HER: an information theoretic alternative for geostatistics

Stephanie Thiesen[1], Diego M. Vieira[2,3], Mirko Mälicke[1], Ralf Loritz[1], J. Florian Wellmann[4], Uwe Ehret[1]

[1]Institute of Water Resources and River Basin Management, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Department for Microsystems Engineering, University of Freiburg, Freiburg, Germany
[3]Bernstein Center Freiburg, University of Freiburg, Freiburg, Germany
[4]Computational Geosciences and Reservoir Engineering, RWTH Aachen University, Aachen, Germany

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

Abstract. Interpolation of spatial data has been regarded in many different forms, varying from deterministic to stochastic, parametric to non-parametric, and purely data-driven to geostatistical methods. In this study, we propose a non-parametric interpolator which combines information theory with probability aggregation methods in a geostatistical framework for stochastic estimation of unsampled points. Histogram via entropy reduction (HER) predicts conditional distributions based on empirical probabilities, relaxing parametrizations and therefore avoiding the risk of adding information not present in data. By construction, it provides a proper framework for uncertainty estimation, since it accounts for both spatial configuration and data values, while allowing to introduce or infer properties of the field through the aggregation method. We investigate the framework using synthetically generated datasets and demonstrate its efficacy in ascertaining the underlying field with varying sample densities and data properties. HER shows comparable performance to popular benchmark models with the additional advantage of higher generality. The novel method brings a new perspective of spatial interpolation and uncertainty analysis to geostatistics and statistical learning, using the lens of information theory.

1 Introduction

Spatial interpolation methods are useful tools for filling gaps in data. Since information of natural phenomena is often collected by point sampling, interpolation techniques are essential and required for obtaining spatially continuous data over the region of interest (Li and Heap, 2014). There is a broad range of methods available that have been considered in many different forms, from simple approaches such as nearest neighbor (NN, Fix and Hodges, 1951) and inverse distance weighting (IDW, Shepard, 1968) to geostatistical and, more recently, machine learning methods.

Geostatistical, stochastic approaches, such as ordinary kriging (OK), have been widely studied and applied in various disciplines since their introduction to geology and mining by Krige (1951), bringing significant results in the context of environmental sciences. However, like other parametric regression methods, it relies on prior assumptions about theoretical functions and, therefore, includes the risk of sub-optimal performance due to sub-optimal user choices (Yakowitz and Szidarovszky, 1985). OK uses fitted functions to offer uncertainty estimates, while deterministic estimators (NN and IDW) avoid function parametrizations at the cost of neglecting uncertainty analysis. In this sense, researchers are confronted with the trade-off between avoiding parametrization assumptions and obtaining uncertainty results (stochastic predictions).

More recently, with the increasing availability of data volume and computer power (Bell et al., 2009), machine learning methods (here referred to 'data-driven' methods) have become increasingly popular as a substitute or complement to established modeling approaches. In the context of data-based modeling in the environmental sciences, concepts and measures from information theory are being used for describing and inferring relations among data (Liu et al., 2016; Thiesen et al., 2019; Mälicke et al., 2020), quantifying uncertainty and evaluating model performance (Chapman, 1986; Liu et al., 2016; Thiesen et al., 2019), estimating information flow (Weijs, 2011; Darscheid, 2017), and measuring similarity, quantity and quality of information in hydrological models (Nearing and Gupta, 2017; Loritz et al. 2018; Loritz et al. 2019). In the spatial context, information-theoretic measures were used to obtain longitudinal profiles of rivers (Leopold and Langbein, 1962), to solve problems of spatial aggregation and quantify information gain, loss and redundancy (Batty, 1974; Singh, 2013), to analyze spatiotemporal variability (Mishra et al., 2009; Brunsell, 2010), to address risk of landslides (Roodposhti et al., 2016), and to assess spatial dissimilarity (Naimi, 2015), complexity (Pham, 2010), uncertainty (Wellmann, 2013), and heterogeneity (Bianchi and Pedretti, 2018).

Most of the popular data-driven methods have been developed in the computational intelligence community and, since they are not built for solving particular problems, applying these methods remains a challenge for the researchers outside this field (Solomatine and Ostfeld, 2008). The main challenges for researchers in hydroinformatics to apply data-driven methods lie in testing various combinations of methods for particular problems, combining them with optimization techniques, developing robust modeling procedures able to work with noisy data, and providing the adequate model uncertainty estimates (Solomatine and Ostfeld, 2008). To overcome these challenges and the mentioned parametrization-uncertainty tradeoff in the context of spatial interpolation, this paper is concerned with formulating and testing a novel method based on principles of geostatistics, information theory and probability aggregation methods to describe spatial patterns and to obtain stochastic predictions. In order to avoid fitting of spatial correlation functions and assumptions about the underlying distribution of the data, it relies on empirical probability distributions to i) extract the spatial dependence structure of the field; ii) minimize entropy of predictions; and iii) produce stochastic estimation of unsampled points. Thus, the proposed histogram via entropy reduction (HER) approach allows non-parametric and stochastic predictions, avoiding the shortcomings of fitting deterministic curves and therefore the risk of adding information not contained in data, but still relying on geostatistical concepts. HER is seen as in-between geostatistics (knowledge-driven) and statistical learning (data-driven) in the sense that it allows automated learning from data, bounded by a geostatistical framework.

Our experimental results show that the proposed method is flexible for combining distributions in different ways and presents comparable performance to ordinary kriging for various sample sizes and field properties (short and long range, with and without noise). Furthermore, we show that its potential goes beyond prediction, since, by construction, HER allows inferring or introducing physical properties (continuity or discontinuity characteristics) of a field under study, and provides a proper framework for uncertainty prediction, which takes into account not only the spatial configuration but also the data values.

The paper is organized as follows. The method is presented in Sect. 2. In Sect. 3, we describe the data properties, performance parameters, validation design and benchmark models. In Sect. 4, we explore the properties of three different aggregation methods, present the results of HER for different samples sizes and data types, compare the results to benchmark models, and, in the end, discuss the achieved outcomes and model contributions. Finally, we draw conclusions in Sect. 5.

2 Method description

Histogram via entropy reduction method (HER) has three main steps: i) characterization of the spatial correlation; ii) selection of aggregation method and optimal weights via entropy minimization; and iii) prediction of the target probability distribution (which uses the two first steps to interpolate conditional distributions for the unsampled targets). The first and third steps are shown in Fig. 1.

In the following sections, we start with a brief introduction of information theoretic measures employed in the method, and then detail all the three method steps.

2.1 Information theory

Information theory provides a framework for measuring information and quantifying uncertainty. In order to extract the spatial correlation structure from observations and to minimize the uncertainties of predictions, two information theoretic measures are used in HER and will be described here: Shannon entropy and Kullback-Leibler divergence. We recommend Cover and Thomas (2006) for further reference.

The entropy of a probability distribution measures the average uncertainty in a random variable. The measure, first derived by Shannon (1948) is additive for independent events (Batty, 1974). The formula of Shannon entropy (H) for a discrete random variable X with a probability $p(x)$, $x \in \chi$, is defined by

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

We use the logarithm to base two so that the entropy is expressed in unit bits. Each bit corresponds to an answer to one optimal yes/no question asked with the intention of reconstructing the data. It varies from zero to $\log_2 n$, where n represents the number of bins of the discrete distribution. In the study, Shannon entropy is used to extract the infogram and correlation length of the dataset (explored in Sect. 2.2).

Besides quantifying the uncertainty of a distribution, it is also possible to compare similarities between two probability distributions p and q using the Kullback-Leibler divergence (D_{KL}). Comparable to the expected logarithm of the likelihood ratio (Cover and Thomas, 2006; Allard et al., 2012), the Kullback-Leibler divergence quantifies the statistical 'distance' between two probability mass functions p and q using the following equation

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

Also referred to as relative entropy, D_{KL} can be understood as a measure of information loss of assuming that the distribution is q when in reality it is p (Weijs et al., 2010). It is nonnegative and is zero strictly if $p = q$. In HER context, Kullback-Leibler divergence is optimized to select the weights for aggregating distributions (detailed in Sect. 2.3). The measure is also used as a scoring rule for performance verification of probabilistic predictions (Gneiting and Raftery, 2007; Weijs et al., 2010). Note that the measures presented by Eqs. (1) and (2) are defined as functionals of probability distributions, not depending on the variable X value or its unit. This is favorable, as it allows joint treatment of many different sources and sorts of data in a single framework.

2.2 Spatial characterization

The spatial characterization (Fig. 1a) is the first step of HER. It consists of quantifying the spatial information available in data and of using it to infer its spatial correlation structure. For capturing the spatial variability and related uncertainties, concepts of geostatistics and information theory are integrated into the method. As shown in Fig. 1a, the spatial characterization phase aims to obtain: Δz probability mass functions (PMFs), where z is the variable under study; the behavior of entropy as a function of lag distance (which the authors denominate 'infogram'); and, finally, the correlation length (range). These outputs are outlined in Fig. 2 and attained in the following steps:

i. Infogram cloud (Fig. 2a): Calculate the difference of the z values (Δz) between pairs of observations; associate each Δz to the Euclidean separation distance of its respective point pair. Define the lag distance (demarcated by red dashed lines), here called distance classes, or simply classes. Divide the range of Δz values into a set of bins (demarcated by horizontal gray lines).

ii. Δz PMFs (Fig. 2b): For each distance class, construct the Δz PMF from the Δz values inside the class (conditional PMFs). Also construct the Δz PMF from all data in the dataset (unconditional PMF).

iii. Infogram (Fig. 2c): Calculate the entropy of each Δz PMF and of the unconditional PMF. Compute the range of the data: this is the distance where the conditional entropy exceeds the unconditional entropy. Beyond this point, the neighbors start becoming un-informative, and it is pointless to use information outside this neighborhood.

The infogram cloud is a preparation to construct the infogram. It contains complete cloud of pair points. The infogram plays a role similar to that of the variogram: through the lens of information theory, we can characterize the spatial dependence of the dataset, calculate the spatial (dis)similarities, and compute its correlation length (range). It describes the statistical dispersion of pairs of observations for the distance class separating these observations. Quantitatively, it is a way of measuring the uncertainty about Δz given the class. Graphically, the infogram shape is the fingerprint of spatial dependence, where the larger the entropy of one class, the more uncertain (disperse) its distribution is. It reaches a threshold (range), where the data no longer show significant spatial correlation. We associate neighbors beyond the range with the Δz PMF of the full dataset. By doing so, we restrict ourselves to the more informative classes and reduce the number of classes to be mapped, thus improving the results and the speed of calculation. Note that in the illustrative case of Fig. 2, we limited the number of classes shown to four classes beyond the range. A complete infogram cloud and infogram is presented and discussed in the method application, Fig. 5 in Sect. 4.1.

Naimi (2015) introduced a similar concept to the infogram called entrogram, which is used for the quantification of the spatial association of both continuous and categorical variables. In the same direction, Bianchi and Pedretti (2018) employed the term entrogram for quantifying the degree of spatial order and ranking different structures. Both works, as well as the present study, are carried out with variogram-like shape, entropy-based measures, and looking for data (dis)similarity, yet with different purposes and metrics. The proposed infogram terminology seeks to provide an easy-to-follow association with the quantification of information available in the data.

Converting the frequency distributions of Δz into PMF requires a cautious choice of bin width, since this decision will frame the distributions used as model and directly influence the statistics we compute for evaluation (D_{KL}). Many methods for choosing an appropriate binning strategy have been suggested (Knuth, 2013; Gong et al., 2014; Pechlivanidis et al., 2016; Thiesen et al., 2018). These approaches are either founded on a general physical understanding and relate, for instance, measurement uncertainties to the binning width (Loritz et. al., 2018) or are exclusively based on statistical considerations of the underlying field properties (Scott, 1979). Regardless of which approach is chosen, the choice of bin width should be communicated in a clear manner to make results as reproducible as possible. Throughout this paper, we will stick to equidistant bins, since they have the advantage of being simple, computationally efficient (Ruddell and Kumar, 2009), and introduce minimal prior information (Knuth, 2013). The bin size was defined based on Thiesen et al. (2018), by comparing the cross entropy ($H_{pq} = H(p) + \mathrm{D}_{\mathrm{KL}}(p||q)$) between the full learning set and subsamples for various bin widths. The selected one shows a stabilization of the cross entropy for small sample sizes, meaning that the bin size is reasonable for small and large sample sizes and analyzed distribution shapes. For favoring comparability, the bins are kept the same for all applications and performance calculations.

Additionally, to avoid distributions with empty bins, which might make the PMF combination (discussed in Sect. 2.3.1) unfeasible, we assigned a small probability equivalent to the probability of a single pair-point count to all bins in the histogram after converting it to a PMF by normalization. This procedure does not affect the results when the sample size is large enough (Darscheid et al., 2018), and it was inspected by result and cross-entropy comparison (as described in the previous paragraph). It also guarantees that there is always an intersection when aggregating PMFs, and that we obtain a uniform distribution (maximum-entropy) in case we multiply distributions where the overlap happens uniquely on the previously empty bins. Furthermore, as shown in the Darscheid et al. (2018) study, for the cases where no distribution is known a priori, adding one counter to each empty bin performed well across different distributions.

Altogether, the spatial characterization stage provides a way of inferring conditional distributions of the target given its observed neighbors without the need, for example, of fitting a theoretical correlation function. In the next section, we describe how these distributions can be jointly used to estimate unknown points and how to weight them when doing so.

2.3 Minimization of estimation entropy

For inferring the conditional distribution of the target z_0 (unsampled point) given its neighbors z_i (where $i = 1, \ldots, n$ are the indices of the sampled points), we use the Δz PMFs obtained at the spatial characterization step (Sect. 2.2). To do so, each neighbor z_i is associated to a class, and hence to a Δz distribution, according to their distance to the target z_0. This implies the assumption that the empirical Δz PMFs apply everywhere in the field, irrespective of specific location, and only depend on the distance between points. Each Δz PMF is then shifted by the z_i value of the observation it is associated with, yielding the z PMF of the target given the neighbor i, denoted by $p(z_0|z_\mathrm{i})$. Assume for instance three observations z_1, z_2, z_3 from the field and that we want to predict the probability distribution of the target z_0. In this case, what we infer at this stage are the conditional probability distributions $p(z_0|z_1)$, $p(z_0|z_2)$, and $p(z_0|z_3)$.

Now, since we are in fact interested in the probability distribution of the target conditioned to multiple observations $p(z_0|z_1, z_2, z_3)$, how can we optimally combine the information gained from individual observations to predict this target probability? In the next sections, we address this issue by using aggregation methods. After introducing potential ways to combine PMFs (Sect. 2.3.1), we propose an optimization problem via entropy minimization for defining the weight parameters needed for the aggregation (Sect. 2.3.2).

2.3.1 Combining distributions

The problem of combining multiple conditional probability distributions into a single one is treated here by using aggregation methods. This subsection is based on the work by Allard et al. (2012), which we recommend as a summary of existing aggregation methods (also called opinion pools), with a focus on their mathematical properties.

The main objective of this process is to aggregate probability distributions coming from different sources into a global probability distribution. For this purpose, the computation of the full conditional probability $p(z_0|z_1, \ldots, z_\mathrm{n})$ – where z_0 is the event we are interested in (target) and z_i, $i = 1, \ldots, n$ is a set of data events (or neighbors) – is obtained by the use of an aggregation operator P_G, called pooling operator, such that

$$p(z_0|z_1, \ldots, z_\mathrm{n}) \approx P_\mathrm{G}\big(p(z_0|z_1), \ldots, p(z_0|z_\mathrm{n})\big). \quad (3)$$

From now on, we will adopt a similar notation to that of Allard et al. (2012), using the more concise expressions $P_\mathrm{i}(z_0)$ to denote $p(z_0|z_\mathrm{i})$ and $P_\mathrm{G}(z_0)$ for the global probability $P_\mathrm{G}\big(P_1(z_0), \ldots, P_\mathrm{n}(z_0)\big)$.

The most intuitive way of aggregating the probabilities $p_1, \ldots, p_\mathrm{n}$ is by linear pooling, which is defined as

$$P_{\mathrm{G_{OR}}}(z_0) = \sum_{i=1}^{n} w_{\mathrm{OR_i}} P_{\mathrm{i}}(z_0), \tag{4}$$

where n is the number of neighbors, and $w_{\mathrm{OR_i}}$ are positive weights verifying $\sum_{i=1}^{n} w_{\mathrm{OR_i}} = 1$. Eq. (4) describes mixture models in which each probability p_{i} represents a different population. If we set equal weights $w_{\mathrm{OR_i}}$ to every probability P_{i}, the method reduces to an arithmetic average, coinciding with the disjunction of probabilities proposed by Tarantola and Valette (1982) and Tarantola (2005), illustrated in Fig. 3b. Since it is a way of averaging distributions, the resulting distribution $P_{\mathrm{G_{OR}}}$ is often multi-modal. Additive methods, such as linear pooling, are related to union of events and to the logical operator OR.

Multiplication of probabilities, in turn, is described by the logical operator AND, and it is associated to the intersection of events. One aggregation method based on the multiplication of probabilities is the log-linear pooling operator, defined by

$$\ln P_{\mathrm{G_{AND}}}(z_0) = \ln \zeta + \sum_{i=1}^{n} w_{\mathrm{AND_i}} \ln P_{\mathrm{i}}(z_0), \tag{5}$$

or equivalently $P_{\mathrm{G_{AND}}}(z_0) \propto \prod_{i=1}^{n} P_i(z_0)^{w_{\mathrm{AND_i}}}$, where ζ is a normalizing constant, n is the number of neighbors, and $w_{\mathrm{AND_i}}$ are positive weights. One particular case consists of setting $w_{\mathrm{AND_i}} = 1$ for every i. This refers to the conjunction of probabilities proposed by Tarantola and Valette (1982) and Tarantola (2005), shown in Fig. 3c. In contrast to linear pooling, log-linear pooling is typically unimodal and less dispersed.

Aggregation methods are not limited to log-linear and linear pooling presented here. However, the selection of these two different approaches to PMF aggregation seeks to embrace distinct physical characteristics of the field. The authors naturally associate the intersection of distributions (AND combination, Eq. (5)) to fields with continuous properties. This idea is supported by Journel (2002) when remarking that a logarithmic expression evokes the simple kriging expression (used for continuous variables). For example, if we have two points z_1 and z_2 with different values and want to estimate the target z_0 at a location between them in a continuous field, we would expect that the estimate z_0 would be somewhere between z_1 and z_2, which can be achieved by an AND combination. In a more intuitive way, if we notice that, for kriging, the shape of the predicted distribution is assumed to be fixed (Gaussian, for example), multiplying two distributions with different means would result in a Gaussian distribution as well, less dispersed than the original ones, as also seen for the log-linear pooling. It is worth mentioning that some methods for modeling spatially dependent data such as Copulas (Bárdossy, 2006; Kazianka and Pilz, 2010) and Effective Distribution Models (Hristopulos and Baxevani, 2020) also use log-linear pooling for constructing conditional distributions.

On the other hand, Krishnan (2008) pointed out that the linear combination, given by linear pooling, identifies a dual indicator kriging estimator (kriging used for categorical variables), which we see as an appropriate method for fields with discontinuous properties. Along the same lines, Goovaerts (1997, p.420) defended that phenomena that show abrupt changes should be modeled as mixture of populations. In this case, if we have two points z_1 and z_2 belonging to different categories, a target z_0 between them will either belong to the category of z_1 or z_2, which can be achieved by the mixture distribution given by the OR pooling. In other words, the OR aggregation is a way of combining information from different sides of the truth, thus, a conservative way of considering the available information from all sources.

Note that, for both linear and log-linear pooling, weights equal to zero will lead to uniform distributions, therefore bypassing the PMF in question. Conveniently, the uniform distribution is the maximum entropy distribution among all discrete distributions with the same finite support. A practical example of the pooling operators is illustrated in the end of this section.

The selection of the most suitable aggregation method depends on the specific problem (Allard et al., 2012), and it will influence the PMF prediction and, therefore, the uncertainty structure of the field. Thus, depending on the knowledge about the field, a user can either add information to the model by applying an a-priori chosen aggregation method or infer these properties from the field. Since, in practice, there is often a lack of information to accurately describe the interactions between the sources of information (Allard et al., 2012), inference is the approach we tested in the comparison analysis (Sect. 4.2). For that, we propose to estimate the distribution P_{G} of a target, by combining $\mathrm{P}_{\mathrm{G}_{\mathrm{AND}}}$ and $P_{\mathrm{G}_{\mathrm{OR}}}$, such that

$$P_{\mathrm{G}}(z_0) \propto P_{\mathrm{G}_{\mathrm{AND}}}(z_0)^{\alpha} \, P_{\mathrm{G}_{\mathrm{OR}}}(z_0)^{\beta}, \tag{6}$$

where α and β are positive weights varying from 0 to 1, which will be found by optimization. Eq. (6) is the choice made by the authors as a way of balancing both natures of PMF aggregation. The idea is to find the appropriate proportion of α (continuous) and β (discontinuous) properties of the field by minimizing relative estimation entropy. Note that, when the weight α or β is set to zero, the final distribution results respectively in a pure OR, Eq. (4), or pure AND aggregation, Eq. (5), as special cases. The equation is based on the log-linear aggregation, as opposed to linear aggregation, since the latter is often multi-modal, which is an undesired property for geoscience applications (Allard et al., 2012). Alternatively, Eqs. (4) or (5) or a linear polling of $P_{\mathrm{G}_{\mathrm{AND}}}(z_0)$ and $P_{\mathrm{G}_{\mathrm{OR}}}(z_0)$ could be used. We explore the properties of the linear and log-linear pooling in Sect. 4.1.

The practical differences between the pooling operators used in this paper are illustrated in Fig. 3, where Fig. 3a introduces two PMFs to be combined, and Figs. 3b,c,d show the resulting PMFs for Eqs (4), (5), and (6), respectively. In Fig. 3b, we use equal weights to both PMFs, and the resulting distribution is the arithmetic average of the bin probabilities. In Fig. 3c, we use unitary PMF weights so that the multiplication of the bins (AND aggregation) leads to a simple intersection of PMFs weighted by the bin height. Fig. 3d shows a log-linear aggregation of the two previous distributions (Figs. 3b,c). In all three cases, if the weight of one distribution is set to one and the other is set to zero (not shown), the resulting PMF would be equal to the distribution which receives all the weight.

The following section addresses the optimization problem for estimating the weights of the aggregation methods.

2.3.2 Weighting PMFs

Scoring rules assess the quality of probabilistic estimations (Gneiting and Raftery, 2007) and, therefore, can be used for estimating the parameters of a pooling operator (Allard et al., 2012). We select the Kullback-Leibler divergence (D_{KL}, Eq. (2)) as loss function for optimizing α and β, Eq. (6), as well as the $w_{\mathrm{OR_k}}$ and $w_{\mathrm{AND_k}}$ weights (Eqs. (4) and (5), respectively), here generalized as w_{k}. The logarithmic score proposed by Good (1952), associated to Kullback-Leibler divergence by Gneiting and Raftery (2007), and reintroduced from an information-theoretical point of view by Roulston and Smith (2002) is a strictly proper scoring rule since it provides summary metrics that address calibration and sharpness simultaneously by rewarding narrow prediction intervals and penalizing intervals missed by the observation (Gneiting and Raftery, 2007).

By means of leave-one-out cross-validation (LOOCV), the optimization problem is then defined in order to find the set of weights which minimizes the expected relative entropy (D_{KL}) of all targets. The idea is to choose weights such that the disagreement of the 'true' distribution (or observation value when no distribution is available) and estimated distribution is minimized. Note that the optimization goal can be tailored for different purposes, e.g., by binarizing the probability distribution (observed and estimated) with respect to a threshold in risk analysis problems or categorical data. In Eqs. (4) and (5), we assign one weight for each distance class k. This means that, given a target z_0, the neighbors grouped in the same distance class will be assigned the same weight. For a more continuous weighting of the neighbors, as an extra step, we linearly interpolate the weights according to the Euclidean distance and the weight of the next class. Another option could be narrowing down the class width, in which case more data are needed to estimate the respective PMFs.

Firstly, we obtained in parallel the weights of Eqs. (4) and (5) by convex optimization, and later α and β by grid search with both weight values ranging from 0 to 1 (steps of 0.05 were used in the application case). In order to facilitate the convergence of the convex optimization, the following constraints were employed: i) for linear pooling, set $w_{\mathrm{OR_1}} = 1$, to avoid non-unique solutions; ii) force weights to decrease monotonically (i.e., $w_{\mathrm{k+1}} \leq w_{\mathrm{k}}$); iii) define a lower bound to avoid numerical instabilities (e.g., $w_{\mathrm{k}} \geq 10^{-6}$); iv) define an upper bound ($w_{\mathrm{k}} \leq 1$). Finally, after the optimization, normalize the weights to verify $\sum_k w_{\mathrm{OR_k}} = 1$ for linear pooling (for log-linear pooling, the resulting PMFs are normalized).

In order to increase computational efficiency, and due to the minor contribution of neighbors in classes far away from the target, the authors only used the twelve neighbors closest to the target when optimizing α and β and when predicting the target. Note that this procedure is not applicable for the optimization of the $w_{\mathrm{OR_k}}$ and $w_{\mathrm{AND_k}}$ weights, since we are looking for one weight w_{k} for each class k, and therefore we cannot risk neglecting classes whose weights we have an interest in. For the optimization phase discussed here, as well as for the prediction phase (next section topic), the limitation of number of neighbors together with the removal of classes beyond range are efficient means of reducing the computational effort involved in both phases.

2.4 Prediction

With the results of the spatial characterization step (classes, Δz PMFs, and range, as described in Sect. 2.2), the definition of the aggregation method and its parameters (Sects. 2.3.1 and 2.3.2, respectively), and the set of known observations, we have the model available for predicting distributions.

Thus, for estimating a specific unsampled point (target), first, we calculate the Euclidean distance from the target to its neighbors (sampled observations). Based on this distance, we obtain the class of each neighbor, and associate to each its corresponding Δz PMF. As mentioned in Sect. 2.2, neighbors beyond the range are associated with the Δz PMF of the full dataset. To obtain the z PMF of target z_0 given each neighbor z_{i}, we simply shift the Δz PMF of each neighbor by its z_{i} value. Finally, by applying the defined aggregation method, we combine the individual z PMFs of the target given each neighbor to obtain the PMF of the target conditional on all neighbors. Fig. 1b presents the z PMF prediction steps a single target.

3 Testing HER

For the purpose of benchmarking, this section presents the data used for testing the method, establishes the performance metrics, and introduces the calibration and test design. Additionally, we briefly present the benchmark interpolators used for the comparison analysis and some peculiarities of the calibration procedure.

3.1 Data properties

To test the proposed method in a controlled environment, four synthetic 2D spatial datasets with grid size 100x100 were generated from known Gaussian processes. A Gaussian process is a stochastic method that is specified by its mean and a covariance function or kernel (Rasmussen and Williams, 2006). The data points are determined by a given realization of a prior, which is randomly generated from the chosen kernel function and associated parameters. In this work, we used rational quadratic kernel (Pedregosa et al., 2011) as the covariance function, with two different correlation length parameters for the kernel, namely 6 and 18 units, to produce two datasets with fundamentally different spatial dependence. For both, short- and long-range fields, a white noise was introduced given by Gaussian distribution with mean 0 and standard deviation equal to 0.5. The implementation was taken from the Python library scikit-learn (Pedregosa et al., 2011). The generated sets comprise: i) a short-range field without noise (SR0); ii) a short-range field with noise (SR1); iii) a long-range field without noise (LR0); and iv) a long-range field with noise (LR1). Fig. 4 presents the field characteristics and their summary statistics. The summary statistics of each field type is included in Supplement S1.

3.2 Performance criteria

To evaluate the predictive power of the models, a quality assessment was carried out with three criteria: mean absolute error (E_{MA}) and Nash–Sutcliffe efficiency (E_{NS}), for the deterministic cases, and mean of the Kullback-Leibler divergence (D_{KL}), for the probabilistic cases. E_{MA} was selected because it gives the same weight to all errors, while E_{NS} penalizes variance as it gives more weight to errors with larger absolute values. E_{NS} also shows a normalized metric (limited to 1) which favors general comparison. All three metrics are shown in Eqs. (7), (8) and (2), respectively. The validity of the model can be asserted when the mean error is close to zero, Nash–Sutcliffe efficiency is close to one, and mean of Kullback-Leibler divergence is close to zero. The deterministic performance coefficients are defined as

$$\mathrm{E_{MA}} = \frac{1}{n}\sum_{i=1}^{n} |\hat{z}_\mathrm{i} - z_\mathrm{i}|, \tag{7}$$

$$\mathrm{E_{NS}} = 1 - \frac{\sum_{i=1}^{n}(\hat{z}_\mathrm{i} - z_\mathrm{i})^2}{\sum_{i=1}^{n}(z_\mathrm{i} - \bar{\mathrm{z}})^2}, \tag{8}$$

where $\hat{z}_\mathrm{i}$ and z_i are, respectively, the predicted and observed values at the ith location, $\bar{\mathrm{z}}$ is the mean of the observations, and n is the number of tested locations. For the probabilistic methods, $\hat{z}_\mathrm{i}$ is the expected value of the predictions.

For the applications in the study, we considered that there is no true distribution (ground truth) available for the observations in all field types. Thus, the D_{KL} scoring rule was calculated by comparing the filling of the single bin where the observed value is located, i.e., in Eq. (2), we set p equal to one for the corresponding bin and compare it to the probability of the same bin in the predicted distribution. This procedure is just applicable to probabilistic models, and it enables to measure how confident the model is in predicting the correct observation. In order to calculate this metric for ordinary kriging, we must convert the predicted PDFs (probability density functions) to PMFs employing the same bins used in HER.

3.3 Calibration and test design

To benchmark and to investigate the effect of sample size, we applied holdout validation as follows. Firstly, we randomly shuffled the data, and then divided it in three mutually exclusive sets: one to generate the learning subsets (containing up to 2000 data points), one for validation (containing 2000 data points), and another 2000 data points (20% of the full dataset) used as test set. We calibrated the models on learning subsets with increasing sizes of 200, 400, 600, 800, 1000, 1500, and 2000 observations. We used the validation set for fine adjustments and plausibility checks. For avoiding multiple calibration runs, the resampling was designed in a way that the learning subsets increased in size by adding new data to the previous subset, i.e., the observations of small sample sizes were always contained in the larger sets. To facilitate model comparison, the validation and test datasets were fixed for all performance analyses, independently of the analyzed learning set. This procedure also avoided variability of results coming from multiple random draws, since, by construction, we improved the learning with growing sample size, and we assessed the results always in the same set. The test set was kept unseen until the final application of the methods, as a 'lock box approach' (Chicco, 2017), and its results were used for evaluating the model performance presented in Sect. 4. See Supplement S1 for the summary statistics of learning, validation, and test subsets.

3.4 Benchmark interpolators

In addition to presenting a complete application of HER (Sect. 4.1), a comparison analysis among the best-known and used methods for spatial interpolation in the earth sciences (Myers, 1993; Li and Heap, 2011) is performed (Sect. 4.2). Covering deterministic, probabilistic, and geostatistical methods, three interpolators were chosen for the comparison, namely nearest neighbor (NN), inverse distance weighting (IDW), and ordinary kriging (OK).

As in HER, all these methods assume that the similarity of two-point values decrease with increasing distance. Since NN simply selects the value of the nearest sample to predict the value at an unsampled point without considering the remaining observations, it was employed as a baseline comparison. IDW, in turn, linearly combines the set of sample points for predicting the target, inversely weighting the observations according with their distance to the target. The particular case where the exponent of the weighting function equals two is the most popular choice (Li and Heap, 2008). It is known as the inverse distance squared (IDS), and it is the one applied here.

OK is more flexible than NN and IDW, since the weights are selected depending on how the correlation function varies with distance (Kitanidis, 1997, p.78). The spatial structure is extracted by the variogram, which is a mathematical description of the relationship between the variance of pairs of observations and the distance separating these observations (also known as lag). It is also described as the best linear unbiased estimator (BLUE) (Journel and Huijbregts, 1978, p.57), which aims at minimizing the error variance, and provides an indication of the uncertainty of the estimate. The authors suggest Kitanidis (1997) and Goovaerts (1997) for a more detailed explanation of variogram and OK, and Li and Heap (2008) for NN and IDW.

NN and IDS do not require calibration. To calibrate HER aggregation weights, we applied LOOCV, as described in Sect. 2.3.2, for optimizing the performance of the left-out sample in the learning set. As loss function, minimization of the mean D_{KL} was applied. After learning the model, we used the validation set for plausibility check of the calibrated model and, eventually, adjustment of parameters. Note that no function fitting is needed to apply HER.

For OK, the fitting of the model was applied in a semi-automated approach. The variogram range, sill and nugget were fitted to each of the samples taken from the four fields individually. They were selected by least squares (Branch et al., 1999). The remaining parameters, namely the semi-variance estimator, the theoretical variogram model, the minimum and the maximum number of neighbors considered during OK were jointly selected for each field type (SR and LR), since they derive from the same field characteristics. This means that for all sample sizes of SR0 and SR1 the same parameters were used, except range, sill, and nugget, which were fitted individually to each sample size. The same applies to LR0 and LR1. These parameters were chosen by expert decision, supported by result comparison for different theoretical variogram functions, validation, and LOOCV. Variogram fitting and kriging interpolation were applied using the scikit-gstat Python module (Mälicke and Schneider, 2019).

The selection of lag size has important effects on HER infogram and, as discussed in Oliver and Webster (2014), on the empirical variogram of OK. However, since the goal of the benchmarking analysis was to find a fair way to compare the methods, we fixed the lag distances of OK and HER in equal intervals of two distance units (three times smaller than the kernel correlation length of the short-range dataset).

Since all methods are instance-based learning algorithms, due to the fact that the predictions are based on the sample of observations, the learning set is stored as part of the model and used in the test phase for the performance assessment.

4 Results and discussion

In this section, three analyses are presented. Firstly, we explore the results of HER using three different aggregation methods on one specific synthetic dataset (Sect. 4.1). In Sect. 4.2, we summarize the results on synthetic datasets LR0, LR1, SR0, SR1 for all calibration sets and numerically compare HER performance with traditional interpolators. For all applications, the performance was calculated on the same test set. For brevity, the model outputs were omitted in the comparison analysis, and only the performance metrics for each dataset and interpolator are shown. Finally, Sect. 4.3 provide a theoretical discussion on the probabilistic methods (OK and HER), contrasting their different properties and assumptions.

4.1 HER application

This section presents three variants of HER, applied to the LR1 field with a calibration subset of 600 observations (LR1-600). This dataset was selected since, due to its optimized weights α and β (which reach almost the maximum value of one suggested for Eq. (6)), it favors to contrast uncertainty results of HER applying the three distinct aggregation methods proposed by Eqs. (4), (5), and (6).

As a first step, the spatial characterization of the selected field is obtained and shown in Fig. 5. For brevity, only the odd classes are shown in Fig. 5b. In the same figure, the Euclidean distance (in grid units) relative to the class is indicated after the class name in interval notation (left-open, right-closed interval). For both z PMFs and Δz PMFs, a bin width of 0.2 (10% of the distance class width) was selected and kept the same for all applications and performance calculations. As mentioned in Sect. 3.4, we fixed the lag distances in equal intervals of two distance units.

Based on the infogram cloud (Fig. 5a), the Δz PMFs for all classes were obtained. Subsequently, 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 in Fig. 5b). This occurs at class 23, corresponding to a Euclidean distance of 44 grid units. In Fig. 5c, it is also possible to notice a steep reduction in entropy (red curve) for furthest classes due to the reduced number of pairs composing the Δz PMFs. A similar behavior is also typically found in experimental variograms (not shown).

The number of pairs forming each Δz PMF and the optimum weights obtained for Eqs. (4) and (5) are presented in Fig. 6. Fig. 6a shows the number of pairs which compose the Δz PMF by class, where the first class has just under 500 pairs and the last class inside the range (light blue) has almost 10 000 pairs. About 40% of the pairs (142 512 out of 359 400 pairs) are inside the range. We obtained the weight of each class by convex optimization as described in Sect. 2.3.2. The dots in Fig. 6b represent the optimized weights of each class. As expected, the weights reflect the decreasing spatial dependence of variable z with distance. Regardless of the aggregation method, LR1-600 models are highly influenced by neighbors up to a distance of 10 grid units (distance class 5). For estimating z PMFs of target points, three different methods were tested:

i. Model 1: AND/OR combination, proposed by Eq. (6), where LR1-600 weights resulted in $\alpha = 1$ and $\beta = 0.95$;
ii. Model 2: pure AND combination, given by Eq. (5);
iii. Model 3: pure OR combination, given by Eq. (4).

The model results are summarized in Table 1 and illustrated in Fig. 7, where the first column of the panel refers to the AND/OR combination, the second column to the pure AND combination, and the third column to the pure OR combination. To assist in visually checking the heterogeneity of z, the calibration set representation is scaled by its z value, with the size of the cross increasing with z. For the target identification, we use its grid coordinates (x,y).

Fig. 7a shows the E-type estimate[a] of z (expected z obtained from the predicted z PMF) for the three analyzed models. Neither qualitatively (Fig. 7a) nor quantitatively (Table 1) is it possible to distinguish the three models based on their E-type estimate or its summary statistics. Deterministic performance metrics (E_{MA} and E_{NS}, Table 1) are also similar among the three models. However, in probabilistic terms, the representation given by the entropy map (Fig. 7b, which shows the Shannon entropy of the predicted z PMFs), the statistics of predicted z PMFs, and the D_{KL} performance (Table 1) reveal differences.

By construction, HER takes into account not only the spatial configuration of data but also the data values. In this fashion, targets close to known observations will not necessarily lead to reduced predictive uncertainty (or vice-versa). This is, e.g., the case of targets A (10,42) and B (25,63). Target B (25,63) is located in between two sampled points in a heterogeneous region (small and large z values, both in the first distance class), and presents distributions with bimodal shape and higher uncertainty (Fig. 7c) especially for model 3 (4.68 bits). For the more assertive models (1 and 2), the distributions of target B (25,63) has lower uncertainty (3.42 and 3.52 bits, respectively). It shows some peaks, due to small bumps in the PMF neighbors (not shown) which are boosted by the $w_{\mathrm{AND_k}}$ exponents in Eq. (5). In contrast, target A (10,42), which is located in a more homogeneous region, with the closest neighbors in the second distance class, shows a sharper z PMF in comparison to target B (25,63) for models 1 and 3, and for all models a Gaussian-like shape.

Targets C (47,16) and D (49,73) are predictions for locations where observations are available. They were selected in regions with high and low z values to demonstrate the uncertainty prediction in locations coincident with the calibration set. For all three models, target C (47,16) presented lower entropy and D_{KL} in comparison to target D (49,73), due to the homogeneity of z values in the region.

Although the z PMFs (Fig. 7c) from models 1 and 2 present comparable shapes, the uncertainty structure (color and shape displayed in Fig. 7b) of the overall field differs. Since model 1 is derived from the aggregation of models 2 and 3, as presented in Eq. (6), this combination is also reflected in its uncertainty structure, lying somewhere in-between models 2 and 3.

---

[a] E-type estimate refers to the expected value derived from a conditional distribution which depends on data values (Goovaerts, 1997, p.341). They differ, therefore, from ordinary kriging estimates, which are obtained by linear combination.

Model 1 is the bolder (more confident) model, since it has the smallest median entropy (3.45 bits, Table 1). On the other hand, due to the averaging of PMFs, model 3 is the more conservative model, verified by the highest overall uncertainty (4.17 bits). Model 3 also predicts smaller minimum and higher maximum of E-type estimate, as well, for the selected targets, it provides the widest confidence interval.

The authors selected model 1 (AND/OR combination) for the sample size and benchmarking investigation presented in the next section. There, we evaluate various models via direct comparison of performance measures.

4.2 Comparison analysis

In this section, the test set was used to calculate the performance of all methods (NN, IDS, OK, and HER) as a function of sample size and dataset type (SR0, SR1, LR0, and LR1). HER was applied using the AND/OR model proposed by Eq. (6). See Supplement S2 for the calibrated parameters of all models discussed in this section.

Fig. 8 summarizes values of mean absolute error (E_{MA}), Nash–Sutcliffe efficiency (E_{NS}) and mean Kullback-Leibler divergence (D_{KL}) for all interpolation methods, sampling sizes, and dataset types. The SR fields are located in the left column and the LR in the right. Datasets without noise are represented by continuous lines and datasets with noise by dashed lines.

E_{MA} is presented in Figs. 8a,b for the SR and LR fields, respectively. All models have the same order of magnitude of E_{MA} for the noisy datasets (SR1 and LR1, dashed lines), with the performance of the NN model being the poorest, and OK being slightly better than IDS and HER. For the datasets without noise (SR0 and LR0, continuous lines), OK performed better than the other models, with a decreasing difference given sample size. In terms of E_{NS}, all models have comparable results for LR (Fig. 8d), except NN in the LR1 field. A larger contrast in the model performances can be seen for the SR field (Fig. 8c), where for SR1, NN performed worst and OK best. For SR0, especially for small sample sizes, OK performed better and NN poorly, while IDS and HER have similar results, with a slightly better performance for HER.

The probabilistic models OK and HER were comparable in terms of D_{KL}, with OK being slightly better than HER, especially for small sample sizes (Figs. 8e,f). An exception is made for OK in LR0. Since D_{KL} scoring rule penalizes extremely confident but erroneous predictions, D_{KL} of OK tended to infinity for LR0 and, therefore, it is not shown in Fig. 8f.

For all models, the performance metrics for LR showed better results when compared to SR (compare left and right column in Fig. 8). The performance improvement given the sample size is similar for all models, as can be seen by the similar slopes of the curves. In general, we noticed a prominent improvement in the performance in SR fields up to a sample size of 1000 observations. On the other hand, in LR fields, the learning process already stabilizes at around 400 observations. In addition to the model performance presented in this section, the summary statistics of the predictions and the correlation of the true value and the residue of predictions can be found in Supplement S3.

In the next section, we discuss fundamental aspects of HER and debate its properties with a focus on comparing it to OK.

4.3 Discussion

4.3.1 Aggregation methods

Several important points emerge from this study. Because the primary objective was to explore the characteristics of HER, we first consider the effect of selecting the aggregation method (Sect. 4.1). Independent of the choice of the aggregation method, the deterministic results (E-type estimate of z) of all models were remarkably similar. In contrast, we could see different uncertainty structures of the estimates for all three cases analyzed, ranging from a more confident method to a more conservative one. The uncertainty structures also reflected the expected behavior of larger errors in locations surrounded by data that are very different in value as mentioned in Goovaerts (1997, p.180, p.261). In this sense, HER has proved effective in considering both spatial configuration of data and the data values regardless of which aggregation method is selected.

As previously introduced in Sect. 2.3.1, the choice of pooling method can happen beforehand in order to introduce physical knowledge to the system or several can be tested to learn about the response of the field to the selected model. Aside from their different mathematical properties, the motivation behind the selection of the two aggregation methods (linear and log-linear) was the incorporation of continuous or discontinuous field properties. The interpretation is supported by Journel (2002), Goovaerts (1997, p.420), and Krishnan (2008) where the former connects a logarithmic expression (AND) to continuous variables, while the latter two associate linear pooling (OR) to abrupt changes in the field and categorical variables.

As verified in Sect. 4.1, the OR (=averaging) combination of distributions to estimate target PMFs was the most conservative (with largest uncertainty) method among all those tested. For this way of PMF merging, all distributions are considered feasible and each point adds new possibilities to the result. Whereas the AND combination of PMFs was a bolder approach, intersecting distributions to extract their agreements. Here, we are narrowing down the range of possible values so that the final distribution satisfies all observations at the same time. Complementarily, considering the lack of information to accurately describe the interactions between the sources of information, we proposed to infer α and β weights (proportion of AND and OR contributions, respectively) using Eq. (6). It resulted in a reasonable tradeoff between the pure AND and the pure OR model and was hence used for benchmarking HER against traditional interpolation models in Sect. 4.2.

With HER, the spatial dependence was analyzed by extracting Δz PMFs and expressed by the infogram, where classes composed by point-pairs further apart were more uncertain (presented higher entropy) than classes formed by point-pairs close to each other. Aggregation weights (Supplement S2, Figs. S2.1 and S2.2) also characterize the spatial dependence structure of the field. In general, as expected, noisy fields (SR1 and LR1) lead to smaller influence (weights) of the closer observations than non-noisy datasets (Fig. S2.1). In terms of α and β contribution (Fig. S2.2), while α received for all sample sizes the maximum weight, β increased with the sample size. As expected, in general, the noisy fields reflected a higher contribution of β due to their discontinuity. For LR0, starting at 1000 observations, β also stabilized at 0.55, indicating that the model identified the characteristic β of the population. The most noticeable result along these lines was that the aggregation method directly influences the probabilistic results and, therefore, the uncertainty (entropy) maps can be adapted according to the characteristics of the variable or interest of the expert.

4.3.2 Benchmarking and applicability

Although the primary objective of this study is to investigate the characteristics of HER, Sect. 4.2 compares it to three established interpolation methods. In general, HER performed comparable to OK, the best performing method among the analyzed ones. The probabilistic performance comparison was only possible between HER and OK, where both methods also produced comparable results. Note that the datasets were generated using Gaussian process (GP) so that they perfectly fulfilled all recommended requisites of OK (field mean independent of location, normally distributed data), thus favoring its performance. Additionally, OK was also favored when converting their predicted PDFs to PMFs, since the defined bin width was often orders of magnitude larger than the standard deviation estimated by OK. However, the procedure was a necessary step for the comparison, since HER does not fit continuous functions for their predicted PMFs.

Although environmental processes hardly fulfill Gaussian assumptions (Kazianka and Pilz, 2010; Hristopulos and Baxevani, 2020), GP allows the generation of a controlled dataset where we could examine the method performances in fields with different characteristic. Considering that it is common to transform the data so that it fits the model assumptions and back-transform it in the end, the used datasets are, to a certain extent, related to environmental data. However, the authors understand that, due to being non-parametric, HER handles different data properties without the need of transforming the available data. And since HER uses binned transformation of the data, it is also possible to handle binary (e.g., contaminated and safe areas) or even, with small adaptations, categorical data (e.g., soil types), covering another spectrum of real-world data.

4.3.3 Model generality

Especially for HER, the number of distance classes and bin width define the accuracy of our prediction. For comparison purposes, bin widths and distance classes were kept the same for all models and were defined based on small sample sizes. However, with more data available, it would be possible to better describe the spatial dependence of the field by increasing the number of distance classes and the number of bins. Although the increase in the number of classes would also affect OK performance (as it improves the theoretical variogram fitting), it would allow more degrees of freedom for HER (since it optimizes weights for each distance class), which would result in a more flexible model and closer reproducibility of data characteristics. In contrast, the degrees of freedom in OK would be unchanged, since the number of parameters of the theoretical variogram does not depend on the number of classes.

HER does not require fitting of a theoretical function, its spatial dependence structure (Δz PMFs, infogram) is derived directly from the available data, while, according to Putter and Young (2001), OK predictions are only optimal if the weights are calculated from the correct underlying covariance structure, which in practice is not the case, since the covariance is unknown and estimated from the data. Thus, the choice of the theoretical variogram for OK can strongly influence the predicted z depending on the data. In this sense, for E-type estimates, HER is more robust against user decisions than OK. Moreover, HER is flexible in the way it aggregates the probability distributions, not being a linear estimator as OK. In terms of number of observations, being a non-parametric method, HER requires sufficient data to extract the spatial dependence structure, while

OK can fit a mathematical equation with fewer data points. The mathematical function of the theoretical variogram provides advantages in respect to computational effort. Nevertheless, relying on fitted functions can mask the lack of observations, since it still produces attractive but not necessarily reliable maps (Oliver and Webster, 2014).

OK and HER have different levels of generality: OK weights depend on how the fitted variogram varies in space (Kitanidis, 1997, p.78), HER weights take into consideration the spatial dependence structure of the data (via Δz PMFs) and the z values of the observations, since they are found by minimizing D_{KL} between the true z and its predicted distribution. In this sense, the variance estimated by kriging ignores the observation values, retaining from the data only their spatial geometry (Goovaerts, 1997, p.180), while for HER, it is additionally influenced by the z value of the observations. This means that HER predicts distributions for unsampled points that are conditioned to the available observations and based on its spatial correlation structure, a characteristic which was first possible with the advent of indicator kriging (Journel, 1983). Conversely, when no nugget effect is expected, HER can lead to undesired uncertainty when predicting the value at or near sampled locations. This can be overcome by defining a small distance class for the first class, changing the binning to obtain a point-mass distribution as prediction, or asymptotically increasing the weight towards infinity as the distance approaches zero. With further developments, the matter could be handled by coupling HER with sequential simulation or using kernels to smooth the spatial characterization model.

4.3.4 Weight optimization

Another important difference is that OK performs multiple local optimizations (one for each target) and the weight of the observations varies for each target, whereas HER performs only one optimization for each one of the aggregation equations, obtaining a global set of weights which are kept fixed for the classes. Additionally, OK weights can reach extreme values (negative or greater than 1), which on the one hand is a useful characteristic for reducing redundancy and predicting values outside the range of the data (Goovaerts, 1997, p.176), but on the other hand can lead to unacceptable results, such as negative metal concentrations (Goovaerts, 1997, p. 174-177) and negative kriging variances (Manchuk and Deutsch, 2007). HER weights are limited to the range of [0,1]. Since the used dataset was evenly spaced, a possible issue of redundant information in the case of clustered samples was not considered in this paper. The influence of data clusters could be reduced by splitting the search neighborhood into equal angle sectors and retaining within each sector a specified number of nearest data (Goovaerts, 1997, p.178) or discarding measurements that contains no extra information (Kitanidis, 1997, p.70). Although kriging weights naturally control redundant measurements based on the data configuration, OK does not account for clusters with heterogeneous data since it presumes that two measurements located near each other contribute the same type of information (Goovaerts, 1997, p.176, p.180; Kitanidis, 1997, p.77).

Considering the probabilistic models, both OK and HER present similarities. The two approaches take into consideration the spatial structure of the variables since their weights depend on its spatial correlation. Just as OK (Goovaerts, 1997, p.261), we verified that HER is a smoothing method since the true values are overestimated in low-valued areas and underestimated in high-valued (Supplement S3, Fig. S3.1). However, HER revealed a reduced smoothing (residue correlation closer to zero) compared to OK for SR0, SR1 and LR1. In particular, for points beyond the range, both methods predict by averaging the available observations. While OK calculates the same weight for all observations beyond the range and proceeds with their linear combination, HER associates Δz PMF of the full dataset to all observations beyond the range and aggregates them using the same weight (last-class weight).

5 Summary and conclusion

In this paper we introduced a spatial interpolator which combines statistical learning and geostatistics for overcoming parametrization with functions and uncertainty tradeoffs present in many existing methods. Histogram via entropy reduction (HER) is free of normality assumptions, covariance fitting, and parametrization of distributions for uncertainty estimation. It is designed to globally minimize the predictive entropy (uncertainty) and uses probability aggregation methods for introducing or inferring (dis)continuity properties of the field and estimating conditional distributions (target point conditioned to the sampled values).

Throughout the paper, three aggregation methods (OR, AND, AND/OR) were analyzed in terms of uncertainty and resulted in predictions ranging from conservative to more confident ones. HER's performance was also compared to popular interpolators (nearest neighbor, inverse distance weighting, and ordinary kriging). All methods were tested under the same conditions. HER and ordinary kriging (OK) were the most accurate methods for different sample sizes and field types. HER has featured some properties: i) it is non-parametric, in the sense that predictions are directly based on empirical distribution, thus bypassing function fitting and therefore avoiding the risk of adding information not available on the data; ii) it allows to incorporate different uncertainty properties according to the dataset and user interest by selecting the aggregation method; iii) it enables the calculation of confidence intervals and probability distributions; iv) HER is non-linear and the predicted conditional distribution depends on both the spatial configuration of the data and the field values; v) the number of parameters to be optimized can be adjusted to the amount of data available; vi) since HER uses binned transformation of the data, it is adaptable to handle binary or even categorical data; and vii) it can be extended to conditional stochastic simulation by directly performing sequential simulation on the predicted conditional distribution.

Considering that the quantification and analysis of uncertainties are important in all cases where maps and models of uncertain properties are the basis for further decisions (Wellmann, 2013), HER proved to be a suitable method for uncertainty estimation, where information theoretic measures, geostatistics, and aggregation method concepts are put together to bring more flexibility to uncertainty prediction and analysis. Additional investigation is required to analyze the method in the face of spatio-temporal domains, categorical data, probability and uncertainties maps, sequential simulation, sampling designs, and handling additional variables (co-variates), all of which are possible topics to be explored in future studies.

Data availability

The source code for an implementation of HER, containing spatial characterization, convex optimization and distribution prediction is published alongside this manuscript via GitHub at https://github.com/KIT-HYD/HER. The repository also includes scripts to exemplify the use of the functions and the dataset used in the case study. The synthetic field generator using Gaussian process is available in scikit-learn (Pedregosa et al., 2011), while the code producing the fields can be found at: https://github.com/mmaelicke/random_fields.

Author contribution

ST and UE directly contributed to the design of the method and test application, to the analysis of the performed simulations, and wrote the manuscript. MM programmed the algorithm of data generation and, together with ST, calibrated the benchmark models. ST implemented HER algorithm, performed the simulations, calibration-validation design, parameter optimization, benchmarking, and data support analyses. UE implemented the calculation of information theory measures, multivariate histograms operations and, together with ST and DV, the PMF aggregation functions. UE and DV contributed with interpretations and technical improvement of the model. DV improved the computational performance of the algorithm, implemented the convex optimization for the PMF weights, and provided insightful contributions to the method and the manuscript. RL brought key abstractions from mathematics to physics, when dealing with aggregation methods and binning strategies. FW provided crucial contributions to the PMF aggregation and uncertainty interpretations.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

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

References

Allard, D., Comunian, A., and Renard, P.: Probability aggregation methods in geoscience, Math. Geosci., 44(5), 545–581, doi:10.1007/s11004-012-9396-3, 2012.

Bárdossy, A.: Copula-based geostatistical models for groundwater quality parameters, Water Resour. Res., 42(11), 1–12, doi:10.1029/2005WR004754, 2006.

Batty, M.: Spatial Entropy, Geogr. Anal., 6(1), 1–31, doi:10.1111/j.1538-4632.1974.tb01014.x, 1974.

Bell, G., Hey, T., and Szalay, A.: Computer science: Beyond the data deluge, Science, 323(5919), 1297–1298, doi:10.1126/science.1170411, 2009.

Bianchi, M. and Pedretti, D.: An entrogram-based approach to describe spatial heterogeneity with applications to solute transport in porous media, Water Resour. Res., 54(7), 4432–4448, doi:10.1029/2018WR022827, 2018.

Branch, M. A., Coleman, T. F., and Li, Y.: A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems, SIAM J. Sci. Comput., 21(1), 1–23, doi:10.1137/S1064827595289108, 1999.

Brunsell, N. A.: A multiscale information theory approach to assess spatial-temporal variability of daily precipitation, J. Hydrol., 385(1–4), 165–172, doi:10.1016/j.jhydrol.2010.02.016, 2010.

Chapman, T. G.: Entropy as a measure of hydrologic data uncertainty and model performance, J. Hydrol., 85(1–2), 111–126, doi:10.1016/0022-1694(86)90079-X, 1986.

Chicco, D.: Ten quick tips for machine learning in computational biology, BioData Min., 10(1), 1–17, doi:10.1186/s13040-017-0155-3, 2017.

Cover, T. M. and Thomas, J. A.: Elements of information theory, 2nd ed., John Wiley & Sons, New Jersey, USA, 2006.

Darscheid, P.: Quantitative analysis of information flow in hydrological modelling using Shannon information measures, Karlsruhe Institute of Technology, 73 pp., 2017.

Darscheid, P., Guthke, A., and Ehret, U.: A maximum-entropy method to estimate discrete distributions from samples ensuring nonzero probabilities, Entropy, 20(8), 601, doi:10.3390/e20080601, 2018.

Fix, E. and Hodges, J. L. Jr.: Discriminatory analysis, non-parametric discrimination, USA School of Aviation Medicine, Project 21-49-004, Report. 4, doi:10.2307/1403797, Texas, 1951.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc., 102(477), 359–378, doi:10.1198/016214506000001437, 2007.

Gong, W., Yang, D., Gupta, H. V., and Nearing, G.: Estimating information entropy for hydrological data: one dimensional case, Water Resour. Res., 1, 5003–5018, https://doi.org/10.1002/2014WR015874, 2014.

Good, I. J.: Rational decisions, J. R. Stat. Soc., 14(1), 107–114, 1952.

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

Hristopulos, D. T. and Baxevani, A.: Effective probability distribution approximation for the reconstruction of missing data, Stoch. Environ. Res. Risk Assess., 34(2), 235–249, doi:10.1007/s00477-020-01765-5, 2020.

Journel, A. G.: Nonparametric estimation of spatial distributions, J. Int. Assoc. Math. Geol., 15(3), 445–468, doi:10.1007/BF01031292, 1983.

Journel, A. G.: Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses, Math. Geol., 34(5), 573–596, doi:10.1023/A:1016047012594, 2002.

Journel, A. G. and Huijbregts, C. J.: Mining geostatistics, 1978.

Kazianka, H. and Pilz, J.: Spatial Interpolation Using Copula-Based Geostatistical Models, in geoENV VII – Geostatistics for Environmental Applications, pp. 307–319, 2010.

Kitanidis, P. K.: Introduction to geostatistics: applications in hydrogeology, Cambridge University Press, Cambridge, United Kingdom, 1997.

Knuth, K. H.: Optimal data-based binning for histograms, 30, doi:abs/physics/0605197, 2013.

Krige, D. G.: A statistical approach to some mine valuation and allied problems on the Witwatersrand, Master's thesis, University of Witwatersrand., 1951.

Krishnan, S.: The tau model for data redundancy and information combination in earth sciences: theory and application, Math. Geosci., 40(6), 705–727, doi:10.1007/s11004-008-9165-5, 2008.

Leopold, L. B. and Langbein, W. B.: The concept of entropy in landscape evolution, U.S. Geol. Surv. Prof. Pap. 500-A, 1962.

Liu, D., Wang, D., Wang, Y., Wu, J., Singh, V. P., Zeng, X., Wang, L., Chen, Y., Chen, X., Zhang, L. and Gu, S.: Entropy of hydrological systems under small samples: uncertainty and variability, J. Hydrol., doi:10.1016/j.jhydrol.2015.11.019, 2016.

Li, J. and Heap, A. D.: A review of spatial interpolation methods for environmental scientists, Canberra Geosci. Aust., 137(2008/23), 154, 2008.

Li, J. and Heap, A. D.: A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors, Ecol. Inform., 6(3–4), 228–241, doi:10.1016/j.ecoinf.2010.12.003, 2011.

Li, J. and Heap, A. D.: Spatial interpolation methods applied in the environmental sciences: A review, Environ. Model. Softw., 53, 173–189, doi:10.1016/j.envsoft.2013.12.008, 2014.

Loritz, R., Gupta, H., Jackisch, C., Westhoff, M., Kleidon, A., Ehret, U., and Zehe, E.: On the dynamic nature of hydrological similarity, Hydrol. Earth Syst. Sci., (22), 3663–3684, doi:10.5194/hess-22-3663-2018, 2018.

Loritz, R., Kleidon, A., Jackisch, C., Westhoff, M., Ehret, U., Gupta, H., and Zehe, E.: A topographic index explaining hydrological similarity by accounting for the joint controls of runoff formation, Hydrol. Earth Syst. Sci. Discuss., 1–22, doi:10.5194/hess-2019-68, 2019.

Manchuk, J. G. and Deutsch, C. V: Robust solution of normal (kriging) equations, 10, Available from: http://www.ccgalberta.com, 2007.

Mälicke, M. and Schneider, H. D. Scikit-GStat 0.2.6: A scipy flavored geostatistical analysis toolbox written in Python. (Version v0.2.6). Zenodo. http://doi.org/10.5281/zenodo.3531816, 2019.

Mälicke, M., Hassler, S., Blume, T., Weiler, M., and Zehe, E.: Soil moisture: variable in space but redundant in time, Hydrol. Earth Syst. Sci. Discuss., 1–28, doi:10.5194/hess-2019-574, 2019.

Mishra, A. K., Özger, M., and Singh, V. P.: An entropy-based investigation into the variability of precipitation, J. Hydrol., 370(1–4), 139–154, doi:10.1016/j.jhydrol.2009.03.006, 2009.

Myers, D. E.: Spatial interpolation: an overview, Geoderma, 62(1–3), 17–28, doi:10.1016/0016-7061(94)90025-6, 1993.

Naimi, B.: On uncertainty in species distribution modelling, Doctoral thesis, University of Twente, 2015.

Nearing, G. S. and Gupta, H. V.: Information vs. Uncertainty as the Foundation for a Science of Environmental Modeling, eprint arXiv:1704.07512, 1–23 [online] Available from: http://arxiv.org/abs/1704.07512, 2017.

Oliver, M. A. and Webster, R.: A tutorial guide to geostatistics: Computing and modelling variograms and kriging, Catena, 113, 56–69, doi:10.1016/j.catena.2013.09.006, 2014.

Pechlivanidis, I. G., Jackson, B., Mcmillan, H., and Gupta, H. V.: Robust informational entropy-based descriptors of flow in catchment hydrology, Hydrol. Sci. J., 61(1), 1–18, doi:10.1080/02626667.2014.983516, 2016.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Blondel, M., Thirion, B., Grisel, O., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Pham, T. D.: GeoEntropy: A measure of complexity and similarity, Pattern Recognit., 43(3), 887–896, doi:10.1016/j.patcog.2009.08.015, 2010.

Putter, H. and Young, G. A.: On the effect of covariance function estimation on the accuracy of kriging predictors, Bernoulli, 7(3), 421–438, 2001.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian processes for machine learning, The MIT Press., 2006.

Roodposhti, M. S., Aryal, J., Shahabi, H., and Safarrad, T.: Fuzzy Shannon entropy: a hybrid GIS-based landslide susceptibility mapping method, Entropy, 18(10), doi:10.3390/e18100343, 2016.

Roulston, M. S. and Smith, L. A.: Evaluating probabilistic forecasts using information theory, Mon. Weather Rev., 130(6), 1653–1660, doi:10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2, 2002.

Ruddell, B. L. and Kumar, P.: Ecohydrologic process net- works: 1. Identification, Water Resour. Res., 45, 1–23, https://doi.org/10.1029/2008WR007279, 2009.

Scott, D. W.: Scott bin width, Biometrika, 66(3), 605–610, doi:10.1093/biomet/66.3.605, 1979.

Shannon, C. E.: A mathematical theory of communication, Bell Syst. Tech. J., 27, 379–423, 623–656, 1948.

Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data, in Proceedings of the 1968 23rd ACM National Conference, pp. 517–524, 1968.

Singh, V. P.: Entropy theory and its application in environmental and water engineering, first edition, John Wiley & Sons, Ltd., 2013.

Solomatine, D. P. and Ostfeld, A.: Data-driven modelling: some past experiences and new approaches, J. Hydroinform., 10(1), 3–22, doi:10.2166/hydro.2008.015, 2008.

Tarantola, A.: Inverse problem theory and methods for model parameter estimation, Philadelphia., 2005.

Tarantola, A. and Valette, B.: Inverse problems = quest for information, J. Geophys., 50, 159–170, 1982.

Thiesen, S., Darscheid, P., and Ehret, U.: Identifying rainfall-runoff events in discharge time series: A data-driven method based on Information Theory, Hydrol. Earth Syst. Sci., 23(2), 1015–1034, doi:10.5194/hess-23-1015-2019, 2019.

Weijs, S. V.: Information theory for risk-based water system operation, Technische Universiteit Delft., 210 pp., 2011.

Weijs, S. V., van Nooijen, R., and van de Giesen, N.: Kullback-Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition, Mon. Weather Rev., 138(9), 3387–3399, doi:10.1175/2010mwr3229.1, 2010.

Wellmann, J. F.: Information theory for correlation analysis and estimation of uncertainty reduction in maps and models, Entropy, 15, 1464–1485, doi:10.3390/e15041464, 2013.

Yakowitz, S. J. and Szidarovszky, F.: A comparison of kriging with nonparametric regression methods, J. Multivar. Anal., 16, 21–53, 1985.

Table 1: Summary statistics and model performance of LR1-600.

	Test set	HER AND/OR (Model 1)	HER pure AND (Model 2)	HER pure OR (Model 3)	True test set
Summary statistics of the E-type estimate of z	mean	-0.98	-0.98	-0.98	-1.00
	standard deviation	0.89	0.89	0.90	1.03
	entropy (H)	4.07	4.04	4.10	4.39
	maximum	1.32	1.26	1.33	2.14
	median	-0.83	-0.82	-0.85	-0.96
	minimum	-2.82	-2.77	-2.92	-3.75
	kurtosis	2.23	2.19	2.27	2.44
	skewness	0.02	0.02	0.03	0.02
Summary statistics of predicted distribution	median entropy	3.45	3.75	4.17	–
	z maximum[a]	2.40	3.20	2.60	–
	z minimum[a]	-4.20	-7.00	-4.80	–
	target (49,73): [95% CI]	[-0.40, 1.60]	[-0.60, 1.60]	[-1.20, 2.20]	–
	mean	0.69	0.66	0.70	1.35
	target (47,16): [95% CI]	[-2.00, -0.20]	[-2.20, 0.00]	[-2.60, 0.20]	–
	mean	-0.99	-1.00	-0.98	-1.02
	target (25,63): [95% CI]	[-2.40, -0.40]	[-2.40, -0.40]	[-4.00, 0.60]	–
	mean	-1.19	-1.33	1.20	-1.34
	target (10,42): [95% CI]	[-3.00, -1.20]	[-3.20, -1.20]	[-3.80, -0.80]	–
	mean	-2.06	-2.06	-2.05	-1.64
Performance	E_{MA}	0.43	0.43	0.44	–
	E_{NS}	0.72	0.72	0.71	–
	mean D_{KL}	3.54	3.58	3.76	–

[a] Considering a 95% confidence interval (CI).

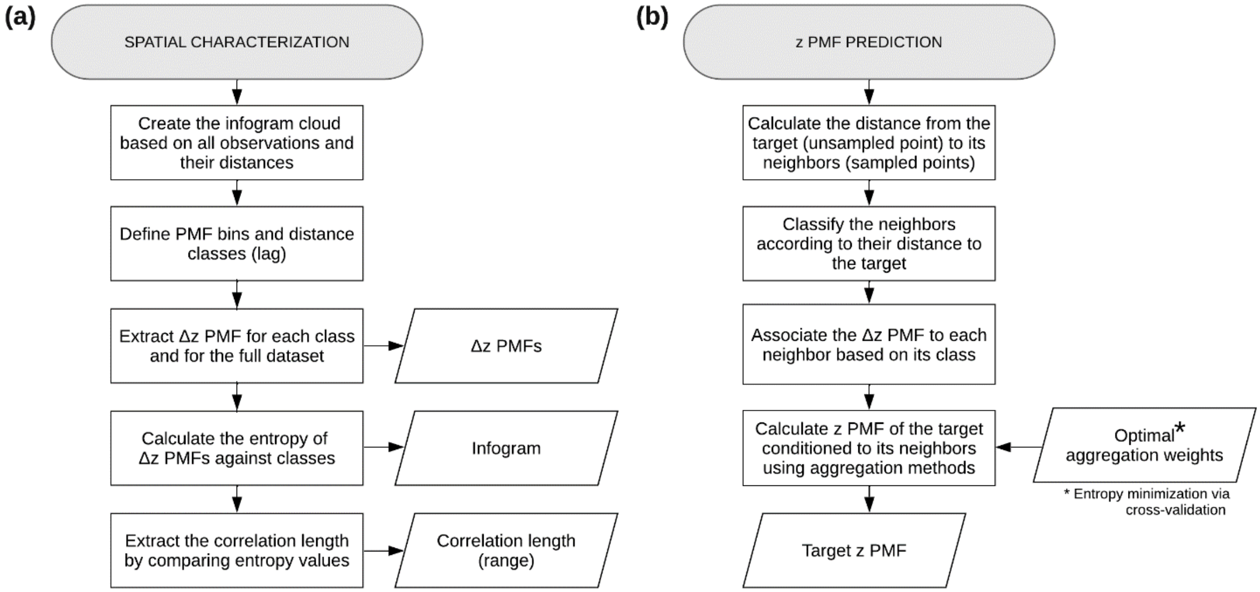

Figure 1: HER method. Flowcharts illustrating: a) spatial characterization and b) *z* probability mass function (PMF) prediction.

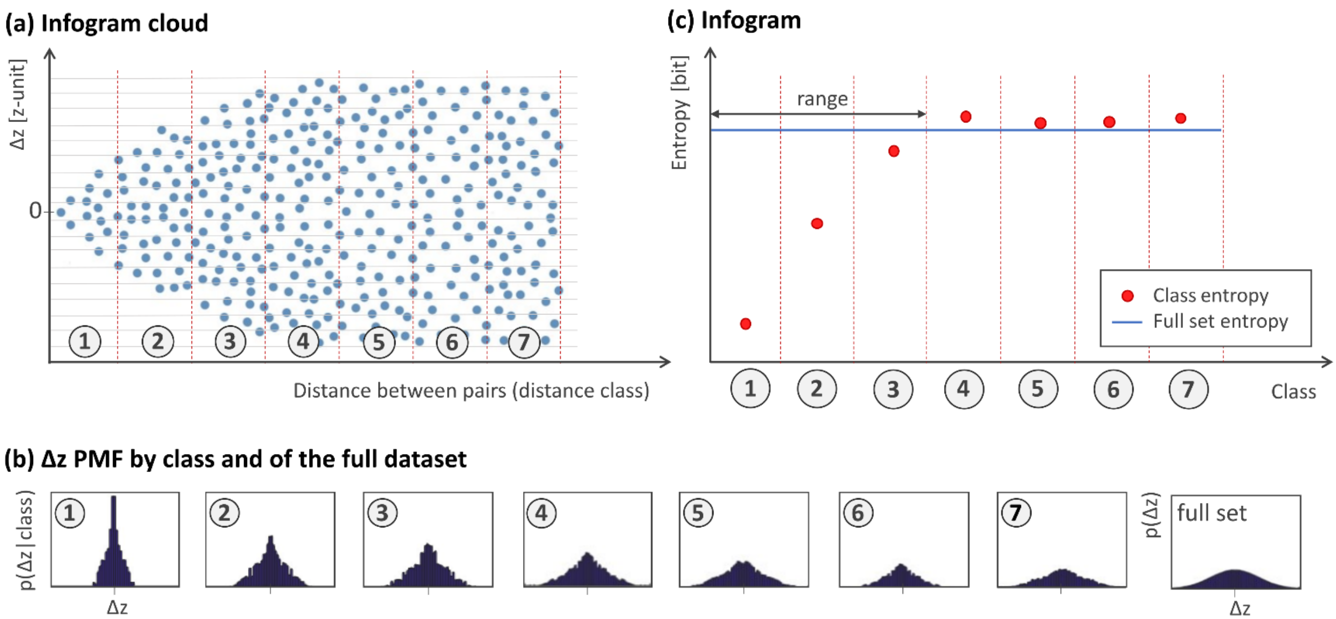

Figure 2: Spatial characterization. Illustration of: a) infogram cloud, b) Δz PMFs by class, and c) infogram.

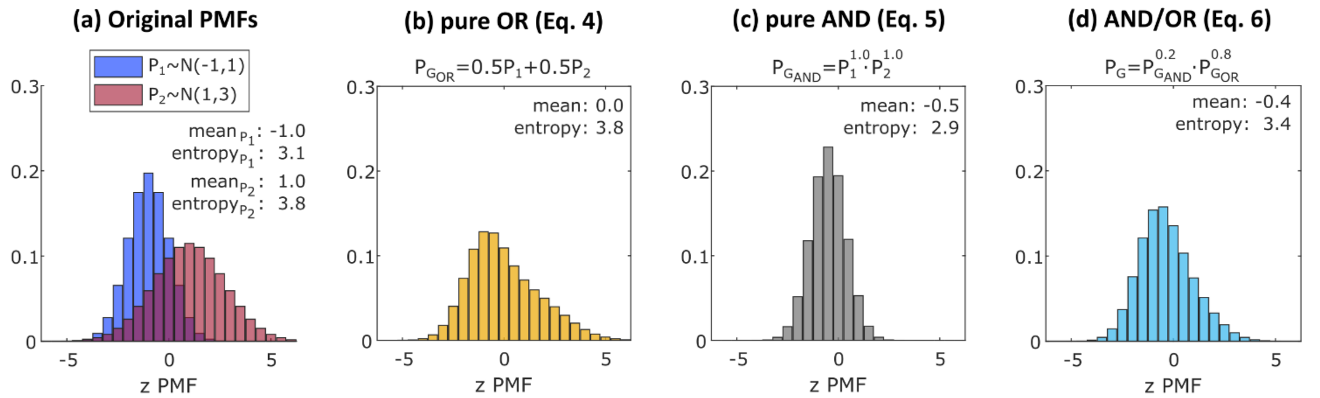

Figure 3: Examples of the different pooling operators. Illustration of: a) Normal PMFs $N(\mu, \sigma^2)$ to be combined; b) linear aggregation of (a), Eq. (4); c) log-linear aggregation of (a), Eq. (5); and d) log-linear aggregation of (b) and (c), Eq. (6).

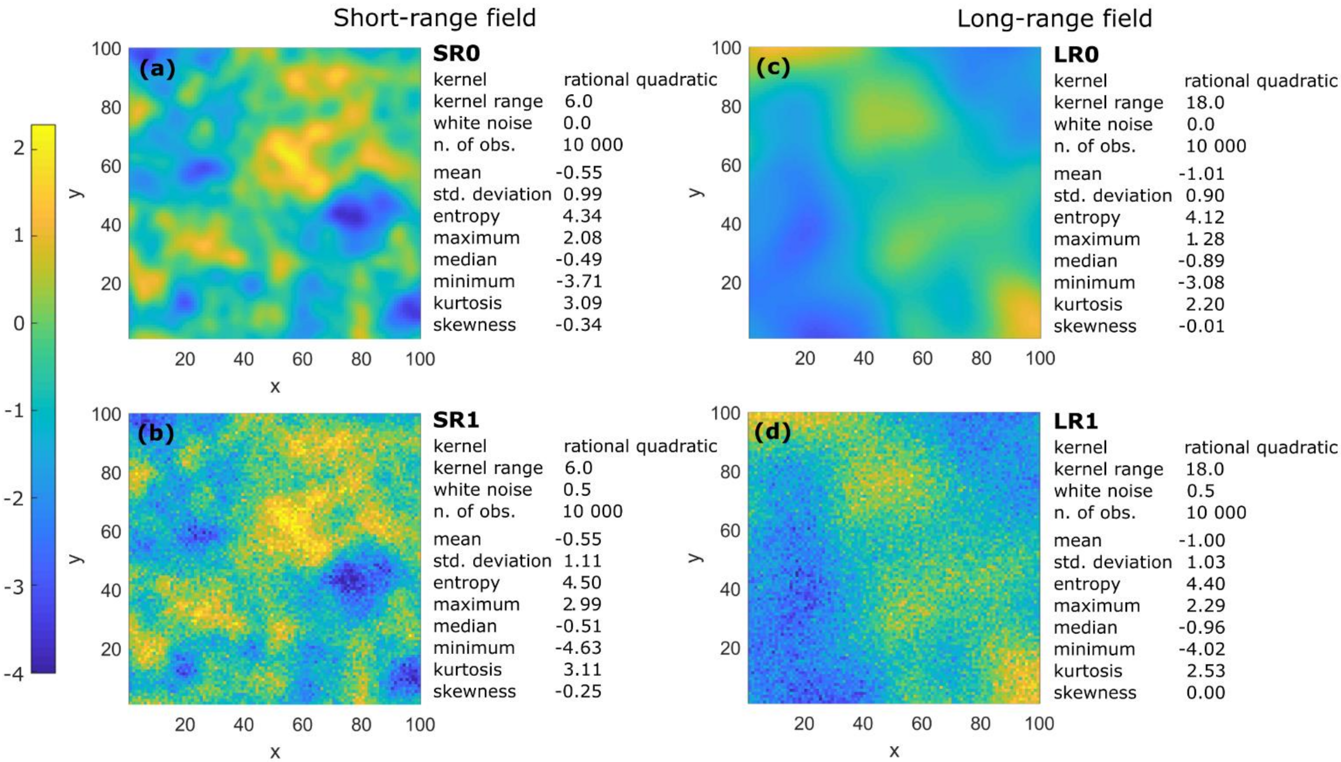

Figure 4: Synthetic fields and summary statistics: a) SR0, b) SR1, c) LR0, and d) LR1.

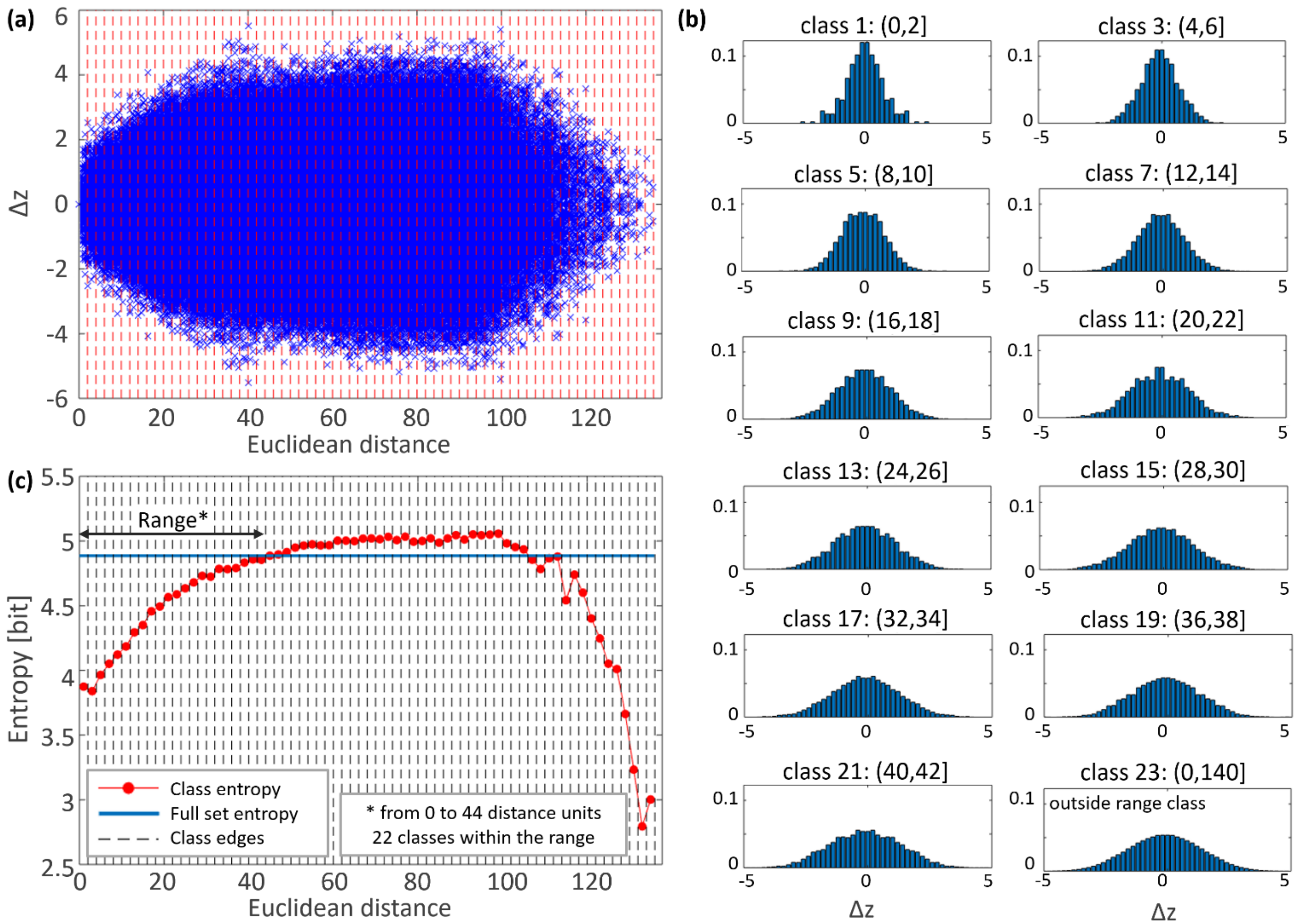

Figure 5: Spatial characterization of LR1-600: a) infogram cloud, b) Δ*z* PMFs by class, and c) infogram.

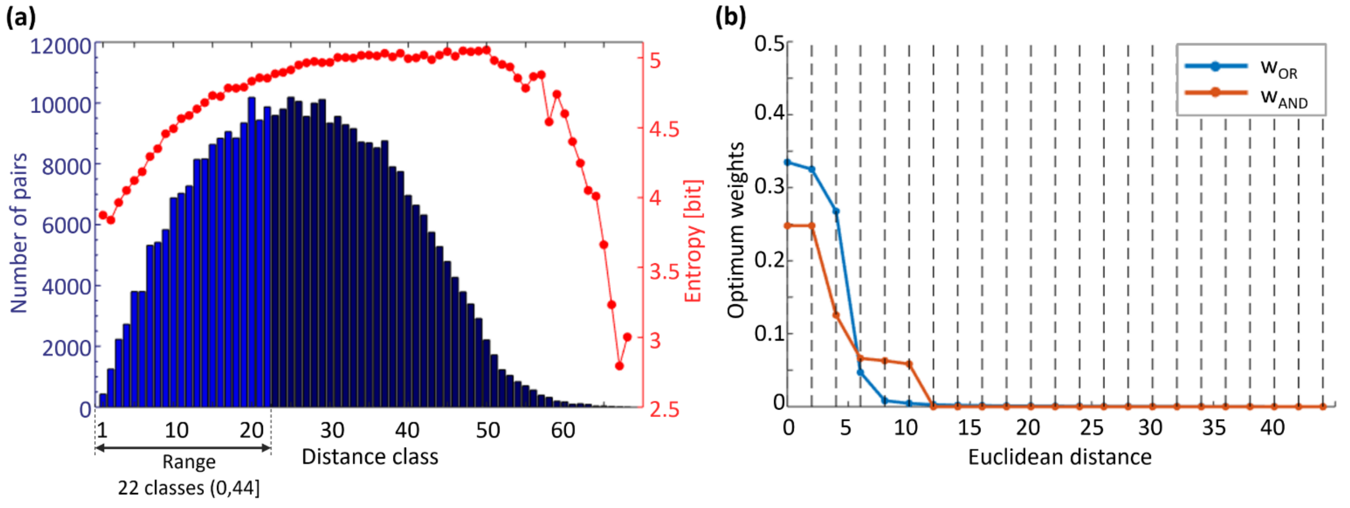

Figure 6: LR1-600: a) class cardinality and b) optimum weights, Eqs. (4) and (5).

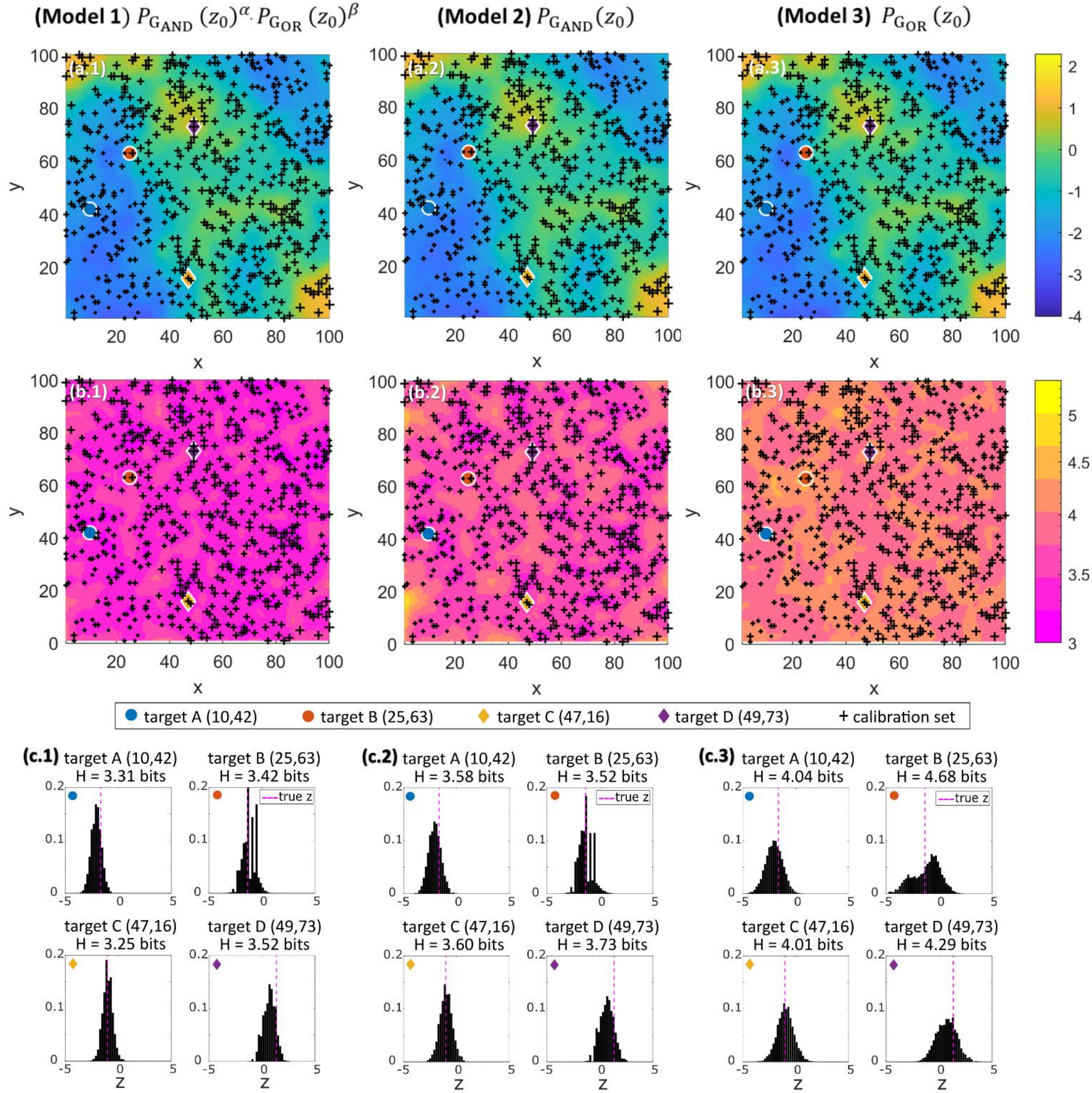

Figure 7: LR1-600 results: a) E-type estimate of *z*, b) entropy map (bit), and c) *z* PMF prediction for selected points. The first, second and third columns of the panel refer to the results of model 1 (AND/OR), model 2 (AND), and model 3 (OR), respectively.

730

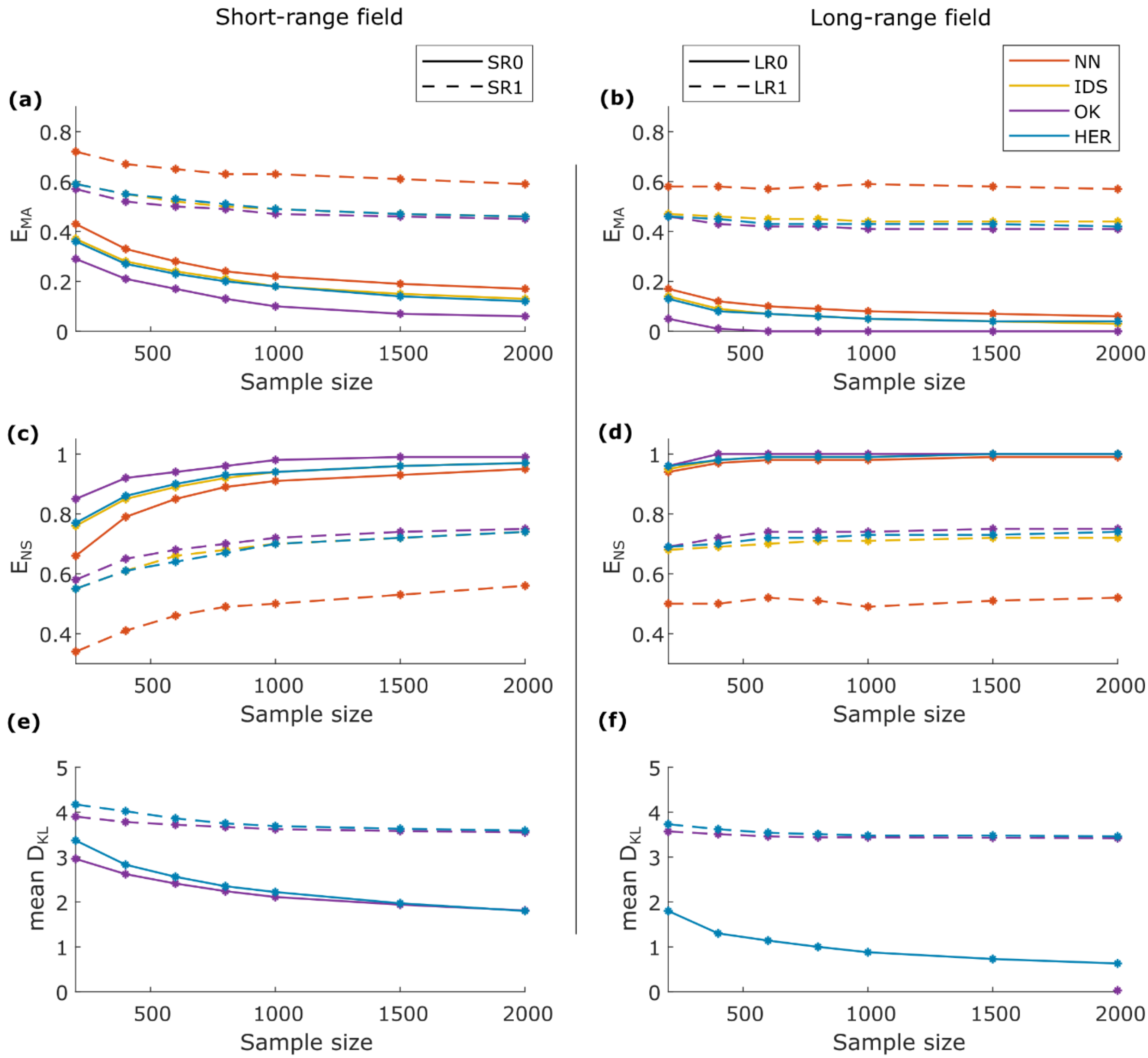

Figure 8: Performance comparison of NN, IDS, OK and HER: a,b) mean absolute error, c,d) Nash–Sutcliffe efficiency, and e,f) Kullback-Leibler divergence scoring rule, for the SR datasets in the left column and the LR datasets in the right. Continuous line refers to datasets without noise and dashed lines to datasets with noise.

735