Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-021-02038-5

ORIGINAL PAPER

Assessing local and spatial uncertainty with nonparametric geostatistics

Stephanie Thiesen[1] · Uwe Ehret[1]

Accepted: 13 May 2021

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 how HER can be used for estimating threshold-exceeding probabilities, it is applied 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 the literature. For the analyzed dataset, IK and HER predictions achieve the best performance and exhibit comparable accuracy and precision. Compared to IK, advantages of HER for uncertainty estimation in a fine resolution are that it does not require modeling of multiple indicator variograms, correcting order-relation violations, or defining interpolation/extrapolation of distributions. Finally, to avoid the well-known smoothing effect when using point estimations (as is the case with both kriging and HER), and to provide maps that reflect the spatial fluctuation of the observed reality, we demonstrate how HER can be used in combination with sequential simulation to assess spatial uncertainty (uncertainty jointly over several locations).

Keywords Nonparametric geostatistics · Non-Gaussian · Conditional distribution · Sequential simulation · Uncertainty analysis · Risk mapping

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 of 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 restrictions of this approach are i) that the distribution of the estimation error is assumed to be normal, and ii) that the variance of the errors is assumed to be independent of the data values, and only dependent 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. Instead, they 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 it applies, all marginal and conditional distributions are Gaussian, and hence the

✉ Stephanie Thiesen
stephanie.thiesen@kit.edu

[1] Institute of Water Resources and River Basin Management, Karlsruhe Institute of Technology, Karlsruhe, Germany

Published online: 15 July 2021

variance of the CPD depends only on the data configuration, not on the data 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 an erroneous parametric model assumption (Fernández-Casal et al 2018). IK, on the other hand, was developed to avoid assuming 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 observations (e.g., from continuous to binary) 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 at unsampled locations. Histogram via entropy reduction (HER) was primarily proposed to bypass fitting spatial correlation functions and assumptions about the underlying distribution of the data. In addition, it is a proper framework for uncertainty estimation since it accounts for both spatial configuration and data values and offers higher generality than 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 incorporating different uncertainty properties by selecting the aggregation method. For the present paper, these primary findings paved the way for the further development of the spatial interpolation framework of HER to assess both i) the local uncertainty when dealing with categorical data and threshold-exceeding probabilities, and ii) the spatial uncertainty by reproducing the spatial fluctuation of the dataset with sequential simulation.

In the context of risk mapping, an important goal of many environmental applications is to delimit zones in the soil containing potentially toxic substances (Goovaerts et al. 1997 p.334). For decision-making in such a context, 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 compare its results to existing benchmark methods. 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 the literature.

Although local estimation methods honor local data, are locally accurate, and have a smoothing effect appropriate for visualizing trends, they are inappropriate for simulating extreme values (Rossi and Deutsch 2014 p.167). In addition, they are suitable for assessing the uncertainty at a specific unsampled location, but not for assessing uncertainty at many locations simultaneously (*spatial uncertainty*; Goovaerts 2001). Therefore, to reproduce the variability observed in the original data and to provide a joint model of uncertainty, HER is expanded using sequential simulation (a version named HERs) which generates stochastic realizations of the field under study. For brevity, in this paper we only demonstrate the feasibility of HERs. Further applications, e.g., for the definition of remediation costs of contaminated areas or the use of transfer functions (Goovaerts 2001) are possible but not included.

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; apply OK, IK, and HER to a real dataset; and compare their estimation and local uncertainty results. Finally, a proof of concept of HERs is presented. In Sect. 4 we discuss results, and in the closing Sect. 5, we summarize the key findings and draw conclusions.

2 Method description

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

2.1 Information theoretic measures employed in HER

To assess the spatial dependence structure of data, minimize estimation uncertainties, and evaluate the quality of probabilistic predictions, we apply two measures of information theory, namely Shannon entropy (H) and Kullback–Leibler divergence (D_{KL}). This section is based on Cover

and Thomas (2006), which we suggest for an introduction to the topic.

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

$$H(X) = -\sum_{x \in \chi} p(x) \log_2 p(x). \quad (1)$$

The logarithm to base two denotes entropy in unit of bits, which is associated to the number of binary questions needed to reconstruct a random variable. This means that, e.g., the entropy of a fair coin toss is 1 bit or, in other words, the answer of one yes–no question (e.g., is it tails?) is enough to identify the toss output. Therefore, the above expression measures the average uncertainty (or the average number of questions) associated with random draws from a given 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

$$D_{KL}(p||q) = \sum_{x \in \chi} p(x) \log_2 \frac{p(x)}{q(x)}. \quad (2)$$

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 distributions. Additionally, it is used as a scoring rule for performance verification of probabilistic predictions.

Note that from now on, instead of x (used to present general information theoretic concepts in this section), we adjust the variable terminology to z and Δz when dealing with spatial problems.

2.2 HER for local uncertainty

The brief introduction to HER presented in the following is based on Thiesen et al. (2020), further details can be found there. HER is a distribution-free interpolator enclosed in a geostatistical framework. It was formulated to describe spatial patterns and solve spatial interpolation problems. In HER, we incorporate 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 referred to as probability mass functions, PMFs). HER comprises three main steps: i) characterization of spatial dependence, ii) selection of an aggregation method and associated optimal weights, and iii) prediction of the target CPD. These steps are explained in the following sections.

2.2.1 Characterization of spatial dependence

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 ($p(z_0)$ is the estimated probability mass function of z at the unsampled location u_0) given its neighbors 1, 2, and 3 (z_1, z_2, and z_3 are available observations sampled at locations u_1, u_2, and u_3). In order to characterize the spatial dependence, we extract the distribution associated to each neighbor and the correlation length (range) in the following actions. First, for each lag distance interval k – also called distance class or simply class – with bounds d_{k-1} and d_k, we calculate the difference of the z-values between all pairs of observations within the interval ($\Delta Z_k = \{ z_i - z_j \mid i \neq j,\ d_{k-1} < |u_i - u_j| \leq d_k \}$) and generate the corresponding Δz PMF ($p_{\Delta Z_k}(\Delta z)$, Fig. 1a).[1] The entropy values of each Δz PMF (one for each distance class k) is visualized as a 2D plot called infogram ($H(\Delta Z_k)$, Fig. 1b). The infogram describes the statistical dispersion of pairs of observations for the distance separating these observations (Thiesen et al. 2020). Quantitatively, it is a way of measuring the uncertainty about Δz given the separation distance of the data, meaning that observations start becoming less informative as the distance increases. Note that in the same figure, the range can be identified as the distance where the entropy of the classes exceeds the full dataset entropy $H(\Delta Z)$, calculated over the difference of z-values between all pairs of observations in the dataset ($\Delta Z = \{ z_i - z_j \mid i \neq j \}$). This range definition is based on the principle that the observations beyond this distance start becoming uninformative, and it is pointless to use information outside of this neighborhood.[2] Finally, we associate to each neighbor the Δz PMF of the corresponding class k, according to its

[1] Note that Z and ΔZ are random variables within the continuous intervals $z \in [z_{\min} - \Delta z_{\max},\ z_{\max} + \Delta z_{\max}]$ and $\Delta z \in [-\Delta z_{\max},\ \Delta z_{\max}]$, respectively, where $\Delta z_{\max} = \max_{i,j} |z_i - z_j|$, $z_{\min} = \min_i z_i$ and $z_{\max} = \max_i z_i$ are calculated over all observations z_i in the calibration dataset.

[2] In the unusual case where the entropy of the classes at large distances does not exceed the entropy of the full dataset, to improve the computational efficiency, we recommend to manually set the range of the infogram by identifying the saturation on the entropy of the classes (similarly to the process done for a variogram fitting).

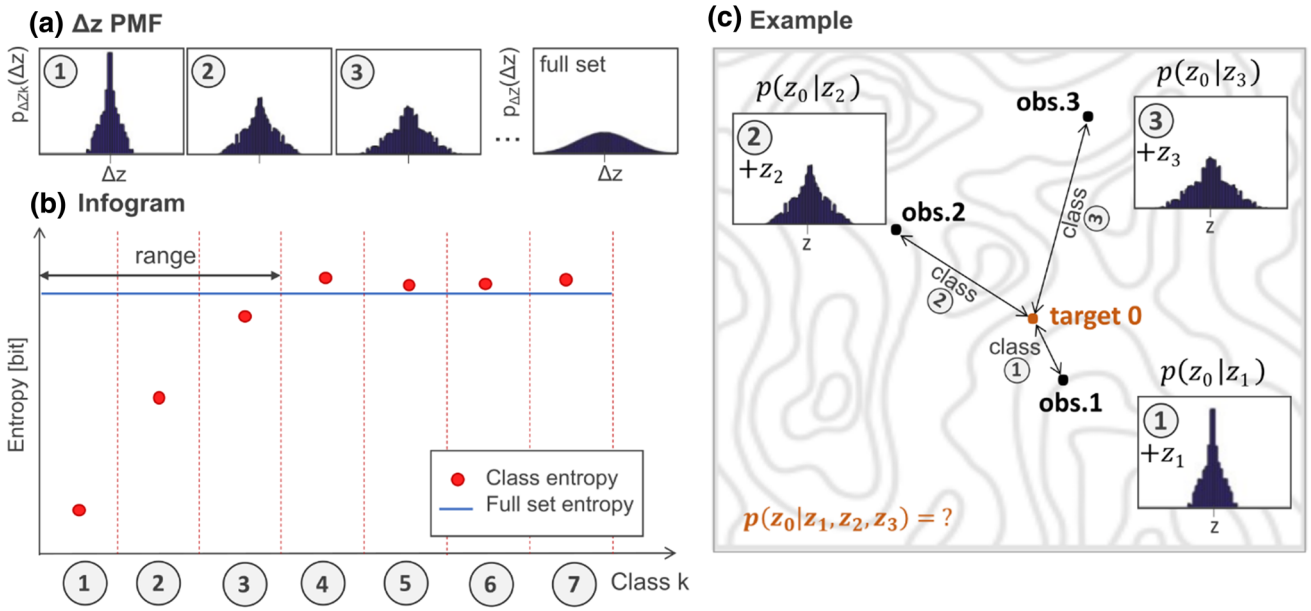

Fig. 1 Schematic of the HER method. **a** Δz PMFs $p_{\Delta Z_k}(\Delta z)$ of the difference in the z-values (Δz) between all pairs of observations within distance class k and Δz PMF $p_{\Delta Z}(\Delta z)$ of the full dataset; **b** infogram, obtained by calculating the entropy $H(\Delta Z_k)$ of PMFs in (a) and plotting them against their respective distance class, with the range determined by the entropy of the full dataset $H(\Delta Z)$; and **c** practical example where the target value to be estimated is z_0 and the available observations are z_1, z_2, and z_3

absolute lag distance from the target, then shift this distribution by its z-value $p(z_0|z_i) = p_{\Delta Z_k}(z_0 - z_i)$, 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 $p(z_0|z_3)$. A practical example using HER is shown in Fig. 13 with more details.

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. The aggregation method is based on work by Allard et al. (2012), which we recommend as a summary of existing aggregation methods. The probability aggregation yields a single, global distribution for the target 0, so that the joint probability $p(z_0|z_1, ..., z_n) \approx P_G(p(z_0|z_1), \ldots, p(z_0|z_n))$, with z_0 being the estimation of the target value (at an unsampled location) and z_i values at neighboring locations, where $i = 1, ..., n$ are the indices of the sampled observations 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), namely linear pooling and log-linear pooling. Linear pooling (Eq. 3) is a way of averaging distributions. It is related to the union of events and associated with the logical operator OR. Multiplication of probabilities, or log-linear pooling in Eq. 4, in turn, is associated with the logical operator AND, and related to the intersection of events. Due to their distinct characteristics, Thiesen et al. (2020) associated the linear aggregation to discontinuous field properties, and the log-linear to continuous ones. The authors exemplified that, if we have two points A and B with different z-values (z_A, z_B) and want to estimate the z-value of a target point X located between both in a continuous field, we would expect that z_X would be somewhere between the z-values of 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 located between both will either belong to the category of A or B, which can be achieved by an OR combination.

The third pooling operator (Eq. 5), which combines $P_{G_{AND}}$ and $P_{G_{OR}}$, was proposed and explored in Thiesen et al. (2020). It optimally expresses continuous and discontinuous properties of a field (controlled by parameters α and β, respectively) by minimizing the relative entropy (D_{KL}) of the estimation and the true data. Since the final distribution of this pooling contains the pure OR (Eq. 3) and the pure AND (Eq. 4) aggregation as special cases, it was recommended by the authors for cases where the field properties are not known a priori.

$$P_{G_{OR}}(z_0) = \sum_{i=1}^{n} w_{OR_i} \; P_i(z_0), \tag{3}$$

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

$$\ln P_{G_{AND}}(z_0) = \ln \zeta + \sum_{i=1}^{n} w_{AND_i} \; \ln P_i(z_0), \tag{4}$$

where ζ is a normalizing constant satisfying $\sum_{z} P_{\mathrm{G_{AND}}}(z) = 1$, n is the number of neighbors, and $w_{\mathrm{AND_i}}$ are positive weights.

$$P_G(z_0) \propto P_{G_{AND}}(z_0)^{\alpha} \; P_{G_{OR}}(z_0)^{\beta}, \tag{5}$$

where α and β 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 (one for each distance class) which minimizes the expected relative entropy (D_{KL}) of all targets. Note that when dealing with single-value observations (or categorical data), this is equivalent to subtracting the probability of the bin containing the true value from one. The D_{KL} evaluation of a single prediction is outlined in Fig. 2a.

In the present study, we propose an adaptation of this loss function (Fig. 2a) to focus on the estimation of threshold-exceeding probabilities (Fig. 2b). Here, instead of optimizing the probability of single bin containing the true observation, we minimize the probability 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' goals were to reduce 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 will simply be referred to as HER.

For both optimization problems (Fig. 2a and Fig. 2b), one optimum weight is obtained for each distance class k and used in Eqs. 3 and 4, referred to as $w_{\mathrm{OR_k}}$ and $w_{\mathrm{AND_k}}$, respectively (here generalized as w_{k}). After that, α and β from Eq. 5 are optimized by grid search, with candidate 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 z-values at sampled locations: within the first distance class, we asymptotically increase the weight towards infinity as the distance approaches zero, by scaling with the inverse of the distance. For all other distance classes, similarly to Thiesen et al. (2020), we linearly interpolate the weights according to the Euclidean distance and the weight of the next class. A practical example of the proposed interpolation is illustrated in Fig. 14b.

2.2.4 PMF prediction

As seen before, to estimate the z-value of the target 0 (i.e., the unknown observation z_0), first we classify its neighbors z_i (sampled observations) according to their distance to the target. Each neighbor is then associated to its corresponding Δz 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 distribution of the target conditioned on all neighbors (z PMF). By construction, the assessed PMF is nonparametric since no prior assumption is made regarding 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

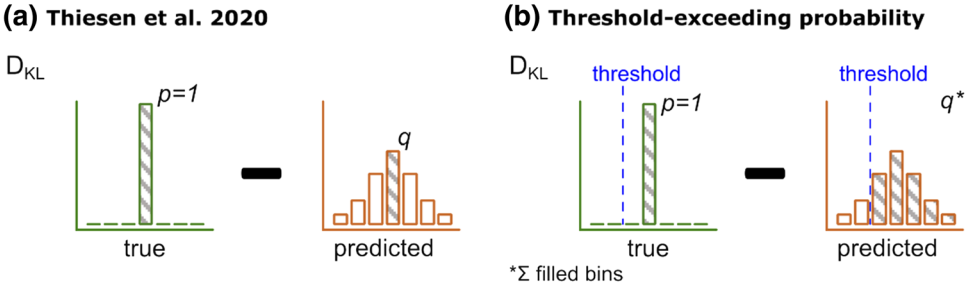

Fig. 2 Optimization problem. **a** Maximizing the probability of the 'true' observation (Thiesen et al. 2020) and **b** maximizing the estimation of threshold-exceeding probability

are associated to the Δz 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 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 simultaneously assess the uncertainty about attribute values at several locations (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 observed reality (Journel 1974; Deutsch and Journel 1998 p.18). 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 conditioned on 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 1998 p.19; Journel 2003). Such equiprobable realizations are known as stochastic images and share the same sample statistics and conditioning data (Gómez-Hernández and Cassiraga 1994).

Sequential simulations with HER are generated by first 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 as explained in Sect. 2.2, ii) randomly draw a single value from this PMF, and iii) assign the value to the grid as an additional observation. With this procedure, we sequentially include the simulated values to the original dataset and use them to condition predictions at the remaining locations. The simulated value (step ii) is derived from 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 1998 p.19; Goovaerts 1999).

Due to the randomness of the path and draws, 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 smooth out the real fluctuation of the attribute due to the missing variability between unsampled locations, HER-based sequential simulation (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 applying them 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 from a 14.5 km^2 area in 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. To simplify benchmarking comparison, the authors decided to use the logarithm to base ten of Pb throughout the paper (the same logarithm base was used 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 the 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

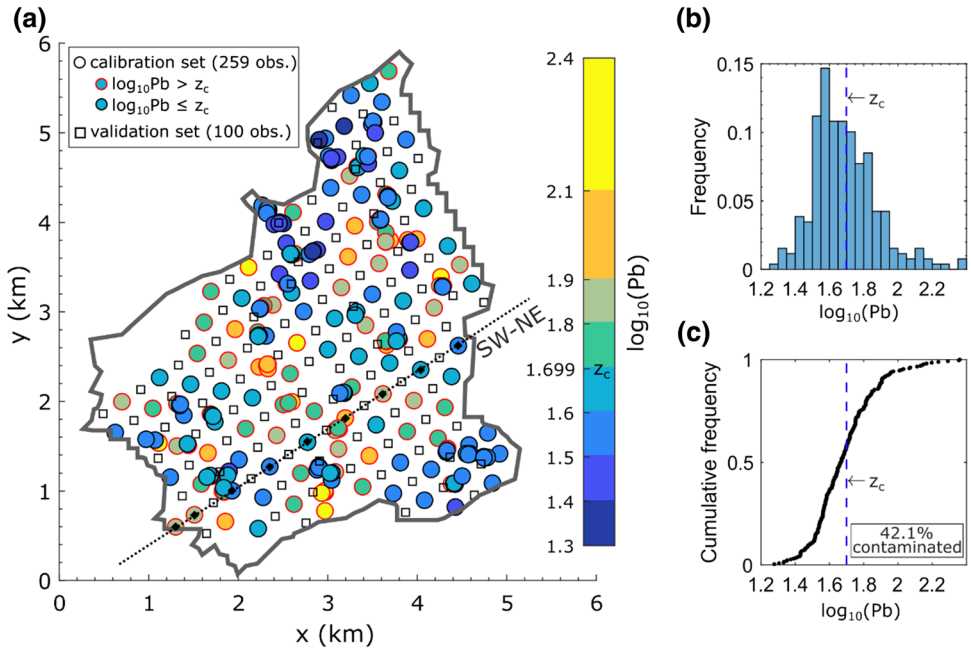

Fig. 3 Calibration set. **a** Concentration values, **b** histogram, and **c** cumulative distribution

Table 1 Summary statistics of $\log_{10}$(Pb) datasets

Statistic	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

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, this limit is exceeded at 42.1% of the calibration set locations, see Fig. 3c. The dotted line in Fig. 3a indicates the transect SW-NE to be discussed in Sect. 3.4.1, which was based on the cross section shown in Goovaerts (1997).

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

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

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

where $\hat{z}_i$ and z_i are, respectively, the expected value of the predictions and observed values at the i-th location, $\overline{z}$ 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. With its limitation to a maximum value of 1, E_{NS} facilitates general comparison.

For verifying the quality of predicted probability distributions, their 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}, 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 predicted probability values. 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). It goes towards infinity as disagreement increases.

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 limiting $(1-p)/2$ 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

$$\overline{\xi}(p) = \frac{1}{n}\sum_{i=1}^{n} \xi(u_i; p) \; \forall \; p \in [0,1], \tag{8}$$

with

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

A distribution is said to be accurate when $\overline{\xi}(p) \geq p$. The cross plot of the estimated $\overline{\xi}(p)$ 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)

$$G = 1 - \frac{1}{L}\sum_{l=1}^{L} w_l \left|\overline{\xi}(p_l) - p_l\right|, \tag{9}$$

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

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

$$\overline{W}(p) = \frac{1}{n\,\overline{\xi}(p)}\sum_{i=1}^{n} \xi(u_i; p) \cdot \left[F^{-1}\left(u_i; \frac{1+p}{2}\right) - F^{-1}\left(u_i; \frac{1-p}{2}\right)\right]. \tag{10}$$

The cross plot of the estimated $\overline{W}(p)$ versus the expected fractions p is referred 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 modeling of local uncertainty will entail the balance of both accuracy and precision.

Overall, the validity of the model can be asserted when the mean error is close to 0, Nash–Sutcliffe efficiency is close to 1, mean of Kullback–Leibler divergence is close to 0, and accuracy (given by the goodness statistic) close to 1. Visually, a goodness statistic equal to 1 corresponds to an 'accuracy plot' with maximum agreement between $\overline{\xi}(p)$ 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 how HER was set up for the described dataset (Sect. 3.1) and briefly describes the two benchmark models, namely ordinary kriging (OK) and indicator kriging (IK). The authors suggest consulting Kitanidis (1997), Goovaerts (1997), and Deutsch and Journel (1998) for a more detailed explanation of the OK and IK methods. For brevity, details of the implemented models were included in Appendix 1.

In OK, the unsampled values 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) lead concentration of the Jura dataset in the literature. Therefore, OK parameters and results were taken directly from Atteia et al. (1994). The fitted variogram parameters are specified in Appendix 1 (Table 4). 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 dataset, and the performance is obtained in 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), we used the explicit assumption of normally distributed estimation errors in this study, which is a common practice for modeling local uncertainty in linear geostatistics (Kitanidis 1997 p.68; Goovaerts 1998) in this study. Finally, to keep the results comparable, we discretized the predicted probability density functions employing the same discretization (bins) as used in HER. This binning scheme is presented and discussed in the next paragraph.

Similar to HER, the objective of IK is to directly estimate the distribution of z at an unsampled location 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 z at the unsampled location being above or below this threshold. When dealing with continuous variables, many cutoffs can be defined so that putting together their probabilities results in a full cumulative distribution. 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 the logarithm of lead concentration of 1.699). We defined the extremes of the distributions predicted by IK as the minimum and maximum Pb concentration of the calibration set (1.278 and 2.361, Table 1) as proposed by Deutsch and Journel (1998 p.238) and Goovaerts (2009). Furthermore, the lag spacing used for the IK variogram was also the same as that used for the HER infogram, namely 70 m (0.07 km). The parameter file used to model IK is shown in Appendix 1 (Fig. 15). Although choosing such a large number of thresholds is not common practice, it facilitates local uncertainty comparison (entropy maps and CPDs).

By using many 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), however at the cost of potentially increasing order relation problems (Rossi and Deutsch 2014 p.160; Goovaerts 1997 p.321). Therefore, results from a more common model referred to as IK_{10} are presented in Appendix 2. Following Goovaerts (1998 and 2001), it was modeled with 10 cutoffs, nine deciles of the calibration histogram plus the critical limit z_c. This is also in line with the recommendation by Rossi and Deutsch (2014 p.160) to use between 8 and 15 cutoff values. Finally, for each target, we linearly interpolate the calculated probabilities and extrapolate the tails to the calibration bounds for obtaining a complete distribution. This procedure is implemented in the AUTO-IK code by Goovaerts (2009), which we used in this paper.

For comparison purposes, we fixed the lag distances of IK and HER at equal intervals of 70 m (0.07 km) and the predicted $\log_{10}(\text{Pb})$ distributions of OK, IK, and HER were equally discretized with evenly spaced intervals of 0.015. We selected 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 around z_c)[3] showed a stabilization of the cross-entropy ($H_{pq} = H(p) + \mathrm{D_{KL}}(p||q)$) when comparing the full calibration set and subsamples for various bin widths. Furthermore, to increase computational efficiency, and due to the minor contribution of faraway neighbors, we 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 1, Figs. 13 and 14). As shown in Fig. 13, the calculated range contains 20 distance classes reaching 1.4 km (roughly a third of the length of the x-domain). Considering the optimization problem proposed in Sect. 2.2.3, the optimum weights (w_{OR} and w_{AND}) obtained for Eqs. 3 and 4 are illustrated in Appendix 1 (Fig. 14b). Both contributions considerably decrease until the sixth class (circa 0.4 km), beyond which 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 are $\alpha = 0.65$ and $\beta = 0$, therefore exclusively intersecting distributions. The spatial characterization, aggregation method, optimal weights, and the set of known observations define the HER model for predicting local distributions.

The general procedures to obtain target estimates, distributions, and the binary probability for the contamination classification are summarized for each method in Table 2. The performance metrics related to each output are also shown.

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. OK maps are offered in Fig. 17 (Appendix 2). In Sect. 3.4.2, the performance of all three interpolators is compared and discussed.

[3] Note that 1.7 ppm is approximately half of the standard deviation of various-sources errors estimated in Atteia et al. (1994) for the lead 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 using the error variance centered on the estimated value. The distribution was then discretized in bins. The Gaussian assumption calls for a kriging variance which is independent of the data values.	The local conditional cumulative distribution of the target is modeled though 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, greater than z_c and less than or equal to z_c.	Same as OK.	Same as OK.	We measured the performance of the classification probability using D_{KL}.

* All distributions are discretized by the same binning scheme

3.4.1 Model application

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 Figs. 4 and 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, a similar 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. 17a (Appendix 2).

Despite the similar trend of E-type values, the local uncertainty (Fig. 5) consistently differs between HER and IK. While IK predictions show generally lower uncertainty (all values are 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 generally lower entropy of the 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), thus reducing the uncertainty of the distribution in terms of entropy. The entropy map and predicted distributions of an IK model with coarse resolution (IK_{10}) are available in Appendix 2 (Figs. 16 and 18, respectively). Although different in magnitude, the same behavior of higher uncertainty in denser areas can be seen in IK_{10} (Appendix 2, Fig. 16). The entropy map of OK is available in Fig. 17b (Appendix 2).

Using the maximum acceptable concentration of lead (z_c), probability maps for exceeding this critical threshold were produced (Fig. 6). These maps were built by cumulating probabilities above z_c. Both methods, HER and IK, show high probability of contamination (in black) in zones of higher Pb concentrations and low probability of contamination (in light gray) in areas of lower concentration. HER shows larger areas in black and light gray than IK, being therefore a bit bolder in its predictions. Note that IK maps in Figs. 6b and 7b do not suffer any negative impact

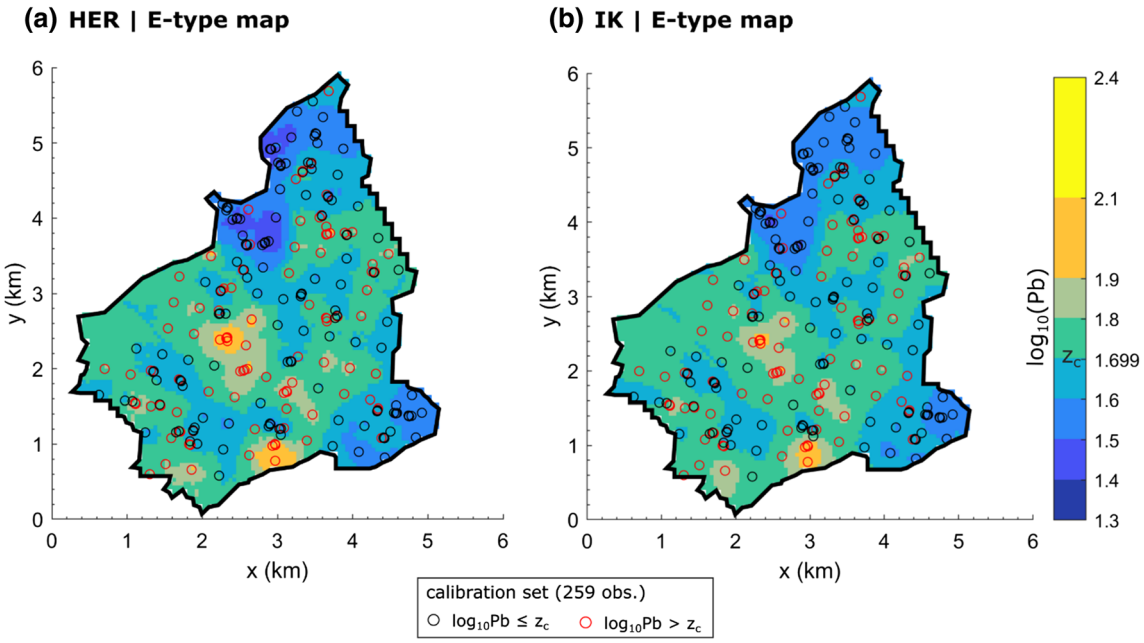

Fig. 4 E-type map. **a** HER method, and **b** IK method

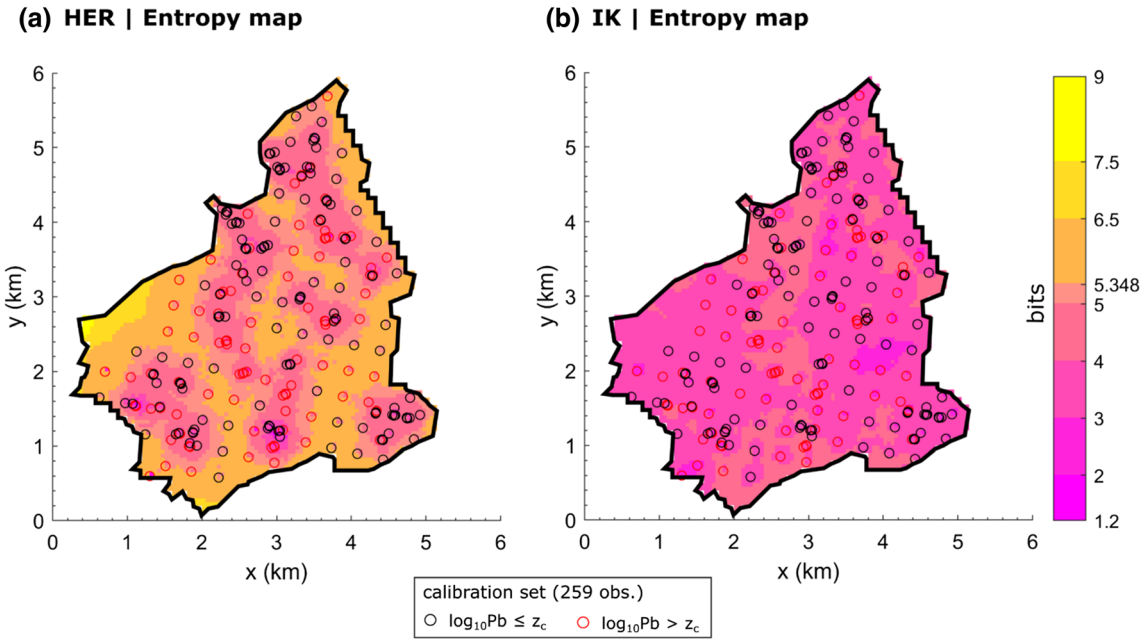

Fig. 5 Entropy map. Local uncertainty in terms of entropy. **a** HER method, and **b** IK method

due to a large number of cutoffs, since only one cutoff (z_c) was used. The probability map of OK is available in Fig. 17c (Appendix 2).

According to Goovaerts (1997 p.362), contaminated areas can be delineated by setting a 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 generally reaches its minimum close to the marginal probability of contamination (Goovaerts 1997 p.366). In the present application, all lead models (OK, IK, and HER) presented the minimum misclassification occurring close to the probability of 0.5 instead of the marginal probability of 0.421 (further discussed in Appendix 2, Fig. 20). However, 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

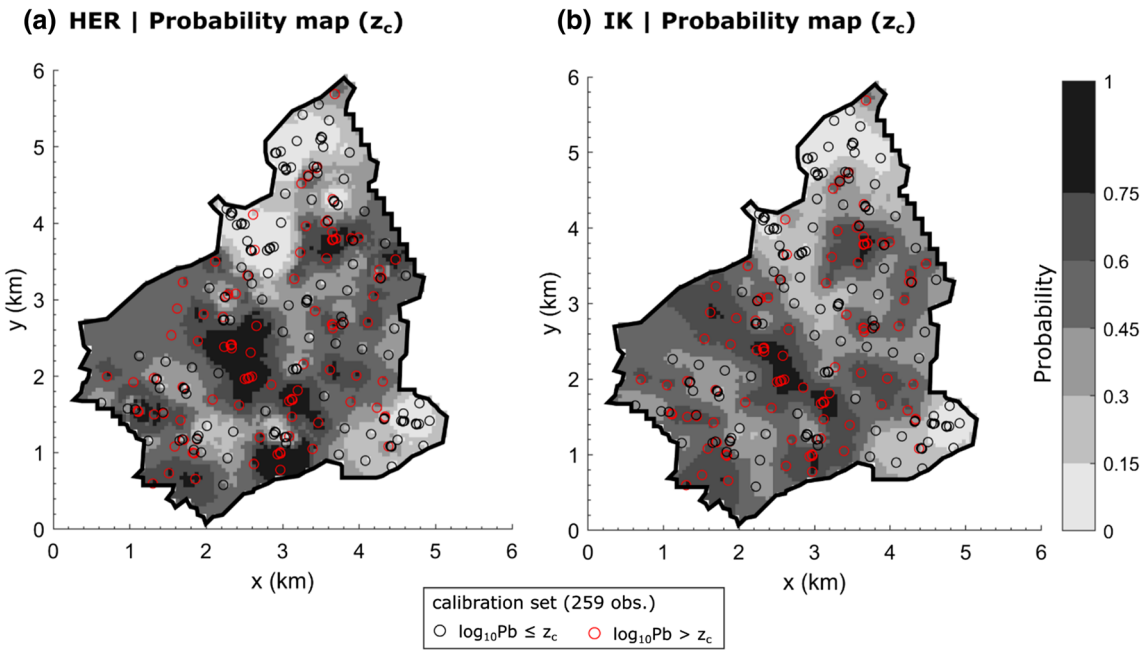

Fig. 6 Probability map. Probability of exceeding the critical threshold ($z_c = 1.699$). **a** HER method, and **b** IK method

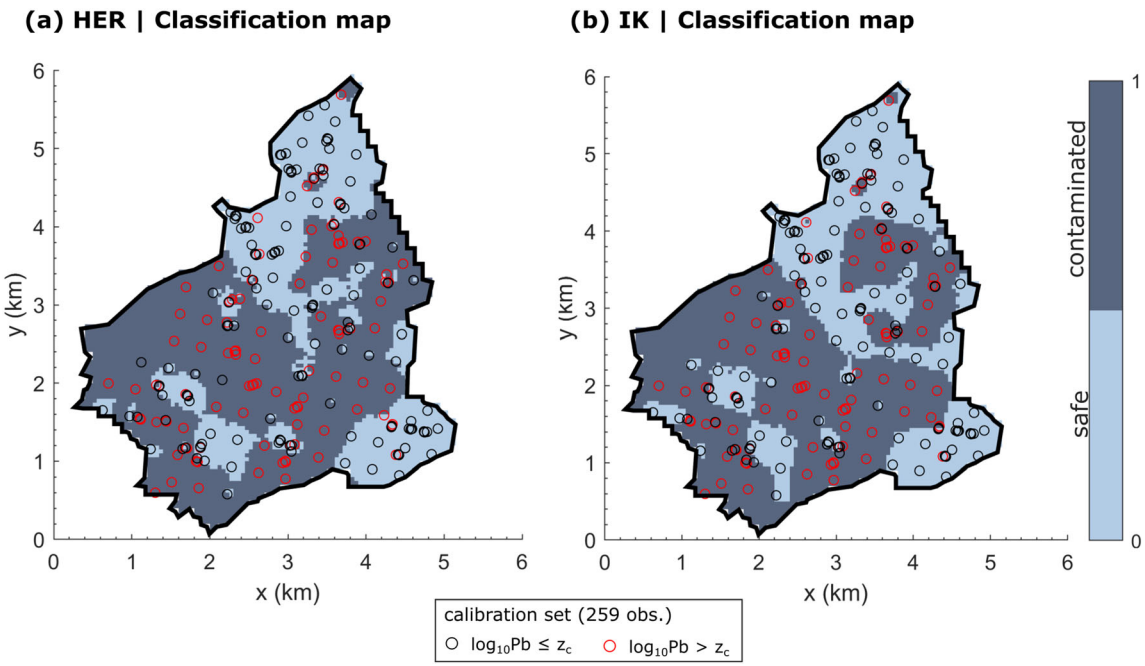

Fig. 7 Classification map. Classification of locations as contaminated by lead on the basis that the probability of exceeding the critical threshold ($z_c = 1.699$) is larger than the marginal probability of contamination (0.421). **a** HER method, and **b** IK method

classification criteria (Goovaerts 1997 p.347, p.362), comparing their differences is not within 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. 17d (Appendix 2).

The classification maps of HER and IK are relatively similar, however areas declared safe by IK are slightly more connected (Fig. 7b). In contrast, contaminated areas are more connected in the HER map (Fig. 7a). The respective OK maps can be found in Appendix 2 (Fig. 17),

revealing a very local influence of each calibration point. For a more detailed theoretical comparison between HER and OK, please refer to Thiesen et al. (2020).

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

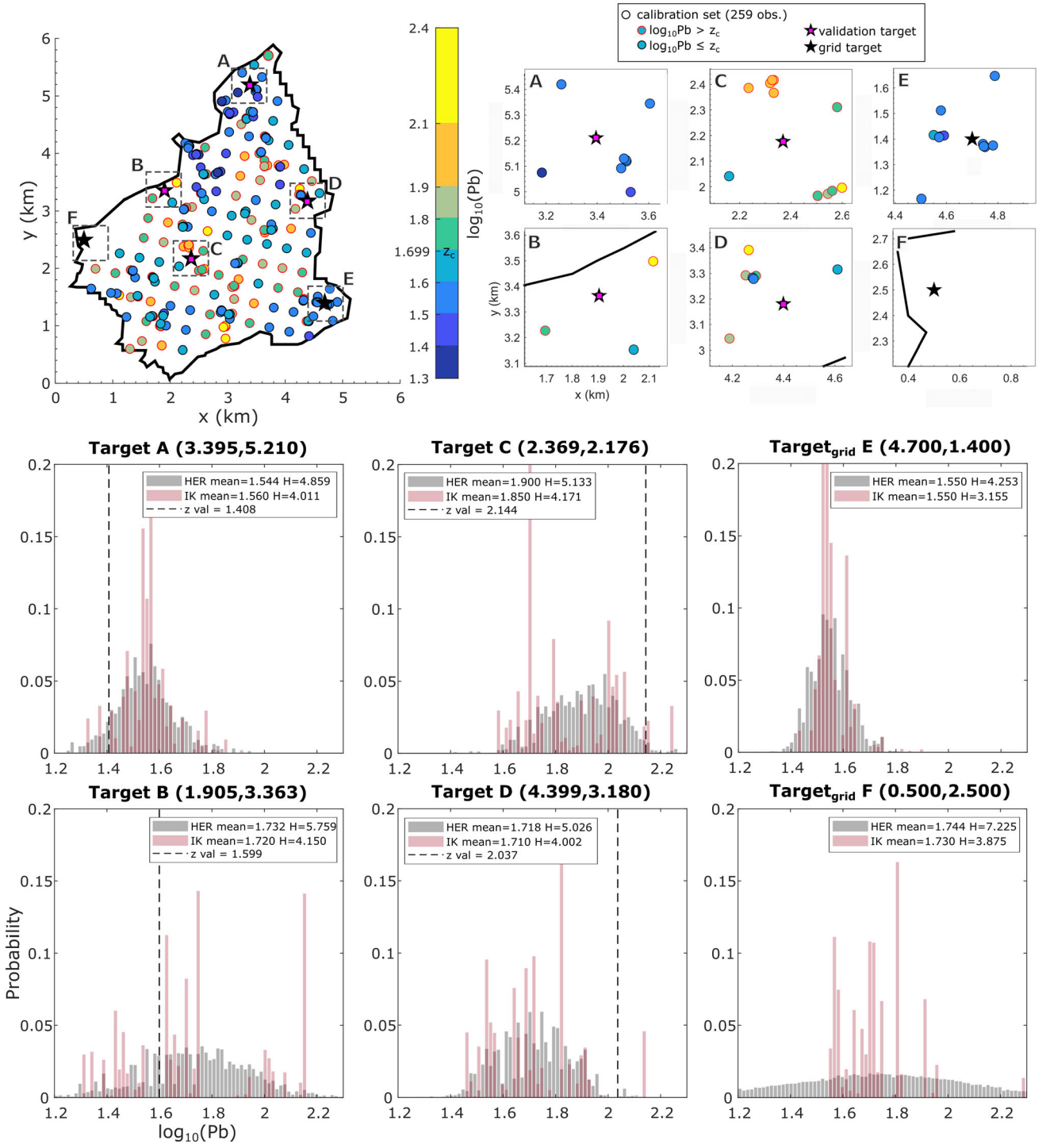

Fig. 8 Local distribution of targets of the validation set (targets A to D) and grid (targets E and F) for HER (gray) and IK (red). Targets are identified by their coordinates (x,y). The location of each target is shown in a buffer of 600 m by 600 m

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 (targets B, D, and F) neighborhood.

In general, all IK distributions (Fig. 8) contain empty bins between sampled values, while by construction, HER offers 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 2, Fig. 18), fewer IK cutoffs were used, and the resolution was artificially increased by linearly interpolating the probability values within each cutoff. Nevertheless, IK and HER show relatively similar shapes and spread for targets A and E, 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, presents a higher entropy when predicted with HER, and a more certain distribution for IK. The local distributions of these targets and the IK_{10} model are available in Appendix 2 (Fig. 18). Neither IK nor IK_{10} achieved the finer resolution of HER.

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 CI 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 as locations where the uncertainty goes to zero (from left to right, 1st, 4th, and 9th black circles). For points not exactly on the cross section, their influence in reducing the uncertainty due to their proximity to the transect is visible. 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 results 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 relatively homogeneous neighbors in the first six distance classes (within a radius of circa 0.4 km), where none presents such high Pb concentration. On the other hand, for the more homogeneous regions (4th, 6th, and 7th red squares), E-type predictions are 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 is used to calculate the performance metrics of OK, IK, and HER. Table 3 summarizes their 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.

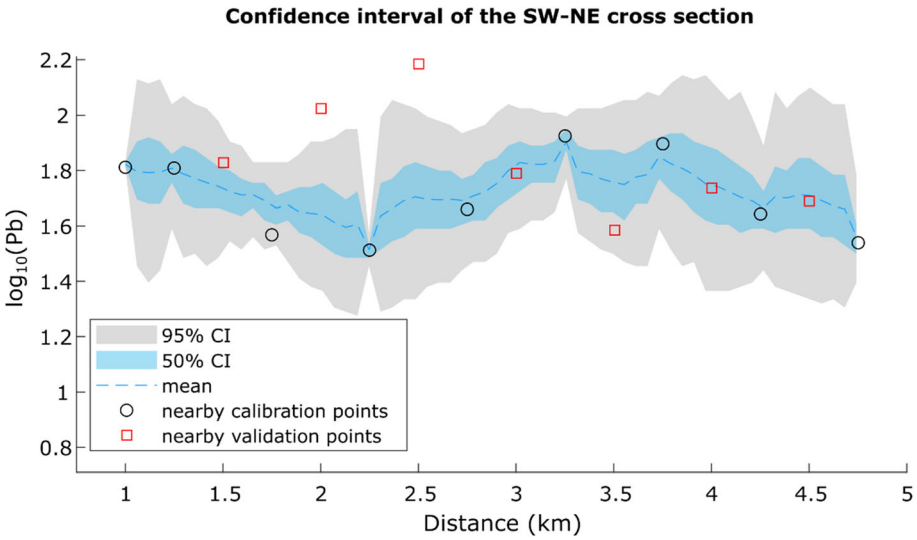

Fig. 9 HER confidence interval (CI) of the SW-NE cross section (shown in Fig. 3a)

Table 3 Cross-validation results for OK, IK, 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
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)

Considering the deterministic metrics (based on the expected value), all models have a comparable E_{MA}. OK presents larger E_{NS} errors than IK and HER (Table 3). IK and HER have similar efficiency E_{NS}. On the other hand, when we cumulate 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 presents the smallest divergence D_{KL} (mean over all validations points) between predicted and true probability, and OK the largest.

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 shown in Fig. 10a. 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 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 $\overline{W}(p)$ 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 predict more precise distributions approximately for $p < 0.40$, HER for $0.40 < p < 0.70$, and IK for $p > 0.70$. Besides 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 in the predicted distributions 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 are available in Appendix 2 (Fig. 19). Even though IK and IK_{10} present similar E_{MA}, E_{NS}, and D_{KL} (Appendix 2, Table 5), IK_{10} linear extrapolation of the distribution tails contributes to its increase in uncertainty (PI-widths as large as OK for large intervals, Fig. 19b), therefore increasing accuracy ($G = 0.960$, Fig. 19a).

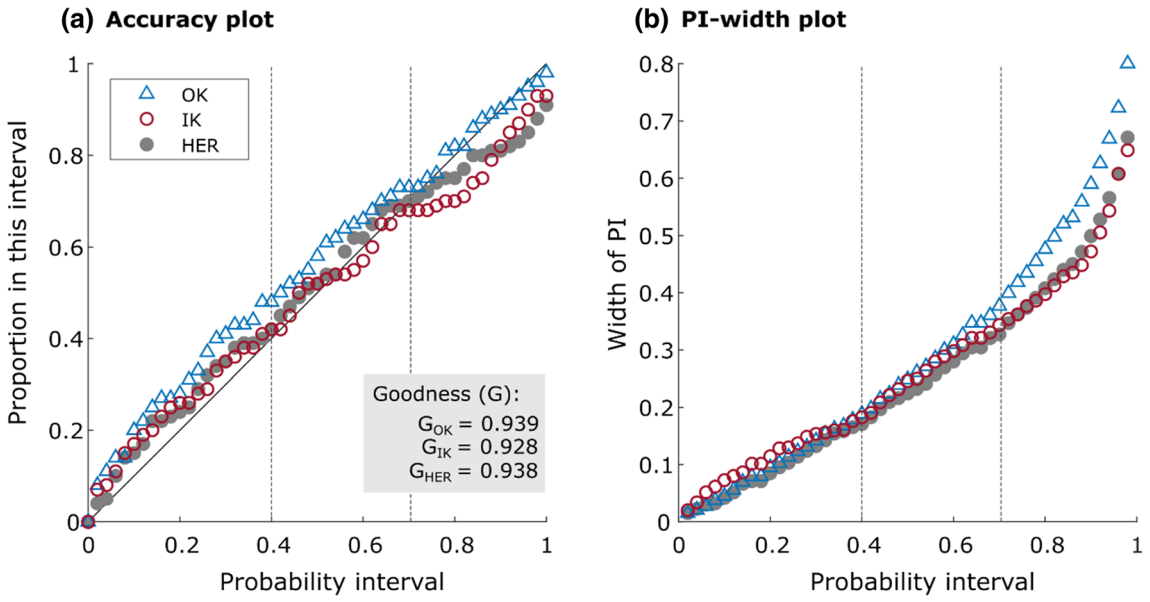

Fig. 10 OK, IK, and HER performance. **a** Proportion of the true lead values falling within the probability intervals (p-PI) of increasing sizes and **b** width of these intervals versus p-PI. The goodness statistic (G) quantify the similarity between the expected and observed proportions in the accuracy 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 the 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 locations and should reflect the statistics deemed consequential for the problem at hand (Goovaerts 1997 p.370).

HERs 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, which 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 one third of 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. 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 1.

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, by calculating the E-type at each location over all 100 realizations, we could also assess its smoothing effect (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 close to the entropy of the calibration (red dot, 5.348 bits), indicating a reasonable reproduction of the uncertainty in the observed data. On the other hand, the mean of the realizations (1.704) is approximately 1% higher than the mean of the calibration set (1.687) and less than 0.25% higher than the mean of the E-type of IK (1.704) and HER (1.700). In this sense, the difference between the mean values of the simulation and the calibration dataset could reflect a bias

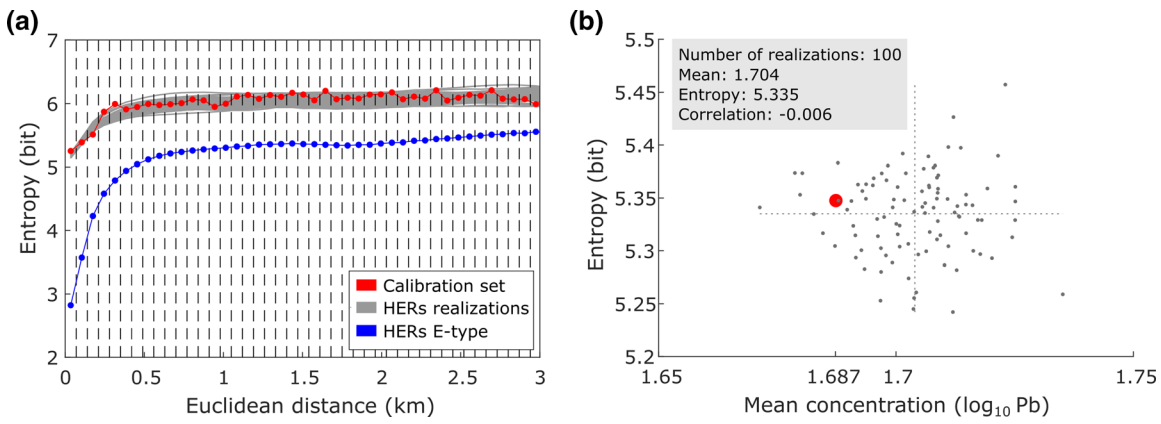

Fig. 11 Ergodic fluctuations of 100 realizations generated with HERs. **a** Infogram and **b** scatterplot of the mean and entropy values

due to spatial clustering of the observations, instead of a bias in the realizations with respect to the true mean of the population (Goovaerts 1997 p.370). 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 (1998 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). For illustration, two arbitrary stochastic images constructed with HERs and the calibration dataset are pictured in Fig. 12.

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 the estimation maps (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 the estimated probabilities are then interpolated to obtain the full distribution. Furthermore, a global set of weights for the classes is obtained with HER, while IK performs multiple local optimizations, one for each target and cutoff. Both methods share similarities: they are nonparametric in

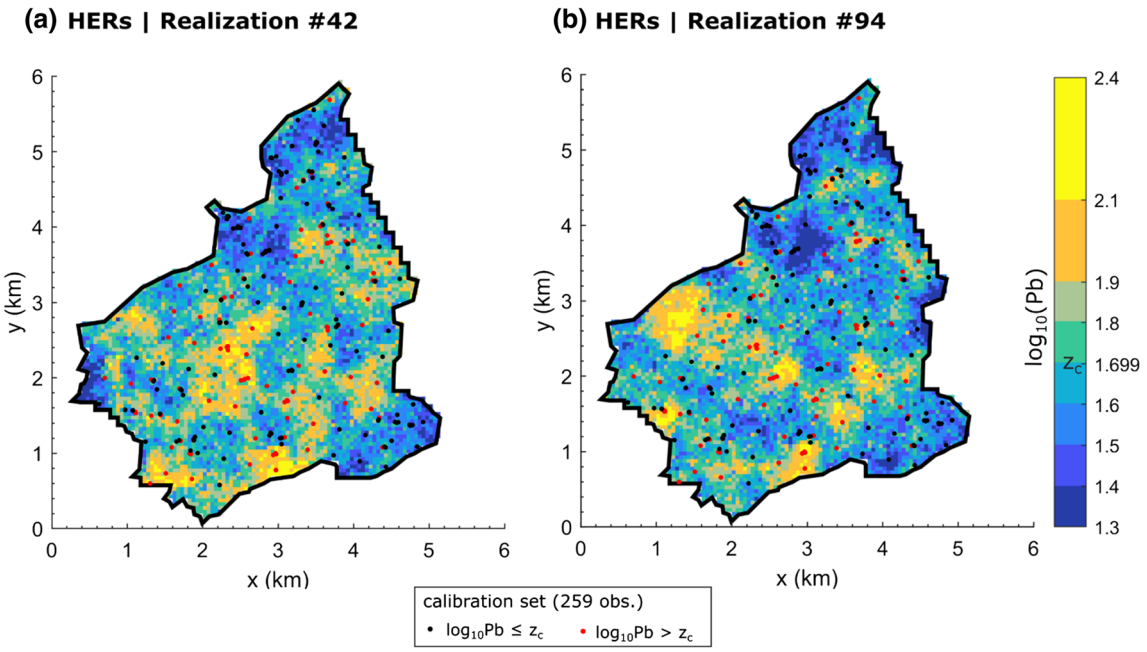

Fig. 12 Realizations generated using HERs. **a** Realization #42 and **b** realization #94. Simulation grid size of 0.05 km × 0.05 km

the sense that no prior assumption about the shape of the distribution being estimated is made, their results are data dependent, and they can be applied to continuous or categorical variables. Such characteristics do not apply to OK, therefore, we focused our analysis on IK and HER. A detailed conceptual discussion comparing OK and HER is available in Thiesen et al. (2020). Although HER is considered nonparametric, two assumptions are implicit in defining the weights used for the PMF aggregation: one in linearly interpolating the optimum weights obtained for each class, and the other in defining the optimization problem (both topics are discussed in Sect. 2.2.3). An analogous interpretation of these assumptions can be applied to IK, where the weights are obtained by minimizing the variance and applied to the linear combination of the observations. The latter step is comparable to the choice of the aggregation method in HER.

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 differs for each cutoff (Goovaerts 1997 p.393), HER relies directly on the dataset to extract one distribution for each distance class (as seen in Fig. 13). 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 inference of the spatial dependence together with the aggregation procedure allows the spread 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, in this study HER and IK show comparable performance in both 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 presented, we should be mindful that there 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 cases, 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_{10}). Regardless of the chosen approach, the risk of suboptimal choices by the user remains. 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 in 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 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 be neither achieved by the IK nor the IK_{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 1998 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 for a 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 number of aggregated distributions (thus, number of neighbors) when intersecting distributions. The stronger the contribution of the AND combination (which is the case here), 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, 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 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 p.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, consequently, can be controlled by the expert according to their wishes (Journel 2003). No model, hence, no uncertainty measure, can ever be objective: the point is to accept this limitation and to 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, since no interpolation/extrapolation assumption was done for predicting continuous distributions in IK in the presence of sparse data and it is limited to the sample dataset values (Goovaerts 2009). For predicting continuous distributions, such interpolation/extrapolation assumptions introduce 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).

HER-based sequential simulation (called HERs) allowed generating 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 HERs 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 shown to be a unique tool for estimating nonparametric conditional distributions with the advantage of i) not presenting problems of order-relation deviations, ii) being free of function assumptions for interpolating probabilities or extrapolating 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 through 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 worth of further research.

Appendix 1: 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 parameters presented in Table 4.

Table 4 Parameters of OK fitted variograms as proposed by Atteia et al. (1994)

$\log_{10}$(Pb)	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 lead dataset using HER (Fig. 13) and the optimum weights obtained to be used in aggregation methods (Fig. 14).

Fig. 13a presents the raw infogram from where the class PMFs (Fig. 13b) and, consecutively, the infogram (Fig. 13c) were obtained. In Fig. 13b, 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. 13b) is numerically verified also in the infogram (red curve, Fig. 13c), which presents increasing entropy (therefore, decreasing spatial dependence or increasing spatial disorder) with distance. As shown in Fig. 13c, 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 dataset (seen as the intersect of the blue and red-dotted lines).

The number of pairs forming each Δz PMF and the optimum weights (w_{OR} and w_{AND}) obtained for Eqs. 3 and 4, respectively, are illustrated in Fig. 14. 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. 14b, 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 $\beta = 0$.

IK and IK_{10}

This section presents the parameters used in AUTO-IK program (developed by Goovaerts 2009) to calibrate the indicator kriging model (called IK) for the paper dataset. The parameter file employed is available Fig. 15 The program AUTO-IK described in Goovaerts (2009) is available on his personal website (https://sites.google.com/site/goovaertspierre/pierregoovaertswebsite/download/indicator-kriging).

Based on this IK model (Fig. 15), 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

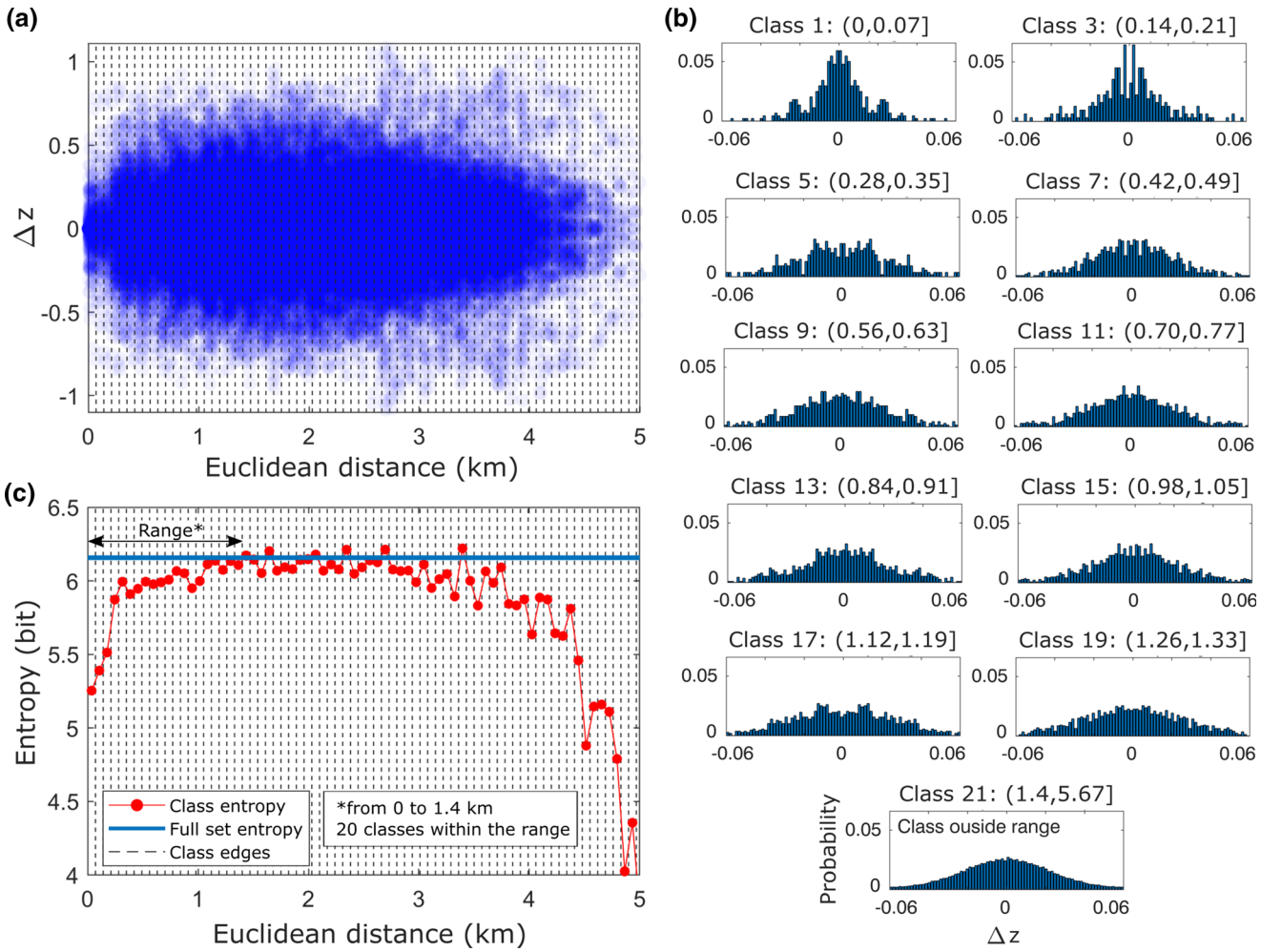

Fig. 13 Spatial characterization of the lead dataset using HER. **a** Infogram cloud, **b** Δz PMFs by class, and **c** infogram

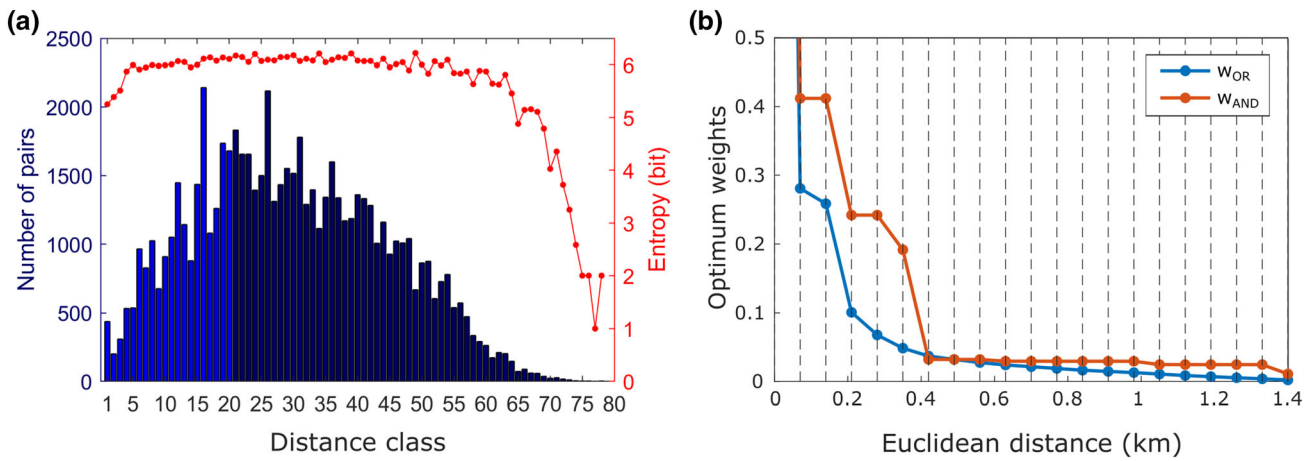

Fig. 14 HER model characteristics of the lead dataset. **a** Class cardinality and **b** optimum weights – Eqs. 3 and 4

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

```
                  Parameters for AUTO-IK
                  **********************

START OF PARAMETERS:
JuraST_calibration_logPb.dat -File with data
1 2 6                        -Column numbers for X & Y coordinates + variable under study
-9999                        -Code for missing value
4                            -Options: 1=user grid, 2=regular grid, 3=Xvalidation, 4=jackknife
JuraST_validation_logPb.dat -File with user grid or jackknife data
1 2 6                        -Column numbers for X & Y node coordinates + observations (jackknife)
121   0    0.05              -nx,xmn,xsiz
121   0    0.05              -ny,ymn,ysiz
69                           -Number of thresholds for indicator kriging
1                            -Choice of thresholds (0=automatic computation,1=user's choice)
1.29 1.305 1.32 1.335 1.35 1.365 1.38 1.395 1.41 1.425 1.44 1.455 1.47 1.485 1.50 1.515
1.53 1.545 1.56 1.575 1.59 1.605 1.62 1.635 1.65 1.665 1.68 1.695 1.699 1.71 1.725 1.74
1.755 1.77 1.785 1.80 1.815 1.83 1.845 1.86 1.875 1.89 1.905 1.92 1.935 1.95 1.965 1.98
1.995 2.01 2.025 2.04 2.055 2.07 2.085 2.10 2.115 2.13 2.145 2.16 2.175 2.19 2.205 2.22
2.235 2.25 2.265 2.28 2.295 -Values of thresholds if specified by the user
0                            -IK options: 0=full IK, 1=median IK
1                            -Kriging types: 0=simple kriging, 1=ordinary kriging
30 .07                       -Number of lags + lag spacing for variogram computation
1 22.5                       -Number of directions (ndir=1 or 4) + 1st azimuth for ndir=4
2                            -Weights for semivariogram modeling
8 32 2.0                     -Minimum & maximum number of observations + search radius
Pblog10_69thresh-variog.txt -Output file for semivariogram values + models
Pblog10_69thresh-IK.out      -Output file for probability estimates(GEO-EAS format)
Pblog10_69thresh-stat.out    -Output file for Ccdf statistics (GEO-EAS format)

Weights option for semivariogram modeling:
1 => constant weight
2 => weight = (Number of data pairs)^0.5/gamma
3 => weight = 1/gamma^2
4 => weight = Number of data pairs
5 => weight = Number of data pairs/log(lag distance)
```

Fig. 15 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

recommend between 8 and 15 cutoff values. Thus, due to its 10 cutoff values, this model is called IK_{10}.

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. 13), same cardinality and similar w_{OR} and w_{AND} (as in Fig. 14), $\alpha = 0.55$ (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 α contribution), the influence in the number of neighbors plays an important role when intersecting distributions and, therefore, we reduced it to seven in HERs. 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

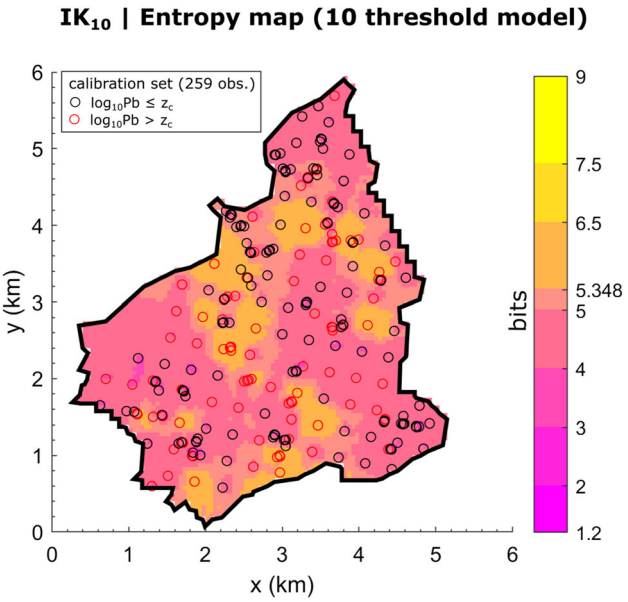

Fig. 16 Entropy map. Local uncertainty in terms of entropy for IK_{10}

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 1000 times, and taking the mean of these repetitions.

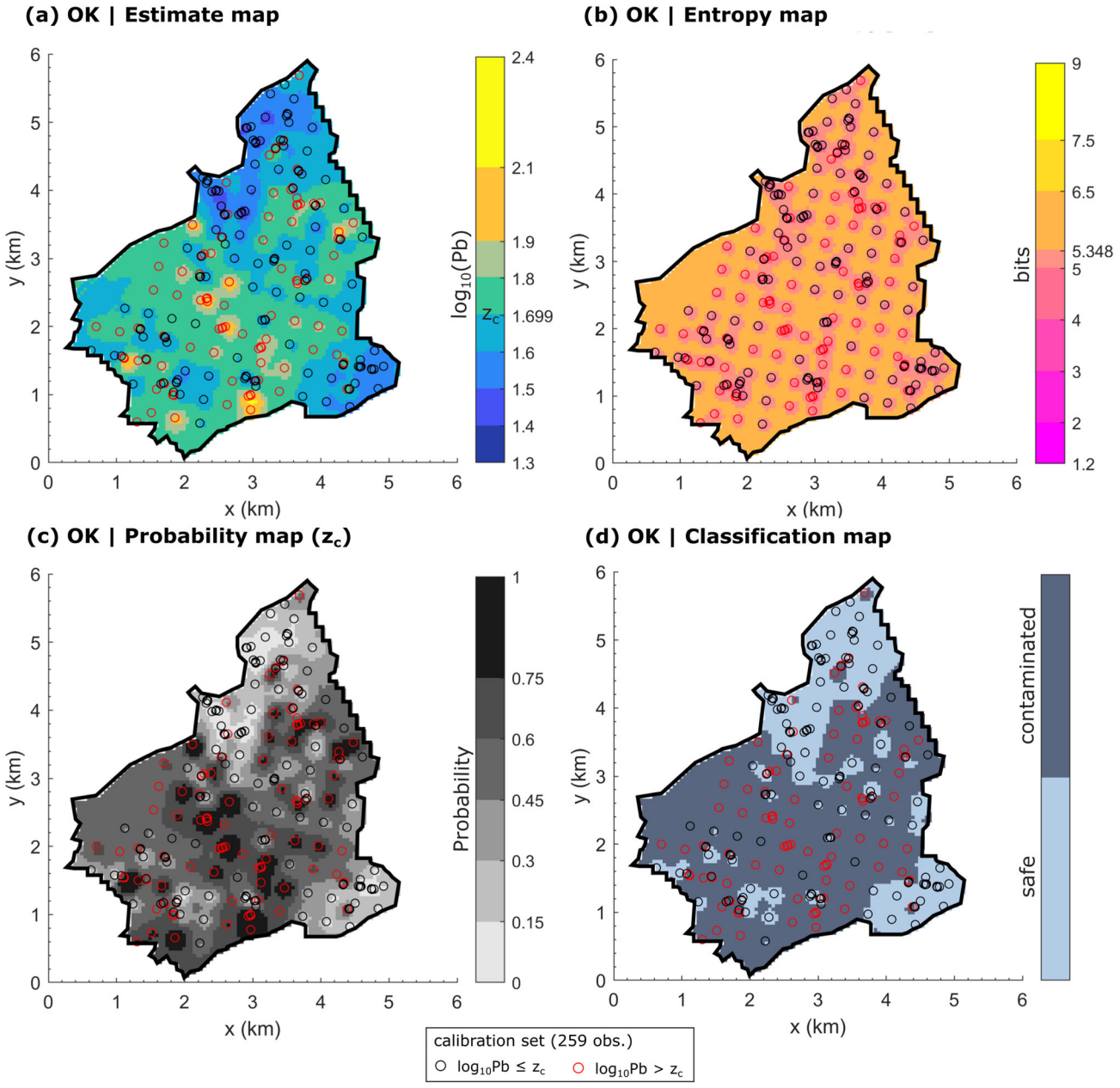

Fig. 17 OK maps for $\log_{10}(\text{Pb})$ dataset. **a** Estimates, **b** local uncertainty in terms of entropy, **c** probability of exceeding the critical threshold $(z_c = 1.699)$, and **d** classification of locations as contaminated by lead on the basis that the probability of exceeding the critical threshold z_c is larger than the marginal probability of contamination (0.421)

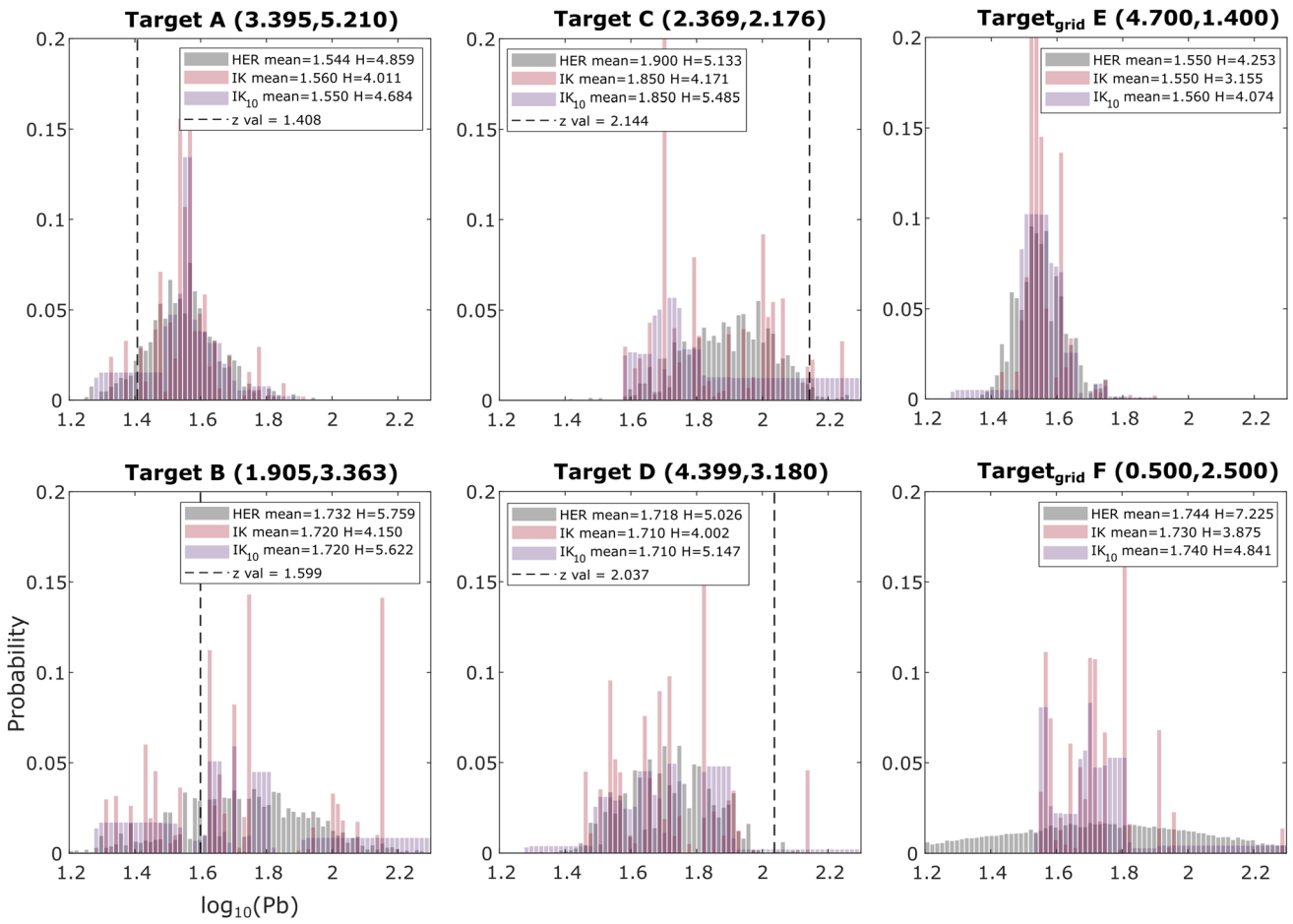

Fig. 18 Local distribution of targets of the validation set (targets A to D) and grid (targets E and F) for HER (gray), IK (red), and IK_{10} (purple)

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 is not straightforward since the obtained value is 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 verified 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 2: Extra results

This section consolidates extra results for the local uncertainty of OK, IK, IK_{10} and HER models. Fig. 16 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.

Fig. 17 displays the local results for the OK model, including estimation, entropy, probability, and

Table 5 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).

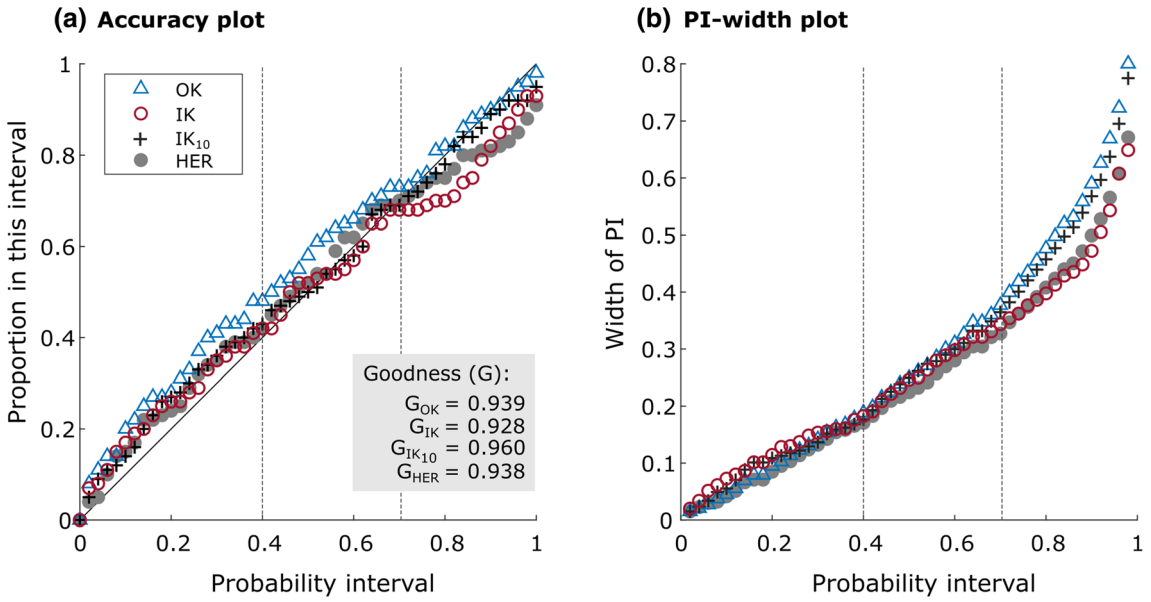

Fig. 19 OK, IK, IK_{10}, and HER performance. **a** Proportion of the true lead values falling within the probability intervals (*p*-PI) of increasing sizes and **b** the width of these intervals versus *p*-PI

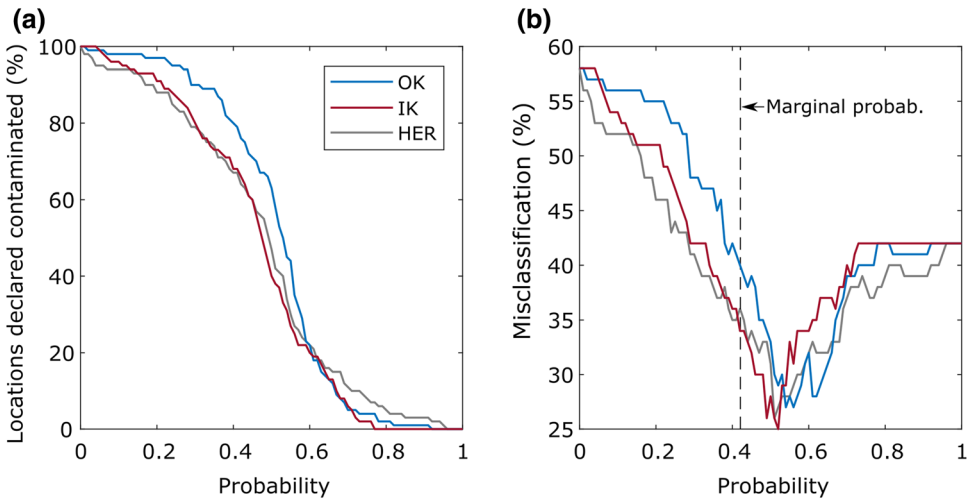

Fig. 20 Proportion of validation locations **a** that are declared contaminated with respect to lead concentration and **b** that are wrongly classified for OK, IK, and HER models

classification maps. Similar to Goovaerts (1997 p.362), the estimation map of OK (Fig. 17a), which is optimal for least-square criterion, tends to overestimate the Pb concentration, leading to most locations being classified as contaminated (Fig. 17d). While the OK estimates (Fig. 17a) and E-type estimates presented in the paper (Fig. 4) are similar, their uncertainty (Figs. 17b and 5) are completely different. The map of OK entropy indicates 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).

The local distributions of IK, IK_{10}, and HER models are displayed in Fig. 18. In this image, we can relate the binfilling effect of the linear interpolation and extrapolation of the distribution assumed in IK_{10} with IK.

Table 5 (performance results) and Fig. 19 (accuracy and PI-width plots) contain information already presented in the paper, with the inclusion of IK_{10}.

The misclassification given different probability cutoffs is shown in Fig. 20. Different than expected, all lead models (OK, IK, and HER) presented the minimum misclassification occurring close to the probability of 0.5 instead of the marginal probability of 0.421 (estimated in Sect. 3.1). This could be explained by the fact that the marginal probability was calculated on the calibration set and we are analyzing the models on the validation set, or by the fact that no declustering of the calibration data was

done before calculating the marginal probability. Although, for all models, misclassification is not minimal at the marginal probability of 0.421, 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.

Acknowledgements The authors acknowledge Diego M. Vieira for the implementation of the optimization problem, mathematical formulation, as well as for the insightful discussions throughout the development of this work and the manuscript review. We also thank 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.

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 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.

Funding Open Access funding enabled and organized by Projekt DEAL. 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).

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/pierre-goovaertswebsite/download/, options 'Jura Data' and 'Automatic Indicator Kriging Program (AUTO-IK)'.

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.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

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

Allard D, Comunian A, Renard P (2012) Probability aggregation methods in geoscience. Math Geosci 44:545–581. https://doi.org/10.1007/s11004-012-9396-3

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. 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

Cover TM, Thomas JA (2006) Elements of information theory, 2nd edn. John Wiley & Sons, New Jersey

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

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

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

Deutsch CV, Journel AG (1998) GSLIB: Geostatistical Software Library and User's Guide. Oxford University Press, New York, second edition.

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/f76faa39-2b27-42f2-be1e-9332f795e324

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

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

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

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

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 (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

Journel AG, Xu W (1994) Posterior identification of histograms conditional to local data. Math Geol 26:. https://doi.org/10.1007/BF02089228

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

Leuangthong O, McLennan JA, Deutsch CV (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

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 C V (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

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

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/10.5194/hess-23-1015-2019

Thiesen S, Vieira DM, Mälicke M, 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/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