

Coherent exceedance probabilities from a local distributional estimator for heterogeneous environmental fields

María Jesús Valenzuela

University of Chile

Alvaro F. Egaña

University of Chile

Gonzalo Díaz

University of Chile

Felipe Navarro

`felipe.navarro@amtc.cl`

University of Chile

Gerson Morales

University of Chile

Alejandro Ehrenfeld

University of Chile

Article

Keywords: Exceedance probability, Coherence Geostatistics, Distributional calibration, Spatial heterogeneity, Anomaly detection, Environmental risk

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10845802/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Coherent exceedance probabilities from a local distributional estimator for heterogeneous environmental fields

María Jesús Valenzuela¹, Alvaro F. Egaña ¹, Gonzalo Díaz¹,
Felipe Navarro ^{1*}, Gerson Morales¹, Alejandro Ehrenfeld ¹

^{1*}Advanced Laboratory of Geostatistical Supercomputing (ALGES),
Advanced Mining Technology Center (AMTC), Universidad de Chile,
Santiago, Chile.

*Corresponding author(s). E-mail(s): felipe.navarro@amtc.cl;

Abstract

A thorough environmental risk assessment involves estimating the probability that a contaminant exceeds a regulatory threshold at unsampled locations. Calculating these exceedance probabilities requires the full predictive distribution at each location, not a point prediction alone. Existing methods either impose a globally stationary distributional model, which is often violated in heterogeneous environmental fields, or estimate exceedance probabilities independently at each threshold level, sacrificing coherence across the resulting distribution. We present *Local Distributional Calibration*, a non-parametric method that infers the full local predictive distribution at unsampled locations from an ensemble of conditional spatial realisations, and derives coherent exceedance probabilities from it. A calibrated widening step corrects the systematic under-dispersion of the raw ensemble, restoring nominal interval coverage without refitting a global model. On synthetic fields the method performs comparably to the parametric approaches under local stationarity, and its advantage grows as the local distribution departs from a global one. On the Meuse and Jura heavy-metal surveys, evaluated against Dutch soil intervention values and Swiss guide values, it attains a lower Brier score than ordinary kriging and indicator kriging at five of the six regulatory thresholds examined, and it eliminates the order-relation violations that affect the large majority of indicator kriging predictions. The same

predictive distribution supports a locally normalised anomaly score, which identifies observations that are extreme only within their neighbourhood, a class of anomaly that global thresholds miss.

Keywords: Exceedance probability Coherence Geostatistics Distributional calibration Spatial heterogeneity Anomaly detection Environmental risk

1 Introduction

The reliable spatial characterisation of environmental and anthropogenic variables underpins much of contemporary environmental management, from the assessment of soil and water contamination to the design of pollution control policies and the planning of land use under sustainability constraints (Yang and Li 2005; Borkowski and Kwiatkowska-Malina 2017; Biswas and Biswas 2024). As environmental variables are usually measured at a finite set of locations, decision-making on this basis —such as determining suitable locations for remediation or monitoring, or establishing exposure boundaries— relies on predictions at unsampled locations. However, point predictions alone are usually not sufficient; it is also necessary to obtain a defensible quantification of the uncertainty associated with these predictions and the plausibility of an individual observation within its local context (Thayer et al. 2003; Henshaw et al. 2004). Geostatistical and related spatial interpolation tools have become essential for translating sparse point measurements into spatially continuous maps. These can inform sustainable management across a variety of domains, including precision agriculture and proximal soil sensing (Shaddad 2019; Rodrigues et al. 2020), river-basin assessment (Yermolaev et al. 2024), the geoecological evaluation of urban systems (Kochurov and Movchan 2024), and the ecological resilience of mining areas (Jiang et al. 2025; Garcés et al. 2025).

A wide methodological repertoire is available for spatial prediction. Deterministic interpolators such as inverse distance weighting, global and local polynomial interpolation, and radial basis functions are valued for their simplicity (Rodrigues 2016; Kumari et al. 2018; Deepika et al. 2020; Rusu and Rusu 2006). Geostatistical methods, including ordinary kriging, empirical Bayesian kriging, co-kriging, and regression kriging, exploit spatial autocorrelation through the variogram. These methods are considered the workhorses of the field, especially when local uncertainty must be quantified (Vavourakis 2018; Yi et al. 2020; Bronowicka-Mielniczuk and Mielniczuk 2023). More recently, machine-learning regressors such as random forests, boosted trees, and support vector machines have been adopted for spatial prediction (Li et al. 2011; Heusler et al. 2025). A particularly active line of research involves the combination of these with classical geostatistics (Tan et al. 2021; Gao et al. 2022; Ohmer et al. 2025).

Despite this richness, an important gap persists. Determining whether a contaminant exceeds a regulatory threshold at an unsampled location —the question underpinning the Dutch soil remediation framework (Ministry of Housing, Spatial Planning and the Environment (VROM) 2009) and analogous frameworks worldwide— requires a calibrated probability $P(Z(\mathbf{x}) > t)$ that neither a point predictor nor its kriging variance supplies.

Indicator kriging and its coherence problem.

Indicator kriging (IK; Journel 1983; Goovaerts 1997) was proposed precisely to answer this question. It builds the predictive conditional distribution by kriging the indicator transforms $\mathbf{1}(Z > t_k)$ at a series of threshold levels t_k and assembling the results into an estimate of the cdf. IK is the natural benchmark for the present work, and the spatial literature on probability mapping and risk assessment has used it extensively (Goovaerts 2001; Deutsch and Journel 1997). IK has one known structural weakness. Because each indicator is kriged independently, the assembled cdf is not guaranteed to be monotone. Therefore, order-relation violations, where the estimated $\hat{F}(t_k)$ decreases

139 as t_k increases, are common in practice (Goovaerts 1997; Emery and Ortiz 2004). These
140 violations mean that the assembled distribution is not a valid probability distribution
141 at those locations, and any probability calculation derived from it is not coherent.
142

143 Order-relation violations are not a marginal defect in practice. They can affect a
144 large share of predicted conditional distributions on real geochemical surveys, making
145 the incoherence described above common rather than exceptional. A further limita-
146 tion is that the estimated cdf is not invariant to the sampling configuration, since the
147 addition of a single sample alters it (Emery and Ortiz 2004). The central methodolo-
148 gical question is therefore whether a coherent joint distribution can be recovered while
149 retaining predictive accuracy comparable to IK.
150
151

152 *The present contribution.*

153 We propose *Local Distributional Calibration* (LDC), a non-parametric method that
154 yields calibrated predictive distributions at unsampled locations, grounded in Spatial
155 Distributional Estimation (SDE) (Egaña et al. 2025). SDE builds on the ensemble
156 machinery of Ensemble Spatial Interpolation (ESI) (Egaña et al. 2021; Navarro et al.
157 2026) and its locally adaptive extension, Adaptive Ensemble Spatial Analysis (ESA)
158 (Egaña et al. 2025). Both generate an ensemble of spatial predictions at each target
159 location but aggregate it into a single point estimate; SDE first showed that this same
160 ensemble can instead be read as a predictive distribution. LDC builds on that reading,
161 testing whether the ensemble yields calibrated exceedance probabilities, a question
162 that SDE left explicitly open. The same local predictive distribution also supports
163 a locally normalised quantile score, which places observations from subdomains with
164 different local distributions on a common footing for anomaly screening. Risk mapping
165 and anomaly screening thus become two applications of one estimated law rather than
166 two separate methods.
167
168

169 Three contributions follow from this. First, the ensemble retains the local distribu-
170 tional shape that a smoothing estimator discards, so the predictive distribution tends
171
172
173
174
175
176
177
178
179
180
181
182
183
184

to stay better calibrated as the local law varies across the domain. Second, we identify a systematic under-dispersion of the raw ensemble and correct it with a calibrated widening step, which restores both interval coverage and the Brier score of $\hat{P}(Z > t)$ (the probability that a measurement exceeds a regulatory threshold t). Third, the resulting coherent predictive distribution yields exceedance probabilities that are free of order-relation violations by construction.

This advantage is bounded by a regime, since how far the local distribution departs from a global one determines when the non-parametric approach is preferable to indicator kriging; on locally stationary fields the two perform comparably.

The remainder of the paper is organised as follows. Section 2 develops the motivating problem and explains why pointwise estimators built on a stationary global model are systematically misleading in the presence of local non-stationarity. Section 3 introduces the SDE framework and the ESA mechanism. Section 4 sets out the LDC methodology in detail. Section 5 specifies the experimental design. Section 6 reports the results. Section 7 discusses the regime of applicability and the limitations of the approach.

2 Interpolation under a breakdown of local stationarity

Traditional geostatistical methods estimate the conditional mean under an assumption of stationarity, describing the field throughout the domain with a single global model, most often a variogram paired with a multi-Gaussian law. In many environmental and geological settings this assumption is untenable, because the local distribution of the variable itself changes across space, and distinct subdomains — geochemical populations, contamination plumes, lithological units — carry markedly different local means, spreads, and shapes. A stationary estimator cannot represent this *breakdown of local stationarity* itself, only its individual symptoms, such as skewness or the nugget.

For the exceedance probability $\hat{P}(Z > t)$ introduced above, the consequence of this breakdown is not a larger prediction error but a systematic miscalibration. A stationary Gaussian model fitted to the global mean and global variance will under-estimate the exceedance probability in subdomains with high local mean and overestimate it in subdomains with low local mean, with no means of self-correction, since it has no representation of the local law. An ensemble of realisations, by contrast, retains the local distributional shape that a single mean and variance discard, so the local predictive law and its tail probabilities can be reconstructed from it without committing to a global distributional form.

A second, compounding source of miscalibration is the nugget effect, the small-scale variability that the primary correlation structure does not capture. It is an intrinsic feature of many environmental fields, and it adds variability below the resolution of the conditioning data that a mean predictor cannot quantify, but that an ensemble of realisations can still represent alongside the local distributional shape discussed above.

3 Spatial distributional estimation

The ensemble representation motivated above is formalised by the Spatial Distributional Estimation (SDE) framework (Egaña et al. 2025), a non-parametric, decentralised approach to inferring the spatial distribution of a random field. Unlike traditional methods that fit a single global model, SDE constructs a local empirical distribution at each site from an ensemble of conditional realisations generated by the Adaptive Ensemble Spatial Analysis (ESA) procedure (Egaña et al. 2025).

3.1 Posterior sample generation

ESA generates an ensemble of predictive realisations through a hierarchical generative model,

$$p(Z, \mathcal{P}) = p(Z \mid \mathcal{P}) \cdot p(\mathcal{P}), \quad (1)$$

where Z is the spatial variable and $\mathcal{P}$ is a random partition of the domain $\mathcal{S}$. Each partition is built by randomly sampling the observed data, and within each cell a local interpolator (adaptive inverse distance weighting) produces a complete spatial realisation. Repeating this over T partitions yields an ensemble $\{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(T)}\}$, from which the set of values at any location $\mathbf{x}$ constitutes a sample from the local predictive distribution $P(Z(\mathbf{x})|D)$, where D denotes the conditioning data.

3.2 Local distributional estimation and implicit copula

At each location $\mathbf{x}_i$ the ensemble samples $\mathcal{Z}_i = \{z_i^{(1)}, \dots, z_i^{(T)}\}$ define a local empirical distribution,

$$\hat{F}_i^{(T)}(z) := \frac{1}{T} \sum_{t=1}^T \mathbb{I}\{z_i^{(t)} \leq z\}. \quad (2)$$

As shown in Proposition 1 of [Egaña et al. \(2025\)](#), the overlapping neighbourhoods used by ESA induce statistical dependence between all local distributions $\{\hat{F}_i^{(T)}\}_{i=1}^N$, where N is the number of target locations. By Sklar’s theorem, this guarantees the existence of a unique copula $\hat{C}_T$ that implicitly governs the joint distribution of the estimated field, ensuring spatial coherence and that the assembled exceedance probabilities are jointly consistent.

4 Methodology

LDC builds on the SDE ensemble and its implicit copula. At every location it infers the full local predictive distribution of the primary variable. From that distribution it derives two outputs, the exceedance probability $\hat{P}(Z > t)$ of Section 1 and a locally normalised quantile score for anomaly screening. The procedure has three steps, summarised in Figure 1. The ensemble is generated, widened into a calibrated local predictive distribution, and read for the two outputs.

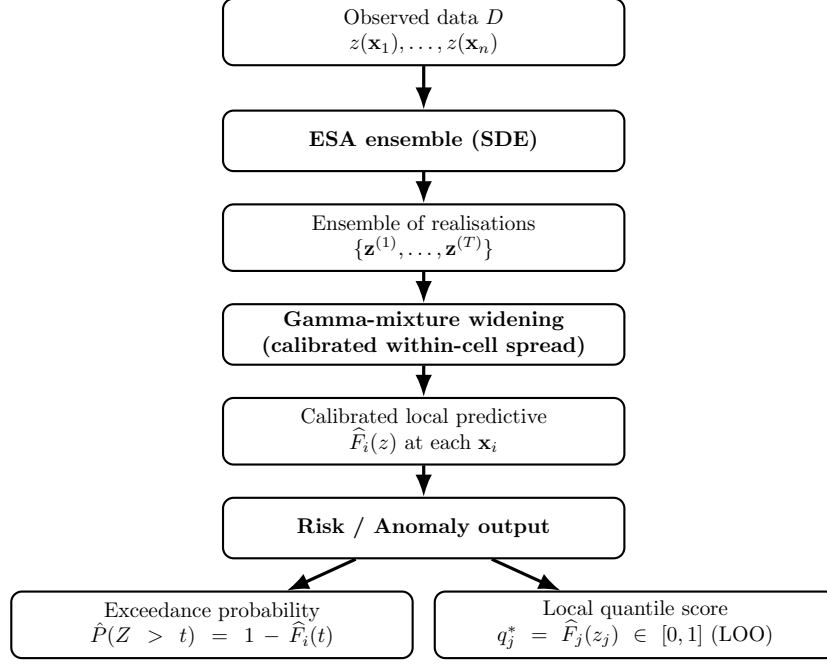


Figure 1: The LDC procedure. The ESA ensemble is widened by a calibrated Gamma mixture into a local predictive distribution, from which both the calibrated exceedance probability and the leave-one-out quantile score are derived.

4.1 Ensemble generation and the local predictive distribution

The ESA ensemble approximates the posterior predictive distribution

$$P(Z(\mathbf{x})|D) = \int P(Z(\mathbf{x})|\mathcal{P}, \theta_Z, D) \cdot P(\mathcal{P}, \theta_{\mathcal{P}}, \theta_Z|D) d\mathcal{P} d\theta_{\mathcal{P}} d\theta_Z \quad (3)$$

by drawing T realisations across the discretised grid. The local empirical distributions $\hat{F}_i^{(T)}(z)$ of Section 3.2 follow directly from this ensemble, and inherit its spatial coherence through the implicit copula $\hat{C}_T$.

In practice the local law is not read off the raw empirical distribution. A Gaussian kernel density estimate smooths the T ensemble values at each location into a continuous predictive distribution, and is fitted in log space for strictly positive variables so that the support remains non-negative. We refer to this estimator as LDC-KDE.

4.2 Widening the ensemble into a calibrated predictive distribution

Each ensemble member is a partition-level estimate, a local average within a cell. The ensemble therefore captures the variability *between* partitions, but compresses the short-range variability *within* a cell. Left uncorrected, the predictive distribution is too narrow and under-covers the truth. To restore the missing spread we treat each ensemble member $z_i^{(t)}$ not as a single draw from the local law but as the mean of a local component distribution. The component is Gamma,

$$z_i^{(t)} = \mathbb{E} \left[\text{Gamma} \left(k_i, \frac{z_i^{(t)}}{k_i} \right) \right], \quad (4)$$

which preserves the member's mean, has positive support, and has a right skew that decreases with the shape k_i . Positive support matches the non-negative environmental concentrations considered here.¹

The mixture of these components over the T members is the widened local predictive distribution, and we refer to this estimator as LDC- Γ . The shape k_i is calibrated at each location so that the variance of the mixture matches a target local variance, estimated from the m nearest observed neighbours. By the law of total variance, this adds the estimated within-cell variance to the between-partition variance already present. Where the ensemble already disperses at least as much as the target, no widening is applied. The correction therefore only ever adds the spread that the partition averaging removed.

The target depends on how many neighbours are averaged, and no single count serves every field. Where the field is stationary, a wide neighbourhood estimates the local spread precisely. Where it is not, the same neighbourhood imports foreign spread across a subdomain boundary, so the preferred count falls as the departure from local

¹For variables that may take negative values, a skew-normal component with location calibrated to preserve the member's mean would serve the same role.

stationarity grows. Selecting the count per dataset is unreliable, because the selection criterion is estimated from the same few dozen observations and is flattest where the choice matters least. We therefore pool the widened draws obtained at several counts in equal shares. The pooled predictive commits to no single scale and has a heavier tail than any of its components. The counts are fixed at $m \in \{6, 8, 20\}$ and held constant across every field, survey and experiment reported below, so pooling introduces no fitted quantity beyond the counts themselves.

4.3 Calibrated exceedance probability

The estimated local distribution $\hat{F}_i$, whether the raw ensemble estimate or its widened form, gives the probability of exceeding a regulatory threshold t directly,

$$\hat{P}(Z(\mathbf{x}_i) > t) = 1 - \hat{F}_i(t). \quad (5)$$

This quantity is defined at every location and requires no model of the indicator variogram at t . It is also free of order-relation violations, because $\hat{F}_i$ is by construction a valid non-decreasing distribution function. Its calibration is assessed by the Brier score (Brier 1950; Gneiting and Ranjan 2011),

$$\text{BS}(t) = \frac{1}{n} \sum_{i=1}^n (\hat{P}(Z(\mathbf{x}_i) > t) - \mathbf{1}(z_i > t))^2, \quad (6)$$

where z_i is the value observed at $\mathbf{x}_i$ and n the number of scored locations. The score is evaluated in a leave-one-out protocol. The reliability diagram, which plots the predicted exceedance probability against the observed exceedance frequency within bins of that prediction, provides a complementary graphical check.

4.4 The local quantile score

The same local predictive distribution also measures how extreme an observation is *for its own location*. A global threshold cannot supply that measure once the local law varies across space, since the same concentration is unremarkable in one subdomain and extreme in another. The local predictive distribution supplies it directly. It is estimated using a leave-one-out construction, so that no observation informs the distribution against which it is judged. For each observed value z_j at location $\mathbf{x}_j$, the local distribution $\hat{F}_j$ is estimated from $D_{-j} = D \setminus \{z_j, \mathbf{x}_j\}$, and the observation is mapped to its predictive quantile,

$$q_j^* = \hat{F}_j(z_j) \in [0, 1]. \quad (7)$$

A value that is extreme for its location maps to a quantile near zero or one regardless of the raw scale. The score therefore places observations from subdomains with different local distributions on a common footing.

5 Experimental design

Section 1 argued that LDC’s advantage over the parametric baselines is bounded by a regime, growing as the local distribution departs from a global one. Four experiments test that claim empirically. Their common object is the *frontier*, the level of departure from local stationarity at which the non-parametric estimate overtakes its parametric competitors. A synthetic experiment with a known ground truth locates the frontier for the predictive distribution as a whole (Experiment 1). A second synthetic experiment examines whether it carries over to the exceedance probability on which a regulatory decision rests (Experiment 2). A third experiment evaluates the anomaly-screening application of the same estimated law (Experiment 3), and a leave-one-out validation

507 checks the frontier against real data (Experiment 4). The last two draw on the real
508 surveys described in Section 5.1.

510 The baselines are ordinary kriging (OK) and sequential Gaussian simulation (SGS),
511 which assume a multi-Gaussian field characterised by a fitted variogram, and indicator
512 kriging (IK), described in Section 1. All three fit their model once for the whole domain,
513 whereas the two LDC estimators of Section 4, LDC-KDE and LDC- Γ , read the local
514 predictive law from the ESA ensemble without committing to a global distributional
515 form.
516

523 5.1 Data

525 Two real datasets anchor the experiments in field conditions. Both are soil geochemical
526 surveys, drawn from different regions and settings, so that a conclusion supported by
527 both is not an artefact of a single site.

530 The Meuse floodplain heavy-metal survey (Bivand et al. 2008) comprises 155 ob-
531 servations of zinc, cadmium, lead, and copper concentrations on the flood plain of the
532 river Meuse near Stein, in the Netherlands, every one of them within a kilometre of
533 the river channel (Figure 2). All four variables are strongly right-skewed, so the para-
534 metric baselines are fitted in log-space, and each method therefore operates under its
535 best-specified configuration.

540 The Dutch soil intervention values of Ministry of Housing, Spatial Planning and
541 the Environment (VROM) (2009) serve as exceedance thresholds: 720 mg kg⁻¹ for zinc,
542 13 mg kg⁻¹ for cadmium, and 530 mg kg⁻¹ for lead (dry soil, standard composition).
543 Copper is treated as a calibration check only, since all 155 observations fall below its
544 intervention value of 190 mg kg⁻¹ and there are no exceedance events to predict. All
545 four metals enter Experiment 4, and zinc is also the metal used in Experiment 3.

550 The Jura heavy-metal survey (Goovaerts 1997) comprises 359 topsoil observations
551 of cadmium, cobalt, chromium, copper, nickel, lead, and zinc concentrations over an
552

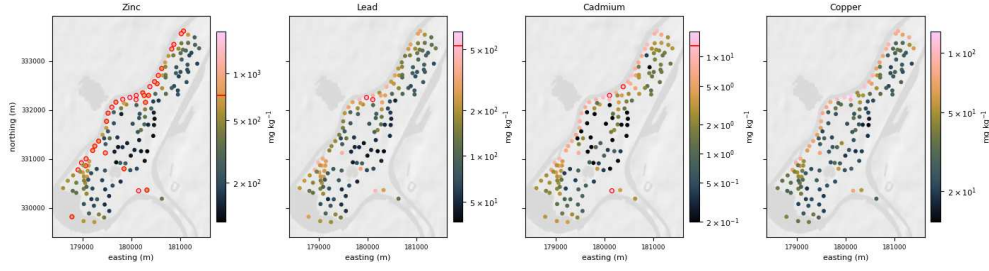


Figure 2: Meuse floodplain heavy-metal dataset (155 observations; [Bivand et al. 2008](#)). The colour scale is logarithmic. Samples above the Dutch soil intervention value ([Ministry of Housing, Spatial Planning and the Environment \(VROM\) 2009](#)) are ringed in red, and the red line on each colourbar marks that value: zinc 720 mg kg^{-1} , lead 530 mg kg^{-1} , cadmium 13 mg kg^{-1} , copper 190 mg kg^{-1} .

approximately $4.4 \text{ km} \times 5.2 \text{ km}$ area near La Chaux-de-Fonds, Switzerland (Figure 3). Concentrations here vary with land use and parent-rock type, both recorded at each location and neither supplied to any method, so that variation reaches every estimator as unexplained heterogeneity. Coordinates are the survey’s published WGS84 longitude/latitude ([Pebesma 2004](#)), reprojected here to the Swiss CH1903+/LV95 grid (EPSG:2056). Two of the seven metals are used. Zinc is the second real field of Experiments 3 and 4, independent of the Meuse in both region and geochemical setting. Cadmium enters Experiment 4 alone, where it carries the two densest exceedance cases of either survey.

The Swiss soil pollution levels of [Swiss Federal Council \(1998\)](#) serve as exceedance thresholds, the zinc guide value of 150 mg kg^{-1} and, for cadmium, both the guide value of 0.8 mg kg^{-1} and the food-crop trigger value of 2 mg kg^{-1} (total content, dry substance).

5.2 Scoring rules and fairness controls

Every method is scored by an identical metric layer. Distributional fidelity is measured by the continuous ranked probability score (CRPS) ([Gneiting and Raftery 2007](#)), a strictly proper scoring rule comparing the whole predictive distribution $\hat{F}$ against the

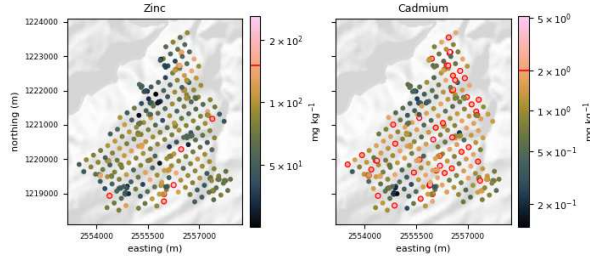


Figure 3: Jura zinc and cadmium fields (359 observations near La Chaux-de-Fonds, Switzerland; [Goovaerts 1997](#)). The colour scale is logarithmic. Samples above the Swiss VBBo level are ringed in red, and the red line on each colourbar marks that level: the zinc guide value of 150 mg kg^{-1} and the cadmium food-crop trigger value of 2 mg kg^{-1} .

realised value z ,

$$\text{CRPS}(\hat{F}, z) = \int_{-\infty}^{\infty} (\hat{F}(y) - \mathbf{1}(y \geq z))^2 dy. \quad (8)$$

Calibration is measured by the empirical coverage of the central predictive intervals and by the probability integral transform (PIT) ([Dawid 1984](#)). For a perfectly calibrated predictor, the PIT values $\hat{F}_i(z_i)$ are uniformly distributed on $[0, 1]$. Therefore, a flat PIT histogram indicates calibration, a central hump indicates over-dispersion, and a U-shape indicates under-dispersion or mis-assigned tails. The histogram is summarised by the PIT-KS distance, the Kolmogorov–Smirnov distance between the empirical distribution of the PIT values and the uniform. It vanishes for a perfectly calibrated predictor and grows with the departure from it, which makes calibration comparable across methods and fields as one number. Every location is scored against its own predictive distribution, so the PIT values are expected to be uniform however the mean and the variance vary in space. The PIT-KS distance therefore tests local validity rather than validity on average. Both the CRPS and the PIT-KS distances are lower-is-better, whereas empirical coverage is better the closer it sits to its nominal level.

The Brier score of the estimated exceedance probability $\hat{p}_i = \hat{P}(Z(\mathbf{x}_i) > t)$, defined in Section 4.3 (Equation 6), is a squared error and so is lower-is-better. A low value is not on its own evidence of a useful exceedance forecast, since a forecast that never leaves the base event rate already scores well at rare-event thresholds. We therefore report it alongside two instruments that separate discrimination from calibration.

The first is the area under the ROC curve (AUC) of $\hat{p}$ against the realised exceedance indicator $\mathbf{1}(z_i > t)$. It equals the probability that a location that truly exceeds t receives a higher $\hat{p}$ than one that does not, and it measures discrimination alone. Any monotone recalibration of $\hat{p}$ leaves it unchanged, so it captures only whether the locations at risk are ranked correctly.

The second is Murphy’s decomposition $\text{BS} = \text{REL} - \text{RES} + \text{UNC}$ (Murphy 1973), which tests whether the reported values themselves hold up. The reliability term REL compares the mean forecast within a bin against the exceedance frequency observed in that bin, and so isolates miscalibration. The resolution term RES measures how far the forecast departs from the base rate, and the uncertainty term UNC depends on the events alone, so it is identical across methods. The decomposition is computed on quantile bins of the forecast, since at a one-percent event rate nearly all forecasts fall in a single equal-width bin and the decomposition degenerates. Reliability diagrams visualise the same comparison.

The same predictive distribution also supplies a design value, the value $\hat{F}_i^{-1}(q)$ read off it at level q . Being a quantile rather than a probability, it is scored by the pinball loss (Gneiting 2011), the proper scoring rule for a quantile, with the realised exceedance rate of that quantile as its calibration counterpart.

Two further controls equalise the conditions under which the methods are compared. Each method is fitted in the space in which it is correctly specified, so the parametric baselines are fitted in log-space on the strongly right-skewed real surveys of Section 5.1, whereas on the synthetic fields below, whose marginal is Gaussian within

each subdomain, every method is fitted in the raw space. LDC-KDE, LDC- Γ , and SGS are given the same ensemble size ($T = 300$ draws), so the ensemble methods operate on an equal information budget.

One variogram is fitted per field and shared by OK, SGS, and IK, so that no baseline is advantaged by a better-tuned structural model. It is estimated over fifteen lags out to 0.35 of the domain extent, and exponential, spherical, and Gaussian models are each fitted with a nugget, retaining the one with highest R^2 . Each SGS realisation set is drawn under its own seed per replicate or fold.

Indicator kriging is estimated at thirteen data quantiles, from the 1st to the 99th percentile, spaced more densely toward the tails than a decile grid. Any fixed threshold that is scored, a regulatory limit or an exceedance level of Experiment 2, is added to that grid, so IK is read where it is estimated rather than interpolated between knots. Its estimated CDF is frequently non-monotone across thresholds, and is repaired by a cumulative maximum before scoring. The repair can only raise the estimated CDF, so it can only lower the exceedance probability the method issues. The violation rate is measured before the repair and reported alongside the scores.

The LDC estimators use the adaptive inverse-distance local interpolator over $T = 300$ partitions. The partition coarseness parameter α is selected using leave-one-out cross-validation. All computations use the `spatialize` library for the ESA ensemble and `gstools` (Müller et al. 2022) for the parametric baselines.

5.3 Experiment 1: locating the frontier under a controlled departure from local stationarity

The first experiment uses synthetic random fields with a known ground truth, so that CRPS, interval coverage, and the PIT are measured at every unsampled node of the field rather than at a handful of held-out observations. Its purpose is to test whether the advantage of the ESA ensemble under non-stationarity, which Egaña et al.

(2025) demonstrated through mean-error metrics, extends to the calibration of the full predictive distribution, and to locate the departure from local stationarity at which that advantage appears.

Field generation.

Each field is defined on a 40×40 regular grid over a square domain. The domain is partitioned into $K = 4$ Voronoi cells around random centres. All cells share one correlated component, which is a Gaussian random field generated by spectral simulation with an exponential covariance of unit sill and a length scale of a fifth of the domain side. The nugget component is added explicitly as independent white noise so that the nugget-to-sill ratio ν is a controlled parameter. A parameter $\eta \geq 0$ governs the departure from local stationarity by scaling how far the cells are allowed to differ from one another. Each Voronoi cell k then receives its own mean offset $\mu_k \sim \mathcal{N}(0, (3\eta)^2)$ and spread multiplier $\sigma_k = \exp(\varepsilon_k)$ with $\varepsilon_k \sim \mathcal{N}(0, (0.8\eta)^2)$, giving

$$Z(\mathbf{x}) = \mu_{r(\mathbf{x})} + \sigma_{r(\mathbf{x})} [\sqrt{1 - \nu} W(\mathbf{x}) + \sqrt{\nu} \epsilon(\mathbf{x})], \quad (9)$$

where $r(\mathbf{x})$ is the Voronoi region of $\mathbf{x}$, W is the correlated component, and $\epsilon \sim \mathcal{N}(0, 1)$ is the nugget. At $\eta = 0$ all cells share the same law and the field is approximately stationary, whereas at $\eta = 2$ the cells carry substantially different local means and spreads. Figure 4 illustrates three levels of η .

Sampling and evaluation.

Each realisation is shifted by a constant so that the field is strictly positive. From it, $n = 120$ locations are drawn at random as observed data. The remaining $40^2 - 120 = 1\,480$ grid nodes are held out as ground truth, against which CRPS, coverage and the PIT are computed. The sweep covers $\eta \in \{0.0, 0.5, 1.0, 1.5, 2.0\}$ and $\nu \in \{0.3, 0.7\}$.

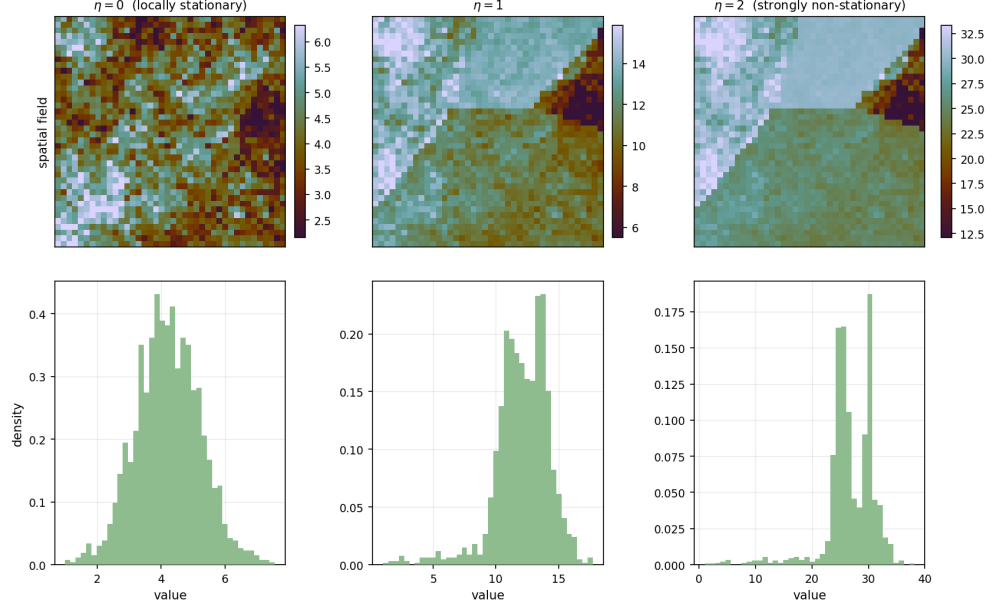


Figure 4: Synthetic fields used in Experiment 1, at three levels of the non-stationarity parameter η (nugget-to-sill ratio $\nu = 0.3$). Top row: spatial field on the 40×40 grid. Bottom row: marginal distribution.

Each (η, ν) combination is repeated over $R = 50$ independent replicates, so reported differences carry standard errors.

Effect of the widening.

LDC-KDE (the raw ensemble) and LDC- Γ (its widened form) are compared within this same sweep. Both are derived from one ESA ensemble per cell, so they share the field, the sample and the ensemble size. They differ only in the widening step of Section 4.2. All methods are scored on identical data over the same $R = 50$ replicates, so the effect of the widening is read from the same tables and figures as the frontier itself.

5.4 Experiment 2: exceedance probability at unsampled	829
locations	830
	831
	832
Experiment 1 scores the predictive distribution as a whole. A risk assessment decision,	833
however, rests on the probability $\hat{P}(Z > t)$ that a regulatory limit t is exceeded at	834
an unmeasured location. The second experiment evaluates the effectiveness of each	835
method at estimating that probability, on synthetic fields drawn from the generator	836
of Experiment 1.	837
	838
	839
	840
Because the field is synthetic, every held-out node carries a known label, so a	841
method's estimates can be confronted with what actually happened. A good estimate	842
must do two things, and each is scored by its own instrument. First, it must assign	843
higher probabilities to the locations that do exceed the threshold than to those that do	844
not, which is <i>discrimination</i> and is measured by the AUC. Second, the probabilities it	845
reports must match the frequencies with which exceedances occur, which is <i>calibration</i>	846
and is measured by the reliability term of the Murphy decomposition.	847
	848
	849
	850
	851
	852
	853
	854
	855
	856
<i>Fields and thresholds.</i>	857
	858
The generator of Experiment 1 is reused with two changes that the question demands.	859
First, the Voronoi geometry is frozen. For each (η, ν) cell, five layouts are drawn once,	860
and ten replicates per layout vary only the correlated component and the nugget,	861
so every comparison is paired on geometry rather than confounded by it. Second,	862
the exceedance thresholds are fixed in advance. Thirty pilot realisations per layout,	863
discarded after use, set the thresholds t_q to the q -quantiles, $q \in \{0.90, 0.95, 0.99\}$,	864
of the pooled pilot values, and the positivity shift is frozen per layout in the same	865
way. A scored realisation is therefore judged against a threshold that no scored data	866
contributed to, and its realised exceedance rate varies around the design rate $1 - q$,	867
as a fixed regulatory limit does. A threshold read from the scored realisation's own	868
	869
	870
	871
	872
	873
	874

875 quantiles would fix that rate by construction and make miscalibration of the event
876 rate unobservable.
877

878

879

880 *Methods and evaluation.*

881

882 The sweep covers the same η and ν grids of Experiment 1, drawing $n = 120$ locations
883 of each realisation as observed values while the remaining grid nodes are held out.

884 Every method estimates $\hat{P}(Z > t)$ at each held-out node, and those estimates are

885

886 scored against the node’s realised exceedance indicator by the Brier score, the AUC,

887

888 the Murphy decomposition, and the reliability curve. The design value $\hat{F}_i^{-1}(q)$ at the
889 same level is scored by the pinball loss against the held-out truth.

890

891 The two LDC estimators are benchmarked against OK, SGS and IK, the last fitted

892

893 at the three scored thresholds directly rather than interpolated between the knots of
894 its quantile grid. Results are aggregated as means with standard errors over a cell’s

895

896 fifty realisations, together with per-realisation paired contrasts of each LDC estimator

897

898 against IK and against OK.

899

900 The sample size is additionally raised to $n = 240$ and $n = 480$ at the two cells where

901

902 departure from local stationarity is strongest, $\eta = 1.5$ and $\eta = 2.0$ at $\nu = 0.3$. Those

903

904 cells place the parametric baselines under the most stress, and the added density tests

905

906 whether they recover. The larger samples are nested, so the sample-size comparison is

907

908 paired within a realisation.

909

910

911 **5.5 Experiment 3: local anomaly detection under local**

912

913 **stationarity and non-stationarity**

914

915 The third experiment evaluates the locally normalised quantile score of Section 4

916

917 directly, as a detector of observations that are extreme for their own location. No real

918

919 survey carries labelled anomalies, so the labels are created by injection into the Meuse

920

920 zinc and Jura zinc fields of Section 5.1.

Anomaly injection.

In each field, $K = 30$ anomalies are injected. Each is displaced away from its local mean by $\delta = 1$ *local* standard deviation,

$$z_j^* = z_j + \text{sign}(z_j - \bar{z}_j^{\text{loc}}) \cdot \delta \hat{\sigma}_j^{\text{loc}}, \quad (10)$$

where $\bar{z}_j^{\text{loc}}$ and $\hat{\sigma}_j^{\text{loc}}$ are the mean and standard deviation of the $k = 12$ nearest observed neighbours of j . Injection is restricted to sites whose displaced value z_j^* stays inside the global $[10, 90]$ percentile band. Every anomaly is therefore extreme for its location yet unremarkable on the global scale, the class a global threshold cannot see. When fewer than K sites satisfy the band, all qualifying sites are used.

Detectors and evaluation.

Every point is scored by five detectors, the LDC-KDE and LDC- Γ local quantiles $|q_j^* - 0.5|$, the analogous IK quantile, and two global standardised scores, one on the raw values z_j and one on their logarithms $\log z_j$. LDC-KDE separates the contribution of the widening from that of the ensemble itself. The logarithmic variant gives the global detector the same skew-aware treatment other models receive, so that any advantage the local scores hold over it cannot be attributed to skew alone. The detectors are compared through the ROC curve and its AUC. They are compared as well through the precision at K , the fraction of the K highest-scored points that are true anomalies, with K the number of anomalies actually injected. The only random element of the experiment is which sites host the anomalies, so the injection is replicated over ten seeds and both metrics are reported as means over the replicates.

5.6 Experiment 4: leave-one-out validation on real data

The fourth experiment tests whether the calibration advantage measured on the synthetic fields survives on real measurements, where the true distributional form is

unknown. It follows a leave-one-out protocol, applied separately to the four metals of the Meuse survey ($n = 155$), namely zinc, cadmium, copper and lead, and to the zinc and cadmium fields of the Jura survey ($n = 359$). Each observation in turn is withheld, the predictive distribution of every method is estimated from the remaining observations, and the held-out value is scored under it. The scores are the CRPS, the empirical coverage of the nominal 80 %, 90 % and 95 % intervals, and the PIT, accumulated over every observation of each metal. The shared variogram is fitted on the full survey rather than refitted per fold, whereas the widening target of LDC- Γ excludes the held-out value.

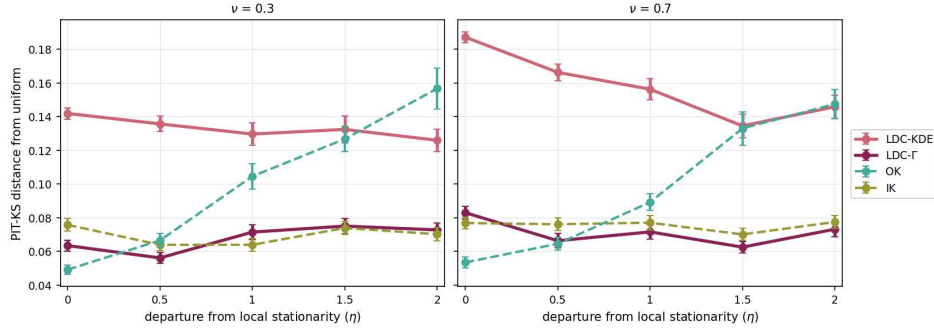
The same leave-one-out predictive distributions are further scored by the Brier score of $\hat{p}$ against each survey’s own regulatory thresholds. The predicted exceedance probabilities are grouped into ten equal-width bins. Each bin contributes its mean prediction against its observed exceedance frequency, and these pairs form reliability diagrams.

6 Results

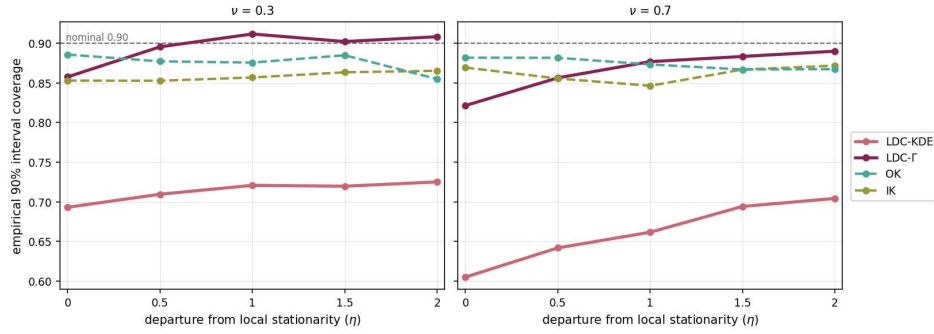
6.1 Experiment 1: the frontier tracks the departure from local stationarity

Figure 5 follows the calibration of LDC-KDE, LDC- Γ , ordinary kriging and indicator kriging, along the departure from local stationarity on three measures, the PIT-KS distance, the empirical coverage of the nominal 90 % interval and the CRPS. Table 1 reports the full grid behind it, with coverage at three nominal levels, and adding SGS to the methods.

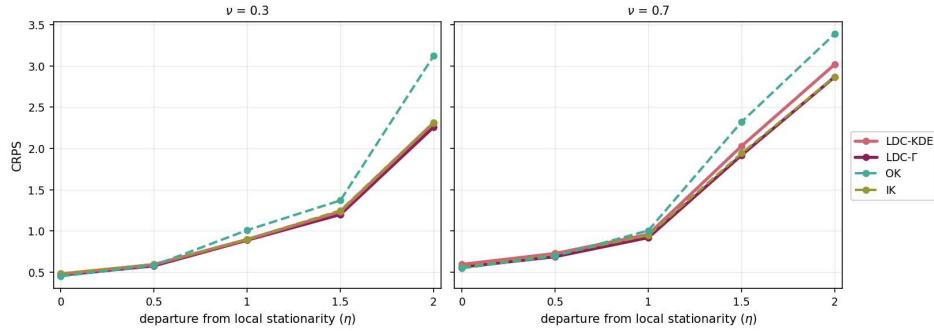
Figure 5a traces the frontier on the PIT-KS distance. On locally stationary fields ($\eta = 0$), ordinary kriging achieves the best calibration under the low nugget-to-sill ratio, with a value of 0.049. As η grows, its single global variogram approach fails to describe the changing local law. Calibration degrades steadily as a result, reaching a



(a) Kolmogorov-Smirnov distance of the probability integral transform from the uniform. Standard errors over $R = 50$ replicates are shown as vertical bars.



(b) Empirical coverage of the nominal 90% predictive interval. The grey rule marks the nominal level.



(c) CRPS.

Figure 5: Calibration against the departure from local stationarity η , on the synthetic fields of Experiment 1, for LDC-KDE, LDC- Γ , ordinary kriging and indicator kriging. In every panel the left plot is the low nugget-to-sill ratio ($\nu = 0.3$) and the right plot the high ratio ($\nu = 0.7$).

Table 1: Calibration performance on the synthetic fields, by nugget-to-sill ratio ν and departure from local stationarity η ($R = 50$ replicates). The best value of each column within a cell is in bold. The order-violation rate applies only to indicator kriging.

ν	η	Method	PIT-KS	Cov ₈₀	Cov ₉₀	Cov ₉₅	CRPS	Order viol.
0.3	0	LDC-KDE	0.142	0.571	0.693	0.776	0.478	—
0.3	0	LDC- Γ	0.063	0.751	0.858	0.914	0.463	—
0.3	0	OK	0.049	0.783	0.886	0.938	0.45	—
0.3	0	SGS	0.051	0.778	0.879	0.931	0.451	—
0.3	0	IK	0.076	0.757	0.853	0.909	0.484	0.925
0.3	0.5	LDC-KDE	0.136	0.588	0.709	0.791	0.592	—
0.3	0.5	LDC- Γ	0.056	0.799	0.895	0.941	0.577	—
0.3	0.5	OK	0.067	0.79	0.877	0.922	0.583	—
0.3	0.5	SGS	0.068	0.786	0.871	0.918	0.585	—
0.3	0.5	IK	0.064	0.758	0.853	0.91	0.596	0.983
0.3	1	LDC-KDE	0.13	0.608	0.721	0.795	0.897	—
0.3	1	LDC- Γ	0.071	0.83	0.912	0.951	0.888	—
0.3	1	OK	0.105	0.807	0.876	0.914	1.009	—
0.3	1	SGS	0.106	0.803	0.872	0.91	1.014	—
0.3	1	IK	0.064	0.768	0.857	0.908	0.899	0.996
0.3	1.5	LDC-KDE	0.132	0.619	0.72	0.789	1.234	—
0.3	1.5	LDC- Γ	0.075	0.821	0.902	0.943	1.199	—
0.3	1.5	OK	0.127	0.827	0.885	0.918	1.37	—
0.3	1.5	SGS	0.125	0.824	0.882	0.915	1.375	—
0.3	1.5	IK	0.074	0.775	0.863	0.911	1.248	0.985
0.3	2	LDC-KDE	0.126	0.626	0.725	0.792	2.311	—
0.3	2	LDC- Γ	0.073	0.832	0.908	0.945	2.259	—
0.3	2	OK	0.157	0.801	0.855	0.888	3.122	—
0.3	2	SGS	0.154	0.798	0.852	0.885	3.13	—
0.3	2	IK	0.07	0.779	0.865	0.912	2.316	0.986
0.7	0	LDC-KDE	0.187	0.487	0.605	0.693	0.595	—
0.7	0	LDC- Γ	0.083	0.707	0.821	0.889	0.562	—
0.7	0	OK	0.054	0.775	0.882	0.937	0.547	—
0.7	0	SGS	0.055	0.771	0.876	0.93	0.549	—
0.7	0	IK	0.077	0.772	0.869	0.93	0.561	0.692
0.7	0.5	LDC-KDE	0.166	0.524	0.642	0.726	0.726	—
0.7	0.5	LDC- Γ	0.066	0.748	0.856	0.915	0.687	—
0.7	0.5	OK	0.065	0.793	0.882	0.927	0.696	—
0.7	0.5	SGS	0.063	0.789	0.876	0.922	0.697	—
0.7	0.5	IK	0.076	0.755	0.855	0.913	0.712	0.934
0.7	1	LDC-KDE	0.156	0.549	0.662	0.739	0.959	—
0.7	1	LDC- Γ	0.072	0.779	0.877	0.925	0.919	—
0.7	1	OK	0.089	0.798	0.873	0.915	1.005	—
0.7	1	SGS	0.091	0.794	0.869	0.91	1.009	—
0.7	1	IK	0.077	0.753	0.846	0.909	0.947	0.974
0.7	1.5	LDC-KDE	0.134	0.595	0.694	0.763	2.029	—
0.7	1.5	LDC- Γ	0.063	0.794	0.883	0.929	1.919	—
0.7	1.5	OK	0.133	0.805	0.867	0.904	2.322	—
0.7	1.5	SGS	0.129	0.802	0.864	0.9	2.33	—
0.7	1.5	IK	0.07	0.776	0.867	0.919	1.941	0.978
0.7	2	LDC-KDE	0.146	0.61	0.704	0.77	3.02	—
0.7	2	LDC- Γ	0.073	0.807	0.89	0.933	2.868	—
0.7	2	OK	0.147	0.81	0.867	0.902	3.39	—
0.7	2	SGS	0.145	0.807	0.864	0.899	3.393	—
0.7	2	IK	0.077	0.78	0.872	0.924	2.87	0.993

value of 0.157 at $\eta = 2$. LDC- Γ and indicator kriging, in contrast, remain more stable and stay between 0.056-0.075 and 0.064-0.076, respectively, across the same range. Both have overtaken ordinary kriging by $\eta = 0.5$. Sequential Gaussian simulation stays within 0.004 of ordinary kriging at every cell of Table 1, so every panel of Figure 5 plots ordinary kriging alone.

Under the high nugget-to-sill ratio, a similar situation occurs. In this case, the crossing happens between $\eta = 0.5$ and $\eta = 1$. That *crossing* is the frontier this experiment sets out to locate, and beyond it LDC- Γ holds the better calibration. LDC- Γ remains slightly under indicator kriging for the higher η values, but the difference is not significant. The frontier of Figure 5a is therefore a frontier against the multi-Gaussian assumption, not against indicator kriging.

LDC-KDE is the variant that fails this test. Its PIT-KS distance is the largest at almost every cell, between 0.126 and 0.187 against the 0.056 to 0.083 of LDC- Γ . Figure 5b, which shows empirical coverage against the 0.9 nominal rate, demonstrates that the raw ensemble is under-dispersed everywhere, and that the deficit is worst where the nugget-to-sill ratio is high. This is the shortfall the widening of Section 4.2 repairs.

Figure 5b shows the repair. The coverage of LDC- Γ rises to between 0.858 and 0.912 under a nugget-to-sill ratio of 0.3 and between 0.821 and 0.890 under a ratio of 0.7, against a nominal 0.90. Under the low nugget-to-sill ratio it is the closest to nominal of any method at every $\eta \geq 0.5$, as Table 1 records. The same *crossing*, or frontier, is therefore present here as well, now against indicator kriging too.

Widening a predictive distribution could recover coverage by discarding the fit to the realised values, so Figure 5c scores both variants by CRPS. The numbers, however, show that it does not. Scored on the same draws as LDC-KDE, LDC- Γ attains the lower CRPS in every cell of Table 1, by margins of 1.0 to 5.9 % small enough to leave the two lines overlapping in the figure. Those margins are small beside the calibration

gap they accompany. At $\eta = 1$ under the low nugget-to-sill ratio, LDC-KDE covers 72 % of its nominal 90 % intervals against 91 % for LDC- Γ , yet only 1.0 % of CRPS separates them. The score integrates the pinball loss over every quantile level with uniform weight, so the centre of the predictive distribution dominates it. An overconfident tail is penalised only weakly as a result. The frontier of Figure 5a is therefore drawn on the PIT.

Against ordinary kriging, the CRPS does separate the methods, in the pattern the frontier predicts. Ordinary kriging holds the lower score at $\eta = 0$, by 2.9 % under the low nugget-to-sill ratio. It then loses that edge, by 13.6 % at $\eta = 1$ and by 38.2 % at $\eta = 2$, both differences significant at $p < 0.001$. Indicator kriging is the only baseline that keeps pace. LDC- Γ scores lower in nine of the ten cells of Table 1, by at most 4.5 %. The CRPS therefore places the frontier where the PIT does, against the multi-Gaussian assumption.

6.2 Experiment 2: exceedance probability at unsampled locations

Table 2 reports discrimination, the ability to rank the held-out locations that exceed the threshold above those that do not, through the AUC. At $\eta = 0$, ordinary kriging is the strongest ranker at all three threshold levels. LDC- Γ , however, improves as the field departs from local stationarity, gaining between 0.05 and 0.16 of AUC from $\eta = 0$ to $\eta = 2$ at every scenario. It has overtaken ordinary kriging by $\eta = 1$ at $q = 0.90$ and 0.95. The crossing thus falls within the same interval where Experiment 1 located the calibration frontier. The one cell it never reaches is the deep tail ($q = 0.99$) under the high nugget-to-sill ratio, where the two methods are comparable at $\eta = 2$, 0.778 for LDC- Γ against 0.767 for ordinary kriging.

Indicator kriging shares the improvement with η but trails LDC- Γ in every cell at every threshold level. It comes closest at $q = 0.90$, within 0.05 at every non-stationary

cell. It falls furthest in the deep tail, roughly 0.15 behind at $q = 0.99$ and $\eta = 2$ under both ratios.

The deep tail also separates the two LDC variants. Since both are built from the same ESA pass, the gap between them isolates what the widening of Section 4.2 contributes to discrimination. At $\eta = 2$ and $q = 0.99$, LDC-KDE attains 0.635 under the low nugget-to-sill ratio and 0.575 under the high, below every other method, against 0.851 and 0.778 for LDC- Γ . The under-dispersed raw ensemble carries too little mass beyond so deep a threshold to grade the rare exceedances.

Calibration is measured by the reliability term of the Murphy decomposition, listed over the grid in Table 3. The term orders LDC- Γ and ordinary kriging in the same direction as the AUC. LDC- Γ stays below 0.01 in every cell. Ordinary kriging's reported probabilities, by contrast, drift furthest from the observed frequencies where the departure is strongest, reaching 0.0281 at $q = 0.90$ and $\eta = 2$ under the low nugget-to-sill ratio against 0.0063 for LDC- Γ .

The smallest terms, however, belong to the forecasts that commit least. Indicator kriging attains 0.0043 in that same cell, and at $q = 0.99$ it joins LDC-KDE below 0.0008 across the sweep. The reliability curves of Figure 6, drawn at $q = 0.90$ under the low nugget-to-sill ratio where indicator kriging's calibration is strongest, show how. Each curve stops at its highest mean forecast, marked by a dotted rule, so a short curve records a forecast that seldom commits. Indicator kriging reaches 0.24–0.35 across the panels against 0.42–0.57 for LDC- Γ . Its calibration is therefore earned near the base rate, where the reliability term costs little. Part of that caution is imposed by the order-relation repair of Section 5.2. Its cost is the deep-tail discrimination deficit of Table 2.

The same predictive distribution also yields a design value, the level $\hat{F}_i^{-1}(q)$ a method expects to be exceeded with probability $1 - q$. The tail pinball loss scores it against the held-out truth (Figure 7). LDC- Γ attains the lowest loss of the three

Table 2: Area under the ROC curve by threshold level q , nugget-to-sill ratio ν and departure from local stationarity η , at $n = 120$. Each entry is the mean over $R = 50$ realisations. The best of each row is in bold.

q	ν	η	LDC-KDE	LDC- Γ	OK	SGS	IK
0.90	0.3	0	0.785	0.789	0.819	0.817	0.727
0.90	0.3	0.5	0.844	0.858	0.867	0.862	0.812
0.90	0.3	1	0.834	0.843	0.836	0.831	0.797
0.90	0.3	1.5	0.844	0.888	0.854	0.847	0.846
0.90	0.3	2	0.827	0.843	0.761	0.754	0.819
0.90	0.7	0	0.653	0.655	0.666	0.654	0.601
0.90	0.7	0.5	0.773	0.780	0.779	0.775	0.735
0.90	0.7	1	0.800	0.806	0.792	0.785	0.789
0.90	0.7	1.5	0.866	0.865	0.842	0.834	0.843
0.90	0.7	2	0.764	0.812	0.763	0.756	0.794
0.95	0.3	0	0.784	0.795	0.841	0.836	0.718
0.95	0.3	0.5	0.814	0.849	0.864	0.853	0.778
0.95	0.3	1	0.801	0.831	0.823	0.816	0.758
0.95	0.3	1.5	0.802	0.889	0.849	0.827	0.803
0.95	0.3	2	0.812	0.868	0.747	0.740	0.807
0.95	0.7	0	0.645	0.661	0.684	0.660	0.592
0.95	0.7	0.5	0.735	0.762	0.767	0.760	0.697
0.95	0.7	1	0.774	0.790	0.781	0.762	0.745
0.95	0.7	1.5	0.816	0.834	0.814	0.798	0.798
0.95	0.7	2	0.713	0.808	0.750	0.740	0.764
0.99	0.3	0	0.728	0.798	0.869	0.852	0.633
0.99	0.3	0.5	0.727	0.834	0.868	0.818	0.696
0.99	0.3	1	0.713	0.828	0.810	0.779	0.660
0.99	0.3	1.5	0.699	0.889	0.885	0.808	0.739
0.99	0.3	2	0.635	0.851	0.777	0.700	0.706
0.99	0.7	0	0.603	0.662	0.713	0.676	0.548
0.99	0.7	0.5	0.654	0.739	0.739	0.707	0.621
0.99	0.7	1	0.658	0.748	0.770	0.701	0.658
0.99	0.7	1.5	0.727	0.799	0.773	0.719	0.692
0.99	0.7	2	0.575	0.778	0.767	0.690	0.629

methods plotted at almost every cell with $\eta \geq 1$. The two exceptions fall under the high nugget-to-sill ratio. Its margin widens with the threshold level. At $\eta = 2$ under the low nugget-to-sill ratio and $q = 0.99$ it reads 0.291 against 0.434 for indicator kriging and 0.567 for ordinary kriging.

Table 3: Murphy reliability term by threshold level q , nugget-to-sill ratio ν and departure from local stationarity η , at $n = 120$. Each entry is the mean over $R = 50$ realisations.

q	ν	η	LDC-KDE	LDC- Γ	OK	SGS	IK
0.90	0.3	0	0.0044	0.0023	0.0022	0.0021	0.0035
0.90	0.3	0.5	0.0043	0.0019	0.0026	0.0027	0.0039
0.90	0.3	1	0.0048	0.0022	0.0033	0.0032	0.0032
0.90	0.3	1.5	0.0055	0.0026	0.0071	0.0067	0.0039
0.90	0.3	2	0.0074	0.0063	0.0281	0.0278	0.0043
0.90	0.7	0	0.0062	0.0028	0.0018	0.0015	0.0027
0.90	0.7	0.5	0.0059	0.0031	0.0033	0.0033	0.0040
0.90	0.7	1	0.0066	0.0066	0.0126	0.0125	0.0050
0.90	0.7	1.5	0.0054	0.0051	0.0209	0.0199	0.0052
0.90	0.7	2	0.0106	0.0042	0.0090	0.0090	0.0046
0.95	0.3	0	0.0018	0.0015	0.0012	0.0012	0.0015
0.95	0.3	0.5	0.0022	0.0019	0.0024	0.0023	0.0026
0.95	0.3	1	0.0016	0.0027	0.0036	0.0036	0.0019
0.95	0.3	1.5	0.0021	0.0017	0.0032	0.0031	0.0020
0.95	0.3	2	0.0030	0.0047	0.0172	0.0172	0.0027
0.95	0.7	0	0.0019	0.0017	0.0009	0.0008	0.0013
0.95	0.7	0.5	0.0021	0.0023	0.0024	0.0025	0.0022
0.95	0.7	1	0.0031	0.0070	0.0135	0.0134	0.0025
0.95	0.7	1.5	0.0038	0.0083	0.0275	0.0204	0.0024
0.95	0.7	2	0.0030	0.0032	0.0059	0.0059	0.0016
0.99	0.3	0	0.0002	0.0006	0.0002	0.0002	0.0003
0.99	0.3	0.5	0.0003	0.0013	0.0013	0.0012	0.0006
0.99	0.3	1	0.0001	0.0014	0.0009	0.0009	0.0002
0.99	0.3	1.5	0.0002	0.0009	0.0005	0.0003	0.0005
0.99	0.3	2	0.0001	0.0008	0.0011	0.0008	0.0004
0.99	0.7	0	0.0001	0.0006	0.0002	0.0001	0.0002
0.99	0.7	0.5	0.0002	0.0015	0.0008	0.0009	0.0003
0.99	0.7	1	0.0003	0.0059	0.0093	0.0092	0.0003
0.99	0.7	1.5	0.0008	0.0097	0.0206	0.0202	0.0004
0.99	0.7	2	0.0002	0.0021	0.0028	0.0027	0.0002

Because every method is scored on the same realisations, the differences can be tested pairwise within a realisation rather than read off the means. Tables 4a and 4b display those contrasts at the strongest departure examined, using ordinary kriging and indicator kriging as baseline, respectively.

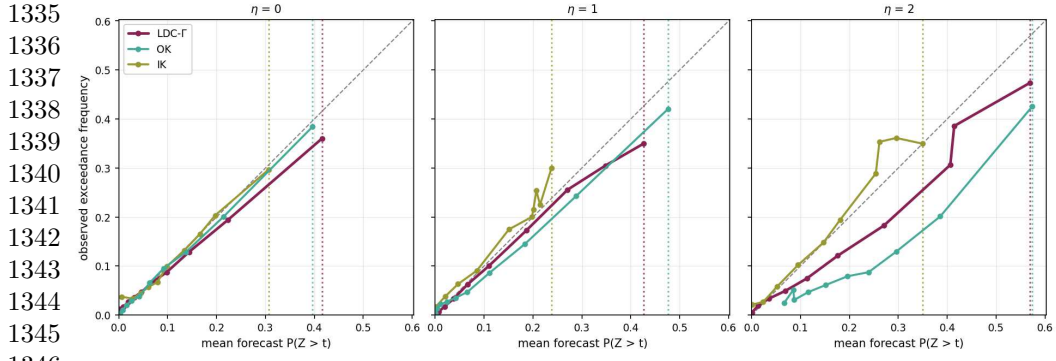


Figure 6: Pooled reliability curves at $q = 0.90$ and $\nu = 0.3$, for LDC- Γ , ordinary kriging and indicator kriging. Predicted exceedance probability on the horizontal axis against the observed exceedance frequency in each bin, with the dashed diagonal marking perfect calibration. The dotted vertical rule in each method's colour marks the mean forecast of that method's highest bin. Departure from local stationarity varies in columns for $\eta \in \{0.0, 1.0, 2.0\}$.

Against ordinary kriging, LDC- Γ holds an advantage that narrows as the nugget-to-sill ratio rises. It reduces the Brier score by 26.9% at $q = 0.90$ and 33.9% at $q = 0.95$ under the low nugget-to-sill ratio, and by 5.1% and 6.5% under the high one, all with $p < 0.001$. Only in the deep tail does the reduction vanish. The contrast at $q = 0.99$ turns to -1.9% under the low ratio ($p = 0.336$). The design value carries the comparison there instead, LDC- Γ improving on the pinball loss of ordinary kriging by 53.1% at that level ($p < 0.001$).

Against indicator kriging, the three instruments disagree. LDC- Γ holds no significant Brier advantage in any cell, and it loses in the deep tail by 1.4% and 6.6% under the two ratios. Ranking and design value go the other way. LDC- Γ leads by 0.145 and 0.149 of AUC at $q = 0.99$ under the two ratios, and by 37.6% of pinball loss under the low one, all with $p < 0.001$. Indicator kriging therefore issues the better-scoring probability at that level, while LDC- Γ identifies the sites at risk and reports the safer design value. That Brier edge is earned near the base rate, where exceedances are rarest and a forecast that seldom commits is penalised least.

(a) Against ordinary kriging

q	ν	Method	AUC diff.	Brier adv. (%)	Reliability adv. (%)	Pinball adv. (%)
0.90	0.3	LDC-KDE	+0.066 (j0.001)	24.3 (j0.001)	213.3 (j0.001)	11.1 (j0.001)
0.90	0.3	LDC- Γ	+0.081 (j0.001)	26.9 (j0.001)	362.0 (j0.001)	33.5 (j0.001)
0.90	0.7	LDC-KDE	+0.000 (0.986)	-3.0 (0.060)	-26.1 (0.037)	6.1 (0.316)
0.90	0.7	LDC- Γ	+0.048 (0.001)	5.1 (j0.001)	98.8 (j0.001)	15.0 (j0.001)
0.95	0.3	LDC-KDE	+0.064 (j0.001)	39.2 (j0.001)	415.6 (j0.001)	2.6 (0.273)
0.95	0.3	LDC- Γ	+0.120 (j0.001)	33.9 (j0.001)	257.9 (j0.001)	42.1 (j0.001)
0.95	0.7	LDC-KDE	-0.038 (0.054)	1.4 (0.145)	30.6 (0.003)	-8.7 (0.001)
0.95	0.7	LDC- Γ	+0.058 (0.007)	6.5 (j0.001)	89.4 (j0.001)	15.3 (j0.001)
0.99	0.3	LDC-KDE	-0.142 (j0.001)	-0.6 (0.804)	508.8 (j0.001)	-32.2 (j0.001)
0.99	0.3	LDC- Γ	+0.074 (0.006)	-1.9 (0.336)	16.2 (0.729)	53.1 (j0.001)
0.99	0.7	LDC-KDE	-0.192 (j0.001)	0.6 (0.003)	976.8 (j0.001)	-45.2 (j0.001)
0.99	0.7	LDC- Γ	+0.011 (0.616)	-0.7 (0.665)	13.4 (0.417)	18.4 (j0.001)

(b) Against indicator kriging

q	ν	Method	AUC diff.	Brier adv. (%)	Reliability adv. (%)	Pinball adv. (%)
0.90	0.3	LDC-KDE	+0.008 (0.191)	-5.5 (0.395)	-40.7 (0.002)	-11.1 (j0.001)
0.90	0.3	LDC- Γ	+0.024 (0.001)	2.0 (0.922)	-23.2 (0.309)	1.4 (0.090)
0.90	0.7	LDC-KDE	-0.031 (j0.001)	-9.5 (j0.001)	-71.3 (j0.001)	-16.4 (j0.001)
0.90	0.7	LDC- Γ	+0.018 (0.029)	-0.8 (0.907)	-24.3 (0.304)	0.2 (0.488)
0.95	0.3	LDC-KDE	+0.004 (0.626)	-2.4 (0.381)	0.4 (0.508)	-19.1 (j0.001)
0.95	0.3	LDC- Γ	+0.061 (j0.001)	-3.0 (0.249)	-38.8 (0.033)	4.5 (0.002)
0.95	0.7	LDC-KDE	-0.051 (j0.001)	-6.8 (j0.001)	-58.7 (j0.001)	-24.7 (j0.001)
0.95	0.7	LDC- Γ	+0.045 (j0.001)	-0.7 (0.005)	-50.8 (0.002)	2.2 (0.919)
0.99	0.3	LDC-KDE	-0.071 (j0.001)	-0.0 (0.427)	133.3 (j0.001)	-42.4 (j0.001)
0.99	0.3	LDC- Γ	+0.145 (j0.001)	-1.4 (0.024)	-46.8 (0.003)	37.6 (j0.001)
0.99	0.7	LDC-KDE	-0.053 (0.017)	-0.1 (0.022)	79.2 (0.005)	-50.4 (j0.001)
0.99	0.7	LDC- Γ	+0.149 (j0.001)	-6.6 (j0.001)	-76.8 (j0.001)	4.6 (0.322)

Table 4: Paired advantage of each LDC variant over OK and IK at the strongest departure examined ($\eta = 2$, $n = 120$), by threshold level q and nugget-to-sill ratio ν . Differences are means for AUC and medians for the percentage columns, positive favouring LDC, with the p -value over $R = 50$ realisations in parentheses.

The percentages are per-realisation statistics; they do not equal the ratios of the cell means reported elsewhere. At $q = 0.99$ roughly one node in a hundred exceeds the threshold, so a Brier contrast becomes unstable; pinball is the dependable tail instrument at that level.

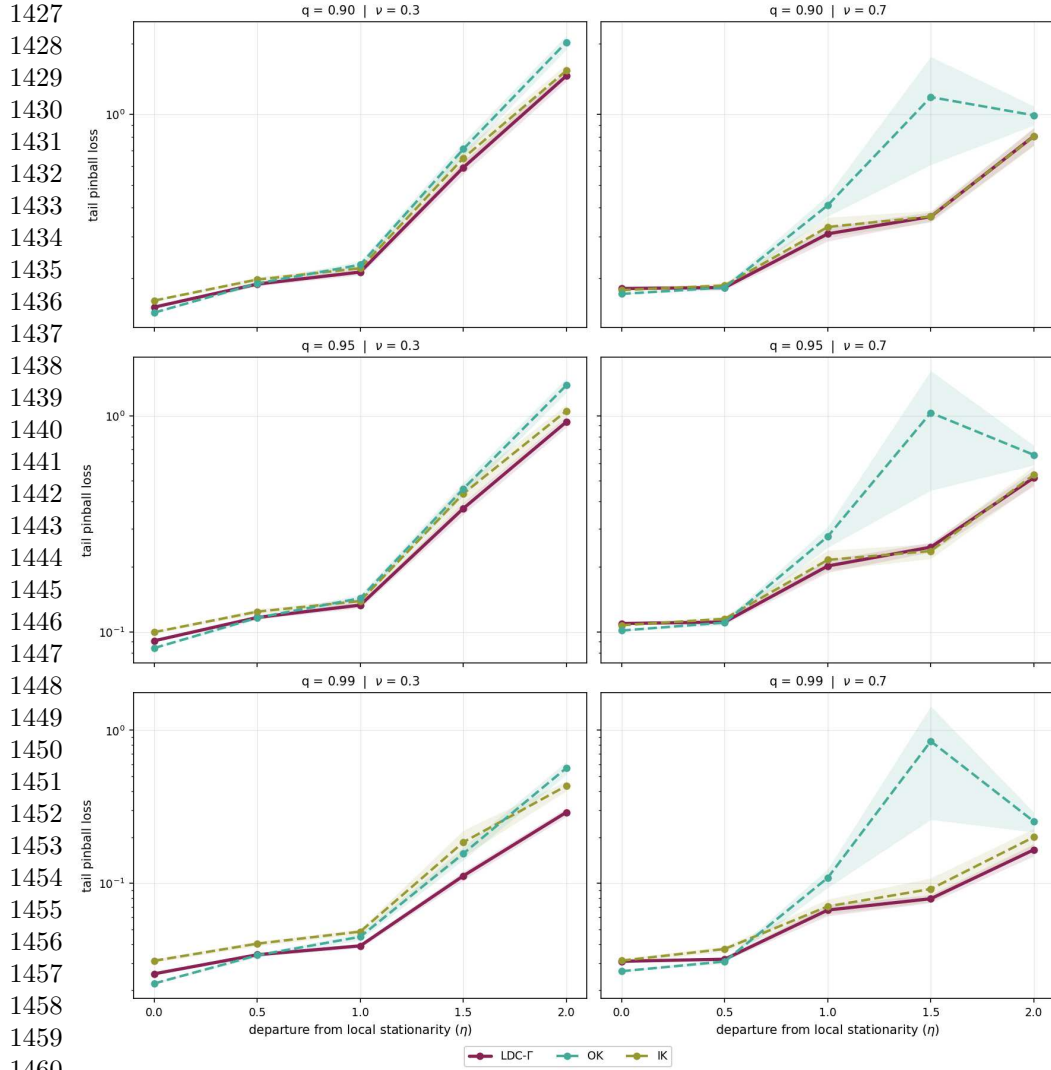


Figure 7: Tail pinball loss of the design value $\hat{F}_i^{-1}(q)$ against the held-out truth (logarithmic scale) for LDC- Γ , OK and IK, plotted against the departure from local stationarity η at $n = 120$. Rows: threshold level $q \in \{0.90, 0.95, 0.99\}$; columns: nugget-to-sill ratio ν . Bands are ± 1 standard error over $R = 50$ realisations.

The sample size is held fixed across the frontier, so a close competitor may be one that is merely short of data. The two cells where indicator kriging comes closest are therefore rerun on nested samples of $n \in \{120, 240, 480\}$, reported in Tables A1 and A2 of Appendix A.

Denser sampling recovers indicator kriging at the moderate threshold alone. It draws level at $q = 0.90$ once n reaches 240. At $q = 0.95$, it closes to a sixth of its initial deficit in AUC without overtaking ($p = 0.026$ at $n = 480$). The deep tail does not follow. LDC- Γ keeps the lead in discrimination and the lowest tail loss at every sample size. What separates the two regimes is the number of observations that fall beyond the threshold, which denser sampling raises at $q = 0.90$ far faster than at $q = 0.99$. Ordinary kriging recovers at no sample size at $\eta = 2$, its deficit belonging to the departure rather than to the data volume.

6.3 Experiment 3: anomaly detection from the same predictive distribution

We now evaluate anomaly detection through the locally normalised quantile score, computed from the same predictive distribution the preceding experiments read through calibration. Table 5 reports the AUC and the precision at K of every detector on both surveys, and Figure 8 the ROC curves behind the AUC.

LDC- Γ attains the best AUC and the best precision at K on both surveys, with an AUC of 0.731 ± 0.041 on Meuse zinc and 0.747 ± 0.042 on Jura zinc. The precision at K orders the five detectors identically on the two fields, with LDC-KDE second, indicator kriging third, and the two global z -scores last. The AUC agrees except at the second place, which LDC-KDE and indicator kriging exchange between the fields by less than a tenth of a standard deviation. This tie indicates that the discrimination margin of LDC- Γ is due to the widening rather than to the ensemble alone. In contrast, both LDC variants outperform indicator kriging on the precision at K , so within the K highest-scored sites the advantage in recovering anomalies does not rest on the widening.

The three local quantiles all reach AUCs well above 0.5, whereas the two global z -scores remain at or below it. This is expected of a score blind to an anomaly that is

globally ordinary by construction. That margin measures what the local reference is worth. Given a randomly chosen anomaly and a randomly chosen ordinary site, LDC- Γ scores the anomaly higher in roughly three cases out of four. In contrast, a global score does no better than a coin toss.

Table 5: Metrics for local anomaly detection on two Zinc surveys. Area under the ROC curve and precision at K injected anomalies are shown as mean $\pm$ standard deviation over 10 replicates. Best per survey in bold.

Dataset	Method	AUC	P@K
Meuse	LDC- Γ	0.731 $\pm$ 0.041	0.417 $\pm$ 0.057
Meuse	LDC-KDE	0.679 $\pm$ 0.045	0.387 $\pm$ 0.082
Meuse	IK	0.678 $\pm$ 0.032	0.220 $\pm$ 0.028
Meuse	z (raw)	0.439 $\pm$ 0.021	0.183 $\pm$ 0.024
Meuse	z (log)	0.434 $\pm$ 0.018	0.113 $\pm$ 0.017
Jura	LDC- Γ	0.747 $\pm$ 0.042	0.143 $\pm$ 0.045
Jura	LDC-KDE	0.697 $\pm$ 0.045	0.123 $\pm$ 0.055
Jura	IK	0.701 $\pm$ 0.037	0.103 $\pm$ 0.046
Jura	z (raw)	0.491 $\pm$ 0.031	0.000 $\pm$ 0.000
Jura	z (log)	0.472 $\pm$ 0.027	0.000 $\pm$ 0.000

When comparing between surveys, the margin of LDC- Γ over indicator kriging in AUC is nearly the same. By contrast, the two differ greatly on the precision at K , but these values are not comparable in practice. Jura has a greater number of sites, so the anomalies represent a smaller proportion of the field, 0.08 against 0.19. That proportion is what a ranking without skill would already recover. It therefore sets a different floor on each survey.

Figure 9a shows the per-replicate spread behind each mean AUC. Replicate noise is as large as the margin between LDC- Γ and indicator kriging, so the two are compared within each draw in Figure 9b. LDC- Γ wins all ten replicates on both surveys, so its lead is not an artefact of a single injection draw. LDC-KDE and indicator kriging, by contrast, trade the wins from draw to draw.

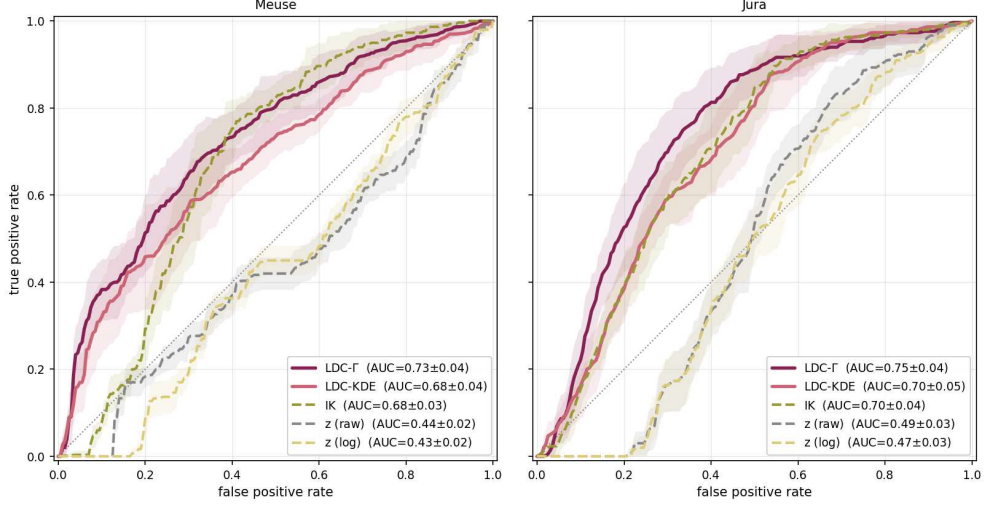
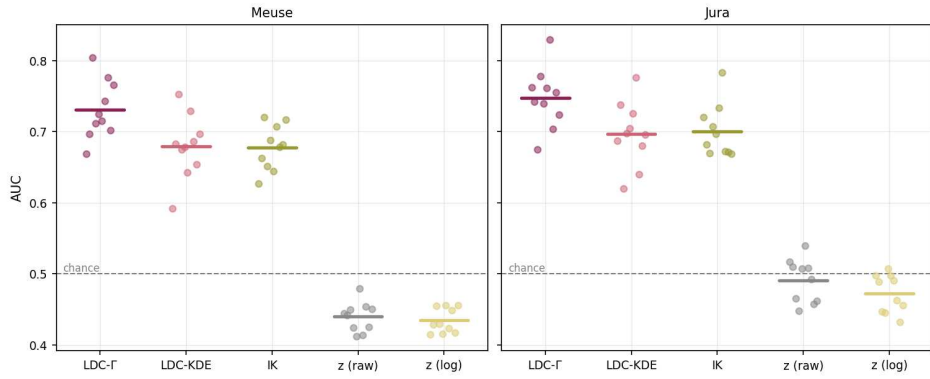
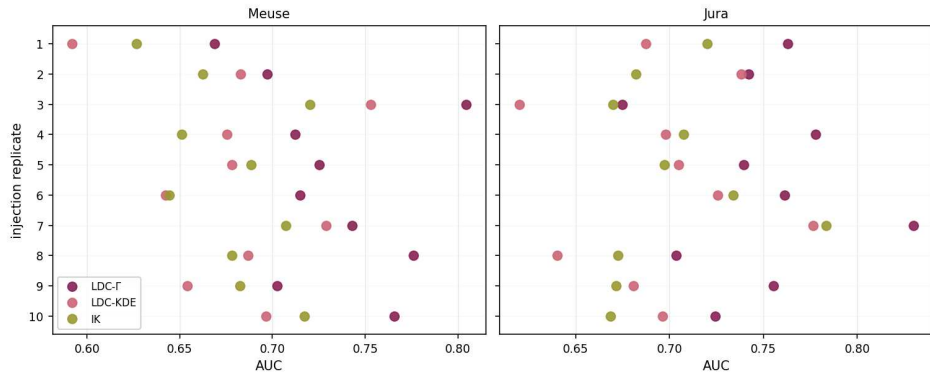


Figure 8: ROC curves of the five anomaly detectors on Meuse zinc (left) and Jura zinc (right), namely the LDC- Γ , LDC-KDE and indicator kriging local quantiles and the raw and log global z -scores. Curves are means over the ten injection replicates, with bands at ± 1 standard deviation.

Finally, Figure 10a maps which anomalies each detector recovers on the Meuse field. Figure 10b shows the same for Jura, where every detector recovers only a small fraction of the injections. The models capture different anomalies, a pattern especially evident on Jura. Jura shows this most clearly because far more sites reach the edge of their local distribution than there are flags. A small difference in ranking therefore decides which of these near-tied candidates each detector shows.

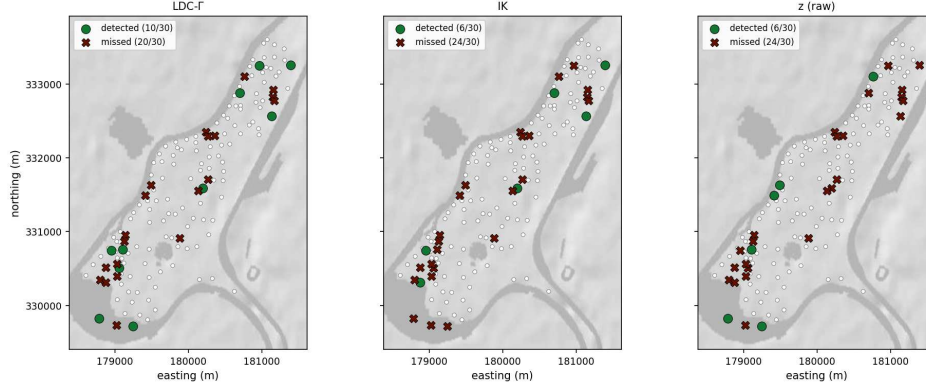


(a) AUC per model (one marker per replicate). Solid rule at the mean over replicates.

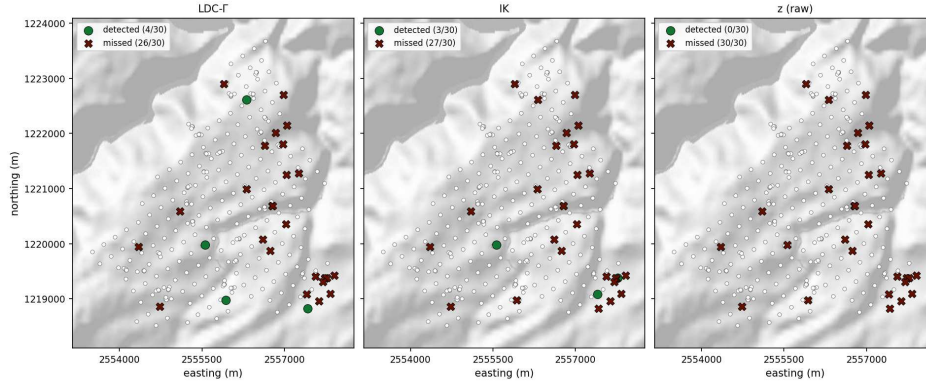


(b) AUC for each injection replicate, one row per replicate.

Figure 9: Per-replicate AUC on Meuse zinc (left of each panel) and Jura zinc (right), over the ten injection replicates of Experiment 3.



(a) Meuse zinc.



(b) Jura zinc.

Figure 10: Injected local anomalies on the Meuse and Jura zinc data and the sites flagged by each detector. Panels, from left to right, show the LDC- Γ local quantile, the indicator kriging local quantile and the global z -score.

1703 6.4 Experiment 4: leave-one-out validation on real data

1704

1705 6.4.1 Distributional fidelity and calibration

1706

1707 Table 6 reports the leave-one-out CRPS, interval coverage and PIT-KS distance on
1708 both surveys. Both LDC variants attain a lower CRPS than every baseline on all six
1709 materials. LDC-KDE is the sharper of the two on all of them but Jura zinc. Measured
1710 against the strongest baseline of each material, the LDC- Γ margin on Meuse runs
1711 from 16.4 % on zinc (86.3 against 103.1 for ordinary kriging) to 25.0 % on cadmium
1712 (0.86 against 1.14 for indicator kriging). That margin narrows on Jura, to 12.8 % on
1713 cadmium and 11.2 % on zinc. Unlike in the previous experiments, the advantage over
1714 the baselines therefore belongs to the local estimator itself, not to the widening. It
1715 survives a change of survey, of metal and of sampling density, which leaves little room
1716 for one favourable field to account for it.

1717 The calibration measures do not follow the accuracy. At the nominal 90 % level,
1718 the two LDC variants over-cover on nine of their twelve rows. The interval holds up to
1719 98.7 % of the held-out values. All three exceptions fall on cadmium, LDC-KDE under-
1720 covering at 0.877 on Meuse and 0.836 on Jura, and LDC- Γ landing on 0.900 exactly
1721 on Jura. The PIT-KS distance admits no such exception. On all six materials it is
1722 larger for both LDC variants than for any baseline, 0.161 for LDC- Γ on Meuse zinc
1723 against 0.077 for the best baseline there.

1724 Figure 11 locates that miscalibration. On the four Meuse metals and on Jura zinc,
1725 the LDC histograms peak in the central bins and thin out at both ends. This is the
1726 signature of an over-dispersed predictive distribution rather than a shifted one, which
1727 would instead pile the mass against one end. On Meuse zinc, no held-out value at
1728 all falls in the lowest decile of the LDC- Γ distribution. 66 % fall in the fourth to the
1729 seventh decile instead, where 40 % would be expected.

1730 Between the two LDC variants, the widening is what moves the calibration. The
1731 LDC- Γ histograms are flatter than the LDC-KDE ones on every material but Meuse
1732

Table 6: Leave-one-out calibration performance on Meuse and Jura surveys. The best value of each column within a material is in bold.

Dataset	Material	Method	CRPS	Cov ₈₀	Cov ₉₀	Cov ₉₅	PIT-KS	Order viol.
Meuse	cadmium	LDC-KDE	0.808	0.806	0.877	0.942	0.124	–
Meuse	cadmium	LDC- Γ	0.856	0.884	0.929	0.968	0.178	–
Meuse	cadmium	OK	1.248	0.832	0.929	0.942	0.102	–
Meuse	cadmium	SGS	1.266	0.852	0.929	0.948	0.118	–
Meuse	cadmium	IK	1.142	0.897	0.929	0.955	0.104	0.897
Meuse	copper	LDC-KDE	6.052	0.955	0.987	0.994	0.16	–
Meuse	copper	LDC- Γ	6.11	0.877	0.948	0.974	0.131	–
Meuse	copper	OK	7.89	0.826	0.948	0.987	0.065	–
Meuse	copper	SGS	7.967	0.826	0.942	0.981	0.063	–
Meuse	copper	IK	8.01	0.768	0.865	0.903	0.09	0.948
Meuse	lead	LDC-KDE	27.294	0.929	0.981	0.981	0.177	–
Meuse	lead	LDC- Γ	28.198	0.91	0.948	0.961	0.169	–
Meuse	lead	OK	34.472	0.832	0.897	0.968	0.085	–
Meuse	lead	SGS	34.492	0.839	0.89	0.955	0.082	–
Meuse	lead	IK	35.958	0.787	0.871	0.897	0.119	0.942
Meuse	zinc	LDC-KDE	83.298	0.935	0.968	0.981	0.181	–
Meuse	zinc	LDC- Γ	86.272	0.948	0.955	0.981	0.161	–
Meuse	zinc	OK	103.144	0.871	0.948	0.968	0.088	–
Meuse	zinc	SGS	103.599	0.845	0.948	0.968	0.091	–
Meuse	zinc	IK	111.152	0.781	0.871	0.91	0.077	0.968
Jura	Cd	LDC-KDE	0.326	0.749	0.836	0.911	0.097	–
Jura	Cd	LDC- Γ	0.335	0.85	0.9	0.944	0.08	–
Jura	Cd	OK	0.387	0.841	0.914	0.955	0.045	–
Jura	Cd	SGS	0.389	0.841	0.914	0.958	0.041	–
Jura	Cd	IK	0.384	0.73	0.891	0.942	0.043	0.942
Jura	Zn	LDC-KDE	11.165	0.922	0.972	0.986	0.155	–
Jura	Zn	LDC- Γ	10.599	0.83	0.911	0.95	0.077	–
Jura	Zn	OK	11.933	0.811	0.894	0.942	0.059	–
Jura	Zn	SGS	12.043	0.819	0.894	0.922	0.064	–
Jura	Zn	IK	12.111	0.766	0.855	0.911	0.045	0.916

The order-violation rate only applies to indicator kriging.

cadmium. The effect is strongest on Jura zinc, where the median decile falls from 24 to 17% of the observations. That is why LDC- Γ is the variant carried through the applications above. The baseline rows stay flatter still on every material of both surveys.

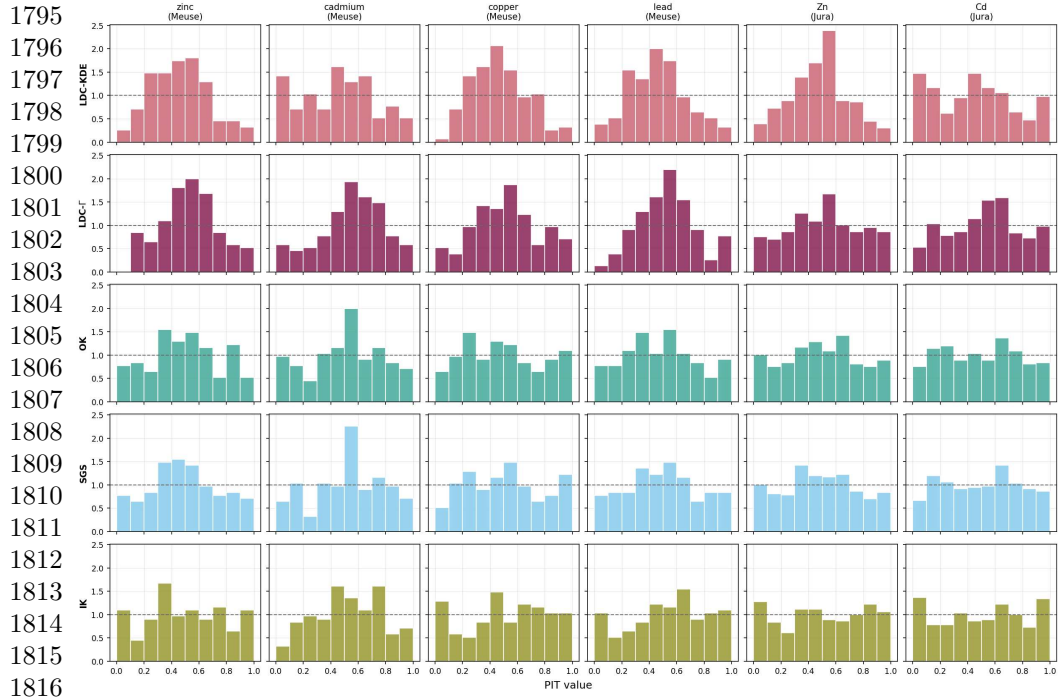


Figure 11: Probability integral transform histograms for the leave-one-out predictions on Meuse and Jura metals, by method (rows) and material (columns). The dashed line marks the uniform density of a perfectly calibrated predictor would produce.

6.4.2 Coherence and calibrated exceedance probabilities

Table 7 gives the Brier score of the same leave-one-out predictive distributions against each survey's regulatory thresholds. LDC-KDE takes the lowest score at four of the six thresholds, and LDC- Γ at a fifth. On the three thresholds carrying more than thirty exceedances, the leading LDC variant sits 9 to 19% below the best baseline. The sixth win goes to indicator kriging. Its seven exceedances place all five methods within 0.3×10^{-2} of one another, so that win separates nothing.

Indicator kriging is the only baseline built to estimate $P(Z > t)$ directly. It nonetheless holds the worst Brier score at every Meuse threshold, and leads only in the one case where the margins are small. The two LDC variants, for their part, are never more

than 0.3×10^{-2} apart. The widening that flattens the PIT histograms of Figure 11 therefore does not cost LDC- Γ much.

The indicator kriging score here is the repaired one throughout. That repair was necessary at 90 to 97 % of locations (Table 6), and it only lowers the probability issued. Neither LDC variant requires it. Its implicit copula structure is coherent by construction (Section 3.2). Furthermore, the two surveys' scores are not comparable to one another, as the Dutch thresholds are remediation limits while the Swiss remediation values lie above the entire Jura survey.

Table 7: Leave-one-out Brier score of $\hat{P}(Z > t)$ ($\times 10^{-2}$) against the regulatory threshold t , with the best per row in bold. The Meuse thresholds are the Dutch intervention values (Ministry of Housing, Spatial Planning and the Environment (VROM) 2009); the Jura thresholds are the Swiss VBBo guide values and, for cadmium, the food-crop trigger value. n_+ is the number of sites exceeding t .

Survey	Metal	t (mg/kg)	n_+	LDC-KDE	LDC- Γ	OK	SGS	IK
Meuse	Zinc	720	33	7.12	7.4	8.81	8.83	9.35
Meuse	Cadmium	13	3	1.87	1.88	1.96	2.04	2.08
Meuse	Lead	530	2	0.76	0.84	1.15	1.11	1.18
Jura	Zinc (guide)	150	7	2.21	2.03	2.01	2.02	1.93
Jura	Cadmium (guide)	0.8	233	14.79	14.59	16.63	16.7	16.07
Jura	Cadmium (trigger)	2	52	10.48	10.78	11.92	11.96	12.01

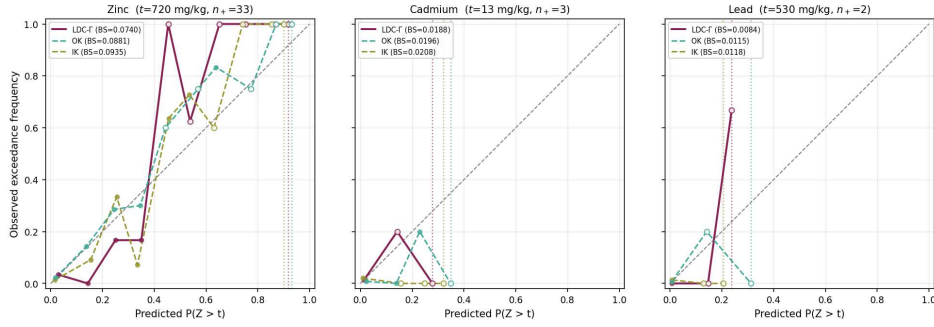
The Brier score records that a method is miscalibrated without saying where or in which direction. Figure 12 shows both, together with the largest probability each method ever issues, marked by the dotted rules. On Meuse zinc (Figure 12a), the LDC- Γ curve lies below the diagonal between $\hat{p} = 0.14$ and $\hat{p} = 0.35$, while ordinary kriging stays on it there. LDC- Γ therefore over-estimates $\hat{P}(Z > t)$ at these sites. This follows from the over-dispersion diagnosed in Figure 11, whose surplus upper-tail mass inflates the exceedance probability wherever the observation lies well below the threshold. Beyond $\hat{p} = 0.4$, however, every point represents fewer than 10 samples, so the three methods cannot be reliably compared over the range in which an exceedance would

1887 actually be declared. The cadmium and lead panels are uninformative altogether,
1888
1889 neither metal exceeding at more than three sites.

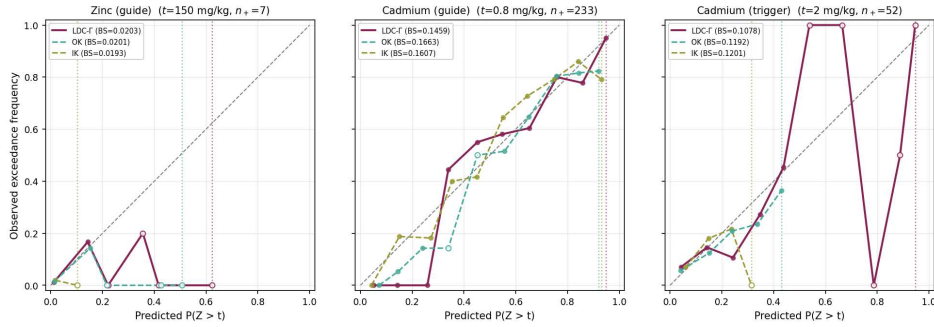
1890 The Jura cadmium thresholds supply the exceedances the Meuse lacks (Fig-
1891
1892 ure 12b). At the guide value, LDC- Γ holds the diagonal up to its rule, whereas ordinary
1893
1894 kriging again falls below it at the top. At the food-crop trigger value, moreover, it is
1895 the rules themselves that separate the methods. Ordinary and indicator kriging stop
1896
1897 at $\hat{p} = 0.43$ and 0.31 respectively, so neither can call a site more likely than not to
1898
1899 exceed. Yet 52 sites do. LDC- Γ alone reaches past one half, if only on hollow points.
1900 The baselines thus obtain part of their calibration by declining to commit, which the
1901
1902 Brier score does not reveal. Indicator kriging's win on Jura zinc has the same origin;
1903
1904 its curve never leaves $\hat{p} = 0.10$.

1905 Figure 13 shows the calibrated $\hat{P}(Z > t)$ as a continuous surface across the entire
1906
1907 estimation grid. It interpolates LDC- Γ 's exceedance probability for zinc across the
1908
1909 Meuse floodplain grid from the original observations, referenced against the Dutch
1910
1911 intervention value of 720 mg kg^{-1} . No ground truth exists away from the sampled sites,
1912 so the map illustrates the method's output rather than adding validation evidence.
1913
1914 That evidence remains Table 7 and Figure 12 above.

1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932



(a) Meuse, against the Dutch soil intervention values.



(b) Jura, against the Swiss VBBo guide values.

Figure 12: Reliability diagrams for $\hat{P}(Z > t)$ on the Meuse and Jura surveys, for LDC- Γ , ordinary kriging and indicator kriging. Predicted exceedance probability is plotted against the observed exceedance frequency, with the dashed diagonal marking perfect calibration. Points resting on fewer than ten sites are drawn hollow. The dotted vertical rules mark the largest exceedance probability that method ever issues.

1979
 1980
 1981
 1982
 1983
 1984
 1985
 1986
 1987
 1988
 1989
 1990
 1991
 1992
 1993
 1994
 1995
 1996
 1997
 1998
 1999
 2000
 2001
 2002
 2003
 2004
 2005
 2006
 2007
 2008
 2009
 2010
 2011
 2012
 2013
 2014
 2015
 2016
 2017
 2018
 2019
 2020
 2021
 2022
 2023
 2024

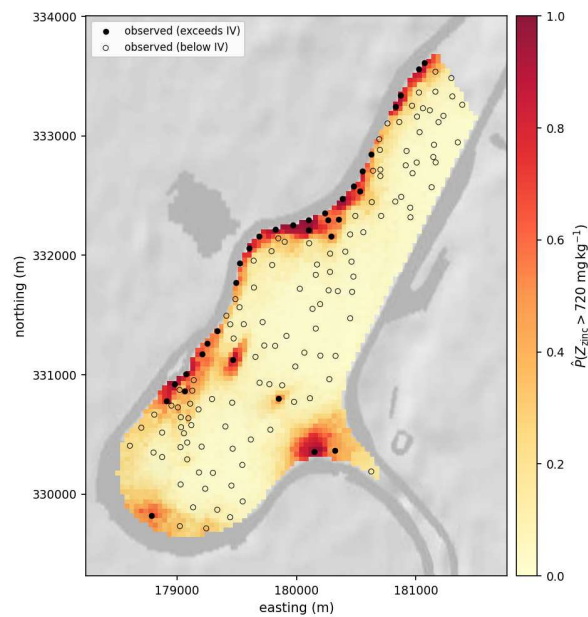


Figure 13: LDC- Γ exceedance probability $\hat{P}(Z > t)$ for zinc across the Meuse flood-plain grid, against the Dutch intervention value of 720 mg kg^{-1} ([Ministry of Housing, Spatial Planning and the Environment \(VROM\) 2009](#)). Illustrative only.

7 Discussion

What LDC adds to the ensemble interpolation line.

The adaptive ensemble interpolator was already shown by Egaña et al. (2025) to match or improve upon kriging in point-prediction error under non-stationarity. The present experiments extend that finding to the full predictive distribution and to the ranking of local anomalies. The locally normalised score outperforms global thresholds on anomalies that are extreme only within their neighbourhood, on both fields examined. It reads the same estimated law that the calibration and exceedance results score. Experiments 1–4 therefore report one predictive distribution serving two applications, rather than four separate findings.

Regime of applicability.

The synthetic experiments delineate the regime in which LDC is useful. The mechanism behind these advantages, however, varies with the metric and the context of application. In the synthetic experiments, it is mostly the widening step that provides the advantage. In the real surveys, meanwhile, it is the locally-adaptive ESI ensemble that carries the benefit. The following points explain each mechanism in more detail.

The effect of the widening across tasks.

In the synthetic experiments, a frontier emerged in which LDC- Γ gains an advantage over the multi-Gaussian assumption as the local law departs from a global one. Metrics such as the PITS-KS distance, the empirical coverage, and the AUC repeat the same pattern, with a *crossing* between ordinary kriging and LDC- Γ occurring around ν values of 0.5 to 1. Ordinary kriging obtains the best performance at lower ν values, where its stationarity assumption is correctly specified. Its performance then declines as ν grows, until it is surpassed by LDC- Γ and even by indicator kriging.

LDC-KDE obtains poor results in most of these frontier-tracing scenarios. The PITS-KS distance and empirical coverage values show it to be under-dispersed,

2071 which attributes the frontier to the widening step rather than to the ensemble. Its
2072 discrimination is also particularly weak at the deep tail ($p = 0.99$).
2073

2074 On a different note, the widening might be expected to trade sharpness for calibra-
2075 tion, since a wider predictive distribution places more mass above any fixed threshold.
2076 The synthetic experiments do not bear this out. LDC- Γ is both the stronger de-
2077 tector and the better calibrated variant, leading LDC-KDE on AUC and precision in
2078 Experiment 3 on both surveys.
2079

2080 Indicator kriging, the one baseline that assumes no global Gaussian law, keeps its
2081 calibration across the whole sweep and is not separated from LDC- Γ by the crossing.
2082 The frontier is therefore drawn against the multi-Gaussian assumption, and the case
2083 against indicator kriging has to be made on other grounds.
2084

2085 This illustrates how the stationarity assumption becomes a source of system-
2086 atic miscalibration as the local law begins to vary across space. The non-parametric
2087 estimation of the local distribution, by contrast, allows LDC- Γ to maintain calibration.
2088

2089 The advantages of the LDC- Γ method established on the synthetic fields carry
2090 over to the real surveys through precision, but not through calibration. Here, the raw
2091 ESI ensemble turned out to be over-dispersed in many cases, leaving no room for
2092 improvement through widening. The advantages in this case are carried instead by
2093 the locally-adaptive ensemble itself, as the following point explains in detail.
2094

2095 ***The effect of the locally-adaptive ensemble across tasks.***
2096

2097 In the Meuse and Jura surveys, the frontier can no longer be tracked, since there is no
2098 way to measure the departure from local stationarity directly. These fields, however,
2099 are mostly heterogeneous, so a high ν value can be assumed. Both LDC variants
2100 accordingly obtain considerably better CRPS and Brier score values than all kriging
2101 variants. LDC-KDE, moreover, outperforms its widened form on these metrics for
2102 the first time. This shows that the distributional-fidelity advantage comes from the
2103 non-parametric, locally-adaptive scheme itself, rather than from the widening.
2104

However, LDC loses whatever calibration advantage was observed in the synthetic fields. The kriging variants always do better than both LDC variants in terms of PIT-SKS distance, and often in terms of empirical coverage as well. The high PIT values, together with empirical coverage that exceeds the target value in 8 of 12 cases, show that the raw ensemble is frequently over-dispersed. The widening fixes this in part, but not as effectively as it fixes under-dispersion in the synthetic cases. That excess spread is all the PIT sees. The CRPS, by contrast, also credits a distribution for sitting close to the observation, which is why LDC loses the calibration comparison but wins the accuracy one.

Where the widening does over-disperse, the extra spread still comes with a gain. LDC- Γ leads LDC-KDE on both the AUC and the precision at K of Experiment 3 on the two surveys. That calibration cost is therefore traded for greater sensitivity to locally extreme observations. This is visible because AUC is invariant to any monotone inflation of the forecast probability.

Coherence as a primary argument.

The order-violation analysis places LDC and IK on genuinely different methodological ground. Every IK score reported above is computed after the cumulative-maximum repair, so the incoherence never inflates its numbers. What the violation rate records instead is that the repair is needed at roughly 90 to 97 % of locations across the metals of both surveys (Table 6). The repair is not free. It can only lower the exceedance probability the method issues, so the deep-tail reticence that the reliability term of Experiment 2 rewards is partly imposed on IK by the correction that makes its output a valid distribution. The correction also acts on each location’s marginal alone, so joint statements drawn from an IK model, for instance whether two locations both exceed the threshold or which is more extreme, remain ungrounded wherever order-relation violations occur. LDC’s coherence follows instead from the shared partition construction and the implicit copula of the SDE framework, and it requires no repair.

2163 Averaged over the injection replicates, this coherence carries no cost in pointwise
2164 ranking. LDC- Γ 's mean AUC exceeds IK's on both surveys.

2166

2167

2168

2169

2170 *Calibration earned by declining to commit.*

2171

2172 Exceedance probabilities were scored by three instruments because a good score can be
2173 earned in two different ways. A forecast that stays near the base rate is seldom far from

2174 the observed frequency, so the Brier score and the reliability term penalise it lightly,
2175 no matter how little it actually detects. Indicator kriging is the clearest instance. Its
2176 reliability curves stop at 0.24 to 0.35 of predicted probability on the synthetic fields,

2177 where LDC- Γ reaches 0.42 to 0.57. Against the Jura food-crop trigger value, it never
2178 passes 0.31, although 52 sites exceed it. It is nonetheless the only baseline built to

2179 estimate $\hat{P}(Z > t)$ directly, and it holds the worst Brier score of the five methods at
2180 every Meuse threshold (Table 7). Proximity to the diagonal is therefore read alongside
2181 the highest probability a method issues. Where a decision must name the sites at

2182 risk or set a design value, discrimination and the tail pinball loss are the instruments
2183 that separate the methods. LDC- Γ leads indicator kriging at the deepest threshold

2184 examined by 0.145 of AUC and 37.6 % of pinball loss.
2185
2186

2187
2188

2189
2190

2191
2192
2193

2194

2195

2196

2197

2198 *What the local reference changes in screening.*

2199

2200 The two local detectors recover largely different anomalies, most visibly on Jura (Fig-

2201 ure 10b). This follows from how each local reference is built. LDC adapts to the data
2202 around each site, while indicator kriging is fitted over the whole survey. The two

2203 therefore rank different sites as locally extreme wherever the local variability departs
2204 from the corresponding model. A screening exercise that reads both should expect

2205 complementary lists rather than a common one.
2206
2207

2208

8 Conclusion

Local Distributional Calibration addresses a practical gap in environmental spatial analysis, namely the need for exceedance probabilities that are simultaneously calibrated and coherent across the prediction domain. The method estimates a single local predictive distribution from a non-parametric ensemble without fitting a global distributional model. It reads both the exceedance probability and a locally normalised anomaly score off that one law. Its advantage over ordinary kriging and sequential Gaussian simulation is regime-dependent, small on locally stationary fields and growing as the local law varies across space. Against indicator kriging, the separation rests on the ranking of sites and on the deep-tail design value rather than on the Brier score. The anomaly score is a second application of the same distribution. It outperforms global thresholds on anomalies that are extreme only within their neighbourhood. The coherence guarantee, which follows structurally from the shared partition construction, distinguishes LDC from indicator kriging regardless of the stationarity regime.

Of the two variants, the results recommend LDC- Γ . The calibrated Gamma-mixture widening restores interval coverage, improves the exceedance probability and the design value at unsampled locations, and strengthens rather than blunts the ranking of sites. It therefore serves risk assessment and anomaly screening alike. LDC-KDE serves as the unwidened reference. Its sharper, unwidened forecasts remain competitive at the regulatory thresholds and take the lowest Brier score at most of them, though at the cost of the calibration advantage that the widening supplies.

The regime-dependent frontier identified here also delineates where the method is *not* needed. On fields whose marginal is well described by a global Gaussian model, the classical estimators remain the appropriate choice. The practical question for any application is therefore not which method is universally preferable but whether the local distributional heterogeneity of the field under study is large enough to place it in the regime where the non-parametric approach is warranted.

2255 **Data availability**

2256

2257 Data sharing is not applicable to this article as no new data were generated in this
2258 study, but is available from the corresponding author upon reasonable request.

2260

2261

2262

2263

2264 **Acknowledgements**

2265

2266

2267 This work was supported in part by the Chilean National Agency for Research and
2268 Development (ANID) under ANID AMTC CIA250010. We acknowledge the computa-
2269 tional resources provided, the useful insights, and the support provided by the whole
2270 ALGES team.

2271

2272

2273

2274

2275

2276

2277 **Declarations**

2278

2279

2280 **Funding.** This work was partially supported by the Chilean National Agency for
2281 Research and Development (ANID) through grant ANID AMTC CIA250010.

2282

2283

2284

2285

2286 **Conflict of interest.** The authors declare that there is no conflict of interest
2287 regarding the publication of this paper.

2288

2289

2290

2291

2292 **Author contribution.** M.J.V. conducted the experiments, analysed the results,
2293 and wrote the original draft. A.F.E. and F.N. conceived the concept and methodology,
2294 supervised the research, and contributed to writing the original draft. G.D. analysed
2295 the results and contributed to writing the original draft. G.M. prepared the dataset
2296 and analysed the results. A.E. contributed to writing the original draft and analysed
2297 the results. All authors reviewed the manuscript.

2298

2299

2300

Appendix A Sample size at the closest cells of Experiment 2

The two cells where indicator kriging comes closest to LDC- Γ , $\eta = 1.5$ and $\eta = 2$ under the low nugget-to-sill ratio, are rerun at $n = 240$ and $n = 480$ with nested samples. Tables A1 and A2 report discrimination and the tail loss of the design value over that axis, at all three threshold levels. Section 6.2 reads them.

Table A1: Area under the ROC curve against the sample size n , at the two cells where the samples are nested, $\eta \in \{1.5, 2.0\}$ under the low nugget-to-sill ratio ($\nu = 0.3$), by threshold level q . Each entry is the mean over $R = 50$ realisations. The best of each row is in bold. The last column is the p -value of the paired test of LDC- Γ against IK over the same realisations, of the mean difference for the AUC and of the median log-ratio for the pinball loss.

q	η	n	LDC-KDE	LDC- Γ	OK	SGS	IK	p vs IK
0.90	1.5	120	0.844	0.888	0.854	0.847	0.846	0.001
0.90	1.5	240	0.878	0.907	0.880	0.877	0.887	0.001
0.90	1.5	480	0.888	0.917	0.895	0.891	0.907	0.001
0.90	2	120	0.827	0.843	0.761	0.754	0.819	0.001
0.90	2	240	0.849	0.876	0.794	0.790	0.875	0.708
0.90	2	480	0.872	0.897	0.819	0.818	0.901	0.120
0.95	1.5	120	0.802	0.889	0.849	0.827	0.803	0.001
0.95	1.5	240	0.843	0.912	0.883	0.865	0.874	0.001
0.95	1.5	480	0.856	0.922	0.902	0.888	0.898	0.001
0.95	2	120	0.812	0.868	0.747	0.740	0.807	0.001
0.95	2	240	0.830	0.891	0.777	0.773	0.870	0.008
0.95	2	480	0.856	0.913	0.803	0.802	0.904	0.026
0.99	1.5	120	0.699	0.889	0.885	0.808	0.739	0.001
0.99	1.5	240	0.721	0.925	0.918	0.839	0.805	0.001
0.99	1.5	480	0.735	0.930	0.944	0.883	0.840	0.001
0.99	2	120	0.635	0.851	0.777	0.700	0.706	0.001
0.99	2	240	0.684	0.870	0.827	0.752	0.791	0.001
0.99	2	480	0.728	0.915	0.869	0.773	0.853	0.003

2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392

Table A2: Tail pinball loss of the design value against the sample size n , at the two cells where the samples are nested, $\eta \in \{1.5, 2.0\}$ under the low nugget-to-sill ratio ($\nu = 0.3$), by threshold level q . Each entry is the mean over $R = 50$ realisations. The best of each row is in bold. The last column is the p -value of the paired test of LDC- Γ against IK over the same realisations, of the mean difference for the AUC and of the median log-ratio for the pinball loss.

q	η	n	LDC-KDE	LDC- Γ	OK	SGS	IK	p vs IK
0.90	1.5	120	0.722	0.592	0.710	0.711	0.649	0.066
0.90	1.5	240	0.649	0.529	0.649	0.651	0.551	0.001
0.90	1.5	480	0.611	0.497	0.614	0.616	0.505	0.010
0.90	2	120	1.669	1.456	2.025	2.024	1.533	0.090
0.90	2	240	1.548	1.286	1.898	1.894	1.271	0.108
0.90	2	480	1.452	1.164	1.790	1.787	1.169	0.637
0.95	1.5	120	0.529	0.372	0.459	0.461	0.436	0.015
0.95	1.5	240	0.470	0.328	0.416	0.419	0.353	0.001
0.95	1.5	480	0.443	0.307	0.393	0.396	0.319	0.001
0.95	2	120	1.258	0.936	1.382	1.385	1.052	0.002
0.95	2	240	1.157	0.818	1.284	1.284	0.838	0.203
0.95	2	480	1.081	0.731	1.197	1.198	0.751	0.137
0.99	1.5	120	0.280	0.112	0.156	0.162	0.186	0.001
0.99	1.5	240	0.243	0.097	0.136	0.143	0.124	0.001
0.99	1.5	480	0.231	0.090	0.127	0.133	0.109	0.001
0.99	2	120	0.721	0.291	0.567	0.584	0.434	0.001
0.99	2	240	0.649	0.253	0.514	0.528	0.349	0.001
0.99	2	480	0.607	0.220	0.454	0.466	0.292	0.001

References	2393
	2394
Biswas, A., Biswas, A.: Geostatistics in soil science: A comprehensive review on past,	2395
present and future perspectives. <i>Journal of the Indian Society of Soil Science</i> (2024)	2396
	2397
	2398
	2399
Borkowski, A.S., Kwiatkowska-Malina, J.: Geostatistical modelling as an assessment	2400
tool of soil pollution based on deposition from atmospheric air. <i>Geosciences Journal</i>	2401
(2017)	2402
	2403
	2404
	2405
Bronowicka-Mielniczuk, U., Mielniczuk, J.: New indices to quantify patterns of relative	2406
errors produced by spatial interpolation models – a comparative study by modelling	2407
soil properties. <i>Ecological Indicators</i> (2023)	2408
	2409
	2410
Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V.: <i>Applied Spatial Data Analysis with</i>	2411
R. Springer, New York (2008)	2412
	2413
	2414
	2415
Brier, G.W.: Verification of forecasts expressed in terms of probability. <i>Monthly</i>	2416
<i>Weather Review</i> 78 (1), 1–3 (1950)	2417
	2418
	2419
Dawid, A.P.: Present position and potential developments: some personal views. <i>Stat-</i>	2420
<i>istical theory. The prequential approach. Journal of the Royal Statistical Society:</i>	2421
<i>Series A</i> 147 (2), 278–292 (1984)	2422
	2423
	2424
	2425
Deutsch, C.V., Journel, A.G.: <i>GSLIB: Geostatistical Software Library and User's</i>	2426
<i>Guide</i> , 2nd edn. Oxford University Press, New York (1997)	2427
	2428
	2429
Deepika, B.V., Ramakrishnaiah, C.R., Naganna, S.R.: Spatial variability of ground	2430
water quality: A case study of Udupi district, Karnataka State, India. <i>Journal of</i>	2431
<i>Earth System Science</i> (2020)	2432
	2433
	2434
	2435
Egaña, Á.F., Díaz, G., Navarro, F.: Spatial Distributional Estimation: a consist-	2436
ent non-parametric framework for random field inference. <i>Companion manuscript</i>	2437
	2438

2439 (2025)
 2440
 2441 Egaña, Á., Navarro, F., Maleki, M., Grandón, F., Carter, F., Soto, F.: Ensemble spa-
 2442 tial interpolation: a new approach to natural or anthropogenic variable assessment.
 2443 Natural Resources Research **30**, 3777–3793 (2021)
 2444
 2445
 2446
 2447 Emery, X., Ortiz, J.M.: Shortcomings of multiple indicator kriging for assessing local
 2448 distributions. Applied Earth Science: Transactions of the Institutions of Mining and
 2449 Metallurgy B **113**(4), 249–259 (2004) <https://doi.org/10.1179/174327504X27242>
 2450
 2451
 2452
 2453 Egaña, Á.F., Valenzuela, M.J., Maleki, M., Sánchez-Pérez, J.F., Díaz, G.: Adaptive
 2454 ensemble spatial analysis. Scientific Reports **15**, 26599 (2025) [https://doi.org/10.](https://doi.org/10.1038/s41598-025-08844-z)
 2455 [1038/s41598-025-08844-z](https://doi.org/10.1038/s41598-025-08844-z)
 2456
 2457
 2458
 2459 Garcés, D., Jiménez-Oyola, S., Sánchez-Palencia, Y., Marcillo-Guillén, J., et al.: As-
 2460 sessment of tailings contamination potential in one of the most important gold
 2461 mining districts of Ecuador. Minerals (2025)
 2462
 2463
 2464
 2465 Gneiting, T.: Quantiles as optimal point forecasts. International Journal of Forecasting
 2466 **27**(2), 197–207 (2011)
 2467
 2468
 2469 Goovaerts, P.: Geostatistics for Natural Resources Evaluation. Oxford University
 2470 Press, New York (1997)
 2471
 2472
 2473 Goovaerts, P.: Geostatistical modelling of uncertainty in soil science. Geoderma **103**(1–
 2474 2), 3–26 (2001)
 2475
 2476
 2477 Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation.
 2478 Journal of the American Statistical Association **102**(477), 359–378 (2007)
 2479
 2480
 2481 Gneiting, T., Ranjan, R.: Comparing density forecasts using threshold- and quantile-
 2482 weighted scoring rules. Journal of Business & Economic Statistics **29**(3), 411–422
 2483
 2484

(2011)	2485
	2486
Gao, B., Stein, A., Wang, J.: A two-point machine learning method for the spatial	2487
prediction of soil pollution. <i>International Journal of Applied Earth Observation and</i>	2488
<i>Geoinformation</i> (2022)	2489
	2490
	2491
	2492
Henshaw, S.L., Curriero, F.C., Shields, T.M., Breyse, P.N., et al.: Geostatistics and	2493
GIS: Tools for characterizing environmental contamination. <i>Journal of Medical</i>	2494
<i>Systems</i> (2004)	2495
	2496
	2497
	2498
Heusler, J., Funk, J., Wagner, A.: Spatial interpolation in applied insect ecology:	2499
A review, including guidelines and datasets for practical use. <i>Journal of Applied</i>	2500
<i>Entomology</i> (2025)	2501
	2502
	2503
	2504
Jiang, Y., Li, H., Zhang, Z., Zhang, J., et al.: Enhancing ecological sustainability in ion-	2505
adsorption rare earth mining areas: A multi-scale model for assessing spatiotemporal	2506
dynamics and ecological resilience. <i>Ecological Modelling</i> (2025)	2507
	2508
	2509
	2510
Journel, A.G.: Nonparametric estimation of spatial distributions. <i>Journal of the</i>	2511
<i>International Association for Mathematical Geology</i> 15 (3), 445–468 (1983)	2512
	2513
	2514
Kochurov, B.I., Movchan, M.A.: Geoecological assessment of the city of Vidnoe using	2515
multifractal analysis. <i>Geography and Natural Resources</i> (2024)	2516
	2517
	2518
Kumari, M.K.N., Sakai, K., Kimura, S., Duminda, D.M.S., et al.: Interpolation meth-	2519
ods for groundwater quality assessment in tank cascade landscape: A study of	2520
Ulagalla cascade, Sri Lanka. <i>Applied Ecology and Environmental Research</i> (2018)	2521
	2522
	2523
	2524
Li, J., Heap, A.D., Potter, A., Daniell, J.J.: Application of machine learning methods	2525
to spatial interpolation of environmental variables. <i>Environmental Modelling and</i>	2526
<i>Software</i> (2011)	2527
	2528
	2529
	2530

2531 Ministry of Housing, Spatial Planning and the Environment (VROM): Circular on tar-
 2532 get values and intervention values for soil remediation. Technical report, Ministerie
 2533 van Volkshuisvesting, Ruimtelijke Ordening en Milieubeheer, The Hague, Neth-
 2534 erlands (2009). Soil Remediation Circular 2009, DGM/BO (reference: DGM/BO
 2535 2009). Annexes contain updated target and intervention values for soil and
 2536 groundwater
 2537
 2538
 2539
 2540
 2541 Müller, S., Schüler, L., Zech, A., Heße, F.: GSTools v1.3: a toolbox for geostatistical
 2542 modelling in Python. *Geoscientific Model Development* **15**(7), 3161–3182 (2022)
 2543 <https://doi.org/10.5194/gmd-15-3161-2022>
 2544
 2545
 2546
 2547 Murphy, A.H.: A new vector partition of the probability score. *Journal of Applied*
 2548 *Meteorology* **12**(4), 595–600 (1973)
 2549
 2550
 2551 Navarro, F., Egaña, A.F., Ehrenfeld, A., Garrido, F., Valenzuela, M.J., Sánchez-
 2552 Pérez, J.F.: Spatialize v1.0: a python/c++ library for ensemble spatial interpolation.
 2553 *Geoscientific Model Development* **19**(10), 4633–4660 (2026) [https://doi.org/10.](https://doi.org/10.5194/gmd-19-4633-2026)
 2554 [5194/gmd-19-4633-2026](https://doi.org/10.5194/gmd-19-4633-2026)
 2555
 2556
 2557
 2558
 2559 Ohmer, M., Doll, F., Liesch, T.: Incorporating spatial information for regionalization
 2560 of environmental parameters in machine learning models. *Mathematical Geosciences*
 2561 (2025)
 2562
 2563
 2564
 2565 Pebesma, E.J.: Multivariable geostatistics in S: the gstat package. *Computers &*
 2566 *Geosciences* **30**(7), 683–691 (2004)
 2567
 2568
 2569 Rodrigues, M.S., Castrignanò, A., Belmonte, A., Trindade Lessa, B.F., et al.:
 2570 Geostatistics and its potential in Agriculture 4.0. *Revista Ciência Agronômica*
 2571 (2020)
 2572
 2573
 2574
 2575 Rodrigues, M.: GIS-based modeling of a rescaled surface of land development pressure
 2576

in the Macaronesian islands. GIScience and Remote Sensing (2016)	2577
	2578
Rusu, C., Rusu, V.: Radial basis functions versus geostatistics in spatial interpolations.	2579
	2580
In: IFIP International Federation for Information Processing (2006)	2581
	2582
	2583
Shaddad, S.M.: Geostatistics and proximal soil sensing for sustainable agriculture. In:	2584
Handbook of Environmental Chemistry, (2019)	2585
	2586
	2587
Swiss Federal Council: Ordinance of 1 july 1998 on the pollution of soil (VBBo). Tech-	2588
nical Report SR 814.12, Swiss Confederation, Bern, Switzerland (1998). Verordnung	2589
über Belastungen des Bodens; Annex 1 lists the guide, trigger and remediation	2590
values for inorganic pollutants	2591
	2592
	2593
	2594
Thayer, W.C., Griffith, D.A., Goodrum, P.E., Hassett, J.M., et al.: Application of	2595
geostatistics to risk assessment. Risk Analysis (2003)	2596
	2597
	2598
	2599
Tan, J., Xie, X., Zuo, J., Zhang, Y., et al.: Coupling random forest and inverse dis-	2600
tance weighting to generate climate surfaces of precipitation and temperature with	2601
multiple covariates. Journal of Hydrology (2021)	2602
	2603
	2604
	2605
Varouchakis, E.A.: Geostatistics: Mathematical and statistical basis. In: Spatiotem-	2606
poral Analysis of Extreme Hydrological Events, (2018)	2607
	2608
	2609
Yang, X.-H., Li, S.-Y.: Geostatistical methods for research on environmental pollution.	2610
Zhongshan Daxue Xuebao / Acta Scientiarum Naturalium Universitatis Sunyatseni	2611
(2005)	2612
	2613
	2614
	2615
Yermolaev, O.P., Sihanova, N.S., Shynbergenov, Y.A., Yantsitov, R.O.: Assessment of	2616
natural resource potential and anthropogenic load in the macroregion of northern	2617
Eurasia based on a basin approach. Geography, Environment, Sustainability (2024)	2618
	2619
	2620
	2621
	2622

2623 Yi, G., Zhang, T., He, Y., Liu, D., et al.: Applicability analysis of four spatial inter-
2624 polation methods for air temperature. Journal of Chengdu University of Technology
2625 (Science and Technology Edition) (2020)
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668