Dear Dr. Kelleher and referees,

The current document consolidates the referee comments, the author responses posted so far, and the actions taken in the major revision of the manuscript. It is organized according to the following color code:

- Black: Referee comments
- Blue: Authors' responses from the previous stage
- Red: Authors' major revisions and responses for the new version of the manuscript

After this section of comments and responses, you can find the manuscript pdf with the tracked changes. Please note that the page numbers and lines in our current responses (red) refer to the revised, tracking-free version of the manuscript.

The new version of the manuscript contains the adjustments requested by the referees during this open discussion, summarized below:

1. Adding a paragraph:
    a. with a practical example of the three aggregation methods (Sect. 2.3.1)
    b. discussing real-world data and GP (Sect. 4.3)
    c. discussing undesired uncertainty when predicting at measurement locations (Sect. 4.3)
    d. discussing redundant measurements (Sect. 4.3)
2. Making the method clearer (e.g., infogram cloud, infogram, etc.)
3. Reviewing and restructuring discussion(Sect. 4.3) and conclusion (Sect. 5)
4. Reviewing math notation
5. Adding Dr. Ralf Loritz as co-author
6. Reviewing all figures (300 dpi quality, readability, and consistency)
7. Improving text readability and fluidity

Best regards,

Stephanie Thiesen, Diego M. Vieira, Mirko Mälicke, Ralf Loritz, J. Florian Wellmann, and Uwe Ehret

Referee #1

Dear Authors,

Thank you very much for this work. I think the work is very interesting but the paper requires some reviews before it can be published.

My main concerns are with respect to the structure of the paper. In its current status, I think the sections are unbalanced in terms of the length and some of them are mixed. E.g. in section 5. Summary and conclusions, the section is missed with the discussion – also i would suggest not including references in the conclusions.

I would suggest to shorten the paper by 1) stating clearly the messages in the paragraphs, 2) rephrasing unclear wording (e.g. lines 31-33, 37-43, etc), 3) removing un necessarily wording (e.g. 87-88 “This section is based on. . .”-> see detains in Cover and thomas, 2006) and/or, 4) including necessary explanations (e.g. line 101, i suggest remove the word “best” or include from what point of view is "best").

I also recommend using math notation consistently, e.g. bold for vectors, capital for random variables, etc.

Please, also:

- add references for nearest neighbors, and inverse distance weighting (see line 25),
- use complete names before using abbreviations (e.g. Sect.2 – the full name should be given before providing any abbreviation),
- have a look at the typos (there are many), and - maybe, good to briefly explain what is the "infogram cloud" - i am not sure that all readers are familiar with it.

In case that the editor asks a revised version of the the manuscript, i am very happy to serve as reviewer of the revised version. Once again, thank you very much for this work.

Kind Regards,

Reviewer

Response: We thank referee #1 for reviewing our manuscript and providing his/her feedback. Since the recommendations encompass a more general aspect, the authors will consider all suggestions (review paper structure, math notation, and abbreviations, insert/exclude references, include clarifications) during a textual revision of the manuscript.
We revised and restructured the discussion (Sect. 4.3) and the conclusion (Sect. 5) sections. Furthermore, we proceeded with all specific suggestions (where the line number was stated by the referee) and endeavored to follow all his/her general suggestions, i.e, a thorough revision of the manuscript considering: i) math notation, ii) misplaced references, iii) unbalanced sections, iv) unclear abbreviations, v) missing references, and vi) lack of clarity.

Referee #2

This paper presents a method called Histogram via Entropy Reduction (HER) for the interpolation of spatial geophysical data. This method is based on information theory measures of entropy and relative entropy, and has advantages over benchmarks of kriging and nearest neighbor methods in terms of its generality and lack of assumptions. The authors present the methodology which determines spatial dependency structure based on observed data points, and estimates optimal weighting parameters used to predict a given variable at a location. An application to several synthetic datasets shows high effectiveness of the method relative to three existing interpolation techniques.

Overall this paper was interesting and clearly written, and the figures are very informative and help to illustrate complex concepts. Although I do not have a background in geostatistics or interpolation of sparse datasets, this paper seems to introduce a promising avenue of how IT measures can be advantageous in this field. Some comments and suggestions listed below, which consist of minor revisions/technical corrections. They mainly highlight places that could use additional explanation or clarification.

Main Comments:

Comment 1: Line 111: on the description of creating the "infogram cloud": at first it was not clear to me whether an "infogram" was an existing technique that I was unfamiliar with, or designed by the authors. I think it is the latter (based on line 134) – either way, this aspect could be made more clear earlier in the subsection, that you have developed this graphical technique called an infogram that shows spatial correlation structure.

Response 1: Indeed, the term "infogram cloud" in L.111 is out of place. The same observation was made by referee #1. To avoid early questions, we propose rephrasing the sentence and including the name of the 3 assets used for the spatial characterization (Infogram cloud, Δz PMFs, and Infogram) when they first appear (L.112-121) as follows:

"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): [...] ii. Δz PMFs (Fig. 2b): [...] iii. Infogram (Fig. 2c)".

As previously proposed, we rephrased the sentence and improved the explanation regarding infogram cloud (Sect. 2.2, p. 4, l. 102-115).

Comment 2: Line 145: Could you add a bit more information on what effect/advantage this has? It seems like attributing a small probability to every category would make a larger difference to some types of distributions than to others.

Response 2: For the application of HER (mainly when using the log-linear aggregation method), it is desirable to assure that all bins of the distribution have a nonzero probability. This guarantees that there is always an intersection when aggregating PMFs. In this way, when the intersection between two PMFs happens only on the previously empty bins, the resulting PMF is a uniform distribution, i.e., the method effectively applies a maximum-entropy approach.

In addition, Darscheid et al. (2018) checked the impact of five alternatives for nonzero probability to a range of typical distributions (uniform, Dirac, normal, multimodal, and irregular) and concluded that, for the cases where no distribution is known a priori, three methods (including the one used in the paper) performed well across analyzed distributions. In order to add more information regarding the nonzero probability, we suggest rewriting the paragraph as follows:

"[...] The bin size was defined based on Thiesen et al. (2018), by comparing the cross entropy (Hpq=H(p)+DKL(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 were kept the same for all applications and performance calculations.

Additionally, to avoid distributions with empty bins which might make the PMF combination (presented in Sect. 2.3.1) unfeasible, as recommended by Darscheid et al. (2018), we assigned a small probability equivalent to the probability of a single point-pair count to all bins in the histogram after converting it to a PMF by normalization. This 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."

We added a brief discussion of the nonzero probabilities and bin selection in the revised version of the manuscript (Sect. 2.2, p. 5, l. 145-152).

Comment 3: Section 2.3.1: For readers less familiar with aggregating probabilities towards a spatial context, I this description could benefit from some sort of illustration or simple example that shows the difference between the different pooling operators in Eqs 4-7. For example, show two measurement points D1 and D2 with a target location A somewhere between them, and show how the measures differ. This actually become more clear to me with the later discussion in Line 445 onward, so maybe some of these aspects could be brought forward earlier.

Response 3: The authors agree that an illustration of the aggregation methods could be beneficial for showing the practical meaning of each one of the aggregation options later explored in the application case. Since the example in the spatial context is given (without illustration) in L.185-201 and L.445-448, we believe that we could explore the practical implication of the methods in the end of the section 2.3.1. We estimate that it will increase the size of the paper in half page (figure plus brief explanation). A preview of the additional figure and explanation is shown below.

"The practical differences between the pooling operators used in this paper are illustrated in Fig. 3, where Fig. 3a introduces the two PMFs to be combined, and Figs. 3b,c,d show the resulting PMFs for Eqs. (5), (4), and (7), respectively. In Fig. 3b, 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. In Fig. 3c, we use equal weights to both PMFs, and the resulting distribution is the arithmetic average of the bin probabilities. 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. Specifically for Eq. (7), this means that the final distribution may result in a pure AND, Eq. (5), or pure OR aggregation, Eq. (4), as special cases."

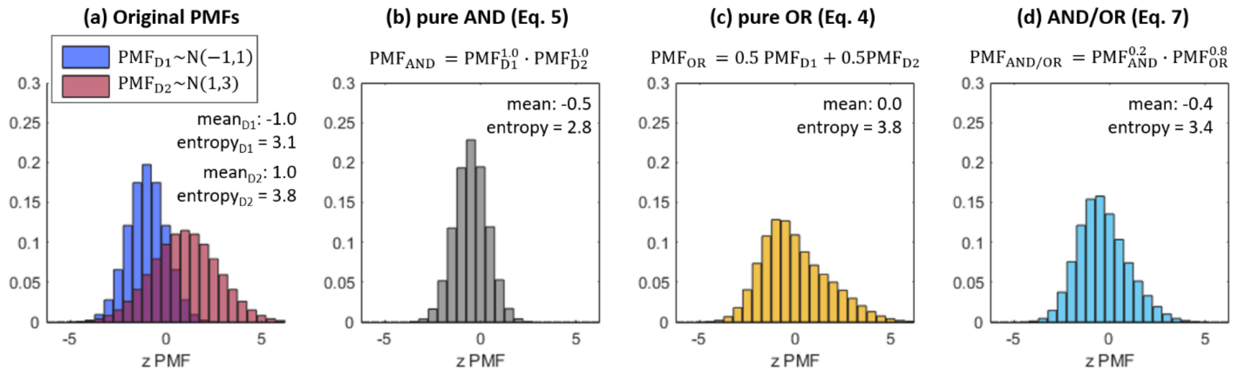


Figure 3: Examples of the different pooling operators. a) Normal PMFs N(μ,σ^2) to be combined; b) log-linear aggregation, Eq. (5); c) linear aggregation, Eq. (4); and d) log-linear aggregation of (b) and (c), Eq. (7).

We included a practical example (figure + discussion) of the three aggregation methods in the revised version of the manuscript (Sect. 2.3.1, p. 8, l. 228-234 + Figure 3 p.28).

Comment 4: Section 3.1: Are there implications to model performance of the Gaussian process used to make the test cases? I wonder how this would compare to a realistic landscape (or whether this is considered very close to a “real world” case). Something is mentioned later about this in the discussion regarding the OK method, but more information would be beneficial here.

Response 4: Thanks for raising this question. The purpose of the paper is to test HER and demonstrate its performance in face of an established geostatistical method, namely OK. Thus, for testing HER, a field generated by Gaussian Process (GP) enables us to have a controlled dataset where we could examine their performance in fields with different characteristics (short and long range, with and without noise, small or large sample size). Since GP datasets fulfill the assumptions of Ordinary Kriging, it allows a fair comparison between the methods. We can say that GP and OK are the inverse of each other. While GP generates a dataset which follows a multivariate gaussian distribution with a known covariance function, OK estimates the stochastic process behind a dataset by fitting a variogram (or covariance function) and assuming that the residuals (i.e., estimation error) follow a gaussian distribution (Kitanidis, 1997, p.95).

It is true that a real-world dataset may not necessarily have the gaussianity properties given by the GP. Therefore it is the role of the geostatistician to guarantee that the data fulfill the method assumptions. . When it is not the case, e.g., it is common to transform the data so that it fits the assumptions, and back-transform it in the end. It is worth mentioning that, while developing the manuscript, the authors tested HER in a real-world case of digital elevation model data. Although HER and OK both performed well, its inclusion would require a proper geostatistical description of the dataset, which would be out of the scope of this paper and therefore we discarded its presentation to keep the paper as short as possible and because GP covered a broader scope of fields. Altogether, this means that the test proposed using GP is also related to a real-world problem.

The authors understand that, due to being non-parametric, HER can deal with different data properties without the need of transforming the available data. HER does not require fitting of a theoretical function for extracting the spatial correlation, because its spatial dependence structure is derived directly from 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, handle categorical data (soil types), covering another spectrum of real data.

Although parts of the above discussion were mentioned in L.470 and L.484, the authors believe that it would be beneficial to review the discussion section (Sect. 4.3) to include the above arguments in the manuscript.

We included the previous discussion regarding real-word data and implications of using GP in the revised version of the manuscript (Sect. 4.3.2, p. 17, l. 486-492).

Comment 5: Line 360: I think this is explaining why the infogram illustration in Figure 2a is very different from that in Figure 4a for the farther distances (which is because with the data, there are fewer pairs that exist at those farthest distances). If this is correctly interpreted, it could be brought up in the description of Figure 2a and the infogram cloud-shape in general.
Response 5: Thanks for pointing it out. As Fig. 2 as a whole merely illustrates what one could expect to get from the spatial characterization part, we decided to show basically the behavior of the first distance classes, where the spatial correlation is stronger. "Ignoring" the last classes is a common practice when analyzing spatial correlation, when the geostatistician defines a distance cutoff (maximum lag) for their analysis. Thus, considering that this omission, although discussed in Fig.4a, is in principle expected, the authors suggest including in L.129 the following clarification "*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. 4.* "

We included the previous clarification in the revised version of the manuscript (Sect. 2.2, p. 5, l. 123-125).

Technical/writing style comments:

Comment 6: Abstract: There are several sentences here with parenthesis for additional context, I would recommend re-writing these without as many parentheses to potentially simplify and help the flow.
Response 6: Thanks. We will adapt the writing style in a revised version of the manuscript considering this point. The authors restructured the whole abstract.
Comment 7: Line 38: applying
Response 7: Ok, thanks. Adjusted.

Comment 8: Line 90: H(X) is upper bounded by infinity in a continuous case, but as you mention in the next sentence and the equation that this case is discrete – the upper bound should be log2(N) where N is the number of bins or categories.
Response 8: We agree that it can cause misunderstanding, we will refine this paragraph. The authors adjusted the paragraph considering entropy in discrete distributions (Sect. 2.1, p. 3, l. 85-86).
Comment 9: Figure 6: It is hard to see the targets in a few of the maps, I think because the markers are in the background behind the observation markers. It would help to make outlines of the marker shapes bolder or colored to show which target is which.
Response 9: Thanks. We will endeavor to improve the visibility of the points. The authors adjusted the previous Figure 6 (Figure 7, p.32).

Comment 10: Line 379: “differentiate between”, instead of differ?

Response 10: Ok, thanks. Adjusted.

Comment 11: Line 429: I found this sentence to be unecessary

Response 11: Thanks. We will consider removing it in the paper writing style revision. Sentence removed.

References:

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.

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

To the editor

We also suggest including in the discussion section (Sect. 4.3) two issues noticed by the authors during the revision process and while testing the method for assessing data uncertainty. The first one is that, since the dataset was evenly spaced, a possible issue of redundant information in case of clustered samples was not considered. Another matter that we wish to briefly discuss and propose theoretical solutions to is that, depending on how the first distance class is chosen, HER can lead to undesired uncertainty when predicting the value at the observations themselves.

We included a brief discussion and proposed theoretical solutions for undesired uncertainty and redundancy issues in the revised version of the manuscript (Sect. 4.3.3, p. 18, l. 518-523 & Sect. 4.3.4, p. 18, l. 531-537).

In addition, we would also like to review the manuscript including some terminologies which are more precise to describe the method and its implications, and they could assist to foster the proper search for scholarly literature, mainly: E-type estimate (for the expected value obtained using HER) and conditional distribution (for the results of the aggregation method and PMFs obtained with HER). Although these terms are implicit in the method and explained, the authors would like to include them explicitly.

The mentioned terminologies were explicitly mentioned throughout the new version of the manuscript.

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~~purely data-driven to geostatistical~~, and purely data-driven to geostatistical~~parametric to non-parametric~~ methods. In this study, we propose a non-parametric ~~stochastic, geostatistical estimator~~ interpolator which combines information theory with probability aggregation methods in a geostatistical framework for stochastic estimation of unsampled points.~~minimizing predictive uncertainty, and predicting distributions directly based on empirical probability.~~ Histogram via entropy reduction (HER) predicts conditional distributions based on empirical probabilities, relaxing ~~es~~ parametrizations and therefore~~,~~ avoiding the risk of adding information not present in data~~ (or losing available information)~~. By construction, ~~I~~it provides a proper framework for uncertainty estimation, since it~~ that takes into~~ accounts for both spatial configuration and data values, while allowing~~ing~~ to ~~infer (or~~ introduce or infer~~)~~ ~~physical~~ properties~~ (continuous or discontinuous characteristics)~~ of the field through the aggregation method. We investigate the framework ~~utility~~ using synthetically generated datasets and demonstrate its efficacy in ascertaining the underlying field with varying sample densities and data properties~~ (different spatial correlation distances and addition of noise)~~. HER shows comparable performance ~~with~~ to popular benchmark models ~~and~~ 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~~s~~ (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 ~~powerful~~ significant results in the context of environmental sciences~~ context~~. However, ~~'kriging',~~ like other parametric regression methods~~ (Yakowitz and Szidarovszky,~~

~~1985)~~, 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). ~~If, on the one hand,~~ OK ~~is probabilistic and therefore~~ uses fitted functions to offer~~s as a result~~ ~~the estimator's~~ uncertainty estimates ~~(through variance)~~, while ~~on the other hand,~~ deterministic estimators (NN and IDW) avoid ~~these~~ 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) ~~methods~~ have become increasingly popular as a substitute or complement to established modeling approaches. ~~However, 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, such as spatial interpolation, apply these methods remains a challenge for the researchers outside the computational intelligence (Solomatine and Ostfeld, 2008). And, even though they enable automated learning from data in stochastic and non-parametric framework; according to Solomatine and Ostfeld (2008), they are not always appreciated for working essentially data-based, replacing the 'knowledge-driven' models describing physical behavior.~~

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~~s~~ (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 ~~the~~ longitudinal profiles of rivers (Leopold and Langbein, 1962), ~~to derive rank-size rule for human settlements (Berry and Garrison, 1958; Curry, 1964), to explore the amount of information in spatial probability distributions for geographical differentiations (Gurevich, 1969),~~ to solve problems of spatial aggregation and~~,~~ ~~analyze~~ quantify ~~spatial redundancy and~~ information gain, ~~and~~ 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 ~~to measure~~ spatial dissimilarity (Naimi, 2015), ~~similarity and~~ complexity (Pham, 2010), ~~to analyze spatial~~ uncertainty (Wellmann, 2013), and ~~to assess the risk of landslides (Roodposhti et al., 2016), and to describe spatial~~ 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). ~~According to Solomatine and Ostfeld (2008), t~~The main challenges for researchers in hydroinformatics to apply data-driven methods lie in testing various combinations of methods for particular ~~water-related~~ problems, ~~in~~ combining them with optimization techniques, ~~in~~ developing robust model~~l~~ing procedures able to work with noisy data, and ~~in developing methods~~ providing the adequate model uncertainty estimates (Solomatine and Ostfeld, 2008). To overcome these challenges ~~in the framework of spatial interpolation~~ and the mentioned parametrization-uncertainty tradeoff in the context of spatial interpolation ~~(parametrization and uncertainty)~~, this paper is concerned with formulating and testing a novel method based on principles of geostatistics, ~~data-based modeling,~~ information theory and probability aggregation methods to describe spatial patterns and to ~~solve~~ obtain ~~spatial~~ stochastic ~~interpolation~~predictions ~~problems~~. In order to avoid fitting of spatial correlation functions and ~~making~~ assumptions about the underlying distribution of the ~~residues or about uncertainty~~data, it relies on empirical~~, discrete~~ 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~~a probabilistic interpolation~~. 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 ~~that is~~ not contained in ~~the~~ data ~~(or losing available information)~~, but still relying on geostatistical concepts. HER is seen as ~~an~~ in-between geostatistics (knowledge-driven) and statistical learning (data-driven) in the sense that it allows automated learning from data, bounded ~~in~~ 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 ~~data~~ 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 ~~of the data, as is the case for geostatistical procedures like kriging (Bárdossy and Li, 2008),~~ 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, ~~and~~ compare the results ~~e them~~ to benchmark models, and, in the end, discuss the achieved outcomes and model contributions. Finally, we draw conclusions in Sect. 5.

2 Method description

~~The core of~~ Histogram via entropy reduction ~~HER~~ 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~~ies~~ (which uses ~~the spatial structure, aggregation method and optimal weights~~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~~y~~ measures employed in the method, and then ~~describe in~~ 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~~. This section~~

~~is based on~~ Cover and Thomas (2006) for further reference~~more details, which we suggest for a more~~. ~~detailed introduction to concepts of information theory.~~

The entropy of a probability distribution ~~can be seen a~~measures~~s a measure of~~ the average uncertainty in a random variable. The measure, first derived by Shannon (1948)~~, varies from zero to infinity and it~~ 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~~-~~ ∈ ~~-~~χ, 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 ~~range (~~correlation length~~)~~ of the dataset (explored ~~in more depth~~ in Sect. 2.2).

Besides quantifying the uncertainty of a distribution, it is also possible to compare similarities between~~of~~ 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

$$\mathrm{D}_{\mathrm{KL}}(p||q) = \sum_{x \in \chi} p(x) \log_2 \frac{p(x)}{q(x)}\,. \tag{2}$$

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 ~~always~~ nonnegative and is zero strictly if $p = q$. In ~~the~~ HER context, Kullback-Leibler divergence is optimized ~~used~~ to select the ~~best~~ weights for aggregating distributions (detailed in Sect. 2.3). The measure ~~was~~ is also used as a scoring rule for performance verification of probabilistic predictions (Gneiting and Raftery, 2007;~~, and~~ 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 ~~incorporated~~integrated into the method. As shown in Fig.~~-~~ 1a, the spatial characterization phase aims~~, through the infogram cloud,~~ 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~~s~~. 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 ~~; calculate~~ the ~~entropy of the~~ unconditional PMF. Compute the range of the data: this is the ~~lag class~~distance where the conditional entropy exceeds the unconditional entropy. Beyond this point, the neighbors start becoming un-informative, and it ~~would~~ is~~be~~ pointless to use information outside this neighborhood.

The infogram cloud is a preparation to construct the infogram. I~~just the previous step to the infogram since i~~t contains complete cloud of pair points. The infogram plays a role similar to that of the variogram: ~~T~~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 ~~structure~~ 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. ~~This procedure~~We associate neighbors beyond the range with the Δz PMF of the full dataset. ~~B, besides guaranteeing less uncertainty in the results~~ By ~~(since~~doing so, we restrict ourselves to ~~are using~~ the more informative ~~relations through the~~ classes~~)~~ and ~~,~~ reduce~~s~~ 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.

~~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. The way we can combine the distributions and the contribution weight of each neighbor are topics of the next section.~~

Converting the frequency distributions of Δz into ~~probability mass function (~~PMF~~)~~ requires a cautious choice of bin width, since this decision will frame the ~~PMFs~~ distributions ~~which will be~~ used as ~~a~~ 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.

~~Furthermore~~Additionally, to avoid distributions with empty bins, which might make the PMF combination (discussed in Sect. 2.3.1) unfeasible, ~~as recommended by Darscheid et al. (2018),~~ 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~~, to assure nonzero probabilities when estimating distributions~~. 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~~The way we can combine the distributions and the contribution weight of each neighbor are topics of the next section~~.

2.3 Minimization of estimation entropy

For inferring the conditional distribution of the target~~ (unknown point)~~ z_0 (unsampled point) given its neighbors ~~each one of the known~~ z_i (~~observations (~~where $i = 1,\ldots,n$ are the indices of the sampled ~~observations~~points), we use~~d~~ the Δz ~~-~~PMFs obtained at the spatial characterization step (Sect. 2.2). To do so, each neighbor z_i ~~(known observation)~~ 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~~ (distance class)~~. 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 $z\cancel{D}_1$, $z\cancel{D}_2$, $z\cancel{D}_3$ from the field and that we want to predict the probability distribution of the target $z_0\cancel{A}$. In this case, what we infer at this stage are the conditional probability distributions $\cancel{P}p(\cancel{A}z_0|z\cancel{D}_1)$, $\cancel{P}p(\cancel{A}z_0|\cancel{D}z_2)$, and $\cancel{P}p(\cancel{A}z_0|\cancel{D}z_3)$.

Now, since we are in fact interested in the probability distribution of the target conditioned to multiple observations $\cancel{P}p(\cancel{A}z_0|\cancel{D}z_1,\cancel{D}z_2,\cancel{D}z_3)$, 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 ~~P_i~~ coming from different sources into a global probability distribution.~~ P_G.~~ For this purpose, the computation of the full conditional probability $\cancel{P}p(\cancel{A}z_0|\cancel{D}z_1,\ldots,\cancel{D}z_\mathrm{n})$ – where $\cancel{A}z_0$ is the event we are interested in (~~the~~ target) and $\cancel{D}z_\mathrm{i}$, $i=1,\ldots,n$ is a set of data events (or neighbors) – is ~~done~~ obtained by the use of an aggregation operator $\cancel{P}P_\mathrm{G}$, called pooling operator, such that

$$\cancel{P}p(\cancel{A}z_0|\cancel{D}z_1,\ldots,\cancel{D}z_\mathrm{n}) \approx \cancel{P}P_\mathrm{G}\left(\cancel{P}p(\cancel{A}z_0|\cancel{D}z_1),\ldots,\cancel{P}p(\cancel{A}z_0|\cancel{D}z_\mathrm{n})\right). \tag{3}$$

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

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

$$P_{\mathrm{G}_\mathrm{OR}}(z_0\cancel{A}) = \sum_{i=1}^{n} w_{\mathrm{OR}_\mathrm{i}}\,\cancel{P}P_\mathrm{i}(z_0\cancel{A}), \tag{4}$$

where n is the number of neighbors, and $w_{\mathrm{OR}_\mathrm{i}}$ are positive weights verifying $\sum_{i=1}^{n} w_{\mathrm{OR}_\mathrm{i}} = 1$. Eq. ~~-~~(4) describes mixture models in which each probability $\cancel{P}p_\mathrm{i}$ represents a different population. If we set equal weights $w_{\mathrm{OR}_\mathrm{i}}$ to every probability $\cancel{P}P_\mathrm{i}$, 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 ~~probability~~ distribution $P_{\mathrm{G}_\mathrm{OR}}$ 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, ~~which is~~ defined by

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

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.

~~$P_{\mathrm{G_{AND}}}(A) \propto \prod_{i=1}^{n} P_i(A)^{w_{\mathrm{AND_i}}}$, (6)~~

~~where Z 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). In contrast to linear pooling, log-linear pooling is typically unimodal and less dispersed.~~

A~~The a~~ggregation 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 ~~D~~z_1 and ~~D~~z_2 with different values and want to estimate the target ~~point~~ z_0~~A~~ at a location between them in a continuous field, we would expect that the estimate z_0 ~~at point A~~ would be somewhere between ~~D~~z_1 and ~~D~~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 ~~the same variance and~~ different means would result in a Gaussian distribution ~~too~~as well, less dispersed than the original ones, as also seen for the log-linear pooling. It is worth mentioning that some methods for model~~l~~ing 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 ~~D~~z_1 and ~~D~~z_2 belonging to different categories, a target

z_0~~A~~ between them will either belong to the category of ~~D~~z_1 or ~~D~~z_2, which can be achieved by the mixture distribution given by the~~an~~ OR ~~combination~~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 ~~for~~ in the comparison analysis (Sect. 4.2). For that, we propose to estimate the distribution ~~of a target~~ P_{G} of a target, by combining $P_{\mathrm{G}_{\mathrm{AND}}}$ and $P_{\mathrm{G}_{\mathrm{OR}}}$,~~ using the log-linear pooling operator,~~ such that

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

where α and β are positive weights varying from 0 to 1, which will be found by optimization. Eq. ~~-~~(6) i~~wa~~s the~~a~~ 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 ~~may~~ 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 ~~combination~~ polling of $P_{\mathrm{G}_{\mathrm{AND}}}(z_0 \cancel{A})$ and $P_{\mathrm{G}_{\mathrm{OR}}}(z_0 \cancel{A})$ could be used. We explore the properties of the ~~pure~~ 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~~ of Eqs. (4), (5) and (7)~~.

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) ~~are~~ is a strictly proper scoring rule~~s (Gneiting and Raftery, 2007)~~ since ~~they~~ it provides summary metrics ~~of performance~~ that address calibration and sharpness simultaneously~~,~~ by rewarding narrow prediction intervals and penalizing intervals missed by the observation (Gneiting and Raftery, 2007)~~Gneiting and Raftery, 2007)~~. ~~According to Gneiting and Raftery (2007), the divergence function associated with the logarithmic score is the Kullback-Leibler divergence (D_{KL}, Eq. (2)), which we used for selecting the proportion of the log-linear and linear pooling (α and β, Eq. (7)), as well as the w_{OR} and w_{AND} weights (Eq. (4) and (5), respectively), here generalized as w.~~

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~~is~~ 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 ~~make~~ the convex optimization~~ more well behaved~~, the following constraints were employed: i) for linear pooling, set w_{OR_1}~~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 $\sum_k w_{\mathrm{OR_k}} = 1$~~$\sum_{i=1}^{k} w_{\mathrm{OR}_i} = 1$~~ for ~~the~~ 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 ~~distance~~ 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 ~~step~~ of the $w_{\mathrm{OR_k}}$ and $w_{\mathrm{AND_k}}$ weights~~using Eqs. (4) and~~

~~(5)~~, since we are looking for one weight w_k for each class k, and therefore we cannot~~can't~~ 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 ~~unknown~~ unsampled point (target), first, we calculate the Euclidean distance from the target to its neighbors (~~known~~ sampled observations). Based on this distance, we obtain the ~~distance~~ 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. ~~For obtaining~~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 ~~predicted~~ PMF of the target conditional on all neighbors. Fig. ~~-~~1b presents ~~a scheme~~the ~~of the main steps for~~ z PMF prediction steps ~~of~~a single ~~one~~ 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 p~~P~~rocess 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, w~~W~~e used rational quadratic kernel (Pedregosa et al., 2011) as the covariance function, with two different correlation length~~s~~ 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 wa~~i~~s taken from the Python library scikit-learn (Pedregosa et al., 2011). The generated sets comprise:~~s~~ 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 ~~Figure 3~~ presents the field characteristics ~~(parameters and image)~~ and their summary statistics. T~~For each field, t~~he summary statistics of each field type ~~for~~ is~~the learning, validation, and test subsets are~~ included in Supplement S1.

3.2 Performance criteria

~~For~~ To ~~elucidating differences in~~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 ~~the~~ mean of the ~~logarithmic score rule, based on 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

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

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

where $\hat{z}_{\mathrm{i}}$ and z_{i} are, respectively, the predicted and observed values ~~observation and the prediction~~ at the ith location, $\bar{\mathrm{z}}$ is the mean of the observations, and n is the number of ~~predicted observations~~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~~ed~~ the predicted PDFs (probability density functions) to PMFs employing the same bins used ~~for~~ in HER.

3.3 Calibration and test design

~~For the purpose of benchmarking~~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 ~~with~~ on learning subsets with increasing sizes of

~~sizes between 200 and 2000 observations in the increments 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, t~~T~~he validation and test datasets were fixed for all performance analyses, independently of the analyzed ~~sample size~~learning set~~, to facilitate model comparison~~. 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 ~~evaluated~~assessed the results always in the same ~~test~~ 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 ~~The~~ summary statistics ~~for the~~of learning, validation, and test subsets~~ are presented in Supplement S1~~.

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) ~~was~~ is performed (Sect. 4.2). Covering deterministic, probabilistic, and geostatistical methods, three interpolators were chosen for the comparison, namely nearest neighbor~~s~~ (NN), inverse distance weighting (IDW), and ordinary kriging (OK).

As in HER, all~~ of~~ 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~~s~~ at an unsampled point~~s~~ without considering the ~~values of 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). I~~, and i~~t is known as the inverse distance squared (IDS), ~~which~~ and it is ~~also~~ 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. ~~For~~ To calibrate~~ting~~ HER aggregation weights, we applied LOOCV, as described in Sect. 2.3.2~~ion 2.3~~, for optimizing the performance of the left-out sample in the learning set. As~~ a~~ loss function, minimization of the mean D_{KL} was applied. After learning the model, we used the validation set for ~~a~~ plausibility check of the calibrated model and, eventually, adjustment ~~ing~~ of parameters. Note that no ~~curve~~ function fitting is needed ~~for~~ to ~~the~~ apply~~ication of~~ 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 ~~(interpolated fields)~~for different theoretical variogram functions, validation, and ~~cross-validation~~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 ~~the~~ HER infogram ~~(HER)~~ 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 ~~distance classes of~~ 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 three distinct models~~ of HER using three different aggregation methods on one specific ~~the~~ synthetic dataset ~~LR1 with learning set of 600 observations~~ (Sect. 4.1). In Sect. 4.2, we summarize the results on synthetic datasets LR0, LR1, SR0, SR1 for all ~~learning~~ calibration sets and numerically compare HER performance with traditional interpolators. For all applications, the performance was calculated on the same test set. ~~For all applications, the test set was used to assess the performance of the methods.~~For brevity, the model outputs were omitted in the comparison analysis, and only the performance metrics~~s~~ for each dataset and interpolator are shown. Finally, Sect. 4.3 ~~discusses~~ provide a theoretical discussion on the probabilistic methods (OK and HER), ~~comparing their performance and~~ contrasting their different properties and assumptions. ~~For all applications, the test set was used to assess the performance of the methods.~~

4.1 HER application

This section presents three variants of HER, ~~models~~ applied to the LR1 field with a ~~learning~~ calibration subset of 600~~-~~ observations (LR1-600). This dataset was selected since, due to its optimized weights ~~results~~ α and β (which reach almost the maximum value of ~~(~~one~~)~~ ~~proposed~~ suggested ~~in~~ for Eq. ~~-~~(6)), it favors to contrast uncertainty results of HER applying the three distinct aggregation methods proposed by Eqs. (4), (5), and (6). ~~For LR1-600, the optimized weights are $\alpha = 1$ and $\beta = 0.95$.~~

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 ~~where the~~the class ~~was extracted~~ 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 ~~range~~ classes were obtained. ~~Then~~ Subsequently, the range was identified as the ~~class~~ point beyond which the class entropy ~~of the class PMFs~~ exceeded the entropy of ~~all data~~ the full dataset~~(class 23 corresponding to a Euclidean distance of 44 grid units)~~, ~~see~~ seen as the intersect of the blue ~~line~~ and ~~the red~~ red-dotted lines in Fig. 5~~Figure 4~~b). 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 ~~. It occurs~~ 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~~s,~~ and the optimum weights obtained for Eqs. (4) and (5) are presented in Fig. 6. Fig. 6~~Figure 5~~a shows the number of pairs which compose the Δz ~~-~~PMF~~s~~ 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.

~~In Figure 5b, w~~We obtained the weight of each class by convex optimization as described in Sect. 2.3.2~~ on the test data set~~. The dots in Fig. 6~~Figure 5~~b 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, ~~the~~ LR1-600 models are highly influenced by neighbors up to a distance of ~~about~~ 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 ~~the~~ LR1-600 ~~optimized~~ 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 ~~results of the~~ AND/OR combination, the second column to the pure AND combination, and the third column to the pure OR combination. ~~In Figure 6a and b, t~~To assist in visually checking the heterogeneity ~~(or homogeneity)~~ of z~~ in the uncertainty maps (Figure 6b)~~, the calibration set representation is scaled by its z value, with the size of the cross increasing with z. ~~In general, f~~For the target identification, we use its grid coordinates (x,y).

Fig. 7~~Figure 6~~a shows the E-type estimate[a1]of z ~~(predicted~~ expected z ~~mean (~~obtained from the predicted z PMF) for the three analyzed models. Neither qualitatively (Fig. 7~~Figure 6~~a) nor quantitatively (Table 1) is it possible to distinguish~~differ~~ the three

[a1] 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.

models based on their ~~mean~~ E-type estimate or its ~~-~~summary statistics ~~of the predicted mean~~. Deterministic performance ~~parameters~~ metrics (E_{MA} and E_{NS}, Table 1) are also ~~quite~~ similar among the three models. However, in probabilistic terms, ~~both~~ the representation given by the entropy map (Fig. 7~~Figure 6~~b, which shows the Shannon entropy of the predicted z PMFs), ~~and~~ the statistics of predicted z ~~-~~PMFs, and ~~together with~~ 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 ~~for~~ 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. 7~~Figure 6~~c) 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 (~~Figure 6c,~~ 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_{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 ~~A~~ (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 ~~its distance to known samples and~~ the homogeneity of z-values in the region.

Although the z ~~provided~~ PMFs (Fig. 7~~Figure 6~~c) from models 1 and 2 present ~~similar~~comparable shapes, the uncertainty structure (color and shape displayed ~~by the colors~~ in Fig. 7~~Fig. 6~~b) of the overall field differs. Since ~~m~~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.

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~~ed~~ smaller minimum and higher maximum of E-type estimate~~mean values of z~~, as well, for the selected targets, ~~and~~ it provides the widest confidence interval.

The authors selected m~~M~~odel~~-~~ 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, t~~HER was applied using the more confident AND/OR model proposed by Eq. (7). T~~he 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 ~~more confident~~ AND/OR model proposed by Eq. ~~-~~(6). ~~T~~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 ~~is presented~~ 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. 8~~Figure 7~~a,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. 8~~Figure 7~~d), except NN in the LR1 field. A larger contrast in the model performances can be seen for the SR field (Fig. 8~~Figure 7~~c), where for SR1, NN performed~~s~~ 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. 8~~Figure 7~~f.

For all models, the performance metrics for LR showed better results when compared to SR (compare left and right column in Fig. 8~~Fig. 7~~). 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~~n~~ 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 ~~given sample sizes~~ already stabilizes at around 400 observations. In addition to the model performance presented in this section, the summary statistics of the predictions and the~~ir residue~~ correlation of the true value and the residue of predictions can be found in Supplement S3.

~~In this section we evaluated various models via direct comparison of performance measures.~~ In the next section, we discuss fundamental aspects of HER~~,~~ and debate its properties with a focus on comparing it to ~~ordinary~~ OK~~kriging~~.

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~~ aggregation ~~pooling~~ method, the deterministic results (~~predicted~~ E-type estimate ~~mean~~ of z) of all models were ~~very similar~~remarkably similar. ~~On the other hand~~In contrast, we could see different uncertainty structures of the estimates for all three cases analyzed, ranging from a more confident method ~~(AND/OR)~~ to a more conservative one ~~(OR)~~. The uncertainty structures also reflected the expected behavior of ~~Considering that~~ larger errors ~~are expected~~ 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 ~~the~~which aggregation method is selected.

As previously introduced in Sect. 2.3.1, the choice of ~~aggregation~~ 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 ~~of the aggregation method~~ is supported by Journel (2002), Goovaerts (1997, p.420), and Krishnan (2008) where the former connects a logarithmic expression (AND) to continuous variables~~ and simple kriging)~~, while the latter two associate~~s~~ linear pooling (OR) to ~~dual indicator kriging and therefore to categorical~~abrupt changes in the field and categorical variables. ~~For example, if we have two points D_1 and D_2 with different values and want to estimate the target point A in a continuous field, we would expect that the estimate at point A would be somewhere between D_1 and D_2, which can be achieved by an AND combination. On the other hand, in the case of categorical data, and D_1 and D_2 belonging to different categories, target A will either belong to the category of D_1 or D_2, which can be achieved by an OR combination.~~

As verified in Sect. 4.1, the OR (=averaging) combination of ~~PMFs~~ distributions to estimate target PMFs was the most conservative (~~not confident~~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 ~~On the other hand,~~ AND combination of PMFs was a bolder approach, ~~where we~~ intersecting distributions to extract their agreements. ~~In other words~~Here, we are narrowing down the range of possible values ~~and~~ 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 ~~turned out to be~~resulted in a ~~good~~ 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~~In~~ 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~~s~~. 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 ~~expert~~ interest of the expert.

4.3.2 Benchmarking and applicability

Although the primary objective of this study ~~was~~ is to investigate the characteristics of HER, Sect. 4.2 compares it to three ~~traditional~~ 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 p~~P~~rocess (GP)~~,~~ so that they perfectly fulfilled ~~remained within the~~all recommended ~~settings~~ requisites of~~for~~ OK (field mean independent of location, normally distributed data ~~and residues~~), 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~~at~~ was a necessary step for the comparison, since HER does not fit ~~PDFs~~ 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, ~~which works with non-parametric PMFs,~~ the number of distance classes and bin width ~~basically~~ define~~s~~ ~~how accurate we want to be in~~ 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 ~~better~~ the spatial dependence of the field ~~in HER~~ by increasing the number of distance classes and the number of ~~PMF~~ 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) ~~are~~ 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~~s~~ the predicted z depending on the data. In this sense, for E-type estimates, HER ~~was~~is more robust against user decisions ~~compared to~~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).

~~Considering the probabilistic models, both OK and HER present similarities. Both approaches take into consideration the spatial structure of the variables, since their weights depend on the spatial correlation of the variable. Just as OK (Goovaerts, 1997, p.261), HER turned out to be a smoothing method, since the true values are overestimated in low-valued areas and underestimated in high-valued areas. However, as verified in Supplement S3 (Fig. S3.1), 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 (weight of the last class).~~

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 ~~very close or~~ 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 ~~it~~ is a useful characteristic for reducing~~e~~ 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). ~~, while~~ HER weights are limited to the~~a~~ 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 (~~areas. However, as verified in~~ Supplement S3, Fig. S3.1). However, ~~(Fig. S3.1),~~ 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 t~~T~~his paper we introduced ~~presented~~ a ~~procedure~~ spatial interpolator which combines statistical learning and geostatistics ~~which aims at~~for overcoming parametrization with functions ~~-~~and uncertainty tradeoffs present in many existing methods ~~for spatial interpolation~~. Histogram via entropy reduction (HER)~~For this purpose, we proposed a new spatial interpolator which~~ is free of normality assumptions, covariance fitting, and parametrization of distributions for uncertainty estimation. ~~Histogram via entropy reduction (HER)~~It is designed to globally minimize the predictive ~~uncertainty~~ entropy (uncertainty) ~~expressed by relative entropy (Kullback-Leibler divergence) between the observation and prediction. More specifically, HER combines measures of information theory with~~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)~~quantifying the available information in the dataset, extracting the structure of the data spatial correlation, relaxing normality assumptions, minimizing the uncertainty of the predictions, and combining probabilities~~.

Throughout the paper, three aggregation methods (~~AND,~~ 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~~s~~, inverse distance weighting, and ordinary kriging). All methods were tested under the same conditions. HER and o~~O~~rdinary k~~K~~riging (OK) were the ~~turned out to be the~~ most accurate methods for different sample sizes and field types. ~~In contrast to OK,~~ HER has featured some ~~advantages~~properties: i) ~~-~~it is non-parametric, in the sense that predictions are directly based on empirical ~~probability~~distribution, thus bypassing function fitting ~~the usual steps of variogram fitting done in OK~~ and therefore avoiding the risk of adding information not available on the data~~ (or losing available information)~~; ~~ii) it is robust against user decisions, i.e., the choice of a theoretical variogram for OK can strongly influences the predicted z values, while HER is less sensitive to the aggregation method for prediction z, and the bin width and distance class definitions for predicting PMF (since it does not change the spatial dependence structure, expressed in the model by~~

~~Δz PMFs and the infogram, which comes directly from the available data); i~~ii) ~~-~~it allows to incorporate different uncertainty properties according to the dataset and user interest by selecting the aggregation method ~~by selecting the aggregation method~~; iii) it enables the calculation of confidence intervals and probability distributions; iv~~v~~) HER is non-linear and the predicted conditional distribution depends on both ~~for uncertainty maps, HER considers not only~~ the spatial configuration of the data and ~~, but also~~ the field values~~, while kriging variance depends on spatial data geometry only (Goovaerts, 1997, p.181~~;~~);~~ ~~v)~~ v)~~it is flexible~~ ~~to increase~~ the number of parameters~~-~~ to be optimized ~~,~~ can be adjusted ~~according~~ to the ~~size of the available~~ amount of data available~~set~~; 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~~, while OK has the number of parameters fixed according to the theoretical variogram~~. ~~On the other hand, being a non-parametric model, HER requires longer runtime and sufficient data to learn the spatial dependence from the data.~~

Considering that the quantification and analysis of uncertainties are~~is~~ 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 ~~multiple-point geostatistics,~~ spatio-temporal domains, categorical data, probability and uncertainties maps, sequential simulation, sampling designs, and handling ~~and analyzing~~ additional ~~observed~~ 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 ~~PMF~~ 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 p~~P~~rocess 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 the 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. ~~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.~~

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.

~~Bárdossy, A. and Li, J.: Geostatistical interpolation using copulas, Water Resour. Res., 44(7), 1–15, doi:10.1029/2007WR006115, 2008.~~

Batty, M.: Spatial Entropy, Geogr. Anal., 6(1), 1–31, doi:~~https://doi.org/~~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.

~~Berry, B. J. L. and Garrison, W. L.: Alternate explanations of urban rank-size relationships, Ann. Assoc. Am. Geogr., 48(1), 83–91, 1958.~~

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.

~~Curry, L.: The random spatial economy: an exploration in settlement theory, Ann. Assoc. Am. Geogr., 54(1), 138–146, 1964.~~

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.

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

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.

~~Gurevich, B. L.: Geographical differentiation and its measures in a discrete system, Sov. Geogr., 10(7), 387–413, doi:10.1080/00385417.1969.10770425, 1969.~~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. and Huijbregts, C. J.: Mining geoestatistics., 1978.~~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.: A review of spatial interpolation methods for environmental scientists, Canberra Geosci. Aust., 137(2008/23), 154, 2008.~~

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.

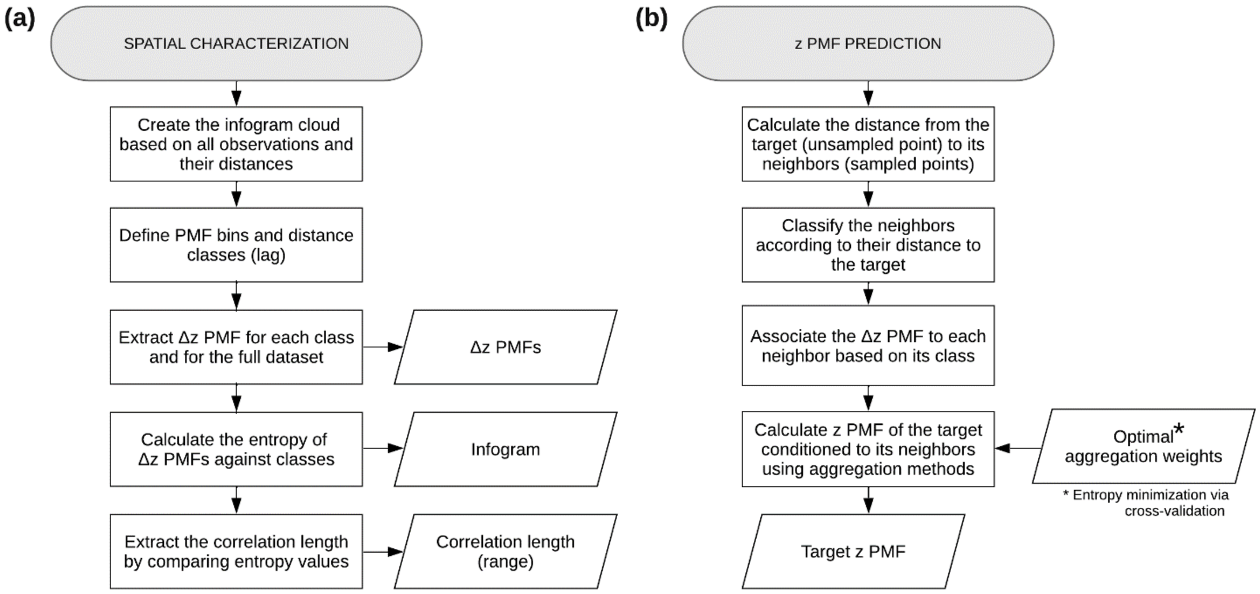
Yakowitz, S. J. and Szidarovszky, F.: A comparison of kriging with nonparametric regression methods, J. Multivar. Anal., 16, 21–53, 1985.

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

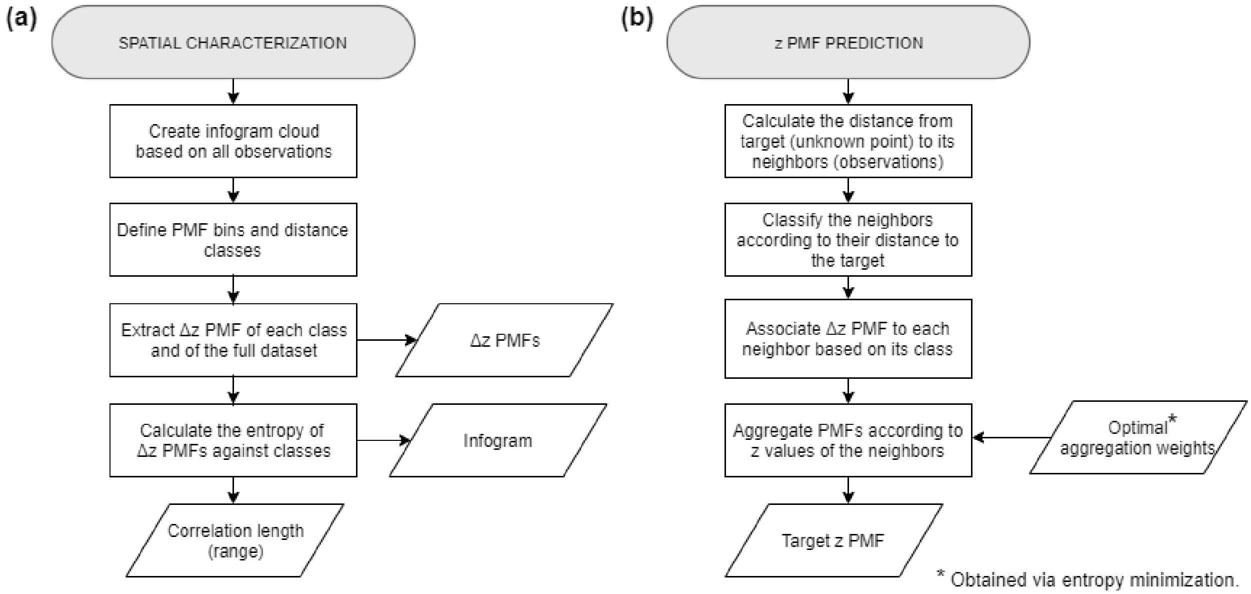
Test set ~~predicted by~~		HER AND/OR (Model 1)	HER pure AND (Model 2)	HER pure OR (Model 3)	True test set
Summary statistics of the E-type estimate of z ~~mean predicted values of z~~	mean	-0.98	-0.98	-0.98	-1.00
	standard deviation	0.89	0.89	0.90	1.03
	entropy (H)	4.07	4.04	4.10	4.39
	maximum	1.32	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 ~~z PMF~~	median entropy	3.45	3.75	4.17	–
	z maximum[~~1~~a]	2.40	3.20	2.60	–
	z minimum[~~1~~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	–

[~~1~~a] Considering a 95% confidence interval (CI).

~~CI: confidence interval.~~

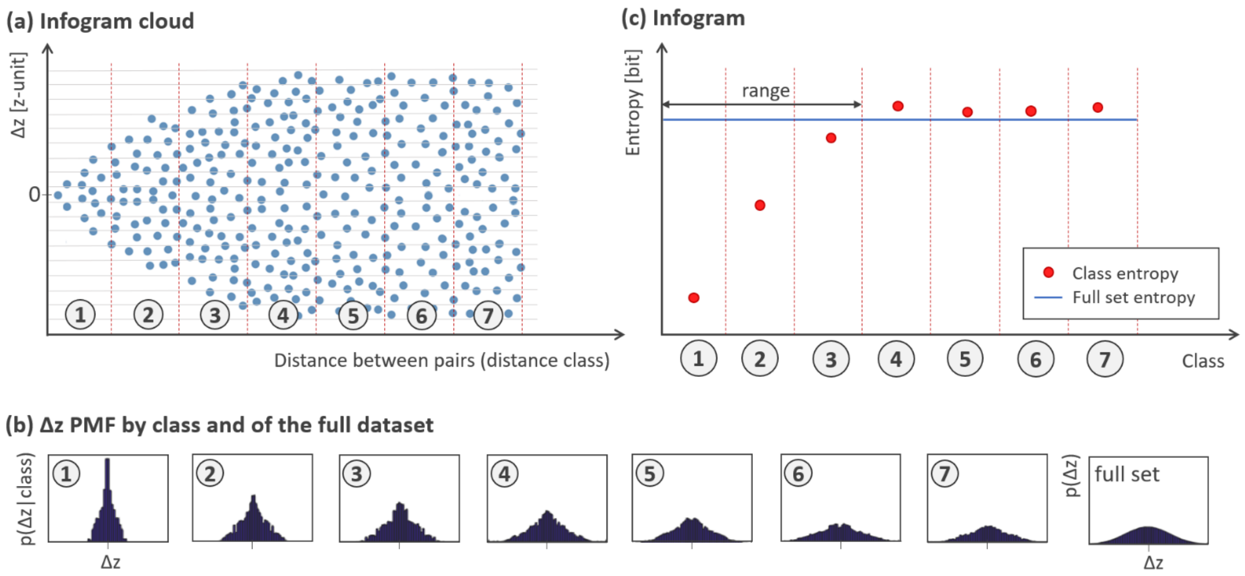


(a)
SPATIAL CHARACTERIZATION
Create the infogram cloud based on all observations and their distances
Define PMF bins and distance classes (lag)
Extract Δz PMF for each class and for the full dataset
Δz PMFs
Calculate the entropy of Δz PMFs against classes
Infogram
Extract the correlation length by comparing entropy values
Correlation length (range)
(b)
z PMF PREDICTION
Calculate the distance from the target (unsampled point) to its neighbors (sampled points)
Classify the neighbors according to their distance to the target
Associate the Δz PMF to each neighbor based on its class
Calculate z PMF of the target conditioned to its neighbors using aggregation methods
Optimal* aggregation weights
* Entropy minimization via cross-validation
Target z PMF



(a)
SPATIAL CHARACTERIZATION
Create infogram cloud based on all observations
Define PMF bins and distance classes
Extract Δz PMF of each class and of the full dataset
Δz PMFs
Calculate the entropy of Δz PMFs against classes
Infogram
Correlation length (range)
(b)
z PMF PREDICTION
Calculate the distance from target (unknown point) to its neighbors (observations)
Classify the neighbors according to their distance to the target
Associate Δz PMF to each neighbor based on its class
Aggregate PMFs according to z values of the neighbors
Optimal* aggregation weights
Target z PMF
* Obtained via entropy minimization.

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



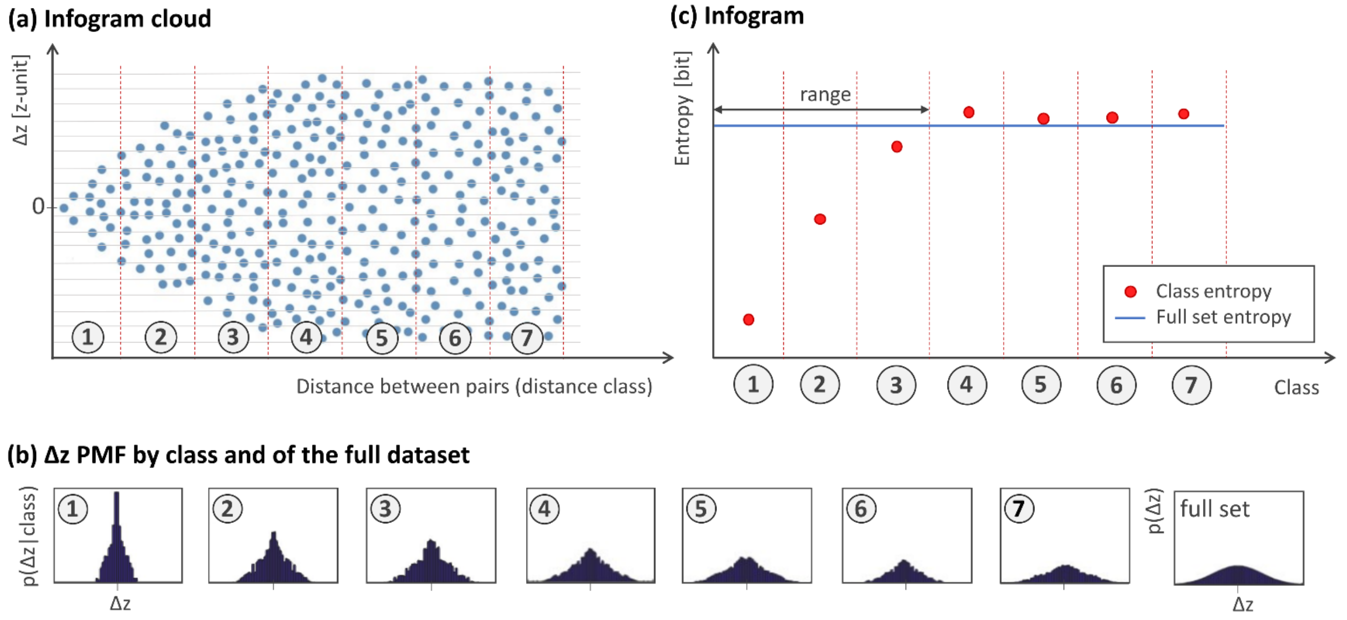


Figure 2: Spatial characterization. Illustration of: ~~the~~ a) infogram cloud, b) *Δz*~~-~~ PMFs by class, and c) infogram.

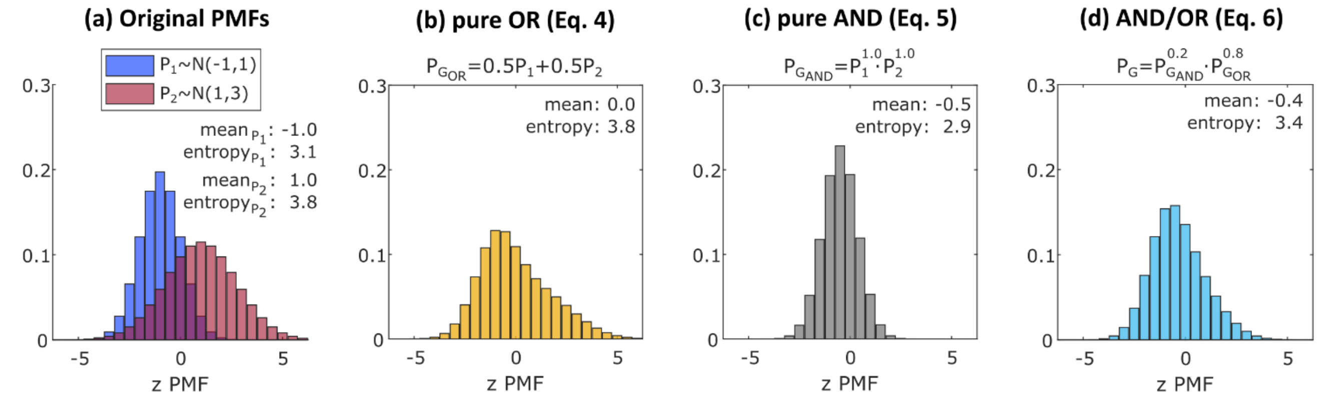
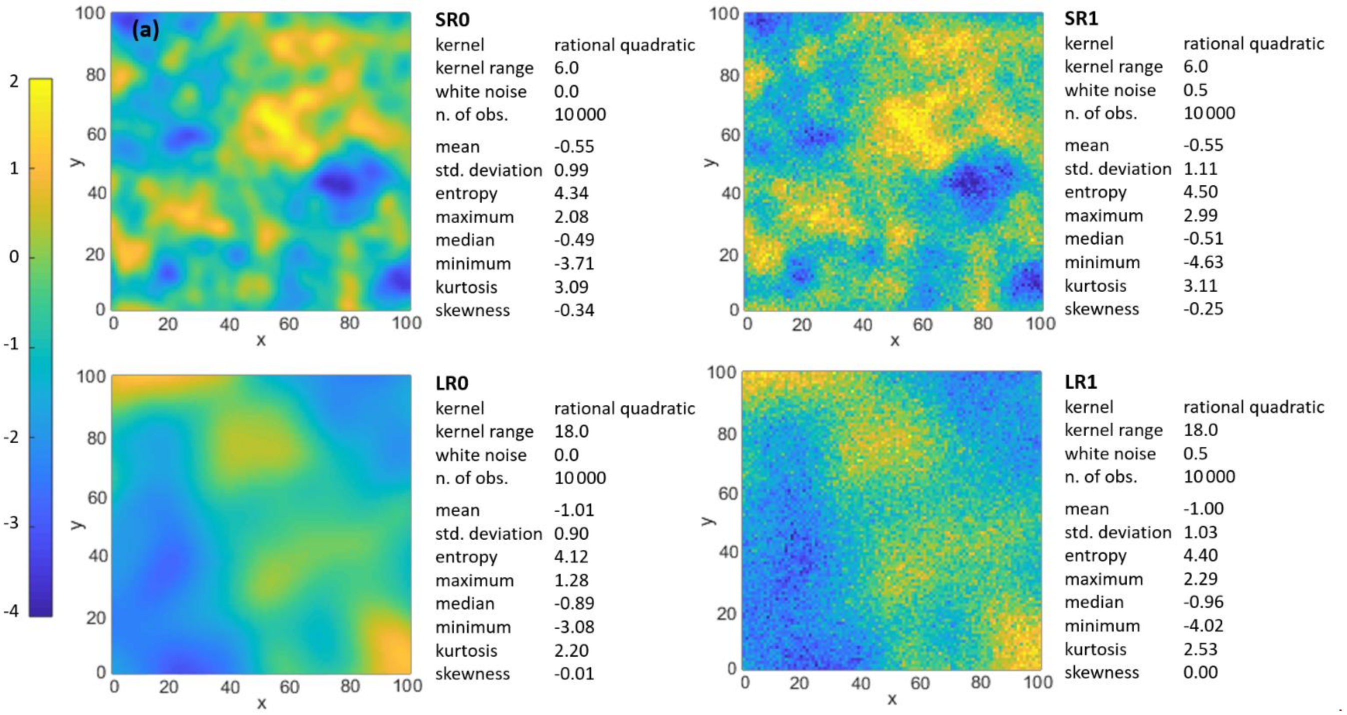


Figure 3: Examples of the different pooling operators. Illustration of: a) Normal PMFs $N(\mu,\sigma^2)$ to be combined; b) linear aggregation of (a), Eq. (4); c) log-linear aggregation of (a), Eq. (5); and d) log-linear aggregation of (b) and (c), Eq. (6).



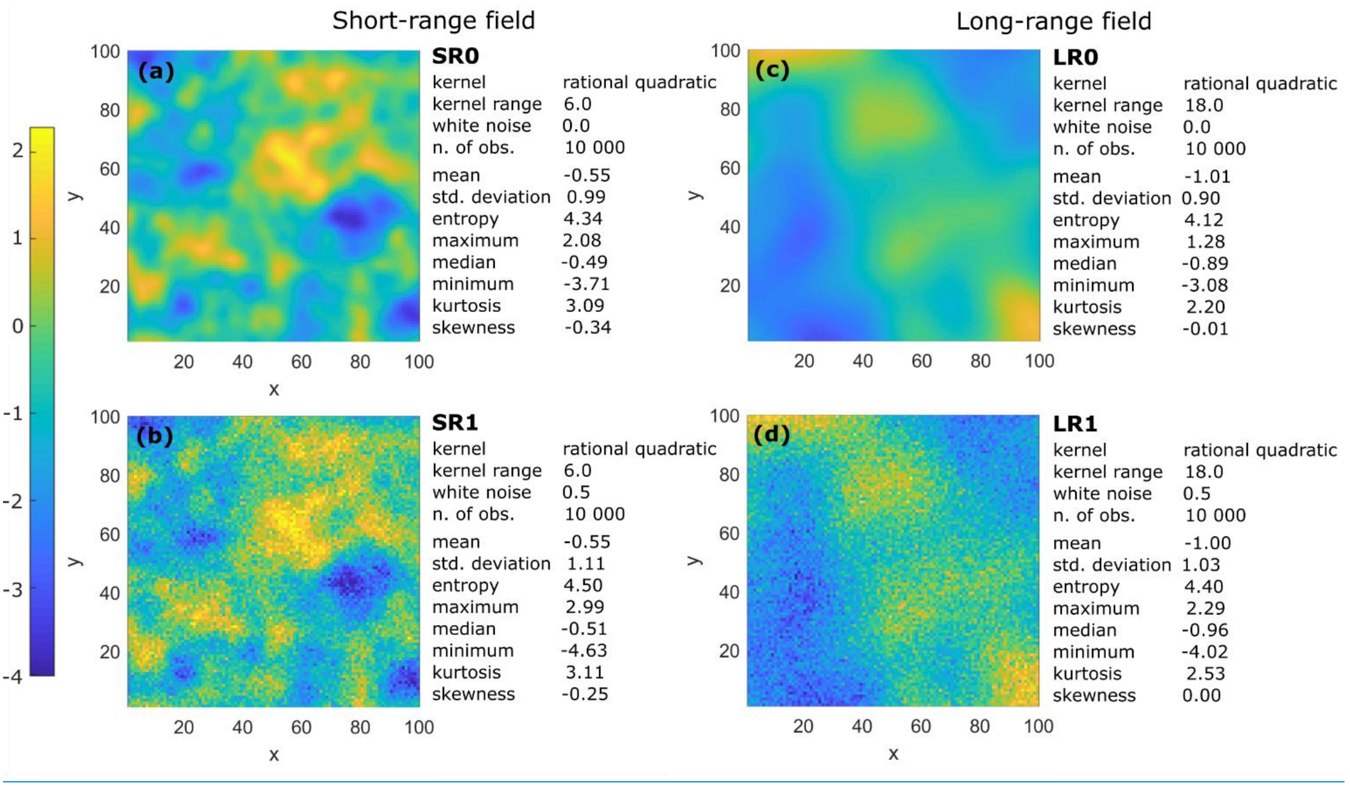
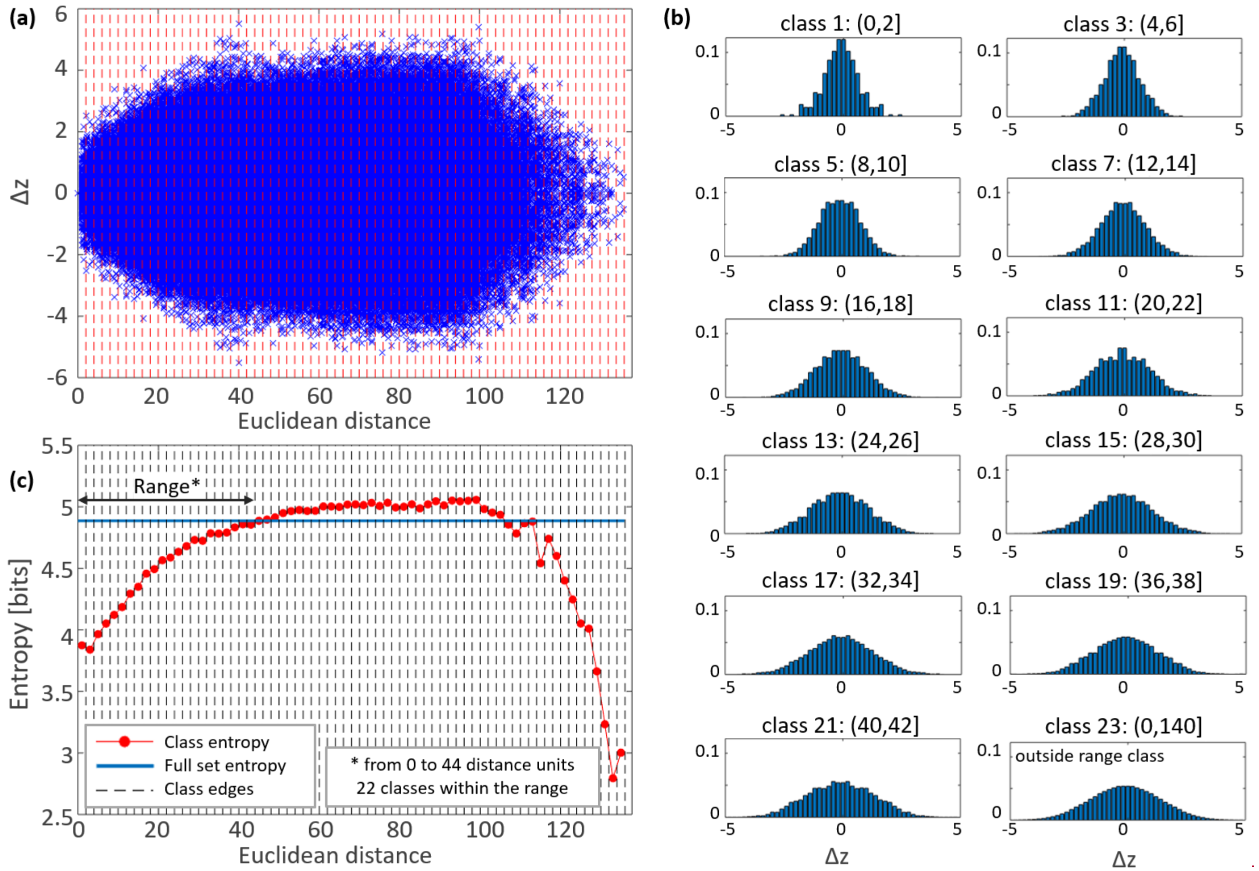


Figure 4: Synthetic fields and summary statistics: a) SR0, b) SR1, c) LR0, and d) LR1.


(a)
Δz
6
4
2
0
-2
-4
-6
0
20
40
60
80
100
120
Euclidean distance
(b)
class 1: (0,2]
class 3: (4,6]
class 5: (8,10]
class 7: (12,14]
class 9: (16,18]
class 11: (20,22]
class 13: (24,26]
class 15: (28,30]
class 17: (32,34]
class 19: (36,38]
class 21: (40,42]
class 23: (0,140]
outside range class
0.1
0
-5
0
5
Δz
Δz
(c)
Entropy [bits]
5.5
5
4.5
4
3.5
3
2.5
Range*
Class entropy
Full set entropy
Class edges
* from 0 to 44 distance units
22 classes within the range
0
20
40
60
80
100
120
Euclidean distance

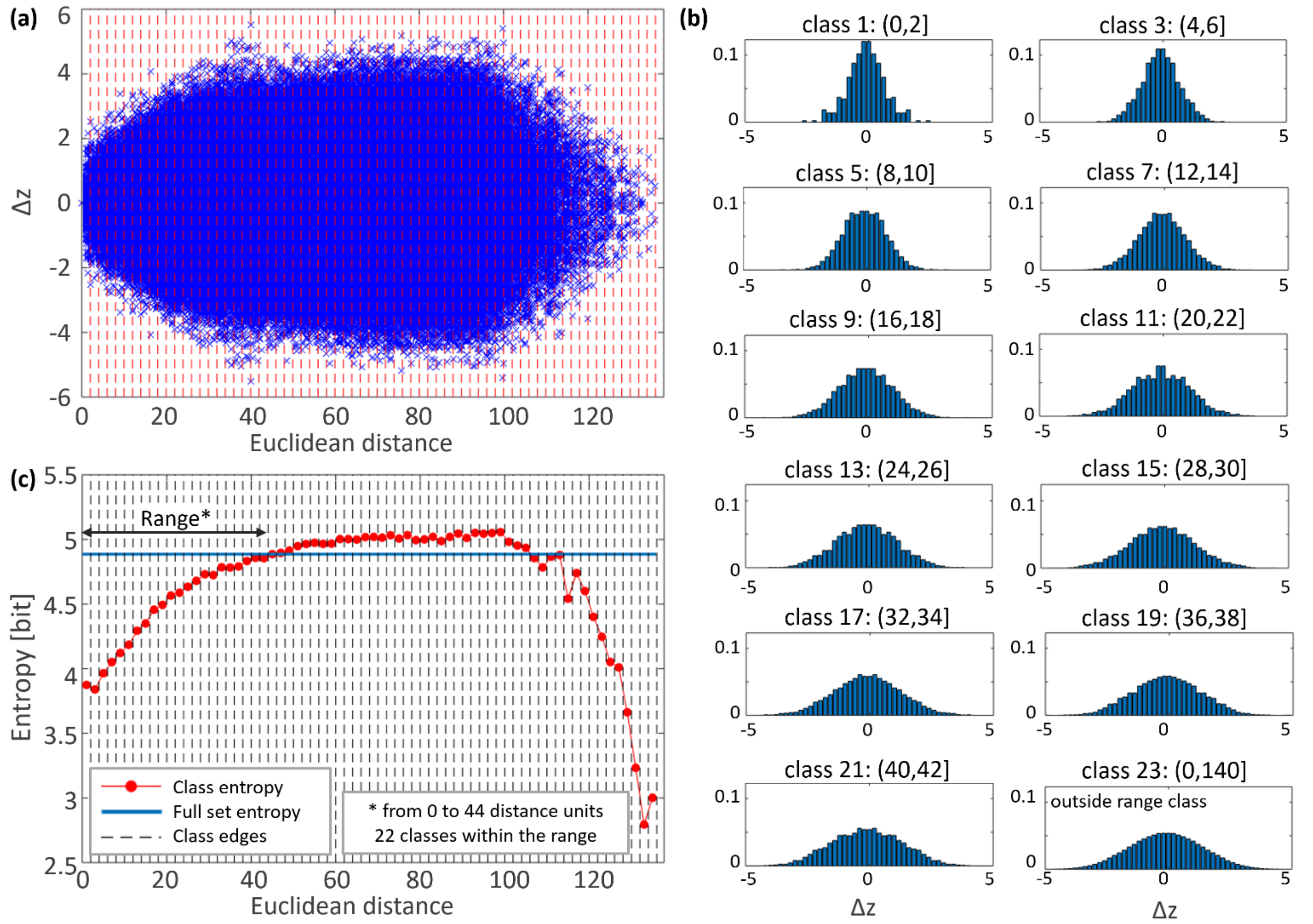
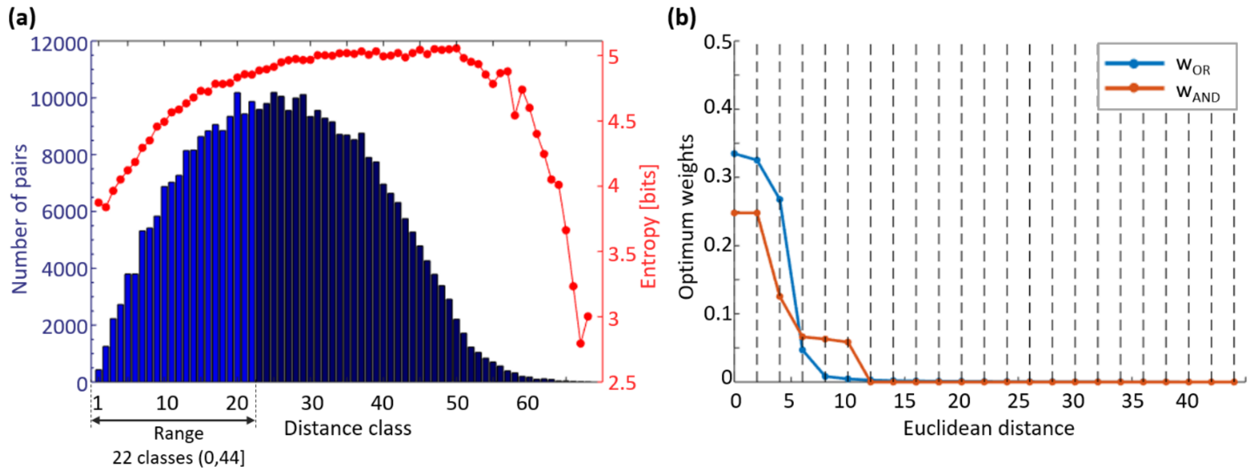


Figure 5: Spatial characterization of LR1-600: a) infogram cloud, b) Δ*z* PMFs by class, and c) infogram.



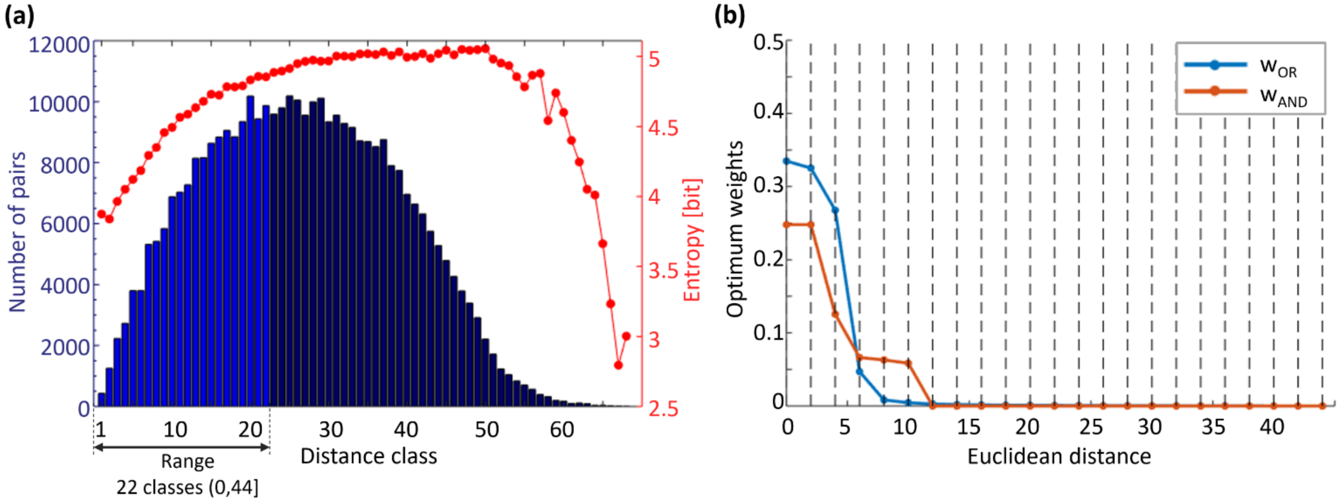
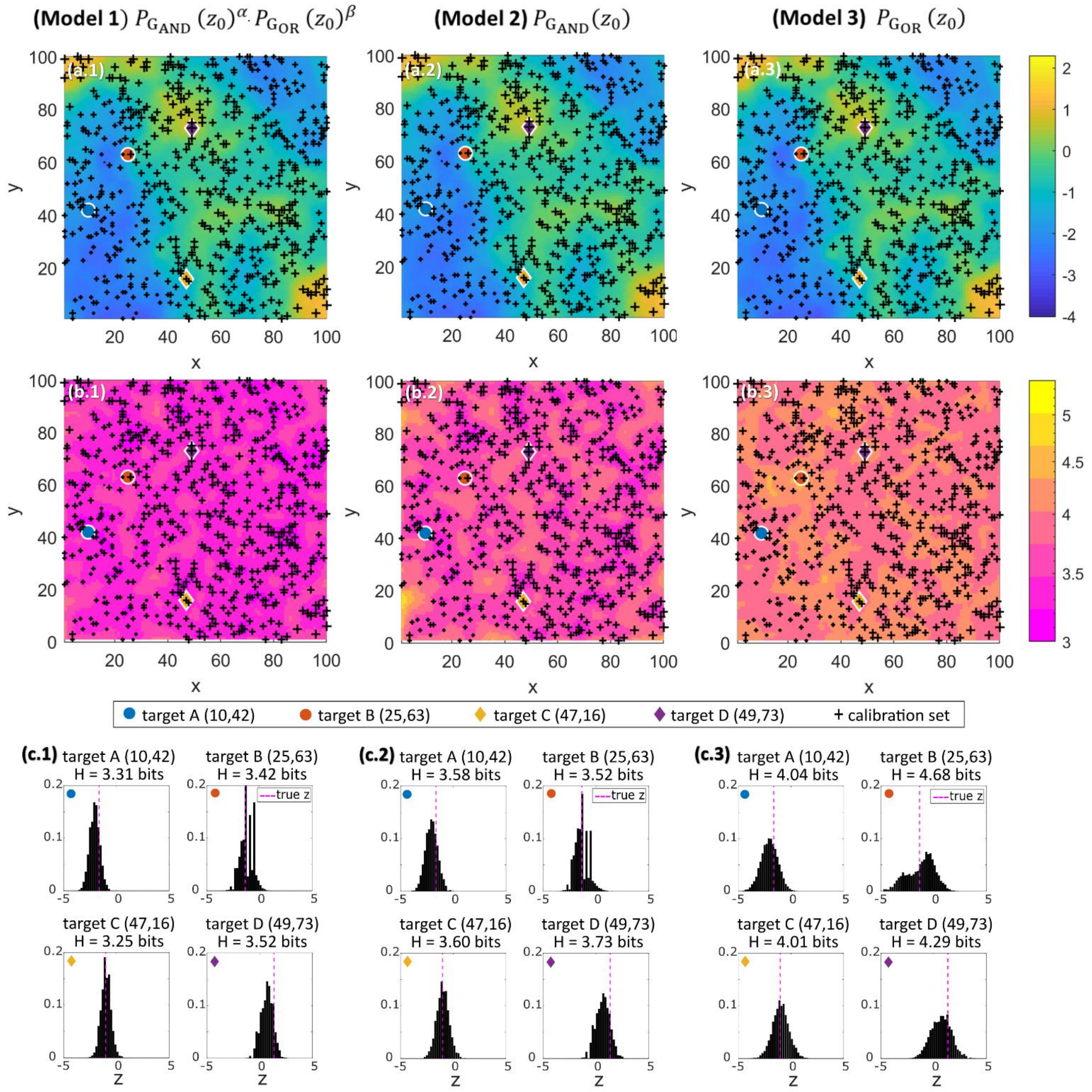


Figure 6: LR1-600: a) class cardinality~~,~~ and b) optimum weights, Eqs. (4) and (5).


(Model 1) $P_{G_{AND}}(z_0)^{\alpha} \cdot P_{G_{OR}}(z_0)^{\beta}$
(Model 2) $P_{G_{AND}}(z_0)$
(Model 3) $P_{G_{OR}}(z_0)$
(a.1)
(a.2)
(a.3)
(b.1)
(b.2)
(b.3)
x
y
0
20
40
60
80
100
2
1
0
-1
-2
-3
-4
5
4.5
4
3.5
3
● target A (10,42)
● target B (25,63)
◆ target C (47,16)
◆ target D (49,73)
+ calibration set
(c.1)
(c.2)
(c.3)
target A (10,42)
H = 3.31 bits
target B (25,63)
H = 3.42 bits
target C (47,16)
H = 3.25 bits
target D (49,73)
H = 3.52 bits
target A (10,42)
H = 3.58 bits
target B (25,63)
H = 3.52 bits
target C (47,16)
H = 3.60 bits
target D (49,73)
H = 3.73 bits
target A (10,42)
H = 4.04 bits
target B (25,63)
H = 4.68 bits
target C (47,16)
H = 4.01 bits
target D (49,73)
H = 4.29 bits
true z
0.2
0.1
0
-5
0
5
z

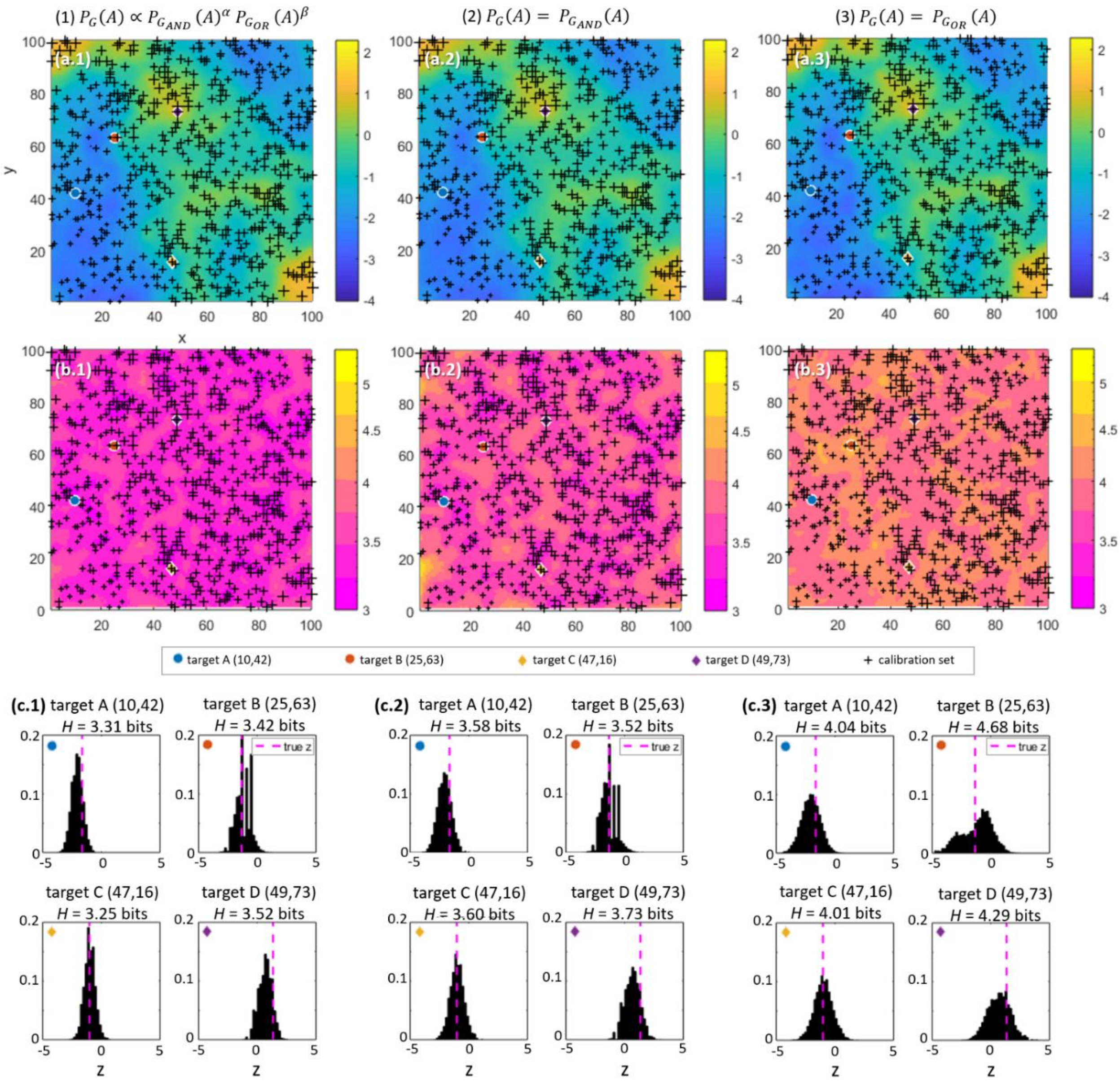
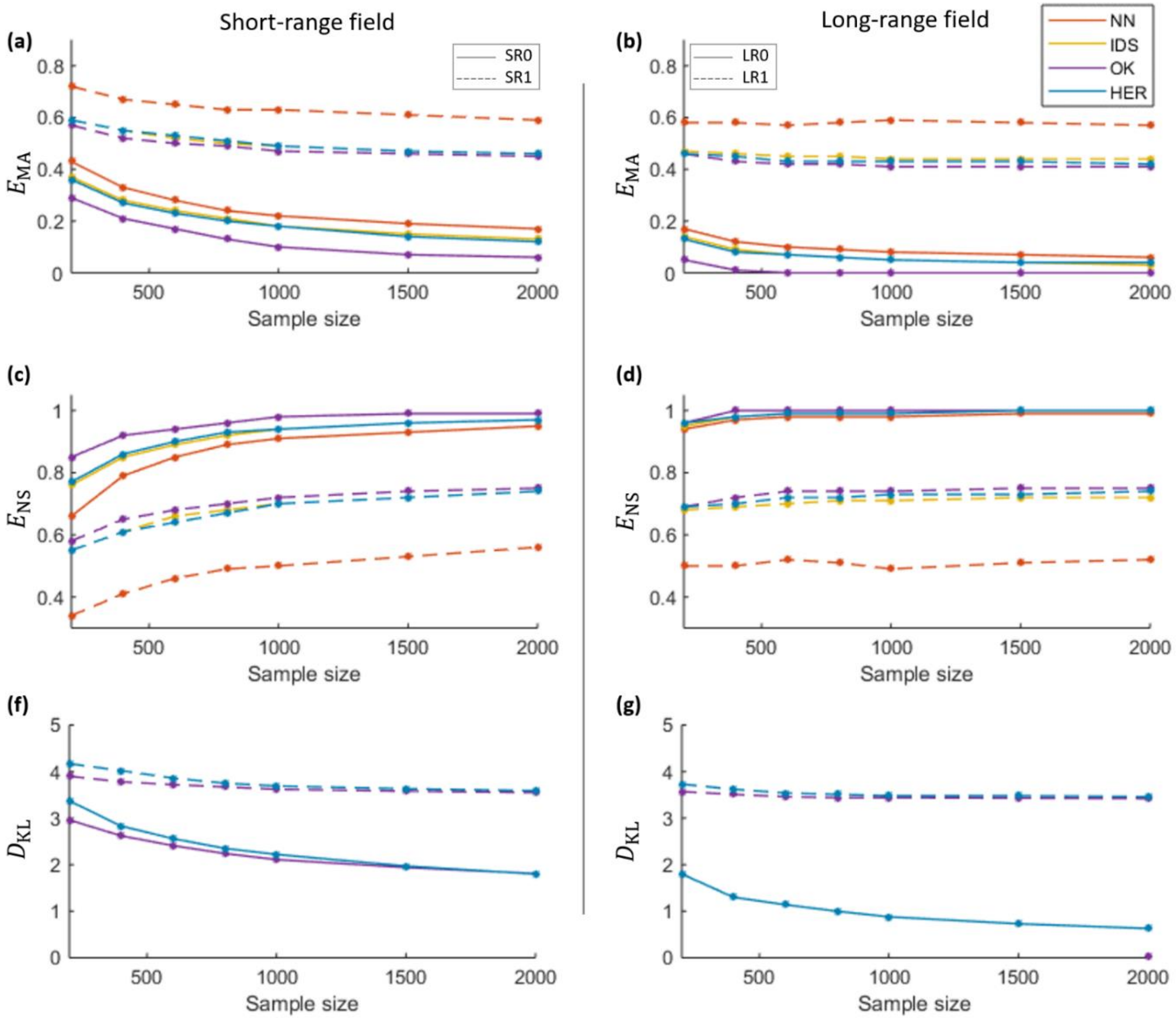


Figure 7: LR1-600 results: a) ~~predicted mean~~E-type estimate of z, b) entropy map (bits), and c) z~~-~~ PMF prediction for selected points. The first, second and third columns of the panel refer to the results of model 1 (AND/OR), model 2 (AND), and model 3 (OR), respectively.


Short-range field
Long-range field
NN
IDS
OK
HER
(a)
SR0
SR1
E_{MA}
0.8
0.6
0.4
0.2
0
500
1000
1500
2000
Sample size
(b)
LR0
LR1
(c)
E_{NS}
1
0.8
0.6
0.4
(d)
(f)
D_{KL}
5
4
3
2
1
0
(g)

845

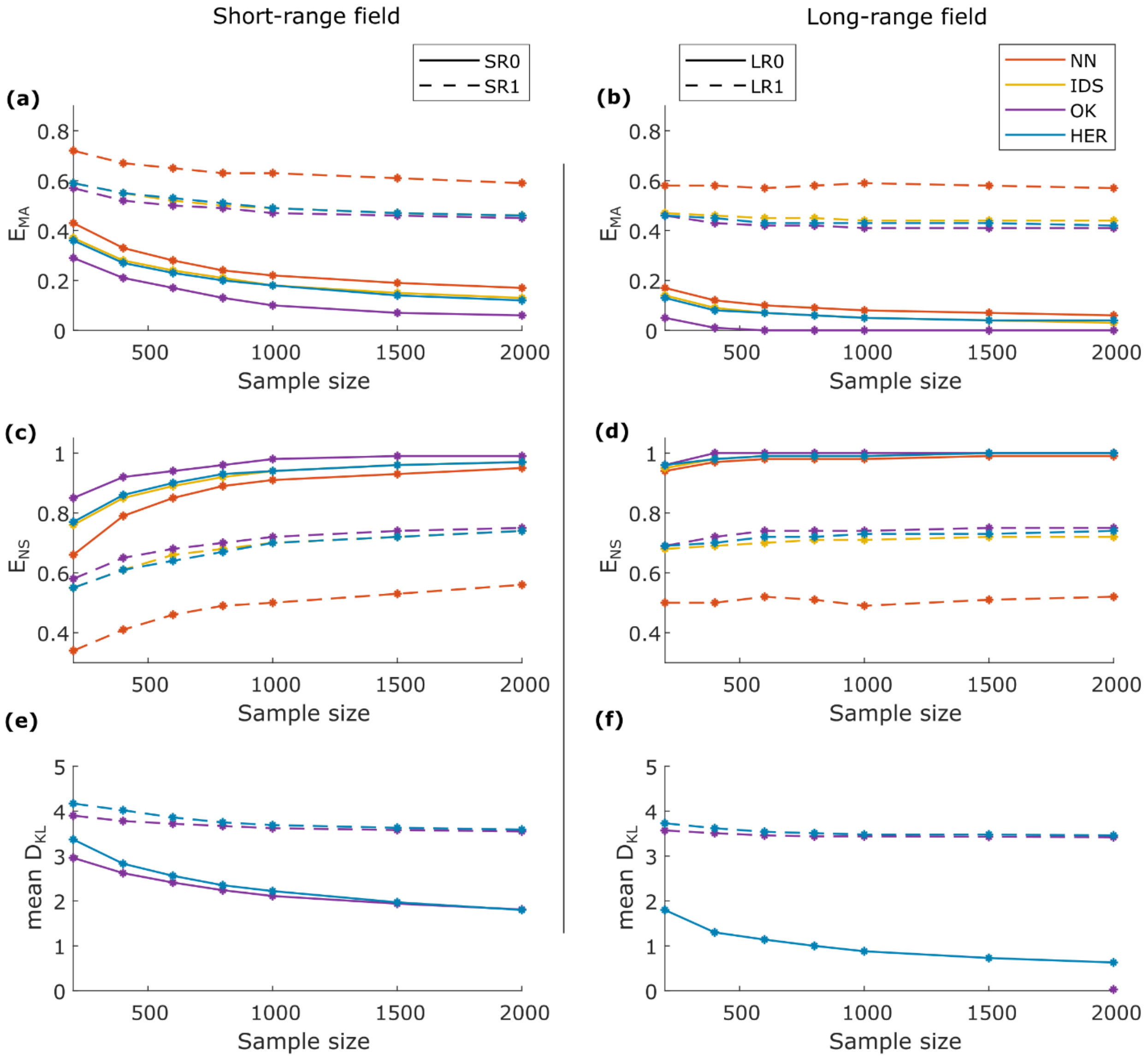


Figure 8: Performance comparison of NN, IDS, OK and HER: a,b) mean absolute error, c,d) Nash–Sutcliffe efficiency, and e,f) -Kullback–-Leibler divergence scoring rule, for the SR datasets in the left column and the LR datasets in the right. Continuous line refers to datasets without noise and dashed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

Supplementary material for

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

Supplement S1: Summary statistics of the resampled datasets

Table S1.1 and Table S1.2 summarize the statistics of the learning, validation, test, and full datasets.

Table S1.1: Summary statistics of the resampled datasets – Short-range dataset (SR0 and SR1).

Sample size	200	400	600	800	1000	1500	2000	2000 (val. set)	2000 (test set)	10 000 (full set)
SR0										
mean	-0.57	-0.59	-0.58	-0.59	-0.59	-0.58	-0.57	-0.53	-0.56	-0.55
sd.	1.05	1.06	1.02	1.01	0.99	0.99	0.99	0.99	1.00	0.99
H	4.27	4.38	4.34	4.33	4.31	4.32	4.32	4.31	4.34	4.34
max.	1.76	1.92	1.92	1.92	1.92	1.92	2.05	2.08	2.02	2.08
median	-0.42	-0.50	-0.51	-0.56	-0.54	-0.52	-0.52	-0.46	-0.50	-0.49
min.	-3.68	-3.68	-3.68	-3.68	-3.68	-3.68	-3.68	-3.67	-3.71	-3.71
kur.	3.21	3.04	3.12	3.15	3.17	3.14	3.12	3.18	3.07	3.09
sk.	-0.62	-0.43	-0.41	-0.35	-0.35	-0.32	-0.30	-0.36	-0.33	-0.34
SR1										
mean	-0.52	-0.54	-0.55	-0.57	-0.57	-0.57	-0.56	-0.54	-0.54	-0.55
sd.	1.17	1.17	1.14	1.12	1.11	1.10	1.10	1.11	1.12	1.11
H	4.46	4.54	4.51	4.50	4.49	4.49	4.49	4.49	4.52	4.50
max.	2.50	2.70	2.70	2.70	2.70	2.70	2.99	2.96	2.86	2.99
median	-0.36	-0.51	-0.51	-0.55	-0.56	-0.54	-0.53	-0.51	-0.48	-0.51
min.	-3.66	-3.66	-3.66	-3.84	-3.84	-4.01	-4.01	-4.63	-4.25	-4.63
kur.	2.82	2.83	2.93	2.94	2.99	3.03	3.04	3.24	3.09	3.11
sk.	-0.40	-0.15	-0.19	-0.19	-0.18	-0.20	-0.20	-0.28	-0.26	-0.25

sd. = standard deviation; *H* = entropy; max. = maximum; min. = minimum; kur. = kurtosis; sk. = skewness.

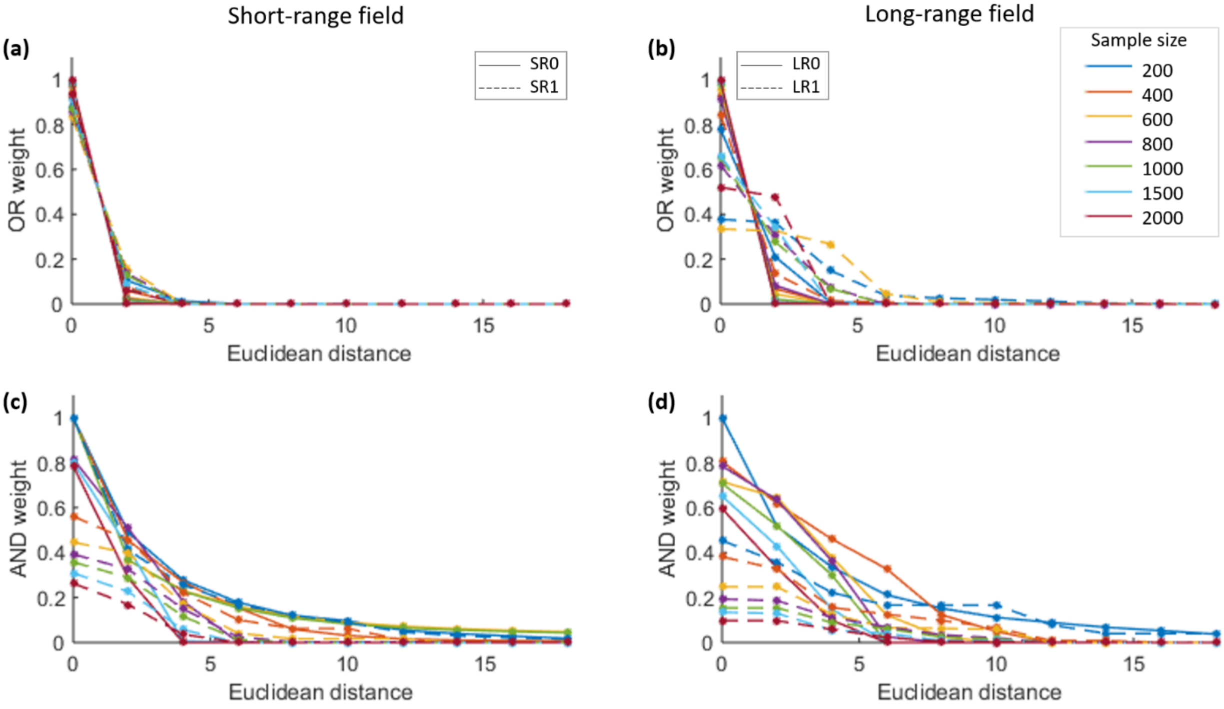
Table S1.2: Summary statistics of the resampled datasets – Long-range dataset (LR0 and LR1).

Sample size	200	400	600	800	1000	1500	2000	2000 (val. set)	2000 (test set)	10 000 (full set)
					LR0					
mean	-0.98	-0.96	-1.03	-1.01	-1.01	-1.01	-1.02	-1.00	-1.02	-1.01
sd.	0.90	0.88	0.89	0.89	0.90	0.91	0.91	0.90	0.91	0.90
H	3.99	4.02	4.07	4.09	4.09	4.11	4.11	4.11	4.12	4.12
max.	1.04	1.15	1.23	1.23	1.23	1.23	1.23	1.28	1.27	1.28
median	-0.77	-0.81	-0.92	-0.92	-0.91	-0.91	-0.92	-0.88	-0.89	-0.89
min.	-2.78	-2.78	-3.07	-3.07	-3.07	-3.08	-3.08	-3.00	-3.07	-3.08
kur.	2.11	2.18	2.26	2.24	2.21	2.16	2.20	2.22	2.16	2.20
sk.	-0.09	-0.07	0.02	0.02	0.03	0.03	0.03	-0.03	0.00	-0.01
					LR1					
mean	-0.92	-0.91	-0.99	-1.00	-1.00	-1.01	-1.01	-1.01	-1.00	-1.00
sd.	0.98	1.00	1.01	1.02	1.03	1.04	1.03	1.05	1.03	1.03
H	4.21	4.31	4.34	4.37	4.38	4.40	4.39	4.41	4.39	4.40
max.	1.40	1.87	1.87	1.87	1.96	1.96	2.00	2.29	2.14	2.29
median	-0.88	-0.91	-0.97	-0.98	-0.99	-0.99	-0.98	-0.98	-0.96	-0.96
min.	-3.19	-3.65	-3.65	-3.74	-3.74	-3.74	-3.95	-4.02	-3.75	-4.02
kur.	2.51	2.67	2.56	2.56	2.59	2.50	2.53	2.59	2.44	2.53
sk.	-0.09	0.02	0.06	0.04	0.06	0.05	0.04	-0.02	0.02	0.00

sd. = standard deviation; *H* = entropy; max. = maximum; min. = minimum; kur. = kurtosis; sk. = skewness.

Supplement S2: Parameter tuning

This supplement consolidates the final parameters used in the models presented in Sect. 4.2. Particularly for HER, Fig. S2.1 presents the final weights optimized for Eqs. (4) and (5). It was limited to 18 grid units (nine distance classes), due to the small contribution of the faraway classes. Similarly, Fig. S2.2 shows α and β weights of Eq. (6~~7~~). Finally, Table S2.1 and Table S2.2 summarize the calibrated parameters obtained for each model (varying method, sample size and dataset type).


Short-range field
Long-range field
(a)
(b)
(c)
(d)
SR0
SR1
LR0
LR1
Sample size
200
400
600
800
1000
1500
2000
OR weight
AND weight
Euclidean distance
0
0.2
0.4
0.6
0.8
1
0
5
10
15

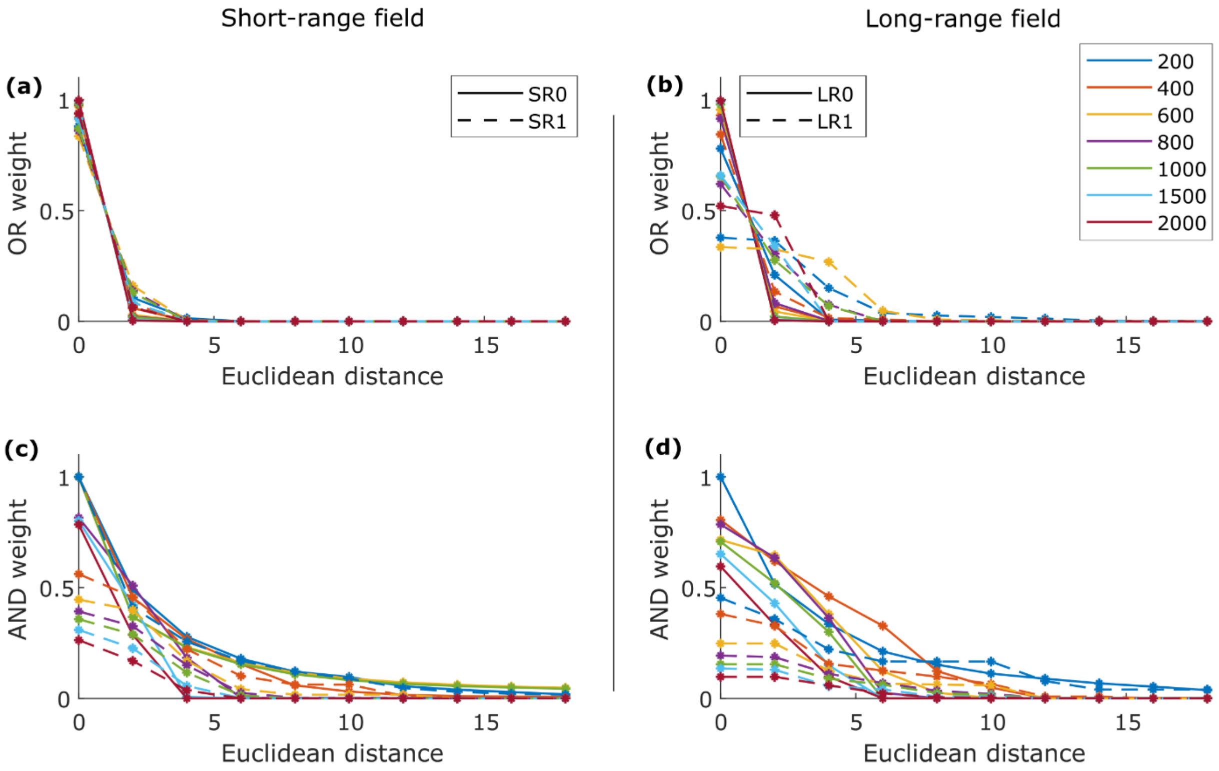
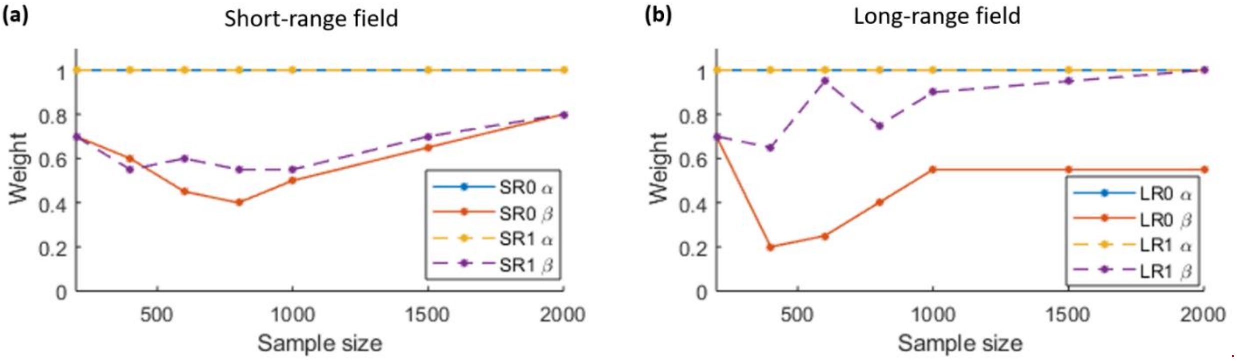

Short-range field
Long-range field
(a)
(b)
(c)
(d)
SR0
SR1
LR0
LR1
200
400
600
800
1000
1500
2000
OR weight
AND weight
Euclidean distance
0
0.5
1
0
5
10
15

Figure S2.1: HER optimized weights by distance class: a,b) w_{OR}, Eq. ~~-~~(4), and c,d) w_{AND}, Eq. ~~-~~(5). SR datasets on the left panel and LR datasets on the right panel. Continuous line refers to datasets without noise and dashed lines to datasets with noise.



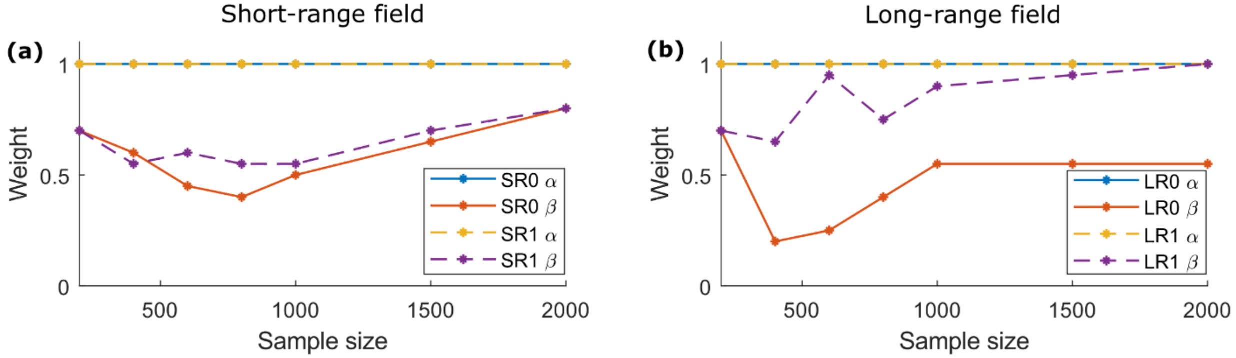


Figure S2.2. HER α and β weights by sample size, Eq. ~~-~~(6~~7~~): a) SR datasets on the left panel~~,~~ and b) LR datasets on the right panel. Continuous line refers to datasets without noise and dashed lines to datasets with noise.

Table S2.1: Method calibration by sample size – Parameters of the models for the short-range dataset (SR0 and SR1).

Model sample size		200	400	600	800	1000	1500	2000
Method	**Parameter~~1~~**				**SR0**			
NN	n.n.	1	1	1	1	1	1	1
IDS	exp.	2	2	2	2	2	2	2
OK	n.n.	12	12	12	12	12	12	12
	lag width	2	2	2	2	2	2	2
	variogram	Spherical	Spherical	Spherical	Spherical	Spherical	Spherical	Spherical
	eff. range	35.99	35.43	33.63	33.50	33.13	33.21	33.65
	nugget	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	sill	1.24	1.28	1.16	1.13	1.11	1.09	1.08
	max. lag	60	60	60	60	60	60	60
	n.n. [min.,max.]	[3,20]	[3,20]	[3,20]	[3,20]	[3,20]	[3,20]	[3,20]
HER	n.n.	12	12	12	12	12	12	12
	class width	2	2	2	2	2	2	2
	bin widths (z, Δz)	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	model range	36.00	24.00	26.00	26.00	26.00	26.00	26.00
	α	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	β	0.70	0.60	0.45	0.40	0.50	0.65	0.80
Method	**Parameter~~1~~**				**SR1**			
NN	n.n.	1	1	1	1	1	1	1
IDS	exp.	2	2	2	2	2	2	2
OK	n.n.	12	12	12	12	12	12	12
	lag width	2	2	2	2	2	2	2
	variogram	Spherical	Spherical	Spherical	Spherical	Spherical	Spherical	Spherical
	eff. range	43.53	35.81	35.43	34.69	32.70	32.18	33.30
	nugget	0.28	0.15	0.18	0.18	0.17	0.17	0.20
	sill	1.29	1.39	1.25	1.22	1.19	1.16	1.12
	max. lag	60	60	60	60	60	60	60
	n.n. [min.,max.]	[3,20]	[3,20]	[3,20]	[3,20]	[3,20]	[3,20]	[3,20]
HER	n.n.	12	12	12	12	12	12	12
	class width	2	2	2	2	2	2	2
	bin widths (z, Δz)	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	model range	38.00	26.00	26.00	26.00	26.00	26.00	26.00
	α	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	β	0.70	0.55	0.60	0.55	0.55	0.70	0.80

~~1~~n.n. = number of neighbors; exp. = exponent of the weighting function; eff. range = effective range; max. = maximum; min. = minimum.

Table S2.2: Method calibration by sample size – Parameters of the models for the long-range dataset (LR0 and LR1).

Model sample size		200	400	600	800	1000	1500	2000
Method	**Parameter~~1~~**	**LR0**						
NN	n.n.	1	1	1	1	1	1	1
IDS	exp.	2	2	2	2	2	2	2
OK	n.n.	12	12	12	12	12	12	12
	lag width	2	2	2	2	2	2	2
	variogram	Gaussian	Gaussian	Gaussian	Gaussian	Gaussian	Gaussian	Gaussian
	eff. range	67.47	66.93	69.10	68.23	69.12	71.82	73.01
	nugget	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	sill	1.06	0.99	1.03	1.03	1.05	1.10	1.10
	max. lag	100	100	100	100	100	100	100
	n.n. [min.,max.]	[3,20]	[3,20]	[3,20]	[3,20]	[3,20]	[3,20]	[3,20]
HER~~2~~	n.n.	12	12	12	12	12	12	12
	class width	2	2	2	2	2	2	2
	bin widths (z, Δz)	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	model range	46.00	48.00	48.00	46.00	46.00	48.00	48.00
	α	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	β	0.70	0.20	0.25	0.40	0.55	0.55	0.55
Method	**Parameter~~1~~**	**LR1**						
NN	n.n.	1	1	1	1	1	1	1
IDS	exp.	2	2	2	2	2	2	2
OK	n.n.	12	12	12	12	12	12	12
	lag width	2	2	2	2	2	2	2
	variogram	Gaussian	Gaussian	Gaussian	Gaussian	Gaussian	Gaussian	Gaussian
	eff. range	81.79	76.14	71.43	69.02	74.43	78.75	78.05
	nugget	0.29	0.31	0.29	0.28	0.30	0.29	0.29
	sill	0.99	0.95	0.98	1.00	1.03	1.10	1.08
	max. lag	100	100	100	100	100	100	100
	n.n. [min.,max.]	[3,20]	[3,20]	[3,20]	[3,20]	[3,20]	[3,20]	[3,20]
HER	n.n.	12	12	12	12	12	12	12
	class width	2	2	2	2	2	2	2
	bin widths (z, Δz)	0.2	0.2	0.2	0.2	0.2	0.2	0.2
	model range	48.00	46.00	44.00	44.00	44.00	46.00	46.00
	α	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	β	0.70	0.65	0.95	0.75	0.90	0.95	1.00

~~1~~n.n. = number of neighbors; exp. = exponent of the weighting function; eff. range = effective range; max. = maximum; min. = minimum.

Supplement S3: Summary statistics of the model predictions

This supplement summarizes the statistics of the deterministic predictions (mean of z) for the test set by method and learning sets (from 200 to 2000 observations). HER outcomes refer to the AND/OR aggregation. The four random fields types are presented from Table S3.1 to Table S3.4. Finally, Fig. ~~-~~S3.1 illustrates their residue correlation (obtained by calculating the Pearson correlation coefficient between the true values and the residue of the predictions).

Table S3.1: Summary statistics of the prediction on test set by model – Short-range dataset without noise (SR0).

Method	Statistics~~[1]~~	200	400	600	800	1000	1500	2000
				SR0				
NN	mean	-0.54	-0.55	-0.56	-0.56	-0.56	-0.56	-0.56
	sd.	1.01	1.03	1.01	1.00	1.00	1.01	1.00
	H	4.17	4.33	4.31	4.31	4.31	4.34	4.33
	max.	1.76	1.92	1.92	1.91	1.91	1.91	1.91
	median	-0.44	-0.47	-0.57	-0.57	-0.53	-0.53	-0.52
	min.	-3.68	-3.68	-3.68	-3.68	-3.68	-3.68	-3.68
	kur.	3.37	3.13	3.06	3.04	3.07	3.08	3.08
	sk.	-0.56	-0.43	-0.36	-0.30	-0.32	-0.30	-0.32
IDS	mean	-0.54	-0.57	-0.58	-0.59	-0.57	-0.57	-0.57
	sd.	0.79	0.88	0.89	0.90	0.91	0.93	0.94
	H	3.96	4.13	4.16	4.19	4.21	4.24	4.26
	max.	1.58	1.80	1.79	1.80	1.80	1.79	1.80
	median	-0.55	-0.53	-0.53	-0.56	-0.53	-0.54	-0.53
	min.	-3.49	-3.49	-3.51	-3.53	-3.54	-3.56	-3.58
	kur.	3.56	3.28	3.27	3.17	3.15	3.13	3.10
	sk.	-0.44	-0.37	-0.37	-0.32	-0.32	-0.30	-0.30
OK	mean	-0.53	-0.56	-0.56	-0.57	-0.56	-0.56	-0.56
	sd.	0.86	0.92	0.93	0.94	0.95	0.97	0.97
	H	4.11	4.21	4.24	4.26	4.27	4.30	4.30
	max.	1.63	1.86	1.90	1.90	1.90	1.90	1.90
	median	-0.47	-0.49	-0.49	-0.52	-0.51	-0.51	-0.51
	min.	-3.60	-3.56	-3.57	-3.63	-3.66	-3.67	-3.67
	kur.	3.46	3.18	3.13	3.09	3.08	3.08	3.08
	sk.	-0.46	-0.41	-0.39	-0.34	-0.35	-0.32	-0.33
HER	mean	-0.54	-0.56	-0.58	-0.57	-0.57	-0.57	-0.57
	sd.	0.87	0.95	0.92	0.96	0.94	0.98	0.98
	H	4.08	4.23	4.21	4.26	4.24	4.31	4.31
	max.	1.70	1.82	1.81	1.83	1.82	1.83	1.86
	median	-0.50	-0.51	-0.54	-0.57	-0.54	-0.53	-0.53
	min.	-3.55	-3.55	-3.57	-3.61	-3.58	-3.59	-3.61
	kur.	3.54	3.18	3.22	3.10	3.13	3.10	3.07
	sk.	-0.54	-0.43	-0.37	-0.31	-0.32	-0.30	-0.31

~~[1]~~sd. = standard deviation; *H* = entropy; max. = maximum; min. = minimum; kur. = kurtosis; sk. = skewness.

Table S3.2: Summary statistics of the prediction on test set by model – Short-range dataset with noise (SR1).

Method	Statistics~~[1]~~	200	400	600	800	1000	1500	2000
				SR1				
NN	mean	-0.50	-0.52	-0.55	-0.55	-0.56	-0.55	-0.56
	sd.	1.15	1.16	1.14	1.14	1.13	1.11	1.11
	H	4.45	4.51	4.49	4.50	4.50	4.48	4.49
	max.	2.50	2.70	2.70	2.70	2.70	2.70	2.99
	median	-0.43	-0.51	-0.53	-0.54	-0.54	-0.53	-0.54
	min.	-3.66	-3.66	-3.66	-3.84	-3.84	-3.84	-4.00
	kur.	2.86	2.79	2.92	2.91	2.90	2.97	2.97
	sk.	-0.27	-0.05	-0.05	-0.09	-0.14	-0.13	-0.18
IDS	mean	-0.49	-0.53	-0.55	-0.58	-0.56	-0.56	-0.56
	sd.	0.85	0.92	0.92	0.95	0.95	0.96	0.96
	H	4.09	4.22	4.24	4.28	4.27	4.29	4.30
	max.	2.19	2.37	2.34	2.28	2.27	2.19	2.07
	median	-0.47	-0.47	-0.50	-0.53	-0.51	-0.53	-0.52
	min.	-3.42	-3.30	-3.29	-3.50	-3.52	-3.59	-3.55
	kur.	3.17	2.84	2.97	2.86	2.91	2.98	2.92
	sk.	-0.23	-0.13	-0.19	-0.21	-0.21	-0.22	-0.23
OK	mean	-0.49	-0.52	-0.54	-0.57	-0.55	-0.56	-0.56
	sd.	0.79	0.90	0.91	0.93	0.93	0.94	0.94
	H	3.99	4.20	4.21	4.24	4.25	4.25	4.25
	max.	1.58	2.30	2.22	2.20	2.21	2.17	1.90
	median	-0.48	-0.46	-0.48	-0.51	-0.49	-0.49	-0.49
	min.	-3.17	-3.16	-3.19	-3.31	-3.44	-3.51	-3.45
	kur.	3.22	2.82	2.84	2.76	2.85	2.94	2.89
	sk.	-0.22	-0.19	-0.24	-0.25	-0.26	-0.27	-0.26
HER	mean	-0.50	-0.53	-0.54	-0.57	-0.55	-0.56	-0.56
	sd.	0.90	0.96	0.98	0.98	0.97	0.97	0.97
	H	4.16	4.28	4.31	4.33	4.31	4.31	4.30
	max.	2.24	2.31	2.35	2.28	2.28	2.26	2.00
	median	-0.47	-0.48	-0.50	-0.54	-0.51	-0.53	-0.52
	min.	-3.32	-3.32	-3.38	-3.46	-3.45	-3.55	-3.54
	kur.	3.11	2.70	2.89	2.82	2.85	2.98	2.89
	sk.	-0.27	-0.13	-0.14	-0.16	-0.20	-0.19	-0.24

~~[1]~~sd. = standard deviation; *H* = entropy; max. = maximum; min. = minimum; kur. = kurtosis; sk. = skewness.

Table S3.3: Summary statistics of the prediction on test set by model – Long-range dataset without noise (LR0).

Method	Statistics~~[1]~~	200	400	600	800	1000	1500	2000
				LR0				
NN	mean	-1.03	-1.02	-1.01	-1.02	-1.02	-1.01	-1.02
	sd.	0.91	0.91	0.91	0.91	0.91	0.91	0.91
	H	3.98	4.06	4.10	4.11	4.11	4.12	4.11
	max.	1.04	1.15	1.15	1.23	1.23	1.23	1.23
	median	-0.92	-0.91	-0.90	-0.90	-0.90	-0.90	-0.90
	min.	-2.78	-2.78	-3.07	-3.07	-3.07	-3.08	-3.08
	kur.	2.10	2.13	2.20	2.18	2.20	2.15	2.16
	sk.	0.00	0.02	0.03	0.02	0.03	0.01	0.00
IDS	mean	-1.04	-1.02	-1.02	-1.02	-1.02	-1.02	-1.02
	sd.	0.85	0.87	0.88	0.89	0.89	0.90	0.90
	H	3.91	3.98	4.05	4.07	4.07	4.08	4.09
	max.	0.99	1.08	1.14	1.15	1.16	1.14	1.14
	median	-0.86	-0.88	-0.89	-0.88	-0.88	-0.88	-0.89
	min.	-2.72	-2.71	-3.01	-3.01	-3.01	-3.02	-3.02
	kur.	1.95	2.01	2.11	2.12	2.12	2.11	2.13
	sk.	-0.12	-0.03	-0.03	-0.01	-0.01	-0.02	-0.01
OK	mean	-1.04	-1.02	-1.02	-1.02	-1.02	-1.02	-1.02
	sd.	0.91	0.91	0.91	0.91	0.91	0.91	0.91
	H	4.11	4.11	4.12	4.12	4.12	4.12	4.12
	max.	1.34	1.28	1.24	1.28	1.27	1.27	1.27
	median	-0.93	-0.88	-0.89	-0.89	-0.89	-0.89	-0.89
	min.	-2.89	-2.97	-3.08	-3.08	-3.07	-3.07	-3.07
	kur.	2.12	2.15	2.17	2.17	2.16	2.16	2.16
	sk.	0.01	0.01	0.01	0.01	0.01	0.00	0.00
HER	mean	-1.04	-1.02	-1.02	-1.02	-1.02	-1.02	-1.02
	sd.	0.88	0.88	0.89	0.90	0.90	0.90	0.91
	H	3.98	4.03	4.07	4.09	4.09	4.09	4.09
	max.	1.02	1.13	1.14	1.22	1.20	1.15	1.15
	median	-0.89	-0.90	-0.90	-0.90	-0.90	-0.90	-0.90
	min.	-2.77	-2.78	-3.06	-3.07	-3.07	-3.08	-3.07
	kur.	2.02	2.09	2.17	2.16	2.16	2.13	2.14
	sk.	-0.05	0.00	0.00	0.00	0.01	-0.01	-0.01

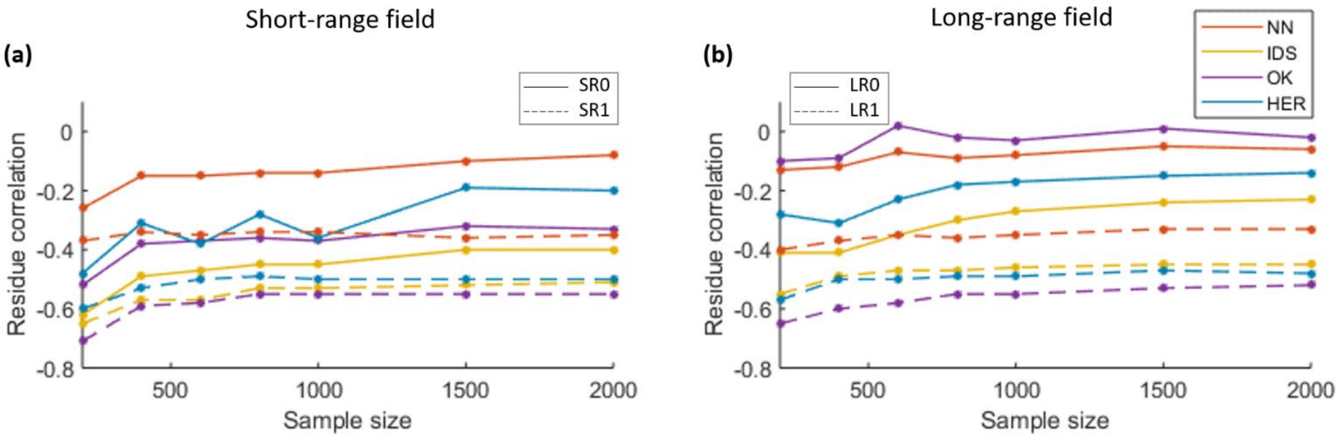
~~[1]~~sd. = standard deviation; *H* = entropy; max. = maximum; min. = minimum; kur. = kurtosis; sk. = skewness.

Table S3.4: Summary statistics of the prediction on test set by model – Long-range dataset with noise (LR1).

Method	Statistics~~[1]~~	200	400	600	800	1000	1500	2000
				LR1				
NN	mean	-1.00	-0.99	-1.00	-1.01	-1.01	-1.00	-1.01
	sd.	1.00	1.02	1.03	1.03	1.04	1.05	1.05
	H	4.23	4.33	4.36	4.35	4.39	4.40	4.40
	max.	1.40	1.87	1.87	1.87	1.87	1.87	1.87
	median	-0.90	-0.94	-0.97	-0.99	-0.99	-0.99	-0.98
	min.	-3.19	-3.65	-3.65	-3.65	-3.65	-3.65	-3.87
	kur.	2.50	2.66	2.56	2.57	2.57	2.51	2.49
	sk.	-0.11	0.03	0.02	0.10	0.08	0.06	0.03
IDS	mean	-0.99	-0.98	-0.99	-1.01	-1.00	-1.01	-1.01
	sd.	0.86	0.90	0.91	0.92	0.92	0.93	0.93
	H	4.04	4.14	4.14	4.16	4.18	4.17	4.16
	max.	1.21	1.76	1.48	1.45	1.61	1.54	1.43
	median	-0.79	-0.85	-0.85	-0.88	-0.90	-0.88	-0.90
	min.	-3.04	-3.12	-3.12	-3.12	-3.05	-3.15	-3.25
	kur.	2.21	2.39	2.28	2.31	2.32	2.26	2.26
	sk.	-0.26	0.01	0.04	0.06	0.05	0.05	0.03
OK	mean	-0.98	-0.96	-0.98	-1.00	-1.00	-1.01	-1.01
	sd.	0.79	0.83	0.85	0.86	0.87	0.88	0.89
	H	3.89	4.01	4.00	4.02	4.02	4.04	4.05
	max.	0.81	1.29	1.25	1.32	1.30	1.14	1.19
	median	-0.78	-0.81	-0.81	-0.84	-0.84	-0.86	-0.88
	min.	-2.85	-2.82	-2.74	-2.76	-2.69	-2.84	-2.92
	kur.	2.28	2.38	2.17	2.18	2.18	2.13	2.13
	sk.	-0.40	-0.10	-0.04	-0.01	-0.01	-0.01	-0.01
HER	mean	-0.99	-0.97	-0.98	-1.01	-1.00	-1.01	-1.01
	sd.	0.85	0.89	0.89	0.90	0.90	0.92	0.91
	H	4.01	4.11	4.07	4.11	4.11	4.12	4.11
	max.	1.20	1.64	1.32	1.33	1.36	1.30	1.30
	median	-0.80	-0.83	-0.83	-0.86	-0.89	-0.89	-0.89
	min.	-3.00	-2.98	-2.82	-2.90	-2.83	-2.98	-3.13
	kur.	2.21	2.46	2.23	2.28	2.27	2.23	2.23
	sk.	-0.28	0.03	0.02	0.05	0.04	0.05	0.02

~~[1]~~sd. = standard deviation; *H* = entropy; max. = maximum; min. = minimum; kur. = kurtosis; sk. = skewness.

Fig. S3.1~~Figure S3.1~~ illustrates for the residue correlation of the models calculated using the test set. The more negative the residue correlation, the greater the tendency of true *z* values being overestimated in low-valued regions of the field and underestimated in high-valued regions.



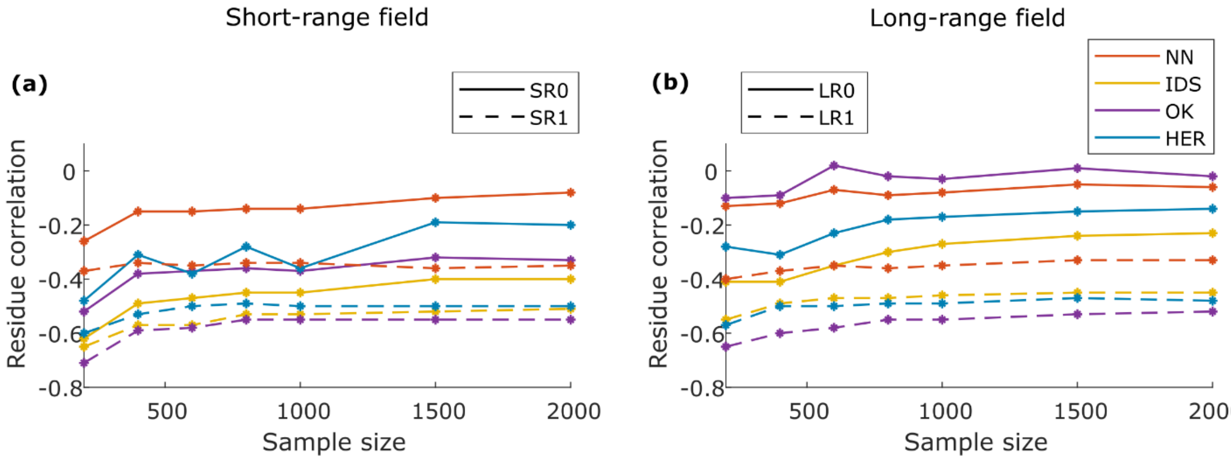


Figure S3.1: Performance comparison of NN, IDS, OK and HER: a) residue~~al~~ correlation for SR datasets~~,~~ and b) residue~~al~~ correlation for LR datasets. Continuous line refers to datasets without noise and dashed lines to datasets with noise.