at short-to-medium lags. This indicates that the underestimation is not merely a byproduct of sill-range correlation, but is largely attributable to the weighting in the covariance structure, as noted in Zimmerman and Ver Hoef (2016); Barbi et al. (2023).

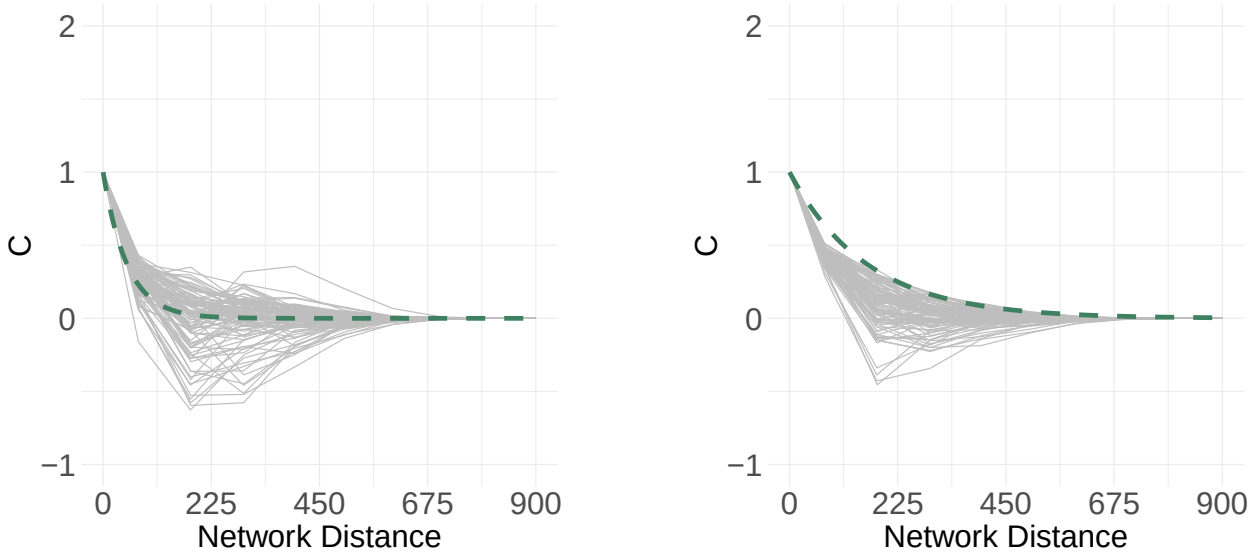
Nevertheless, comparing these curves with the Euclidean "flat line," it is evident that the proposed methodology—despite the regularization bias—effectively captures the functional form and the decay scale of the process, which is the primary requirement for accurate spatial interpolation.

C.3 Global Reconstruction of the Covariance Structure

Beyond the estimation of individual parameters, valid spatial prediction (Kriging) requires that the entire covariance matrix Σ is accurately reconstructed. To assess the global goodness-of-fit, we compared the covariance matrices implied by the estimated parameters against the true data-generating matrix. For the proposed framework, the matrix $\hat{\Sigma}$ was constructed using Equation (9) of the main text with the estimated $\hat{\theta}_s$ and $\hat{\theta}_r$. For the Euclidean benchmark, $\hat{\Sigma}_{eucl}$ was constructed using the standard isotropic exponential kernel.

We evaluated the reconstruction accuracy using two complementary metrics:

1. **Frobenius Norm:** Measures the element-wise distance between matrices: $\|\Sigma - \hat{\Sigma}\|_F = \sqrt{\sum_{i,j} |\Sigma_{[i,j]} - \hat{\Sigma}_{[i,j]}|^2}$.



(a) True Range = 50 km

(b) True Range = 162 km

Figure A7: Diagnostic estimation with fixed sill ($\hat{\theta}_s = \theta_s$). The gray lines represent the empirical estimates, while the green dashed line is the truth. The persistence of the downward bias highlights the effect of the regularization penalty.

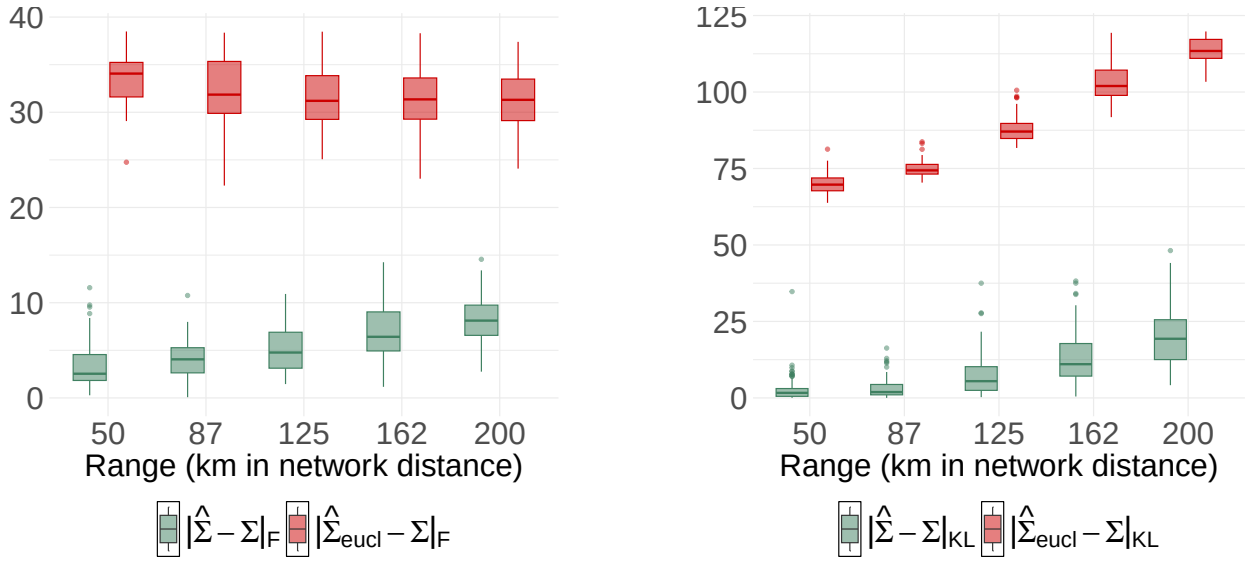
2. **Kullback–Leibler (KL) Divergence:** Measures the information loss when approximating the true multivariate Gaussian distribution $\mathcal{N}(\mathbf{0}, \Sigma)$ with the estimated one $\mathcal{N}(\mathbf{0}, \hat{\Sigma})$.

Figure A8 displays the box-plots of these errors across the simulation replicates. The proposed framework yields reconstruction errors that are consistently lower and less dispersed than those of the Euclidean model. Specifically, the Euclidean approach exhibits high KL divergence, indicating a significant distortion of the probabilistic structure of the field—a direct consequence of its inability to model the physical barriers and flow-directed correlations. Conversely, our framework maintains a low divergence across all range scenarios, confirming that the estimated covariance matrix preserves the essential information required for reliable inference.

C.4 Out-of-Sample Predictive Performance

Finally, we assessed the ultimate goal of the geostatistical analysis: the ability to reconstruct unobserved locations via spatial interpolation. For each simulation replicate, we performed Simple Kriging on the hold-out test set (20% of sites), using the parameters and covariance structures estimated in the training phase. We quantified the accuracy using the Mean Squared Error (MSE).

Figure A9 displays the distribution of MSE values across the $M = 100$ replicates. The results demonstrate the significance of the difference between the two approaches. The Euclidean model consistently yields an MSE approximately equal to the process variance (fixed at $\theta_s = 1$). This confirms that, having estimated a pure nugget effect (as shown in Figure A4), the Euclidean Kriging collapses to the trivial predictor (i.e., predicting the global mean everywhere), failing to exploit any spatial information from neighboring data points.



(a) Frobenius Norm

(b) Kullback–Leibler Divergence

Figure A8: Distribution of reconstruction errors between the true covariance matrix and the estimates. Left: Frobenius Norm (element-wise accuracy). Right: Kullback–Leibler Divergence (distributional accuracy).

In contrast, the proposed physics-informed framework achieves substantially lower prediction errors across all scenarios. By correctly modeling the connectivity induced by the currents, the network-based Kriging effectively leverages information from upstream and downstream neighbors, significantly reducing the predictive uncertainty. Furthermore, the relative advantage of the proposed method scales with the spatial range: as the true correlation length increases, the potential for information transfer across the domain grows. Our framework successfully captures this long-range dependence, leading to a widening performance gap relative to the Euclidean baseline, which remains blind to the spatial structure regardless of the theoretical range.

D A particular case: convolution process over stream networks

This section demonstrates analytically that the proposed physics-informed framework encompasses the classical stream network models as a particular case. Specifically, we show that when the domain is restricted to a simple dendritic river network (a directed binary tree), our formulation is consistent with the *tail-up* model introduced by Ver Hoef et al. (2006) and further developed by Ver Hoef and Peterson (2010).

D.1 The Tail-Up Stream Topology

A stream network is topologically defined as a directed acyclic graph where flow direction is unambiguous. Let $(V, \mathcal{L})$ be this network. Water flows downstream from multiple sources towards a single outlet without bifurcating. The fundamental distinction between this dendritic structure

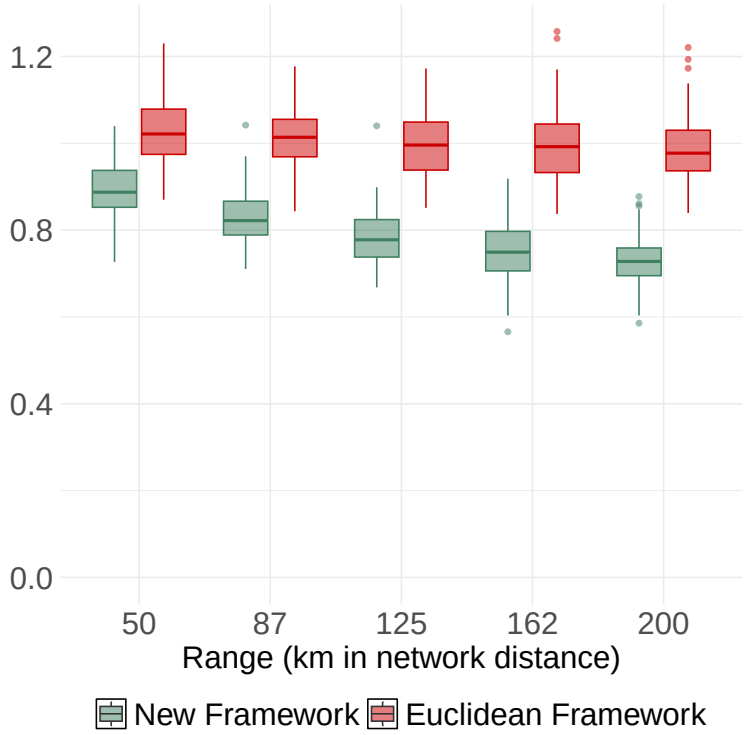


Figure A9: Distribution of Mean Squared Prediction Error (MSE) on the hold-out test set. The Euclidean model (red/right boxplots) stays close to the process variance (1.0), indicating no predictive gain over the mean. The proposed framework (blue/left boxplots) significantly reduces the error.

and the marine environment (defined in Section 3 of the main text) is the property of path uniqueness: for any pair of hydrologically connected points x and y, there exists exactly one directed path $\mathcal{P}(x, y) = \{p(x, y)\}$ connecting them. In the formulation of Ver Hoef et al. (2006), the spatial process relies on a weighting scheme to ensure stationarity. Let $\nu_{[a,b]} \in (0, 1]$ be the weighting corresponding to the edge $l_{[a,b]}$ of the network. The stationarity condition requires that the weights of all upstream segments merging at b sum to unity:

$$\sum_{a \in V} \nu_{[a,b]} = 1. \tag{A11}$$

The resulting "tail-up" spatial covariance between two connected points is given by the product of the square roots of these stream weights along the unique path, multiplied by a valid 1D covariance function $C_0(h)$: for two points x, y, connected by the path $p(x, y)$,

$$Cov_{stream}(x, y) = \left(\prod_{l \in p(x,y)} \sqrt{\nu_l} \right) C_0(|p(x, y)|).$$

D.2 Mapping the Proposed Framework to the Stream Case

In our random walk interpretation, dynamic is governed by the transition probability matrix π. In a stream network where flow does not split (no bifurcations downstream), each vertex a has at most one outgoing edge to a vertex b. Consequently, the transition probability becomes deterministic:

$$\pi_{[a,b]} = \begin{cases} 1 & \text{if } a \to b \text{ is the unique downstream edge,} \\ 0 & \text{otherwise.} \end{cases} \tag{A12}$$

This implies that the random variable T introduced in the main text to handle directional uncertainty at junctions becomes degenerate (i.e., the path is deterministic). Notably, we have that $U(x) = 1$ for all $x \in V$. Hence we omit the quantity in the covariance formulation.

Let $\nu_{[a,b]} = 1/(\sum_{k \in V} \pi_{[k,b]})$, and note that the stationarity condition is satisfied. The normalization constants β's defined in Equation (5) of the main text can be thus expressed in terms of ν:

$$\beta_{p(v,x)} = \prod_{l_{[a,b]} \in p(v,x)} \sum_{k} \pi_{[k,b]} = \prod_{l_{[a,b]} \in p(v,x)} \frac{1}{\nu_{[a,b]}}.$$

Given the specification of the transition probability and of the normalization constants, the weight in the covariance model for a specific path $p(x, y)$ is

$$w_{p(x,y)} = \frac{\mathbb{P}(T(v,x) = p(v,x))}{\beta_{p(v,x)}} = \prod_{l \in p(x,y)} \sqrt{\nu_l}. \tag{A13}$$

Since $\mathcal{P}(x, y)$ contains at most one path $p(x, y)$, the sum for the covariance model reduces to a single term. Thus, for connected points x and y, the covariance becomes:

$$Cov(Z_x, Z_y) = \left(\prod_{l \in p(x,y)} \sqrt{\nu_l} \right) C(|p(x,y)|), \tag{A14}$$

which is identical to the tail-up covariance structure of Ver Hoef et al. (2006).

This derivation confirms that the proposed framework is a consistent generalization of the stream network methodology. The key advancement lies in the ability to handle scenarios where $\pi_{[a,b]} < 1$ (flow splitting) and $|\mathcal{P}(x, y)| > 1$ (multiple paths/cycles), features that are essential for marine domains but absent in river systems.

D.3 The Tail-Down Stream Topology

The stream models of Ver Hoef et al. (2006); Ver Hoef and Peterson (2010) provide a double interpretation and a consequent different model when looking at the flow direction in the opposite sense, yielding the so called "Tail-Down" model. In this section, we demonstrate that our proposed framework yields a consistent covariance structure regardless of the chosen flow direction.

While the physical topology of the domain remains unchanged, the representation of the network dynamics is reversed. The property of path uniqueness between connected locations still holds.

However, because flow can diverge when moving in the upstream direction (i.e., network splits), the transition probability matrix now contains entries strictly less than 1.

Let π be the transition probability matrix for the Markov chain with state space V, and let the entries of π be defined as $\pi_{[a,b]} = \nu_{[b,a]}$. Note that this is a probability matrix, given the condition of stationarity enforced in Equation (A11).

Consider an upstream vertex of a given stream segment. In a strictly branching stream network, this vertex can be reached from exactly one downstream node. Hence, $\sum_k \pi_{[k,b]} = \pi_{[a,b]}$ where $a \in V$ is the only downstream vertex of b connected with it. The full specification of the βs is

$$\beta_{p(x,y)} = \prod_{l_{[a,b]} \in p(x,y)} \sum_k \pi_{[k,b]} = \prod_{l_{[a,b]} \in p(x,y)} \pi_{[a,b]}$$

Summing all up, the weight in the covariance model is

$$w_{p(x,y)} = \prod_{l_{[a,b]} \in p(v,x)} \frac{\pi_{[a,b]}}{\sqrt{\pi_{[a,b]}}} = \prod_{l_{[a,b]} \in p(v,x)} \sqrt{\pi_{[a,b]}}$$

Substituting $\pi_{[a,b]} = \nu_{[b,a]} = \nu_l$, the full covariance mode is expressed as

$$Cov(Z_x, Z_y) = \left(\prod_{l \in p(v,x)} \sqrt{\nu_l} \right) C(|p(x,y)|). \tag{A15}$$

This confirms that the covariance model remains structurally identical under both interpretations of flow direction, unifying the tail-up and tail-down paradigms within our generalized spatial framework.

MOX Technical Reports, last issues

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

38/2026 Clemente, A.; Arnone, E.; Mateu, J.; Sangalli, L.M.
Nonparametric estimators over metric graphs

39/2026 Patanè, G.; Menafoglio, A.; Krauth, A.; Fechner, P.; Dede', L.; Colosimo, B.M.; Nicolussi, F.
K-Models: a Flexible and Interpretable Method for Ordinal Clustering with Application to Antigen-Antibody Interaction Profiles

37/2026 Centofanti, E.; Ziarelli, G.; Scacchi, S.; Pavarino, L.F.
A Neural Latent Dynamics Approach for Solving Inverse Problems in Cardiac Electrophysiology

36/2026 Botti, M.; Mascotto, L.; Mosconi, M.
A nonconforming method for a generalized Darcy-Forchheimer model

35/2026 Caon, B.; Corti, M.; Bonizzoni, F.; Antonietti, P.F.
High-fidelity and Network-based Spatio-temporal Mathematical Models of Alzheimer's Disease Progression and their Validation Against PET-SUVR Imaging Data

34/2026 Mancinelli, F. M.; Torzoni, M.; Maisto, D.; Donnarumma, F.; Corigliano, A.; Pezzulo, G.; Manzoni, A.
Multi-Agent Digital Twins for strategic decision-making using Active Inference

33/2026 Franzoni, G.; Mirabella, S.; Dabek, A.; Ferro, N.; Antona, A.; Carlessi, M.; Cinquemani, S.; Matteucci, M.; Cocetta, G.; Perotto, S.
Integrating Environmental Control and Hyperspectral Imaging to Assess Light and Nutrient Effects on Lettuce Post-Harvest Quality in Vertical Farming

Franzoni, G.; Mirabella, S.; Dabek, A.; Ferro, N.; Antona, A.; Carlessi, M.; Cinquemani, S.; Matteucci, M.; Cocetta, G.; Perotto, S.
ntegrating Environmental Control and Hyperspectral Imaging to Assess Light and Nutrient Effects on Lettuce Post-Harvest Quality in Vertical Farming

32/2026 Antonietti, P.F.; Bonizzoni, F.; Perugia, I.; Verani, M.
A Multilevel Monte Carlo Virtual Element Method for Uncertainty Quantification of Elliptic Partial Differential Equations

31/2026 Guastamacchia, C.; Piersanti, R; Giardini, F.; Coppini, R.; Ferrantini C.; Dede' L.; Sacconi L.; Regazzoni F.
The functional impact of myofiber macroscopic organization and disarray in computational

models of the murine heart