Assessing local and spatial uncertainty with nonparametric geostatistics

Stephanie Thiesen\textsuperscript{1a}, Uwe Ehret\textsuperscript{1b}

\textsuperscript{1} Institute of Water Resources and River Basin Management, Karlsruhe Institute of Technology, Karlsruhe, Germany
\textsuperscript{a} orcid.org/0000-0001-7501-2223, \textsuperscript{b} orcid.org/0000-0003-3454-8755

Correspondence to: Stephanie Thiesen (stephanie.thiesen@kit.edu)

Abstract. Uncertainty quantification is an important topic for many environmental studies, such as identifying zones where potentially toxic materials exist in the soil. In this work, the nonparametric geostatistical framework of histogram via entropy reduction (HER) is adapted to address local and spatial uncertainty in the context of risk of soil contamination. HER works with empirical probability distributions, coupling information theory and probability aggregation methods to estimate conditional distributions, which gives it the flexibility to be tailored for different data and application purposes. To explore the method adaptation for handling estimations of threshold-exceeding probabilities, it is used to map the risk of soil contamination by lead in the well-known dataset of the region of Swiss Jura. Its results are compared to indicator kriging (IK) and to an ordinary kriging (OK) model available in literature. For the analyzed dataset, IK and HER achieved the best performance and exhibited comparable accuracy and precision of their predictions. When compared to IK, HER has shown to be a unique approach for dealing with uncertainty estimation in a fine resolution, without the need of modeling multiple indicator variograms, correcting order-relation violations, or defining interpolation/extrapolation of distribution. Finally, to avoid the well-known smoothing effect when using point estimations (this is the case with kriging, but also with HER) and to provide maps that reflect the spatial fluctuation of the revealed reality, we demonstrate how HER can be used in combination with sequential simulation to assess spatial uncertainty (uncertainty jointly over several locations).

Key words. Nonparametric geostatistics, Non-Gaussian, Conditional distribution, Sequential Simulation, Uncertainty analysis, Risk mapping.

Acknowledgments. The authors acknowledge the support received by the Deutsche Forschungsgemeinschaft (DFG) and the Open Access Publishing Fund of Karlsruhe Institute of Technology (KIT), Diego M. Vieira for the implementation of the optimization problem as well as for the insightful discussions throughout the development of this work and the manuscript review, and Dr. Pierre Goovaerts for his availability in discussing general topics of geostatistics and for clarifying specific matters regarding the Jura dataset and the use of Auto-IK. The first author also acknowledges the support of the Graduate Funding from the German States program (Landesgraduiertenförderung).
1 Introduction

Modeling the uncertainty about the unknown is of crucial importance for evaluating the risk involved in any decision-making process. The traditional approach of modeling the uncertainty with respect to geostatistical interpolation consists in computing a kriging estimate and its attached error variance, and explicitly assuming a Gaussian distribution for assessing the confidence interval (Goovaerts 1997 p.261; Kitanidis 1997 p.68; Bourennane 2007). The major restriction by doing so is that one assumes normality of the distribution of the estimation error and that the variance of the errors is independent of the data values and depends only on the data configuration (Kitanidis 1997 p.68; Goovaerts 1997 p.261). These Gaussian and homoscedastic assumptions are unfortunately rarely fulfilled for environmental attributes and soil variables, which often display skewed distributions (Bourennane 2007; Goovaerts 1997 p.261).

More rigorous approaches such as multivariate-Gaussian model (MGM) and indicator kriging (IK) address the problem of modeling local uncertainty through conditional probability distributions (CPD). Different from the traditional approach, in these CPD models, first the uncertainty about the unknown is assessed and then an estimate optimal in some appropriate sense is deduced (Goovaerts 1997 p.262). MGM is widely used thanks to its mathematical simplicity and easy inference (Goovaerts 1997 p.265; Gómez-Hernández and Wen 1998). However, under the multi-Gaussian spatial law, all marginal and conditional distributions are Gaussian, hence the variance of the CPD depends only on the data configuration, not on their values (Goovaerts 1997 p.284; Ortiz et al. 2004). Likewise, due to its strong distribution hypothesis, it is unfeasible to check the normality of multiple-point (in contrast to two-point) experimental CPD (Goovaerts 1997 p.284) and it might produce inadequate results caused by a wrong parametric model assumption (Fernández-Casal et al. 2018). IK on the other hand was developed to suppress any particular shape or analytical expression of the CPD. Although it is a nonparametric model, when a complete CPD is needed as output, its shortcomings lie in the need to fit multiple indicator variograms (one per cutoff), to correct order-relation violations, and to interpolate and extrapolate the CPD. Furthermore, due to the indicator transform of the neighbors (from continuous to binary, e.g.) it loses information available in data (Fernández-Casal et al. 2018).

Recently, for avoiding the risk of adding information not present in data, Thiesen et al. (2020) proposed combining information theory with probability aggregation methods in a geostatistical framework as a novel nonparametric method for stochastic estimation of unsampled points. Histogram via entropy reduction (HER) was primarily proposed to bypass fitting of spatial correlation functions and assumptions about the underlying distribution of the data. HER proved to be a proper framework for uncertainty estimation, since it accounts for both spatial configuration and data values, and presents a higher generality in comparison to ordinary kriging (OK). HER uses binned transformation of the data and optimization of the information content, which gives some flexibility to adapt the method to handle different kinds of data and problems. Furthermore, it allows to incorporate different uncertainty properties by selecting the aggregation method.

In the context of risk mapping, an important goal of many environmental applications is to delimit zones with the presence of potentially toxic substances in the soil (Goovaerts et al. 1997 p.334). For decision-making of this sort, it is often more pertinent to calculate the risk of exceeding regulatory limits (risk of contamination) rather than deriving a single value estimate
(Goovaerts 1997 p.333). Thus, the purpose of this paper is to extend HER to evaluate the probability or risk, given the data, that a pollutant concentration exceeds a critical threshold at a particular location of interest, and comprehensively compare its results to a method typically used for this purpose. To do so, we tailor HER’s optimization problem for dealing with threshold-exceeding probabilities and investigate the framework using the established Swiss Jura dataset (Atteia et al. 1994; Webster et al. 1994). The estimation and local uncertainty results of HER are then compared to IK, the most widely employed approach to estimate exceeding probabilities (Fernández-Casal et al. 2018), and to an OK model available in literature.

Although the estimation honors local data, is locally accurate, and has a smoothing effect appropriate for visualizing trends, it is inappropriate for simulating extreme values (Rossi and Deutsch 2014 p.167). Local uncertainty, on the other hand, allows us to assess the uncertainty at any specific unsampled location but not the uncertainty when many locations must be considered simultaneously (spatial uncertainty) (Goovaerts 2001). Therefore, to reproduce the original variability observed in the data and to address the joint model of uncertainty, HER is expanded using sequential simulation to generate stochastic realizations of the field under study (HERs). Although using simulated fields for definition of remediation costs of contaminated areas and application in transfer functions are usual (Goovaerts 2001), this paper is limited to demonstrate the feasibility of HERs.

The paper is organized as follows. HER method and its adaptations are presented in Sect. 2. In Sect. 3, we describe the dataset, performance criteria, and benchmark models. At the end of this section, we apply OK, IK and HER to real data and present the comparison analysis for the estimation and local uncertainty results, and a proof of concept of HERs. In Sect. 4, we discuss the achieved outcomes and model contributions. Finally, the main conclusions are summarized in Sect. 5.

2 Method description

In the following sections, we start with a short presentation of information theoretic measures employed in the HER method (Sect. 2.1), then we introduce HER three main steps (Sect. 2.2), and propose an adaptation of the minimization problem tailored to estimate local threshold-exceeding probabilities (Sect. 2.2.3). Finally, in Sect. 2.3, we expand HER for spatial uncertainty analysis.

2.1 Information theoretic measures employed in HER

To assess the spatial dependence structure from the data, minimize the uncertainties, and evaluate the quality of the probabilistic predictions, we apply two measures of information theory, namely Shannon entropy (H) and Kullback-Leibler divergence (DKL). This section is founded on Cover and Thomas (2006), which we suggest for an introduction to the topic.

For a discrete random variable \( X \) with a probability \( p(x) \), \( x \in \chi \), the Shannon entropy equation is defined as

\[
H(X) = - \sum_{x \in \chi} p(x) \log_2 p(x).
\]

(1)

The logarithm to base two denotes entropy in unit of bits. The above expression measures the average uncertainty of a probability distribution. HER uses Shannon entropy to evaluate the spatial dependence of the dataset and its correlation length.
Kullback-Leibler divergence (or relative entropy) compares similarities between two probability distributions $p$ and $q$:

$$D_{KL}(p||q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)}.$$  

(2)

Expressed in bits, it measures the statistical 'distance' between two distributions, where one ($p$) is the reference, and the other ($q$) a model thereof. Kullback-Leibler divergence is nonnegative and it is equals zero if and only if $p = q$. It can be used i) to quantify the information loss of assuming that the distribution is $q$ when really it is $p$ and ii) as a performance metric for probabilistic predictions (Gneiting and Raftery 2007; Weijs et al. 2010). In this study, $D_{KL}$ is applied for two purposes. Primarily, it defines the optimization problem of HER (its loss function), which minimizes the information loss when aggregating the neighbor distributions. Additionally, it is used as a scoring rule for performance verification of probabilistic predictions.

2.2 HER for local uncertainty

The present introduction of HER is based on Thiesen et al. (2020), which we suggest for a more details of the method. HER is a distribution-free interpolator enclosed in a geostatistical framework, formulated to describe spatial patterns and solve spatial interpolation problems. It incorporates concepts from information theory and probability aggregation methods for globally minimizing uncertainty and predicting conditional probability distributions (CPD) directly based on empirical discrete distributions (also known as probability mass functions, PMFs). HER comprises three main steps: i) characterization of spatial correlation, ii) selection of an aggregation method and associated optimal weights, and iii) prediction of target CPD. These steps are explained in detail in the following sections.

2.2.1 Characterization of spatial correlation

Let us consider the situation illustrated in Fig. 1c, where $z$ is the attribute under study and we are interested in inferring the $z$ PMF of the target 0 ($z_0$, unsampled point) given its neighbors 1, 2, and 3 ($z_1$, $z_2$, and $z_3$, sampled points). In order to characterize the spatial correlation, we extract the distribution to be associated to each neighbor and the correlation length (range) in the following actions. First, for each lag distance (also called distance class or simply class), we calculate the difference of the $z$-values ($\Delta z$) between pairs of observations and generate the corresponding $\Delta z$ PMF (Fig. 1a). The entropy values of the $\Delta z$ PMF as a function of lag distance is visualized as a 2D plot called infogram (Fig. 1b), in which the range is identified as the distance where the entropy of the classes exceeds the full dataset entropy. Finally, we associate to each neighbor the $\Delta z$ PMF according to its distance from the target, then shift this distribution by the $z$ value of the neighbor (1, 2, or 3), as outlined in Fig. 1c. In the end of this first step, we have inferred the conditional PMFs $p(z_0|z_1)$, $p(z_0|z_2)$, and $p(z_0|z_3)$. 


2.2.2 Probability aggregation

For the second step of the method, the individual conditional distributions obtained in the previous step are combined by using probability aggregation methods. This yields a single, global distribution for the target $0$, so that the joint probability

$$p(z_0 | z_1, \ldots, z_n) \approx P_G(p(z_0 | z_1), \ldots, p(z_0 | z_n)),$$

with $z_0$ being the target (unsampled point) and $z_i$ its neighbors, where $i = 1, \ldots, n$ are the indices of the sampled points and $z$ is the variable under study. For brevity, from now on we use $P_i(z_0)$ to denote $p(z_0 | z_i)$ and $P_G(z_0)$ for the global probability $P_G(P_1(z_0), \ldots, P_n(z_0))$.

Two basic aggregation methods were discussed by Thiesen et al. (2020), the linear pooling and the log-linear pooling. The linear pooling, Eq. (3), is a way of averaging distributions. It is related to union of events and identified to the logical operator OR. Multiplication of probabilities, or log-linear pooling in Eq. (4), in turn, is described by the logical operator AND, and it is associated to the intersection of events. Due their distinct characteristics, Thiesen et al. (2020) associated the linear aggregation to discontinuous field properties and the log-linear to continuous. The authors exemplified that, if we have two points A and B with different values and want to estimate the target point $X$ between both in a continuous field, we would expect that the estimate at point $X$ would be somewhere between A and B, which can be achieved by an AND combination.

On the other hand, in the case of categorical data (or abrupt changes, Goovaerts 1997 p.420), considering A and B belonging to different categories, a target $X$ between both will either belong to the category of A or B, which can be achieved by an OR combination.

A third pooling operator, which combines $P_{G\text{AND}}$ and $P_{G\text{OR}}$, was proposed and explored in Thiesen et al. (2020) work, Eq. (5), as a way of finding the proportion of $\alpha$ (continuous) and $\beta$ (discontinuous) properties of the field by minimizing relative
estimation entropy \( (D_{KL}) \). Since final distribution of this pooling may result in a pure OR, Eq. (3), or pure AND, Eq. (4), aggregation, as special cases, it was recommended for inferring field properties given the model when there is lack of information to accurately describe the interactions between the sources of information.

\[
P_{GOR}(z_0) = \sum_{i=1}^{n} w_{OR_i} P_i(z_0),
\]

where \( n \) is the number of neighbors, and \( w_{OR_i} \) are positive weights verifying \( \sum_{i=1}^{n} w_{OR_i} = 1 \).

\[
\ln P_{GAND}(z_0) = \ln \zeta + \sum_{i=1}^{n} w_{AND_i} \ln P_i(z_0),
\]

where \( \zeta \) is a normalizing constant, \( n \) is the number of neighbors, and \( w_{AND_i} \) are positive weights.

\[
P_G(z_0) \propto P_{GAND}(z_0)^\alpha P_{GOR}(z_0)^\beta,
\]

where \( \alpha \) and \( \beta \) are positive weights varying from 0 to 1.

### 2.2.3 Entropy minimization

After selecting the appropriate aggregation method, we address the optimization problem for estimating the weights of the pooling operators. In Thiesen et al. (2020), the authors were interested in comparing HER results with OK estimates. Therefore, by means of leave-one-out cross-validation, they chose a global set of weights such that the disagreement of the 'true' observation (left-out measurement) and the estimated probability of the bin containing the true observation was minimized. For doing so, the optimization problem was tailored to find the set of weights which minimizes the expected relative entropy \( (D_{KL}) \) of all targets. The \( D_{KL} \) evaluation of a single prediction is outlined in Fig. 2a.

(a) Thiesen et al. 2020

(b) Threshold-exceeding probability

![Fig. 2 Optimization problem. (a) Maximizing the probability of the 'true' observation (Thiesen et al. 2020) and (b) maximizing the estimation of threshold-exceeding probability.](image)

In the present study, we propose an adaptation of the loss function to focus on the estimation of threshold-exceeding probabilities (Fig. 2b). Here, instead of optimizing the bin containing the true observation, we minimize the disagreement (relative entropy, \( D_{KL} \)) of the binarized left-out measurement (above or below \( z_c \) threshold) and the cumulative probability of
the estimated distribution (also binary, above or below \( z_c \) threshold). With this adaptation, the optimization problem focusses on selecting weights which maximize the probability of the target matching the true classification. The authors’ intentions were to reduce therefore the risk that an unsampled site is declared 'safe' when in reality the soil is 'toxic' and vice versa, and to open the possibility of working with categorical data. The method adaptation proposed in Fig. 2b will be used throughout the paper and called simply HER.

For both models (Fig. 2a and Fig. 2b), one weight is assigned to each distance class \( k \) in Eqs. (3) and (4), referred to as \( w_{OR_k} \) and \( w_{AND_k} \), respectively (here generalized as \( w_k \)). After that, \( \alpha \) and \( \beta \) from Eq. (5) are calculated by grid search with both weight values ranging from 0 to 1 (steps of 0.05 were used in the application case).

Particularly for the present study, another adaptation was done to avoid undesired non-zero uncertainty when predicting values at sampled locations. For that, within the first distance class, we asymptotically increase the weight towards infinity as the distance approaches zero. For all but the first distance classe, similarly to Thiesen et al. (2020), we linearly interpolate the weights according to the Euclidean distance and the weight of the next class.

### 2.2.4 PMF prediction

As seen before, to estimate an unknown point \( z_0 \) (target), first, we classify its neighbors \( z_i \) (sampled observations) according to their distance to the target. Each neighbor is then associated to its corresponding \( \Delta z \) PMF and shifted by its \( z_i \) value. Finally, by applying the selected aggregation method and its optimum weights, we combine the individual \( z \) PMFs of the neighbors to obtain the \( z \) PMF of the target conditional on all neighbors (\( z \) PMF). By construction, such PMF (and its uncertainty) assessment is nonparametric, since no prior assumption is made about the shape of the distribution of possible values.

In order to increase computational efficiency, we do not use classes beyond the range (neighbors beyond the range are associated to the \( \Delta z \) PMF of the full dataset) and, due to the minor contribution of neighbors in classes far away from the target, the authors only used the closest 30 neighbors when estimating the target. Knowledge of the (conditional) local distribution obtained here allows a straightforward assessment of the uncertainty about the unknown point value, independently of the choice of a particular estimate for it (Goovaerts 1997 p.333).

### 2.3 HER for spatial uncertainty

So far, we proposed modeling distributions to obtain estimates of values and related uncertainties at specific locations (local uncertainty) using the HER method. However, these single-point PMFs do not allow to assess the uncertainty about attribute values at several locations taken simultaneously (Goovaerts 1997 p.262). Simply multiplying CPDs of several locations to obtain their joint probability would assume independence between the data, a case of little interest (Goovaerts 1997 p.372). Therefore, we address multiple-point - or spatial - uncertainty by combining HER with sequential simulation (HERs). Stochastic simulation was introduced in the early 1970's to correct for the smoothing effect of kriging and to provide maps...
that reflect the spatial fluctuation of the revealed reality (Journel 1974; Deutsch and Journel 1997). Geostatistical simulation generates a model of uncertainty that is represented by multiple sets of possible values distributed in space, one set of possible outcomes is referred to as a realization (Leuangthong 2004). Different yet equiprobable realizations, all conditional to the same dataset and reflecting the same dispersion characteristics, can be produced to be used for numerical and visual appreciation of spatial uncertainty (Journel and Huijbregts 1978; Deutsch and Journel 1997 p.19; Journel 2003). Such equiprobable realizations are known as stochastic images and match the same sample statistics and conditioning data (Gómez-Hernández and Cassiraga 1994).

Sequential simulations with HER can be generated by firstly establishing a random path along all nodes in the grid network. Then, for each node, and in the order of the random path we i) derive the PMF of the node using HER with the steps as explained in Sect. 2.2, ii) randomly draw a single value from this PMF, and iii) assign the value to the grid as data. By this procedure, we sequentially include the simulated values to the original dataset and used them to condition predictions for the remaining locations. The simulated value (step ii) is derived as a result of a Monte Carlo simulation (Metropolis and Ulam 1949), where we randomly draw a p-value uniformly distributed between 0 and 1 and obtain the z value from the estimated PMF. Equiprobability is ensured by triggering each realization by one random seed drawn from a uniform distribution (Deutsch and Journel 1997; Goovaerts 1999).

Because the path is random, repetitions of the stochastic process will yield different realizations, but all will honor the data and model statistics. Thus, for assessing the spatial uncertainty, multiple realizations can be used to calculate the joint probability of a set of locations simultaneously rather than one at a time. Therefore, while HER as well as OK and IK methods smooth out the real fluctuation of the attribute due to the missing variability between unsampled points, its sequential simulation variety (HERs) reproduces the spatial variability of the sample data. In this study we are interested in developing and presenting the realizations generated by HERs as a proof of concept.

3 Application to real data

3.1 Jura dataset

We evaluate HER (Sect. 2.2) and HERs (Sect. 2.3) by applications to the well-known Jura dataset, which is often used as benchmarking in the geostatistical literature, e.g., Atteia et al. (1994), Webster et al. (1994), Goovaerts (1997), Goovaerts et al. (1997), Bel et al. (2009), Allard et al. (2011), Loquin and Dubois (2010), Dabo-Niang et al (2016), Bandarian et al. (2018).

The data were collected by the Swiss Federal Institute of Technology at Lausanne, concerning 14.5 km² of the Swiss Jura region. A comprehensive description of the sampling, field, and laboratory procedures is available in Atteia et al. (1994) and Webster et al. (1994), and a detailed exploratory data analysis can be found in Goovaerts (1997).

The data contain topsoil concentrations of seven heavy metals, including lead (Pb), which is used in the present study. Lead concentrations were sampled at 359 locations scattered in space, and are available in two mutually exclusive sets: a calibration set of 259 observations and a validation set of 100 observations. Lead concentrations are expressed in parts per million (ppm,
S.I. units = mg kg\(^{-1}\)) or their logarithm transform. In order to avoid adjustments of the HER-bin boundaries (removing negative concentrations) and to simplify benchmarking comparison, the authors decided to use the logarithm to base ten of Pb throughout the paper (the same logarithm base can be seen for the Pb model in Atteia et al. 1994).

Fig. 3 illustrates the \(\log_{10}(\text{Pb})\) concentrations at the locations of the calibration set, the locations of the validation set, and histogram and cumulative distribution of the calibration set. Table 1 presents the summary statistics of \(\log_{10}(\text{Pb})\) for all datasets. The Swiss federal ordinance defined the regulatory threshold used as the tolerable maximum for healthy soil (FOEFL 1987): Locations with lead concentrations above the critical threshold \((z_c)\) of 50 mg kg\(^{-1}\) (or \(z_c = 1.699\) in its logarithm transform) are considered contaminated. For the available dataset, the concentration at 42.1\% of the calibration set locations exceed this limit as indicated in Fig. 3c. The dotted line in Fig. 3a indicates the transect SW-NE to be discussed in Sect. 3.4.1, based on the cross section found in Goovaerts (1997).

**Fig. 3** Logarithm of Pb of the calibration set. (a) Concentration, (b) histogram, and (c) cumulative distribution.
Table 1  Summary statistics of log_{10} Pb datasets.

|                      | log_{10} Pb         |
|----------------------|---------------------|
|                      | Calibration set | Validation set | Full dataset |
| n                    | 259               | 100             | 359          |
| mean                 | 1.687             | 1.689           | 1.688        |
| entropy*             | 5.348             | 5.167           | 5.453        |
| std. deviation       | 0.184             | 0.214           | 0.193        |
| variance             | 0.034             | 0.046           | 0.037        |
| cv                   | 0.109             | 0.127           | 0.114        |
| maximum              | 2.361             | 2.477           | 2.477        |
| median               | 1.667             | 1.672           | 1.670        |
| minimum              | 1.278             | 1.271           | 1.271        |
| kurtosis             | 4.328             | 4.891           | 4.651        |
| skewness             | 0.854             | 1.038           | 0.931        |

* Evenly spaced bins, with intervals of 0.015 (more in Sect. 3.3).

Regulatory threshold: \( z_c = 1.699 \).

3.2 Performance criteria

The quantitative evaluation of the predictive power of the models was carried out with two criteria for the deterministic results, namely, mean absolute error \( (E_{MA}) \) and Nash–Sutcliffe efficiency \( (E_{NS}) \), and another two for the probabilistic outcomes, i.e., Kullback-Leibler divergence \( (D_{KL}) \) and goodness statistic \( (G) \). These metrics are presented in Eqs. (6), (7), (2) and (9), respectively.

The deterministic performance metrics are defined as

\[
E_{MA} = \frac{1}{n} \sum_{i=1}^{n} |\hat{z}_i - z_i|, \tag{6}
\]

\[
E_{NS} = 1 - \frac{\sum_{i=1}^{n} (\hat{z}_i - z_i)^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2}, \tag{7}
\]

where \( \hat{z}_i \) and \( z_i \) are, respectively, the expected value of the predictions and observed values at the \( i \)th location, \( \bar{z} \) is the mean of the measurements, and \( n \) is the number of tested locations. \( E_{MA} \) was selected because it gives the same weight to all errors, while \( E_{NS} \) penalizes variance as it gives more weight to errors with larger absolute values. \( E_{NS} \) also shows a normalized metric (limited to 1) which favors general comparison.

For verifying the quality of the predicted probability distribution, its accuracy and precision will be calculated for the validation set (where a 'true' measurement is available). While precision is a measure of the narrowness of the distribution, accuracy measures if the true value is contained in some fixed symmetric probability \( p \)-probability intervals (PI), e.g., interquartile range (Deutsch 1997). For evaluating accuracy and precision together, we assess the Kullback-Leibler divergence \( (D_{KL}, \text{Eq. (2)}) \) between the binary probability distribution (above–below threshold) and the true measurement (as shown in Fig. 2b) and take the mean over all validation points. \( D_{KL} \) is more than a measure of accuracy, since it does not need the definition of a probability cutoff to classify the binary distribution as hit or misclassification, and it is dependent on the probability values.
predicted. A maximum agreement $D_{KL} = 0$ is obtained when all binary PMFs are very precise (probability of 1) and accurate (correct prediction) in predicting the true (above or below threshold), and it goes to infinity when a maximum disagreement is met.

Additionally, the accuracy and precision of the full distribution (without binarization) is quantified by analyzing different symmetric $p$-PI. For the predicted conditional probability distribution (CPD) at location $u$, a series of symmetric $p$-PI can be constructed by identifying the limits $(1 - p)/2$ and $(1 + p)/2$ quantiles. For example, 0.5-PI is bounded by the first and third quantiles. In this case, a probability distribution is said to be accurate if there is a 0.5 probability that the true $z$-value at the target location falls into that interval or, equivalently, that over the study area, 50% of the 0.5-PI include the true value (Goovaerts 2001; Deutsch 1997). The fraction of true values falling into the symmetric $p$-PI is computed as:

$$\bar{\xi}(p) = \frac{1}{n} \sum_{i=1}^{n} \xi(u_i; p) \quad \forall \, p \in [0,1],$$

with $\xi(u_i; p) = \begin{cases} 1 & \text{if } F^{-1} \left( \frac{1-p}{2} \right) < z_i \leq F^{-1} \left( \frac{1+p}{2} \right), \\ 0 & \text{otherwise} \end{cases}.$

A distribution is said to be accurate when $\bar{\xi}(p) \geq p$. The cross plot of the estimated $\bar{\xi}(p)$ versus expected fractions $p$ is referred to as an 'accuracy plot'. To assess the closeness of the estimated and theoretical fractions and, consequently, the associated measure of accuracy of the distribution, Deutsch (1997) proposed the following goodness statistic ($G$):

$$G = 1 - \frac{1}{K'} \sum_{k=1}^{K'} w_k \left| \bar{\xi}(p_k) - p_k \right|,$$

where $w_k = 1$ if $\bar{\xi}(p_k) > p_k$, and 2 otherwise. $K'$ represents the discretization level of the computation, i.e., the number of $p$-PI. Twice as much penalization is given to deviations when $\bar{\xi}(p_k) < p_k$ (inaccurate case). Maximum goodness $G = 1$ is obtained when $\bar{\xi}(p_k) = p_k$, and $G = 0$ (the worst case) when no true values are contained in any $p$-PI, hence $\bar{\xi}(p_k) = 0$.

To visualize the spread of the CPD and therefore the precision of the distribution, Goovaerts (2001) averages the width of the PIs that include the true values for a series of probabilities $p$, as follows:

$$\bar{W}(p) = \frac{1}{N} \sum_{i=1}^{N} \xi(u_i; p) \cdot \left[ F^{-1} \left( \frac{1+p}{2} \right) - F^{-1} \left( \frac{1-p}{2} \right) \right].$$

The cross plot of the estimated $\bar{W}(p)$ versus expected fractions $p$ is referred to as an 'PI-width plot'. To be legitimate, uncertainty cannot be artificially reduced at the expense of accuracy (or achieve accuracy at the expense of precision) (Goovaerts 1997 p.435), therefore a correct modelling of local uncertainty would entail the balance of both, accuracy and precision.

Overall, the validity of the model can be asserted when the mean error is close to zero, Nash–Sutcliffe efficiency is close to one, mean of Kullback-Leibler divergence is close to zero, and accuracy (given by the goodness statistic) close to one. Visually, a goodness statistic equal to 1 corresponds to an 'accuracy plot' with maximum agreement between $\bar{\xi}(p)$ and $p$-PI. Note that
the precision is only visually verified throughout the 'PI-width plot', where the narrower the width of the PI (y-axis) the better. In Sect. 3.4.2, we discuss with real examples how these two plots (Fig. 10) interact.

3.3 Benchmark models and setup of HER

This section presents HER model decisions for the described dataset (Sect. 3.1) and briefly describes the two models used in the comparison analysis (Sect. 3.4), namely ordinary kriging (OK) and indicator kriging (IK). The authors suggest consulting Kitanidis (1997), Goovaerts (1997), and Deutsch and Clayton (1998) for a more detailed explanation of the OK and IK methods. For brevity, details of the implemented models were included in Appendix A.

In OK, the unsampled points are estimated by a linear combination of the available data, which are weighted according to a spatial variability function (variogram) fitted to the data. It was selected for the comparison analysis due to the availability of a complete model for the logarithm base of Pb for the Jura dataset in the literature. Therefore, OK parameters were taken directly from Atteia et al. (1994) and its results were used as a baseline for comparison. The fitted variogram parameters are specified in Appendix A (Table A1). It is noteworthy that Atteia et al. (1994) estimated the model parameters by training on the full dataset (calibration plus validation set) while for all other models used in this paper, parameters are estimated by training exclusively on the calibration data set and the performance is assessed for the validation set only. Since the uncertainty of OK models ignores the observation values, retaining only the spatial geometry from the data (Goovaerts 1997 p.180), the explicit assumption of normality, i.e., Gaussian distribution of the estimation errors, is a common practice for modeling local uncertainty in linear geostatistics (Kitanidis 1997 p.68; Goovaerts 1998) which was also assumed in this study. Finally, in order to keep the results comparable, we discretized the predicted probability density functions employing the same discretization (bins) used in HER. This binning scheme is presented and discussed in the next paragraph.

Similarly to HER, the objective of IK is to directly estimate the distribution of the unsampled point without assuming a predefined uncertainty shape. For that, considering a defined cutoff value, an indicator transform (above–below cutoff) of the available data is combined with kriging weights to assess the probability of the unsampled point being above or below this threshold. When dealing with continuous variables, many cutoffs can be defined so that putting together their probability results a full cumulative distribution can be accessed. Since we are dealing with continuous lead concentrations, for a fair comparison between HER and IK, the IK cutoffs were defined to coincide with the bins of HER. Therefore, in total, 69 cutoff values were specified, varying from 1.290 to 2.295 in steps of 0.015 (plus the critical limit \( z_c \) for lead concentration 1.699), covering the calibration set except for the bins containing the minimum and maximum values. We defined the extremes of the distributions predicted by IK as the minimum and maximum Pb concentration of calibration set (1.278 and 2.361, Table 1) as proposed by Deutsch and Journel (1998) and Goovaerts (2009). Furthermore, the lag spacing used for the IK variogram was also the same as that used for the HER infogram computation, namely 70 meters (0.07 km). The parameter file used to model IK is shown in Appendix A (Fig. A3). Although such a large number of thresholds is not a common practice, this decision favored local uncertainty comparison (entropy maps and CPDs).
By using a large number of thresholds, the impact of the linear modeling for the interpolation (within class probabilities) and extrapolation (upper and lower tails) of the distribution is reduced (Goovaerts 2009) at the cost of potentially increasing order relation problems (Rossi and Deutsch 2014 p.160; Goovaerts 1997 p.321). Therefore, results from a more usual model are presented in Appendix B and remarked along the comparison analysis as IK\textsubscript{10}. It was modeled with 10 cutoffs, nine deciles of the calibration histogram plus the critical limit $z_c$, as also employed by Goovaerts (1998 and 2001) for the Jura dataset and following the recommendation by Rossi and Deutsch (2014 p.160) of using between 8 and 15 cutoff values. Finally, for each target point, we linearly interpolate the calculated probabilities and extrapolate it to the calibration bounds for obtaining a complete distribution. This procedure is implemented by Goovaerts (2009), in his AUTO-IK code, script therefore used in this paper for IK modeling and predicting.

For comparison purposes, we fixed the lag distances of IK and HER at equal intervals of 70 meters (0.07 km) and the predicted log\textsubscript{10}(Pb) distributions of OK, IK, and HER were equally discretized with evenly spaced intervals of 0.015. We select this bin width for HER according to Thiesen et al. (2019), in which the size of 0.015 (equivalent to a concentration difference of 1.7 ppm\textsuperscript{1} around $z_c$) showed a stabilization of the cross-entropy ($H_{pq} = H(p) + D_{KL}(p||q)$) when comparing the full calibration set and subsamples for various bin widths, meaning that the bin size is reasonable for the dataset in question.

Furthermore, in order to increase computational efficiency, and due to the minor contribution of neighbors in classes far away from the target, the authors used only the 30 neighbors closest to the target. With the lag (or class), bin width, and number of neighbors defined, it was possible to assess the spatial characterization and, consequently, to proceed with the weight optimization (both available in Appendix A, Fig. A1 and Fig. A2). As shown in Fig. A1, the calculated range included 20 distance classes, therefore reaching 1.4 km (circa three times smaller than the length of the x-domain of about 4 km).

Considering the optimization problem proposed in Sect. 2.2.3, the optimum weights ($w_{\text{AND}}$ and $w_{\text{OR}}$) obtained for Eqs. (3) and (4) are illustrated in Appendix A (Fig. A2). Both contributions considerably decrease until the sixth class (circa 0.4 km), beyond they stabilize and decrease almost linearly until reaching the range (1.4 km, class 20). The optimum contributions obtained for AND and OR aggregation in Eq. (5) were $\alpha = 0.65$ and $\beta = 0$, therefore exclusively intersecting distributions. The spatial characterization, aggregation method, optimized weights, and the set of known observations define the HER model for predicting local distributions.

Finally, the general procedures to obtain target estimates, distributions and the binary probability for the contamination classification was summarized for each method in Table 2. The performance metrics related to each output were also identified in the last column.

\textsuperscript{1} Note that 1.7 ppm is approximately half of the standard deviation of various-sources errors estimated in Atteia et al. (1994) for the Pb dataset.
Table 2 Summary of the method procedures and associated performance metrics.

| Target results | OK | IK | HER | Performance metric |
|----------------|----|----|-----|--------------------|
| Estimate       | With OK, we first obtained the estimate of the target and the associated error variance. | The expected value is obtained from the target distribution. It is particularly called E-type estimate because it comes from a conditional distribution. | Same as IK. | We measured the performance of the estimates using $E_{MA}$ and $E_{NS}$. |
| Distribution*  | With an explicit Gaussian assumption, we derived the target distribution centered on the estimated value and using the error variance. The distribution was then discretized in bins. | The local conditional cumulative distribution of the target is modeled through a series of cutoffs, interpolated when required, and converted to a conditional probability distribution (CPD) discretized in bins. | We directly calculated the local conditional probability distribution (CPD) of the target already discretized in bins. | We measured the accuracy of the distributions using $G$ and the 'accuracy-plot', and its precision by the 'PI-width plot'. |
| Probability of being above or below $z_c$ | To obtain the probability of the target being above $z_c$, we cumulate the probability of the distribution in two bins, above or equal $z_c$, and below $z_c$. | Same as OK | Same as OK | We measured the performance of the classification probability using $D_{KL}$. |

* As discussed previously in this section, all distributions are discretized by the same binning scheme.

3.4 Results from local estimation with HER, IK and OK

Considering the similarities between HER and IK (both nonparametric methods with data dependent distributions), Sect. 3.4.1 focuses on presenting the local predictions of these two methods. Complementarily, OK maps are offered in Appendix B. In Sect. 3.4.2, the performance of all three interpolators previously introduced are compared and discussed.

3.4.1 Model application

Due to their conditional distribution results, this section presents maps and distributions produced by IK and HER, using exclusively the Jura calibration set in their logarithm transform. Hereafter, we omit its logarithm form and refer to the data and results simply as lead (Pb) concentrations. For comparison purposes, an identical color range was used for maps presenting the same information. Additionally, the color bars of Fig. 4 and Fig. 5 discriminate, respectively, the $z_c$ threshold of lead concentration (1.699) and the entropy of the calibration set (5.348 bits, Table 1). All maps were developed using a grid with size of 0.05 km by 0.05 km.

In Fig. 4, we show the expected values (E-type) of lead concentrations. In general, an analogous trend (given by the color shapes) for HER and IK can be seen, with similar low and high pollutant concentration areas. HER is slightly bolder in predicting extremely low (below 1.5) and high (above 2.1) concentrations, presenting larger areas in dark blue and yellow. The estimate map of OK is available in Fig. B2a (Appendix B).
Despite this similar trend of E-type values, their local uncertainty (Fig. 5) is consistently different. While IK estimates lower uncertainty as a whole (with all values below the calibration set entropy of 5.348 bits), HER shows a broader range of entropy values. As expected, HER modeled a higher uncertainty to the west of the study area (Fig. 5a), where no nearby measurements are available, and lower uncertainty in the regions with a higher density of observations. Conversely, IK presents higher entropy in these denser areas.

The general lower entropy of IK map can be attributed, in this case, to the resolution of the local PMF, which is given by the numbers of cutoffs used for modeling. Although supporting the comparison analysis, the use of a finer resolution resulted in local distributions with empty bins (visible in Fig. 8), reducing therefore the uncertainty of the distribution in terms of entropy.

The entropy map and predicted distributions of an IK model with coarse resolution (IK10) are available in Appendix B (Fig. B1 and Fig. B3, respectively). Although in different magnitude, the same behavior of higher uncertainty in denser areas can be seen in IK10 (Appendix B, Fig. B1). The entropy map of OK is available in Fig. B2b (Appendix B).
Considering the maximum acceptable concentration of lead ($z_c$), the probability map for exceeding this critical threshold was produced (Fig. 6). These maps were built by cumulating probabilities above $z_c$. Both methods, HER and IK, present high probability of contamination (black color) in zones of higher Pb concentrations and low probability of contamination (light gray) in areas of lower concentration. HER presented larger areas in black and light gray than IK, being therefore a bit more assertive in its prediction. Note that IK maps Fig. 6 and Fig. 7 do not suffer any considerable impact due to the number of cutoffs selected, since only one cutoff ($z_c$) was used. The probability map of OK is available in Fig. B2c (Appendix B).
According to Goovaerts (1997 p.362), the delineation of contaminated areas can be reached by setting the location as 'contaminated' if the probability of exceeding the tolerable maximum \( z_c = 1.699 \) is larger than the marginal probability of contamination (0.421, estimated in Sect. 3.1) and 'safe' otherwise. The proportion of wrongly classified points is generally minimal close to the marginal probability of contamination (Goovaerts 1997 p.366). Although we use this marginal probability to perform a local classification, in the validation set, all lead models presented the minimum misclassification occurring close to the probability of 0.5 (Appendix B, see Fig. B5). Considering that there are several ways to account for uncertainty in the decision-making process, and therefore greatly different results may be reached depending on the classification criteria (Goovaerts 1997 p.347, p.362), comparing their differences is not on the scope of this work.

Thus, based on the probability map for \( z_c \) (Fig. 6) and the marginal probability of contamination (0.421), we binarize the probabilities to classify the results in 'contaminated' and 'safe' areas. HER and IK results are shown in Fig. 7, and OK in Fig. B2d (Appendix B). The classification maps of HER and IK are relatively similar, however it is possible to see that for this marginal probability the areas declared as safe (related to below threshold pollutant concentrations, black circles) are slightly more connected in the IK map (Fig. 7b). In contrast, we can visually identify a larger region declared as contaminated for HER (Fig. 7a). The OK maps can be found in Appendix B (Fig. B2), where a very local influence of each calibration point can be seen. For a more detailed theoretical comparison between HER and OK, please refer to Thiesen et al. (2020).
Fig. 7  Classification of locations as contaminated by Pb on the basis that the probability of exceeding the critical threshold $z_c$ is larger than the marginal probability of contamination (0.421). (a) HER method, and (b) IK method.

Finally, six locations were selected to be explored in more detail. Four of them are from the validation set, and so present a ground truth (targets A to D, Fig. 8), and two of them were selected from the grid, considering their distance to neighbors and homogeneity (targets E and F, Fig. 8). The target locations, neighbors, and results are presented in Fig. 8.

The locations were chosen with the goal to encompass targets with low (targets A and B) and high (targets C and D) concentration as ground truth, and a more homogeneous (targets A, C, and E) and a more heterogeneous neighborhood (targets B, D, and F).

In general, all IK distributions (Fig. 8) presented empty bins between sampled values, while by construction, HER offered a higher resolution in the sense that the estimated CPD is more continuous. As a trade-off for these empty bins, in IK$_{10}$ (Appendix B, Fig. B3) fewer IK cutoffs were used and the resolution was artificially increased by linearly interpolating the probability values within each cutoff. Nevertheless, IK and HER showed relatively similar shapes and spread for targets A and E, point locations with more homogeneous neighbors. Although their uncertainty differs, the expected values are also comparable, being equal for target E. Despite the homogeneity of their neighborhood, the expected values of targets A and C are not equal to their true value. One reason for this is that just a few (or no) nearby calibration points have a concentration as low (target A) or as high (target C) as their true value. The same applies to target D, although it is in a heterogeneous neighborhood. At last, target F, which is located far from the calibration set, presented a higher entropy when predicted with HER and a more certain distribution for IK model. The local distributions of these targets are available in Appendix B (Fig. B3), where it is possible to check the results of the coarse model IK$_{10}$ with linear interpolation of cutoffs. The finer resolution presented by HER was not achieved by neither IK nor IK$_{10}$.
Fig. 8  Local distribution of targets of the validation set (targets A to D) and grid (targets E and F) for HER (gray) and IK (red). Targets are identified by their coordinates (x,y). The location of each target is shown in a buffer of 600 m by 600 m.
Finally, Fig. 9 depicts the mean and two confidence intervals (CI) of the SW-NW cross section exclusively for the HER model. The SW-NW cross section location and its neighborhood are shown in Fig. 3a. The image also contains nine points from the calibration set (black circles), and seven points from the validation set (red squares), all of them located close to the cross section.

Some of the calibration points exactly match the SW-NE cross section. They can be identified in Fig. 9 when the uncertainty goes to zero (from left to right, 1st, 4th, and 9th black circles). For the ones that are not touching the cross section, it possible to visualize their influence in reducing the uncertainty due to their proximity to the transect. In particular, the 3rd and 4th calibration points (black circles, Fig. 9) are in contrasting situations. The 3rd one is in a region with homogeneous calibration points close by, which result in a narrower uncertainty band, while the 4th one presents an abrupt uncertainty reduction since it is located exactly in the transect but its surrounding is rather heterogenous, which explains the wider CI in its surrounding.

Validation points of high Pb concentrations (2nd and 3rd red squares, Fig. 9) are outside the 95% CI. This happens due to a relative homogeneity of the neighbors in the first six distance classes (within a radius of circa 0.4 km), none of which present such high Pb concentration. On the other hand, for the more continuous regions (4th, 6th, and 7th red squares), E-type predictions were close to the true values. Note that despite their continuous vicinity (with an increasing or decreasing tendency), these three validation points present different uncertainty band sizes. It is wider for 6th and 7th, since they are located in a more heterogeneous region.
3.4.2 Performance comparison

In this section, the validation set was used to calculate the performance metrics of all three methods previously presented, namely, OK, IK, and HER. Table 3 summarizes their values of mean absolute error ($E_{MA}$), Nash–Sutcliffe efficiency ($E_{NS}$), Kullback–Leibler divergence ($D_{KL}$), and goodness statistic ($G$). Accuracy and precision are shown in Fig. 10.

Considering the deterministic metrics (based on the expected value), all models have a comparable $E_{MA}$, although OK presents larger errors than IK and HER according to its lowest $E_{NS}$ (Table 3). IK and HER have similar efficiency $E_{NS}$. On the other hand, when we accumulate the predicted distributions for the validation set in two bins (above and below threshold $z_{c}$) and compare its results to the true observation (as in Fig. 2b), HER presented the smallest divergence $D_{KL}$ (mean over all validations points) between predicted and true probability, and OK the largest.

| Method | $E_{MA}$ | $E_{NS}$ | $D_{KL}$ | $G$ |
|--------|----------|----------|----------|-----|
| OK     | 0.139    | 0.199    | 0.858    | 0.939 |
| IK     | 0.135    | 0.233    | 0.840    | 0.928 |
| HER    | 0.134    | 0.232    | 0.808    | 0.938 |

$E_{MA}$ mean absolute error (best: 0), $E_{NS}$ Nash-Sutcliffe efficiency (best: 1), $D_{KL}$ Kullback-Leibler divergence (best: 0), $G$ goodness statistic (best: 1)

With respect to the Goodness statistic, OK and HER obtained the best $G$ (Table 3). This reflects their accuracy in estimating distributions. Accuracy results are also visible in Fig. 10a and show that the nonparametric models (IK and HER) present points below the 45° line, which indicates the inaccuracy of these probabilistic models for large $p$-PI (mainly $p > 0.70$). The lower $G$ of IK can be attributed to the goodness statistic, Eq. (9), penalizing inaccurate predictions, which shows as points further away from the bisector line around 0.80-PI (Fig. 10a) in comparison to OK and HER. Since a high $G$ can be obtained by distributions with large spread, we used Fig. 10b to evaluate the precision of the models. The PI-width plot shows the estimated $\bar{W}(p)$ versus expected fractions $p$.

Considering that the smaller the PI-width (y-axis) the narrower (more precise) the distribution, Fig. 10b indicates that HER and OK predicted more precise distributions approximately for $p < 0.40$, HER for $0.40 < p < 0.70$, and IK for $p > 0.70$. Although being the model with narrower predicted distributions until $p < 0.70$ (Fig. 10b), HER points in Fig. 10a are above the bisector line being, therefore, considered accurate. On the other hand, for intervals of $p > 0.70$, HER and IK are considered more precise than OK (Fig. 10b), but at the cost of increasing their inaccuracy (Fig. 10a), i.e., their narrowness may cause the proportion of true values falling into these intervals to be smaller than for the OK model.

The accuracy and PI-width plots of the coarse model IK$_{10}$ with linear interpolation of cutoffs is available in Appendix B (Fig. B4). Even though IK and IK$_{10}$ present similar $E_{MA}$, $E_{NS}$, and $D_{KL}$ (Appendix B, Table B1), IK$_{10}$ linear extrapolation of the tails
of the distribution contributes to its increase in uncertainty (precision as large as OK for large intervals, Fig. B4b), therefore increasing accuracy \( G = 0.960 \), Fig. B4a

![Accuracy plot and PI-width plot](image)

**Fig. 10** OK, IK, and HER performance. (a) Proportion of the true Pb values falling within the probability intervals \( p\)-PI of increasing sizes, and (b) width of these intervals versus \( p\)-PI. The goodness statistic \( G \) quantify the similarity between the expected and observed proportions in the accuracy plots.

3.5 **Results from spatial simulation with HERs**

Smooth interpolated maps, such as the ones produced by IK and HER, although locally accurate on average and appropriate for visualizing trends (Rossi and Deutsch 2014 p.167), fail to reproduce clusters of large concentrations, and consequently, should not be used for applications sensitive to the presence of extreme values and their patterns of continuity (Goovaerts 1997 p.370). Therefore, in this section, we show results from applying HER in combination with sequential simulation (HERs, detailed in Sect. 2.3) for generating multiple realizations of the Pb concentration that match the calibration statistics and conditioning data. By construction, all these realizations honor the calibration values at their location, and should reflect the statistics deemed consequential for the problem at hand (Goovaerts 1997 p.370).

The HERs model was calibrated such that the statistical fluctuations of the realizations were reasonable and unbiased (Leuangthong et al. 2005). The statistical fluctuations due to a finite domain size are referred to as ergodic fluctuations (Leuangthong et al. 2005). They mainly happen due to the size of the domain relative to the correlation length. We can expect these statistical fluctuations for anything less than an infinite domain (Leuangthong et al., 2005). In HER and HERs case, the correlation length reaches 1.4 km, i.e., circa three times smaller than the x-domain length. Additionally, Rossi and Deutsch (2014 p.168) argue that between 20 and 50 simulations are generally sufficient to characterize the range of possible values for the simulated values, but we used 100 realizations to match the number of simulations done by Goovaerts (1997) for the Jura
dataset. The fluctuation analysis of one hundred realizations is presented in Fig. 11, where we show their discrepancies in relation to the calibration infogram and marginal distribution. The challenges faced during the model calibration and details about the entropy calculation due to finite sample can be found in Appendix A.

Fig. 11 Ergodic fluctuations of 100 realizations generated with HERs. (a) Infogram and (b) scatterplot of the mean and entropy values.

As desired, the fluctuations of the infogram of the 100 realizations (gray curves in Fig. 11a) are unbiased in relation to the calibration infogram (red curve), spreading above and below it. This means that the spatial variability of the calibration set is reproduced by the realizations (although with some fluctuation). Departures between the calibration statistics and realizations are expected, due to the finite domain and density of conditioning data (Goovaerts 1997 p.372), and important, since they allow one to indirectly account for the uncertainty of the sample statistics (Goovaerts 1997 p.427). Furthermore, artificially eliminating it by removing realizations with fluctuations in relation to calibration set is assuming some certainty. Just for illustration, calculating the E-type at each location over all 100 realizations, we could also assess its spatial correlation (blue curve). As expected (Goovaerts 1997 p.372), the HERs E-type infogram (blue curve) depicts much smaller uncertainty in relation to the calibration infogram (red curve), which reflects the underestimation of the short-range variability of Pb values. It presents also similar shape and magnitude in relation to the infogram of HER E-type (not shown).

Fig. 11b depicts that the entropy of the realizations (gray dots) is above and below the entropy of the calibration set (red dot), and that the mean entropy of the realizations (5.335 bits, represented by the gray dashed line) is near the entropy of the calibration (red dot, 5.348 bits), indicating therefore a reasonable reproduction of the uncertainty in the observed data. On the other hand, the mean of the realizations is overestimated in relation to the calibration set. Although it was not done here, when the simulated PMF is deemed too different from the target PMF an adjustment of the simulated PMFs is possible (Goovaerts 1997 p.427). According to Deutsch and Journel (1997 p.134), any realization can be postprocessed to reproduce the sample histogram; hence the sample mean and variance. To do so, Journel and Xu (1994) proposed a posterior identification of the histogram, which allows improving reproduction of the target PMF while still honoring the conditioning data and without
significant modification of the spatial correlation patterns in the original realization. For the sake of brevity, the improved reproduction of PMFs is beyond the scope of this paper. We should bear in mind that verifying the quality of the reproduction does not provide an indication on the goodness of the set of realizations as a whole, because unlike models of local uncertainty that have true observations to be compared, there is no reference spatial distribution of values to be used in models of spatial uncertainty (Goovaerts 2001).

Two arbitrary stochastic images constructed with HERs and the calibration dataset are pictured in Fig. 12.

![Fig. 12](image-url)  
**Fig. 12** Realizations generated using HERs. (a) Realization #42 and (b) realization #94. Simulation grid size of 0.05 km x 0.05 km.

One can notice that the generated stochastic images (Fig. 12) do not smooth out details of the spatial variation of the Pb concentration as in an estimation map (e.g., Fig. 4). And compared to interpolation techniques like OK, IK, and HER, the variability of the simulated maps is higher due to the incorporation of variability between unsampled points. A comparison between the E-type and simulation variability in space is available in Fig. 11a.

In general, both images present low concentration zones (blue) to the North and Southeast of the study area, which are derived from the low uncertainty and the tendency of low concentration previously verified in the regions (Fig. 5a and Fig. 4a, respectively). Similarly, the zone with high concentration and low uncertainty (around x=2.5 and y=2.5, Fig. 4a and Fig. 5a) presents, in both realizations, high Pb concentrations. On the other hand, regions with higher uncertainty (due to the heterogeneity of the sample data or because they are far away from sample data) present a more variable concentration when comparing both images.
4 Discussion

In general, IK and HER are conceptually different in their modeling. HER relies on empirical probability distributions to describe the spatial dependence of the study area and uses aggregation methods to combine distributions. IK estimates a number of probabilities for a series of cutoffs, for each of which an indicator variogram is modeled to describe the spatial continuity of the study area, and then these probabilities are interpolated to obtain the full distribution. Furthermore, whereas a global set of weights for the classes are obtained with HER, IK performs multiple local optimizations, one for each target and cutoff. On the other hand, both methods have similarities, in that they are nonparametric in the sense that no prior assumption about the shape of the distribution being estimated is needed, in that their results are data dependent, and in that they can be applied to continuous or categorical variables. Such characteristics do not apply for OK, therefore we focus our analysis on IK and HER.

A detailed conceptual discussion comparing OK and HER is available in Thiesen et al. (2020). IK and HER are distance models between any two pair of points, with different forms of inference. While in IK the spatial variability of the attribute values can be fully characterized by a single covariance function, which differ for each cutoff (Goovaerts 1997 p.393), HER relies directly on the dataset to extract one PMF for each distance lag (as seen in Fig. A1). The stationarity assumption behind the inference is a model decision (and not a characteristic of the physical phenomenon) and can be deemed inappropriate if its consequences do not allow one to reach the goal of the study (Goovaerts 1997 p.438). The spatial characterization inference together with the aggregation procedure allows the range of local distributions in HER as well as the simulated values of HERs to naturally reach values beyond the calibration set (both above the maximum and below the minimum). For IK, this is only possible if the user imposes extremes beyond the calibration set. Likewise, the extremes of HER distributions can be restricted by the user according to their interest.

Interestingly, despite their conceptual differences, HER and IK presented comparable performance metrics for the analyzed dataset in deterministic and probabilistic terms (Table 3 and Fig. 10). One exception is the Kullback–Leibler divergence ($D_{KL}$), for which HER was able to classify 'contaminated' and 'safe' areas with higher precision and accuracy. Such accomplishment may be explained by the fact that the HER optimization problem was built around this metric (Sect 2.2.3), although this does not guarantee the best performance in the validation set. Regardless of the performance comparison performed, we should be mindful that here is no unique, best, or true model for modeling uncertainty (Journel 2003). Consequently, there can be several alternatives that depend on the user decision to model the uncertainty which can be more suitable to the problem at hand.

When applying IK, two major issues arise, namely, inconsistent (negative) probabilities when estimating distributions and the choice of interpolation/extrapolation models to increase the resolution of the estimated distribution (Goovaerts 1997 p.441, p.319, p.326, Goovaerts 2009). The first is known as order relation deviations, and is typically treated by a posteriori correction of the estimated probabilities, which imposes nonnegative slopes to the cumulative distribution (Goovaerts 2009). For the latter, there are different ways of achieving a finer resolution of the distribution. Increasing the number of cutoffs leads to cumbersome inference and modeling of multiple indicator variograms (one for each cutoff), which consequently increases the likelihood of order relation deviations due to the empty cutoff classes (Goovaerts 1997 p.326; Rossi and Deutsch 2014 p.160).
As an alternative to that, multiple interpolation and extrapolation models are available in the literature. In such case, where interpolation/extrapolation models are used, besides the arbitrariness of the model selection (Goovaerts 2009), distribution statistics such as the mean or variance may overly depend on the modeling of the upper and lower tails of the distribution (Goovaerts 1997 p.337). Therefore, due to the trade-off between increasing the number of thresholds and using models to derive continuous distributions, both alternatives were discussed in this paper (IK and IK\textsubscript{10}). Nevertheless, regardless of the chosen approach, their potential pitfall lays on their subjection to the risk of suboptimal choices by the user. Conversely, HER avoids imposing these corrections to the distributions and multiple variogram fitting, but its parameter choices (such as distance class size, bin width, number of neighbors, and aggregation type) are also subjective. Yet, for both methods HER and IK, parameter decisions can be based on performance metrics via leave-one-out cross-validation, for example.

Both IK and HER estimated remarkably similar values of Pb concentration (E-type map, Fig. 4). On the other side, the maps associated with the probabilistic results (entropy map in Fig. 5, probability of exceeding the critical threshold in Fig. 6, and classification map Fig. 7) are distinct, with increasing uncertainty of HER in data sparse regions. We noticed that when dealing with sparse data, there is not enough data to fill each cutoff in IK, which, due to the resulting empty bins, decreases the uncertainty (entropy). The opposite happens in the denser regions, where more data is available and the chances of more bins being filled is higher, increasing therefore the entropy for heterogeneous regions. As discussed in Sect. 3.4.1 (Fig. 8), both methods reflected the expected behavior of larger errors in locations surrounded by data that are very different in value (as expected and argued by Goovaerts 1997 p. 180). However, in terms of PMF resolution, the greater computational and inference cost of HER in comparison to IK is balanced by a finer resolution of the distributions, which could not be achieved neither by the IK model nor without interpolation of probability values in IK\textsubscript{10} model. The lack of resolution in IK is particularly severe when using indicator-related algorithms with only a few cutoff values such as the nine deciles of the sample (Deutsch and Journel 1997 p.134). In this case, the loss of information available in continuous data is more accentuated in IK than in HER, due to the indicator transform of the data (Fernández-Casal et al 2018) and few cutoffs. In contrast, the resolution of HER distributions is given by the selected bin width and, consequently, an indicator transform would only be needed as a post-processing step (such as probability analysis of exceeding a critical threshold or a classification map).

In terms of simulation, HERs has proven to be difficult to calibrate. Many parameters were tested until the entropy (variability) of the realizations converged to the entropy of the calibration dataset. In the sensitivity analysis performed (not shown), the authors verified a strong impact of the amount of aggregated distributions (thus, number of neighbors) when intersecting distributions. The stronger the contribution of the AND combination (which is the case for the studies in this paper) and the higher the homogeneity of the data, the more sensitive the spatial variability of HERs is to the number of neighbors. Therefore, in general, intersecting too many equal (homogeneous) PMFs would result in a very narrow output (deflation of the spatial variability) whereas too few could inflate it. Although a first analysis of the simulation procedure and results of HERs was introduced in this paper, and with promising results, further investigations considering the influence of different data properties, implementation of strategies (such as search neighborhood and multiple-grid simulation available in Goovaerts 1997 p.378-379 ), and the addition of transfer functions are needed.
Finally, we should bear in mind that uncertainty arises from our lack of knowledge about the phenomenon under study and, therefore, it is not an intrinsic property of the phenomenon (Goovaerts 1997 p.441). Uncertainty is data-dependent and, most importantly, model-dependent, and hence can be controlled by the user according to their wishes (Journel 2003). No model, hence, no uncertainty measure, can ever be objective: the point is to accept that limitation and document clearly all its aspects (Goovaerts 1997 p.441; Journel 2003). Thus, despite the uncertainty differences between IK and HER and our attempt to quantify their performances, IK and HER presented legitimate results, which exhibited similar accuracy and precision performances.

5 Summary and conclusion

Maps derived from local uncertainty estimates can be used for various decision-making processes, including the assessment for additional data (Journel 1989 p.30). Particularly for concentrations of toxic or nutrient elements, which are rarely known with certainty, decisions are most often made in the face of uncertainty (Goovaerts 1997 p.347). There are various ways to assess uncertainty, such as mapping the probability of exceeding a critical threshold or generating sets of realizations of the spatial distribution of the phenomenon under study. In this paper, we addressed the issue of uncertainty assessment of the continuous attribute of lead concentration in soil by adapting the HER method (histogram via entropy reduction, Thiesen et al. 2020) to deliver local and spatial uncertainty. HER results were compared to two different benchmarking models, namely ordinary kriging (OK) and indicator kriging (IK), with a focus on the latter due to its similarity to HER in terms of being nonparametric and predicting conditional distributions. In general, OK presented the worst performance. IK and HER presented legitimate results, which exhibited comparable accuracy (similarity to the true value) and precision (narrowness of the distribution). One exception was the performance of HER when dealing with the probability of exceeding a critical threshold ($z_c$), which presented a higher accuracy and precision when binarizing the distributions according to $z_c$ and considering the local probability of each point being above or below this threshold. This may be explained by the way that the optimization problem was tailored.

Visually contrasting IK and HER, they presented quite similar maps of expected values (E-type map) while their local uncertainty (Entropy map) presented different shapes, and different magnitudes (depending on how IK was modeled, with more or fewer cutoffs). An interesting aspect verified in the visual comparison was the lack of resolution of the predicted distributions of IK in relation to HER, which happened when no interpolation/extrapolation assumption was done for predicting continuous distributions in IK, and in the presence of sparse data, and it is limited to the sample dataset values (Goovaerts 2009). For predicting continuous distributions, such assumption introduces the risk of suboptimal user choices and of adding information not available in the data (IK case), while its lack turns the model computationally demanding and changes the form of inference (HER case).

The sequential simulation using HER (called HERs) allowed the generation of realizations that reproduced the spatial variability of the sample set. The quality of the realizations was verified in terms of their statistical fluctuation in relation to
the sample set. However, no further analyses of the results (such as benchmarking comparison or adding transfer functions) were carried out, due to the typical absence of a spatial distribution of values to be used as a reference (Goovaerts 2001). HER and its adaptation allow nonparametric estimation and stochastic predictions, avoiding the shortcomings of fitting any kind of deterministic curves and, therefore, the risk of adding information that is not contained in the data (or losing available information), but still relying on two-point geostatistical concepts. In relation to IK, HER has been shown to be a unique tool for estimating nonparametric conditional distribution with the advantage of i) not presenting problems of order-relation deviations, ii) being free of function assumptions for interpolating probabilities or extrapolations tails of distributions, iii) not requiring the definition of various cutoffs and, consequently, their respective indicator variogram modeling, iv) displaying a finer resolution of the predicted distribution, v) avoiding strong loss of information due to data binarization, and vi) bringing more flexibility to uncertainty prediction though the different aggregation methods and optimization strategies. Finally, due to the growing use of stochastic simulation algorithms for uncertainty assessment in soil science and the potential improvement of results given the consideration of soft variables (secondary data), the authors believe that additional investigations of HERs and model adaptations of HER are topics for further research.

6 References

Allard D, D'Or D, Froidevaux R (2011) An efficient maximum entropy approach for categorical variable prediction. Eur J Soil Sci 62:381–393. https://doi.org/10.1111/j.1365-2389.2011.01362.x

Atteia O, Dubois JP, Webster R (1994) Geostatistical analysis of soil contamination in the Swiss Jura. Environ Pollut 86:315–327. https://doi.org/10.1016/0269-7491(94)90172-4

Bandarian EM, Mueller UA, Fereira J, Richardson S (2018) Transformation methods for multivariate geostatistical simulation - Minimum/Maximum autocorrelation factors and alternating columns diagonal centres. Adv Appl Strateg Mine Plan 371–394. https://doi.org/10.1007/978-3-319-69320-0_24

Bel L, Allard D, Laurent JM, Cheddadi R, Bar-Hen A (2009) CART algorithm for spatial data: Application to environmental and ecological data. Comput Stat Data Anal 53:3082–3093. https://doi.org/10.1016/j.csda.2008.09.012

Bourennane H, King D, Couturier A, Nicoullaud B, Mary B, Richard G (2007) Uncertainty assessment of soil water content spatial patterns using geostatistical simulations: An empirical comparison of a simulation accounting for single attribute and a simulation accounting for secondary information. Ecol Modell 205:323–335. https://doi.org/10.1016/j.ecolmodel.2007.02.034

Dabo-Niang S, Ternynck C, Yao AF (2016) Nonparametric prediction of spatial multivariate data. J Nonparametr Stat 28:428–458. https://doi.org/10.1080/10485252.2016.1164313

Deutsch CV (1997) Direct assessment of local accuracy and precision. Geostatistics Wollongong'96 1:115–125

Deutsch CV, Journel AG (1997) GSLIB: Geostatistical Software Library and User's Guide Second Edition Preface to the Second Edition. 369
Fernández-Casal R, Castillo-Páez S, Francisco-Fernández M (2018) Nonparametric geostatistical risk mapping. Stoch Environ Res Risk Assess 32:675–684. https://doi.org/10.1007/s00477-017-1407-y

FOEFL (Swiss Federal Office of Environment, Forest and Landscape) (1987). Commentary on the Ordinance relating to Pollutants in Soil (VSBo of June 9, 1986). FOEFL, Bern. https://op.europa.eu/en/publication-detail/-/publication/f76fa39-2b27-42f2-be1e-9332f795e324

García-Soidán P, Menezes R (2012) Estimation of the spatial distribution through the kernel indicator variogram. Environmetrics 23:535–548. https://doi.org/10.1002/env.2151

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. 102:359–378. https://doi.org/10.1198/016214506000001437

Goovaerts P (1997) Geostatistics for natural resources evaluation, Oxford Uni. New York

Goovaerts P (1998) Geostatistics in soil science: State-of-the-art and perspectives. Geoderma 89:1–45. https://doi.org/10.1016/S0016-7061(98)00078-0

Goovaerts P (1999) Impact of the simulation algorithm, magnitude of ergodic fluctuations and number of realizations on the spaces of uncertainty of flow properties. Stoch Environ Res Risk Assess 13:161–182. https://doi.org/10.1007/s004770050037

Goovaerts P (2001) Geostatistical modelling of uncertainty in soil science. Geoderma 103:3–26. https://doi.org/10.1016/S0016-7061(01)00067-2

Goovaerts P (2009) AUTO-IK: a 2D indicator kriging program for the automated non-parametric modeling of local uncertainty in earth sciences. Comput Geosci 23:1–7. https://doi.org/10.1038/jid.2014.371

Goovaerts P, Webster R, Dubois JP (1997) Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. Environ Ecol Stat 4:31–48

Gómez-Hernández JJ, Wen XH (1998) To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology. Adv Water Resour 21:47–61. https://doi.org/10.1016/S0309-1708(96)00031-0

Gómez-Hernández JJ, Cassiraga EF (1994) Theory and practice of sequential simulation. In: Armstrong M, Dowd PA (eds) Geostatistical Simulations. pp 111–124

Journel AG (1974) Geostatistics for conditional simulation of ore bodies. Econ Geol 69:673–687. https://doi.org/10.2113/gsecongeo.69.5.673

Journel AG (1994) Modeling uncertainty: some conceptual thoughts. In: Dimitrakopoulos R (ed) Geostatistics for the Next Century. pp 30–43

Journel AG (1989) Fundamentals of Geostatistics in Five Lessons. American Geophysical Union, Washington, D.C.

Journel AG (2003) Multiple-point Geostatistics: A State of the Art. Stanford Cent Reserv Forecast 1–52

Journel AG, Huijbregts CJ (1978) Mining geostatistics. Academic Press, London, UK

Leuangthong O, McLennan JA, Deutsch C V. (2004) Minimum acceptance criteria for geostatistical realizations. Nat Resour Res 13:131–141. https://doi.org/10.1023/B:NARR.0000046916.91703.bb
Leuangthong O, McLennan JA, Deutsch CV (2005) Acceptable ergodic fluctuations and simulation of skewed distributions. Appl Comput Oper Res Miner Ind - Proc 32nd Int Symp Appl Comput Oper Res Miner Ind APCOM 2005 c:211–218. https://doi.org/10.1201/9781439833407.ch27

Kitanidis PK (1997) Introduction to geostatistics: applications in hydrogeology. Cambridge University Press, Cambridge, United Kingdom

Loquin K, Dubois D (2010) Kriging and epistemic uncertainty: A critical discussion. Stud Fuzziness Soft Comput 256:269–305. https://doi.org/10.1007/978-3-642-14755-5_11

Metropolis N, Ulam S (1949) The Monte Carlo method. J Am Stat Assoc 44:335–341

Ortiz JC, Leuangthong O, Deutsch CV (2004) A MultiGaussian Approach to Assess Block Grade Uncertainty. Cent Comput Geostatistics Annu Rep Pap 1–12

Rossi ME, Deutsch CV. (2014) Mineral Resource Estimation. Springer, London

Thiesen S, Darscheid P, Ehret U (2019) Identifying rainfall-runoff events in discharge time series: A data-driven method based on Information Theory. Hydrol Earth Syst Sci 23:1015–1034. https://doi.org/https://doi.org/10.5194/hess-23-1015-2019

Thiesen S, Vieira DM, Mälicke M, Vieira, Loritz R, Wellmann JF, Ehret U (2020) Histogram via entropy reduction (HER): an information-theoretic alternative for geostatistics. Hydrol Earth Syst Sci 24:4523–4540. https://doi.org/https://doi.org/10.5194/hess-24-4523-2020

Thiesen S., Vieira DM, Ehret, U (2021): KIT-HYD/HERs: version v1.0, Zenodo, https://doi.org/10.5281/zenodo.4501328.

Webster R, Atteia O, Dubois JP (1994) Coregionalization of trace metals in the soil in the Swiss Jura. Eur J Soil Sci 45:205–218. https://doi.org/10.1111/j.1365-2389.1994.tb00502.x

Weijs SV, van Nooijen R, van de Giesen N (2010) Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. Mon Weather Rev 138(9). https://doi.org/10.1175/2010mwr3229.1

7 Declarations

7.1 Funding

The authors acknowledge the support received by the Deutsche Forschungsgemeinschaft (DFG) and the Open Access Publishing Fund of Karlsruhe Institute of Technology (KIT). The first author also acknowledges the support of the Graduate Funding from the German States program (Landesgraduiertenförderung).

7.2 Competing interests

The authors declare that they have no conflict of interest.
7.3 Availability of data and material

The Jura dataset and AUTO-IK (Goovaerts 2009) script were obtained directly on Goovaert's personal website, namely sites.google.com/site/goovaertspierre/pierregoovaertswebsite/download/, options 'Jura Data' and 'Automatic Indicator Kriging Program (AUTO-IK)'.

7.4 Code availability

The source code of the adapted version of HER and its sequential simulation (HERs), containing spatial characterization, convex optimization and distribution prediction, is published alongside this manuscript via GitHub at https://github.com/KIT-HYD/HERs (Thiesen et al. 2021). The repository also includes scripts to exemplify the use of the functions and the dataset used in the case study.

7.5 Authors' contributions

All authors contributed to the study conception and design. Material preparation, data selection and analysis were mainly performed by ST, who also provided the first draft of the manuscript. ST implemented the HERs and performed the simulations. All authors contributed with the interpretation of the models and commented on previous versions of the manuscript. All authors read and approved the final manuscript. ST led the results analysis and manuscript preparation and revisions.
Appendix A: Model parameters

This section presents complementary material regarding the calibration of the models analyzed in the paper, namely, ordinary kriging (OK), indicator kriging (IK), histogram via entropy reduction (HER), and its sequential simulation version (HERs).

OK

Due to the availability of an OK model for the logarithm base of the Jura dataset in the literature, OK was parametrized according to Atteia et al. (1994). It was modeled with two spherical variograms, with the following parameters:

|                      | Nugget | Sill  | Range (km) |
|----------------------|--------|-------|------------|
| spherical model 1    | 0.0096 | 0.0228| 0.287      |
| spherical model 2    | 0.0131 | -     | 2.605      |

HER

This section presents the spatial characterization of the data using HER (Fig. A1) and the optimum weights obtained to be used in aggregation methods (Fig. A2).

Fig. A1a presents the raw infogram from where the class PMFs (Fig. A1b) and, consecutively, the infogram (Fig. A1c) and were obtained. In Fig. A1b, the Euclidean distance (in km) relative to the class is indicated after the class name in interval notation (left-open, right-closed interval) and, for brevity, only the odd classes are shown. The visual increasing of the spread of the Δz PMFs given the distance class (Fig. A1b) is numerically verified also in the infogram (red curve, Fig. A1c), which presents increasing entropy (therefore, decreasing spatial dependence or increasing spatial disorder) with distance. As shown in Fig. A1c, the calculated range included 20 classes, reaching 1.4 km (circa three times smaller than the x-domain length of about 4 km). The range was identified as the point beyond which the class entropy exceeded the entropy of the full data set (seen as the intersect of the blue and red-dotted lines).

The number of pairs forming each Δz PMF and the optimum weights ($w_{\text{AND}}$ and $w_{\text{OR}}$) obtained for Eqs. (3) and (4), respectively, are illustrated in Fig. A2. About 30% of the pairs (20 294 out of 66 822 pairs) are inside the range, where the first class has just under 500 pairs and the last class inside the range (light blue) has above 1500 pairs. Decreasing contribution of the weight with the distance is seen in Fig. A2b, with strong influence of the first six classes (until about 0.4 km). Furthermore, the optimum contribution of AND and OR aggregation, Eq. (5), for this model was $\alpha = 0.65$ and $\beta = 0$. 
Fig. A1  HER spatial characterization of the data. (a) Infogram cloud (b) $\Delta z$ PMFs by class, and (c) infogram.

Fig. A2  HER model characteristics. (a) Class cardinality and (b) optimum weights - Eqs. (3) and (4).
This section presents the parameters used in AUTO-IK program (developed by Goovaerts 2009) to calibrate the indicator kriging model for the dataset presented in paper dataset. The parameter file employed is available Fig. A3. The program AUTO-IK described in Goovaerts’ (2009) publication is available on his personal website https://sites.google.com/site/goovaertspierre/pierregoovaertswebsite/download/indicator-kriging.

Fig. A3 Parameter file used for geostatistical analysis of log_{10} Pb required by AUTO-IK.exe. Indicator semivariograms for thresholds corresponding to 68 equally spaced cutoffs plus $z_c$ threshold, are computed using 30 lags of 0.07 km. The models are fitted automatically and used to perform full ordinary indicator kriging using up to the 32 closest observations located within a radius of 2 km.

Similarly to this model, the authors also generate a model using 10 cutoffs, of which nine are equally spaced p-quantiles of the sample histogram and one is the $z_c$ threshold, i.e., [1.488, 1.543, 1.576, 1.619, 1.667, 1.699 ($z_c$), 1.709, 1.752, 1.816, 1.907].

The decision was based on Goovaerts (1997 p.285), who recommends using $z_c$ as a cutoff to avoid the later interpolation of its probability and argues that cutoff values beyond the first and ninth decile of the calibration set may be inappropriate, since they depend on the spatial distribution of a few pairs of points. In general, Rossi and Deutsch (2014, p.160) also recommend between 8 and 15 cutoff values. Thus, due to its 10 cutoff values, this model is called IK_{10}. In both models we have the inclusion of the $z_c$ as a cutoff, which is recommended by Goovaerts (1997 p. 285) a good practice to avoid that the PMF value at this cutoff needs to be interpolated or extrapolated.
**HERs**

For the sequential simulation model (HERs), we verified the quality of the reproduction of the realizations similarly to the work of Goovaerts (1997) and Leuangthong et al. (2005). The final optimum weights were practically the same as HER model, with the identical infogram and PMF of the classes of HER (as in Fig. A1), same cardinality and similar \( w_{\text{OR}} \) and \( w_{\text{AND}} \) (as in Fig. A2), \( \alpha = 0.55 \) (intersecting PMFs) and \( \beta = 0 \) (averaging PMFs). The small changes on the optimum weights (automatically obtained) happened since the number of neighbors used for HERs was set to seven (instead of 30 used for HER).

Although HER and HERs models resulted both in a pure intersection of PMFs (since we have just \( \alpha \) contribution), the influence in the number of neighbors plays an important role when intersecting distributions and, therefore, we reduced it to 7 in HERs (instead of 30 used for HER). As explored in Thiesen et al. (2020), the higher the number of (similar) distributions to be intersected, the smaller the uncertainty of the resultant distribution. Consequently, due to the sequential procedure of HERs – in which for each iteration we artificially add an extra sample to the data to condition the next prediction – the number of distributions to be intersected greatly increase in relation to the validation set. Thus, to balance this decrease in the entropy (uncertainty), the authors have chosen to reduce the number of neighbors. This implementation decision (number of neighbors) was done by simply checking the infogram of each realization, until it was unbiased in relation to the sample set (Fig. 11a). This is how we also validate the model regarding ergodic fluctuations.

It is important to note that estimating entropy via a finite sample have the tendency to be underestimated (Darscheid 2017). Therefore, considering the great discrepancy in the amount of data between the calibration set (259 observations) and realizations (grid with more than 10,000 targets), we introduced a bias in the realizations so that they could be compared to the calibration set (Fig. 11b). This was conducted by drawing 259 points from each realization (with no replacement), calculating their entropy, repeating it 1,000 times, and taking the mean of this repetitions. Although the bias of the calibration set could be estimated (as proposed by Steck and Jaakkola 2004; Darscheid 2017), a bias correction of the entropy of the calibration set was not straightforward since the obtained value was just a reference to bound the maximum bias and not its exact value. Conversely, adding a bias to the realizations allowed the comparison of the entropy of the calibration set and of the realizations.

Additionally, the authors verify the existence of connectivity of extremely high and small concentration values using indicator variograms for the deciles of 0.2 and 0.8 and different realizations (not shown). The results pointed out no destructuration effect (also known as maximum entropy property, Goovaerts 1997 p.272 p.393), e.g., for the realizations #42 and #94 (Fig. 12), due to the similarity of the indicator variogram of the calibration set and simulated realizations for the different deciles. Therefore, HERs present itself as an appropriate method for cases where extreme values are spatially correlated.
Appendix B: Extra results

This section consolidates extra results for the local uncertainty of OK, IK, IK$_{10}$ and HER models. Fig. B1 displays the entropy map of IK$_{10}$. It is noteworthy that the E-type, probability, and classification maps were not included for IK$_{10}$ due to their similarity to the ones produced to the refined IK model.

**IK$_{10}$ | Entropy map (10 threshold model)**

Fig. B1  Local uncertainty via Entropy map for IK$_{10}$.

Fig. B2 displays the local results for OK model, including estimation, entropy, probability and classification maps. Similar to Goovaerts’ (1997 p.362) results, the estimation map of OK (Fig. B2a) which is optimal for least-square criterion, tends to overestimate the Pb concentration, leading to most locations being classified as contaminated (Fig. B2d). While the OK (Fig. B2a) and E-type estimates presented in the paper (Fig. 4) are similar, their uncertainty (OK in Fig. B2b and Fig. 5) are completely different. The maps of OK entropy indicated greater uncertainty where data are sparse, whereas the uncertainty is smallest near data locations. Such effect is expected since OK ignores the observation values, retaining only the spatial geometry from the data (Goovaerts 1997 p. 180).
Fig. B2  Maps of ordinary kriging for $\log_{10} \text{Pb}$ dataset. (a) OK estimates, (b) local uncertainty in terms of information, (c) probability of exceeding the critical threshold $z_c$, and (d) classification of locations as contaminated by Pb on the basis that the probability of exceeding the critical threshold $z_c$, is larger than the marginal probability of contamination (0.421).

The local distributions of IK, IK$_{10}$ and HER models are displayed in Fig. B3. In this image, we can relate the bin-filling effect of the linear interpolation and extrapolation of the distribution assumed in IK$_{10}$ with IK.
Fig. B3  Local distribution of targets of the validation set (targets A to D) and grid (targets E and F) for HER (gray), IK (red), and IK$_{10}$ (purple).

Table B1 (performance results) and Fig. B4 (accuracy and PI-width plots) contain information already presented in the paper, with the inclusion of IK$_{10}$.

**Table B1** Cross-validation results for OK, IK, IK$_{10}$, and HER method.

| Method | $E_{MA}$  | $E_{NS}$  | $D_{KL}$ | $G$  |
|--------|-----------|-----------|----------|------|
| OK     | 0.139     | 0.199     | 0.858    | 0.939|
| IK     | 0.135     | 0.233     | 0.840    | 0.928|
| IK$_{10}$ | 0.135  | 0.230     | 0.840    | 0.960|
| HER    | 0.134     | 0.232     | 0.808    | 0.938|

$E_{MA}$ mean absolute error (best 0), $E_{NS}$ Nash-Sutcliffe efficiency (best 1), $D_{KL}$ Kullback-Leibler divergence (best 0), $G$ goodness statistic (best 1)
Fig. B4  OK, IK, IK\(_{10}\), and HER plots of the (a) proportion of the true Pb values falling within the probability intervals (p-PI) of increasing sizes and (b) the width of these intervals versus p-PI.

Fig. B5  Proportion of validation locations (a) that are declared contaminated with respect to Pb concentration and (b) that are wrongly classified for OK, IK, and HER models.

The misclassification given different probability cutoffs is shown in Fig. B5. Although, for all models, misclassification is not minimal at the marginal probability of 0.421 (estimated in Sect. 3.1), they have a similar monotonic tendency of decreasing its
values until the minimum (at about 0.5). IK_{10} presented similar misclassification in comparison to IK, which was not plotted to avoid interference with the visualization.

**Additional references to the appendices**

Darscheid P (2017) Quantitative analysis of information flow in hydrological modelling using Shannon information measures. Master thesis. Karlsruhe Institute of Technology

Steck H, Jaakkola TS (2004) Bias-Corrected Bootstrap and Model Uncertainty. In: Thrun S, Saul L, Schölkopf B (eds) Advances in Neural Information Processing Systems. MA: MIT Press, Cambridge, p 8