INFORMATION THEORY FOR NONPARAMETRIC LEARNING AND PROBABILISTIC PREDICTION

APPLICATIONS IN EARTH SCIENCE AND GEOSTATISTICS

STEPHANIE THIESEN

Karlsruhe 2021

INFORMATION THEORY FOR NONPARAMETRIC LEARNING AND PROBABILISTIC PREDICTION

APPLICATIONS IN EARTH SCIENCE AND GEOSTATISTICS

Zur Erlangung des akademischen Grades einer

DOKTOR-INGENIEURIN
(Dr. -Ing.)

von der Fakultät für
Bauingenieur-, Geo- und Umweltwissenschaften

des Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von
Ing. Stephanie Thiesen, M. Sc.
aus Rio do Sul, Brasilien

Tag der mündlichen Prüfung:
6. Dezember 2021

REFERENT: Dr. Uwe Ehret
KORREFERENT: Prof. Dr. Olivier Eiff
KORREFERENT: Prof. Dr. J. Florian Wellmann

Karlsruhe 2021

To Mami, Papi, and Oma,
who inspired it and will not read it

[...] the cosmos is also within us.
We are made of star stuff.

— *Carl Sagan*

ACKNOWLEDGMENTS

Gratitude to the role of randomness in my life, in the specific way the stars have organized themselves bringing me immeasurable luck, invaluable persons, and immense opportunities in this infinite universe of possibilities.

To Uwe, for the great synergy, intrinsic motivation, and unconditional support. For listening, dreaming, and believing. Unquestionably (0 bits) a key person in setting the initial conditions of my PhD trajectory.

To my professors (not exclusively but especially Erwin, Florian, and Uwe), for providing me with solid theoretical foundations and for making a big difference in my path.

To my colleagues at KIT, for their unique scientific enthusiasm and for the not-always-scientific conversations.

To KIT, for the excellent academic infrastructure and funding of my PhD.

To my beloved friends. You know who you are :)

In these wonderful coincidences of randomness, to my greatest present ever Diego – my partner, my home. Thanks for reducing the entropy of our (so far) 2-person system and bringing more purpose to it.

Aos coautores das coisas mais importantes da minha vida. Mami, Papi, Oma, Nati, Thiago, Cauã, Allan, Gabi, Rita, e Indio, obrigada por brilharem do meu lado nesse pálido ponto azul.

To the universe, to the stars, to life.

[...] que façamos destes dias,
dias muito divertidos e inesquecíveis.

— *Mami*

ABSTRACT

Interestingly but challenging, Earth systems are often complex and their problems underdetermined. The lack of a complete understanding of relevant subsystems (complexity issue) and the impossibility of observing everything, everywhere, all the time (underdeterminism issue) lead to a considerable inferential and predictive uncertainty. In fact, uncertainty is part of Earth system science problems, and its quantification is, consequently, an essential aspect of geoscientific analysis and prediction. Additionally, ignoring uncertainty by deterministic models or strong parametric assumptions increases rigidity in the model (as a counterpoint to generality). As a consequence, rigid models can result in overly constrained and overconfident solutions, and therefore, in a suboptimal use of available data. In this fashion, to deal with uncertainty arising from the lack of knowledge or data, probabilistic inference and uncertainty quantification play a central role when modeling or analyzing such complex and underdetermined systems. Uncertainty and information can be objectively quantified by information theory, which, when combined with nonparametric probabilistic modeling, provides a proper framework for evaluating the information content of data and models. In addition, it helps to overcome the issue of using rigid models that, to a certain degree, ignore uncertainty and add information not present in data (or lose available information).

This thesis is motivated and framed by exactly this quest: to propose and validate a nonparametric, probabilistic framework for Earth science problems firmly rooted in concepts from information theory. For that, predictive relations are expressed by multivariate, empirical probability distributions directly derived from data, and information theory is used to explicitly calculate and compare the information content from various sources in a universal unit. Three typical Earth science problems are revisited through the lens of information theory. The testbed problems comprise descriptive and inferential levels, and deal with different data types (continuous or categorical), domains (spatial or temporal observations), sample sizes, and spatial dependence properties. First of all, a nonparametric approach for rainfall-runoff event identification is proposed, tested on a real-world dataset, and compared to a physically-based model (chapter 2). The findings of this study have contributed to propose a distribution-free framework for geostatistics in chapter 3, whose properties are tested on a synthetic dataset and compared to ordinary kriging. Finally, in chapter 4, the proposed nonparametric geostatistical method is adapted to handle categorical data and to simulate field properties. It is tested on a real-world dataset for classifying the risk of soil contamination by lead, and its characteristics are compared to indicator kriging.

Each testbed application addresses particular topics of long-standing geoscientific interest while sharing the overarching problems of underdeterminism and complexity. Several findings emerge from the three studies displayed in this thesis. The proposed nonparametric framework rooted in information theory (i) avoids the introduction of undesirable side information or erasing existing information; (ii) enables to directly quantify uncertainty and information content of datasets, and to analyze patterns and data-relations; (iii) describes the drivers of a system; (iv) allows the selection of the

most informative model according to the dataset availability; (v) relaxes assumptions and minimizes uncertainties; (vi) enables to deal with categorical or continuous data; and (vii) addresses any kind of data-relations.

Due to the advances in computational power and sophisticated instrumentation available these days, the combination of Earth science with allied areas is rapidly increasing. As a special case, the integration of probability and information theory, framed in a nonparametric context, on the one hand, entails the generality and flexibility needed to handle any kind of data-relations and limitations in data volume while, on the other hand, provides a tool for interpretation in terms of information content or its counterpart of uncertainty. This intrinsic interdisciplinarity also allows for more versatility to the modeling in terms of purpose and degrees of freedom. This means that given enough data to build data-driven models, their potential lies in the way they learn and exploit data unconstrained by functional or parametric assumptions and choices. Beyond that, the use of the proposed framework as presented in this thesis explores only particular examples among many potential applications. Overall, this thesis paves the way for enhancing our ability to make realistic predictions. It contributes with a novel framework to avoid conceptualization and compression of data-relations, helping to preserve the information content of the data while allowing an honest account of the related uncertainties. In a broader context, it offers a change of perspective in expressing and using geoscientific knowledge through the lens of information theory.

ZUSAMMENFASSUNG

Interessant, aber herausfordernd: Erdsysteme sind oft komplex und ihre Probleme unterbestimmt. Lückenhaftes Verständnis relevanter Teilsysteme (Komplexitätsfrage) und die Unmöglichkeit, alles, überall und zu jeder Zeit beobachten zu können (Unterbestimmtheitsfrage), führen zu einer erheblichen inferentiellen und prädiktiven Unsicherheit. Tatsächlich ist diese Unsicherheit eines der Probleme der Erdsystemforschung, und ihre Quantifizierung ist folglich ein wesentlicher Aspekt der geowissenschaftlichen Analyse und Prognose. Zusätzlich erhöht das Nichtberücksichtigen von Unsicherheit durch deterministische Modelle oder starke parametrische Annahmen die Starrheit des Modells (als Gegenpol zur Allgemeinheit). Infolgedessen können starre Modelle zu sowohl übermäßig eingeschränkten als auch übermäßig zuversichtlichen Lösungen und damit einer suboptimalen Nutzung der verfügbaren Daten führen. Um vor diesem Hintergrund mit der Unsicherheit, die sich aus dem Mangel an Wissen oder Daten ergibt, umzugehen, spielen probabilistische Inferenz und Unsicherheitsquantifizierung eine zentrale Rolle in der Modellierung oder Analyse solcher komplexen und unterbestimmten Systeme. Unsicherheit und Information können durch Maße aus der Informationstheorie objektiv quantifiziert werden, die in Verbindung mit nichtparametrischer probabilistischer Modellierung einen geeigneten Rahmen für die Bewertung des Informationsgehalts von Daten und Modellen bietet. Außerdem hilft es, das Problem der Verwendung starrer Modelle zu überwinden, die zu einem gewissen Grad Unsicherheiten ignorieren, nicht in den Daten vorhandene Informationen hinzufügen, oder verfügbare Informationen verlieren.

Diese Doktorarbeit befasst sich mit der oben skizzierten Fragestellung: Einen nichtparametrischen und probabilistischen Rahmen für geowissenschaftliche Probleme vorzuschlagen und zu validieren, der auf den Konzepten der Informationstheorie aufbaut. Prädiktive Beziehungen werden durch multivariate und empirische Wahrscheinlichkeitsverteilungen ausgedrückt, die direkt aus Daten abgeleitet werden. Die Informationstheorie wird verwendet, um den Informationsgehalt aus verschiedenen Quellen in einer universellen Einheit explizit zu berechnen und zu vergleichen. Drei typische geowissenschaftliche Probleme werden durch die Sichtweise der Informationstheorie neu betrachtet. Die Testumgebungen umfassen deskriptive und inferentielle Problemstellungen und befassen sich mit unterschiedlichen Datentypen (kontinuierlich oder kategorial), Domänen (räumliche oder zeitliche Daten), Stichprobengrößen und räumlichen Abhängigkeitseigenschaften. Zunächst wird ein nichtparametrischer Ansatz zur Identifikation von Niederschlags-Abfluss-Ereignissen entwickelt, an einem realen Datensatz getestet und mit einem physikalisch basierten Modell verglichen (Kapitel 2). Die Ergebnisse dieser Studie (Kapitel 3) bilden die Grundlage für die Entwicklung eines verteilungsfreien Ansatzes für geostatistische Fragestellungen, dessen Eigenschaften an einem synthetischen Datensatz getestet und mit Ordinary Kriging verglichen werden. Schließlich wird in Kapitel 4 die vorgeschlagene Methode für den Umgang mit kategorischen Daten und für die Simulation von Feldeigenschaften angepasst. Sie wird an einem realen Datensatz zur Klassifizierung des Bodenkontaminationsrisikos durch Blei getestet und ihre Eigenschaften mit Indicator Kriging verglichen.

Jede Testanwendung befasst sich mit bestimmten Themen, die seit langem von geowissenschaftlichem Interesse sind, und beinhaltet gleichzeitig die übergreifenden Probleme der Unbestimmtheit und Komplexität. Aus den drei in dieser Arbeit vorgestellten Anwendungen ergeben sich mehrere Erkenntnisse. Der vorgeschlagene nichtparametrische Rahmen aus Basis der Informationstheorie (i) vermeidet die Einführung unerwünschter Nebeninformationen oder den Verlust vorhandener Informationen; (ii) ermöglicht die direkte Quantifizierung der Unsicherheit und des Informationsgehalts von Datensätzen sowie die Analyse von Mustern und Datenbeziehungen; (iii) beschreibt die Einflussfaktoren eines Systems; (iv) ermöglicht die Auswahl des informativsten Modells je nach Verfügbarkeit des Datensatzes; (v) reduziert die Notwendigkeit für Annahmen und minimiert Unsicherheiten; (vi) ermöglicht den Umgang mit kategorischen oder kontinuierlichen Daten; und (vii) ist anwendbar auf jede Art von Datenbeziehungen.

Aufgrund der Fortschritte in der Rechenleistung und der hochentwickelten Instrumentierung, die heutzutage zur Verfügung stehen, nimmt die Verknüpfung der Geowissenschaften mit verwandten Disziplinen deutlich zu. Die Integration von Wahrscheinlichkeits- und Informationstheorie in einem nichtparametrischen Kontext garantiert einerseits die nötige Allgemeinheit und Flexibilität, um jede Art von Datenbeziehungen und Begrenzungen des Datenumfangs zu handhaben, und bietet andererseits ein Werkzeug für die Interpretation in Bezug auf den Informationsgehalt oder auf sein Gegenstück, die Unsicherheit. Diese inhärente Interdisziplinarität ermöglicht auch eine größere Flexibilität bei der Modellierung in Bezug auf die Zielgröße und die Freiheitsgrade. Beim Vorhandensein genügender Daten liegt das Potential datengetriebener Modellierungsansätze darin, dass sie ohne große Einschränkungen durch funktionale oder parametrische Annahmen und Entscheidungen auskommen. Die in dieser Arbeit vorgestellten Anwendungsbeispiele für den vorgeschlagenen Rahmen sind nur einige von vielen möglichen Anwendungen. Insgesamt trägt diese Doktorarbeit mit dem darin vorgeschlagenen Rahmen dazu bei, Konzeptualisierung und Komprimierung von Datenbeziehungen bei der Modellbildung zu vermeiden, wodurch der Informationsgehalt der Daten erhalten wird. Gleichzeitig ermöglicht er eine realistischere Berücksichtigung der damit verbundenen Unsicherheiten. In einem erweiterten Kontext bietet er einen Perspektivenwechsel bei der Darstellung und Nutzung von geowissenschaftlichem Wissen aus Sicht der Informationstheorie.

We have to live with a certain uncertainty.
— *Karl Popper*

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

INTRODUCTION

1 INTRODUCTION

1.1 MOTIVATION AND OVERVIEW

Modeling Earth systems is challenging, as their systems are often complex and their problems underdetermined (Perdigão et al., 2020; Reichstein et al., 2019). Complex on account of the multitude of nonlinear and interrelated processes, acting across a wide range of spatial and temporal scales; and underdetermined as we usually lack exhaustive measurements of system properties, initial, and boundary conditions, such that identification of system properties or model parameters can be afflicted by limited data (Ehret et al., 2018).

Earth systems are complex and their problems underdetermined.

Knowledge is limited.

This unfortunate situation is mitigated, according to Ehret et al. (2018), by the fact that no Earth system and related problem setting is truly unique, and insights gained in other, similar systems and problems can be used to inform the problem at hand. This is typically done by applying model structures developed in systems deemed similar to the one under analysis (e.g., conceptualization of physical processes as discussed by Klemeš, 1983), and sometimes also by applying parameters from models calibrated in similar systems (a process known as regionalization; Blöschl and Sivapalan, 1995; Merz et al., 2006a).

Insights gained in similar systems can mitigate issues of complexity and underdeterminism.

Taking these steps means that different sources of information are combined, without however explicitly keeping track of the particular uncertainties associated with each of them (Ehret et al., 2018). There are many forms of uncertainty in Earth system models (Reichstein et al., 2019), e.g., uncertainties due to limited observations or due to only partial agreement of the chosen model structure and the system at hand. Tracking sources of information, or uncertainty, is often further hampered by the use of deterministic models (Ehret et al., 2018; Nearing and Gupta, 2017), which offer no direct way to account for uncertainty (Neuper and Ehret, 2019). Taken together, the absence of a complete understanding of relevant subsystems (complexity issue) and the impossibility of observing everything, everywhere, all the time (underdeterminism issue) lead to considerable inferential and predictive uncertainty. As a matter of fact, uncertainty is part of Earth system science problems and, consequently, modeling them with probabilistic and statistical methods will continue to play a crucial role in the field (Perdigão et al., 2020; Reichstein et al., 2019).

It is difficult to backtrack the sources of uncertainties when knowledge is transferred.

Uncertainty is part of Earth system science problems.

Modeling uncertainties, and hence probabilistic inference, is essential for geoscientific problems.

Rigid models may lead to overconstrained and overconfident solutions.

Additionally, the use of deterministic models or strong parametric assumptions could result in a suboptimal use of the available data. That is because such rigidity in the model can lead to overly constrained and overconfident solutions (Nearing and Gupta, 2017). In this sense, data-driven methods have become increasingly popular as a substitute for or a complement to established modeling approaches (Bel et al., 2009; Reichstein et al., 2019), as they present higher generality in contrast to unique model settings and avoid the risk of adding information not present in data or losing available information (Neuper and Ehret, 2019).

This thesis aims to develop and validate a nonparametric, probabilistic framework for Earth system problems using IT.

Specifically, the framework seeks to increase model generality, make better use of data, and change the way of using geoscientific knowledge.

In this context, this thesis is motivated and framed by the need of a more generalized framework to deal with complex systems and interactions of different sources of information, data, and model uncertainties while moving away from strong parametric assumptions. This means avoiding as much as possible conceptualizations and compressions of data-relations to help to preserve their full information content while permitting an honest accounting of the related uncertainties. In that regard, this thesis is framed by exactly this quest of suggesting and demonstrating a nonparametric probabilistic framework based on concepts of information theory (IT), in which predictive relations are expressed by empirical probability distributions directly derived from data, and IT is used to explicitly calculate and compare the information and uncertainty content from data and models.

Fundamentals of information theory

IT provides a compelling framework for information and uncertainty quantification.

Information theory provides a compelling framework for information and uncertainty quantification. Its fundamental quantity is called *entropy*. As I will discuss, it has many properties that agree with the intuitive notion of what a measure of information should be (Cover and Thomas, 2006, p. 13). Entropy $H(X)$ is described as a measure of the uncertainty of a random variable X and is defined as the expected value of the negative logarithm of the probabilities $p(x)$ of all events contained in X:

$$H(X) = -\sum_{x \in X} p(x) \log_2 \mathrm{p}(x) = \mathbf{E}\left[I(x)\right]. \tag{1.1}$$

Entropy directly measures the uncertainty of a distribution, and conversely, its information content.

In this context, the logarithm is to base two so that the entropy is expressed in bits. Information can be represented as bits. One bit of information enables us to select between two equally probable alternatives. In other words, each bit of information corresponds to an answer to one optimal yes-no question asked with the intention of reconstructing the data. For example, the entropy of a fair coin toss is 1 bit, i.e., the answer of the question "is it

tails?" is enough to identify the toss output. The entropy of a random variable is also interpreted as a measure of the uncertainty of the random variable, and it measures the amount of expected *information* $I(x)$ required to describe the random variable X.

Besides quantifying the uncertainty of a distribution, it is also possible to compare (dis-)similarities between two distributions p and q over the same variable using *Kullback–Leibler divergence* (D_{KL}). This measure helps one to determine the difference between two distributions, i.e., it quantifies the statistical "distance" between two probability distributions p and q, such that:

$$D_{\mathrm{KL}}(p||q) = \sum_{x \in X} p\,(x)\ \log_2\ \frac{p(x)}{q(x)}. \tag{1.2}$$

Also referred to as *relative entropy*, D_{KL} can be understood as a measure of the information loss of assuming that the distribution is q when the true distribution is p (Cover and Thomas, 2006, p. 19). For example, when the distribution of the underlying data is originated from a limited sample, we work with an approximation of the distribution instead of its true shape. In this case, it would be needed $H(p) + D_{\mathrm{KL}}(p||q)$ bits (therefore, ask more questions) on average to reconstruct the random variable. This means that the information attached to a variable is estimated based on imperfect premises at the cost of increasing uncertainty. The entropy of the true distribution plus this increase of uncertainty due to an imperfect distribution assumption is called *cross entropy*, $H_{\mathrm{pq}}(p||q) = H(p) + D_{\mathrm{KL}}(p||q)$. Additionally, the D_{KL} measure is also used as a *scoring rule* for performance verification of probabilistic predictions (Gneiting and Raftery, 2007; Weijs et al., 2010).

Relative entropy quantifies the divergence (or missing information) from an approximate distribution q to the true distribution p.

For measuring the dependence between two different random variables, or how significant an explanatory variable is with respect to the target (or dependent) variable, measures such as conditional entropy and mutual information can be used. The *conditional entropy* can be described as the entropy of a random variable conditional on the (prior) knowledge of another random variable. The conditional entropy $H(X|Y)$ of a pair of discrete random variables (X, Y) is defined as:

Conditional entropy quantifies how much information a variable tells us about another.

$$H(X|Y) = -\sum_{y \in Y} p\,(y) \sum_{x \in X} p(x|y)\ \log_2\ p(x|y). \tag{1.3}$$

Conditional entropy is a generic measure of statistical dependence between variables (Sharma and Mehrotra, 2014), which can be used to compare competing model hypotheses and select the best among them. For example, suppose we want to predict rainfall-runoff events (target X) and for that we have information

from discharge and precipitation time series (predictors Y_1 and Y_2, respectively). The relation between the target and predictor can be expressed by their conditional distribution, $H(X|Y_1)$ and $H(X|Y_2)$. Based on that relation, conditional entropy can be used to measure the amount of information that each predictor brings to the target or, conversely, their reduction in uncertainty by considering both predictors simultaneously, i.e., $H(X|Y_1, Y_2)$. Note that the interchange of the terms information gain and uncertainty reduction is possible as they are two sides of the same coin. In doing so, it is possible to take advantage of other variables to help us understand the target and measure how much Y tells us about X by means of conditional entropy.

The reduction in uncertainty due to another variable is called mutual information.

As a matter of fact, the reduction in uncertainty due to another random variable is called *mutual information* $I(X;Y)$, which is equal to $H(X) - H(X|Y)$. Note that if X and Y are independent, knowing Y does not contribute to the uncertainty reduction of X and, therefore, the conditional entropy is exactly equal to the entropy of the target, $H(X|Y) = H(X)$. The opposite happens when two variables share the same information about the target, i.e., when they are completely redundant, resulting in $H(X|Y) = 0$. Thus, the mutual information of a random variable with itself is the entropy of the random variable, $I(X;X) = H(X)$. This is the reason why entropy is sometimes referred to as *self-information* (Cover and Thomas, 2006, p. 21).

IT is a general framework, which allows to objectively measure uncertainty and express it in a universal unit, "bit".

Note that the uncertainty measured by Eqs. 1.1 to 1.3 is defined as a function of the probability distribution of the variable, not on its value, category, or unit. This is convenient, as it allows joint treatment of many different sources and sorts of data in a single framework and in the same universal unit of bits. IT is extending beyond its original field since its proposal by Claude E. Shannon in 1948 (Shannon, 1948) to address data compression and transmission within the context of communication engineering. IT provides an attractive framework for analytical or predictive purposes, bringing an intuitive interpretation of information and uncertainty while allowing a direct way of quantifying them. Due to its universality of concepts, IT is being increasingly applied in a variety of disciplines, including Earth science.

The use of information theory in Earth science

Data-based modeling and measures from IT are being increasingly applied in Earth science.

In the context of data-based modeling, concepts and measures from IT are gaining ground in Earth science and being employed to investigate data series patterns and relations, as well as to quantify and compare the performance of models. Thanks to the generality and the multitude of interpretations of entropy, it has been used in a wide range of applications: for describing and inferring relations among data (Liu et al., 2016; Sharma and

Mehrotra, 2014), quantifying uncertainty and evaluating model performance (Chapman, 1986; Liu et al., 2016), estimating information flow (Darscheid, 2017; Weijs, 2011), analyzing similarity and redundancy (Ehret et al., 2020), describing catchment flow (Pechlivanidis et al., 2016), predicting precipitation (Neuper and Ehret, 2019), and measuring the quantity and quality of information in hydrological models (Loritz et al., 2018, 2019; Nearing and Gupta, 2017). Particularly in the spatial context, information-theoretic measures have been used to solve problems of spatial aggregation and quantify information gain, loss, and redundancy (Batty, 1974; Singh, 2013), to analyze spatio-temporal variability (Brunsell, 2010; Mälicke et al., 2020; Mishra et al., 2009), and to assess spatial dissimilarity (Naimi, 2015), complexity (Pham, 2010), uncertainty (Wellmann, 2013), and heterogeneity (Bianchi and Pedretti, 2018).

IT is gradually gaining ground also in the spatial context.

Besides using an information perspective of Earth science problems and having an objective way to measure uncertainty and information, these approaches have in common the advantage of using a framework which offers a very general language, which at the same time allows explicitly calculating and comparing information from various sources in a single currency, *bit*. Therefore, the scope of this Ph.D. thesis is to express insights about relations among data for analytical or predictive purposes by discrete, multivariate probability distributions derived from data. For analysis of the strength and generality of data-relations, concepts and measures from IT such as entropy, conditional entropy, mutual information, and Kullback-Leibler divergence are applied.

The scope of this thesis is to address typical Earth science problems with multivariate, empirical distributions and IT.

In this thesis, I develop and validate a nonparametric probabilistic framework to express and apply geoscientific knowledge firmly rooted in concepts from information theory. The work is divided in three testbed problems. Each one addresses individual topics of long-standing geoscientific interest while sharing the overarching problems of underdeterminism and complexity previously outlined. The three problems allow learning relations between data unconstrained by functional or strong parametric assumptions. The idea is, hence, to use these topics as a testbed for the envisaged framework and to evaluate its generality while looking at typical Earth science problems under a new perspective, i.e., through the lens of IT. General properties defining the data-relations include, apart from the data themselves, their attributes (or meta data) such as the type of data (continuous or categorical), the domain of observations (spatial or temporal), the distinction of data into predictor and target, and the size of the dataset; all of which are discussed throughout the three testbed problems.

The three testbed applications look afresh at problems ranging from temporal to spatial domain.

Earth science data are typically distributed in space and/or time (Goovaerts, 1997, p. 3). I start the thesis with the analysis of

time series in the context of rainfall-runoff events in chapter 2. Here, I consider the effects of learning from limited data and investigate the predictive power of models and variables under the perspective of information theory. As a case study, the quality of rainfall-runoff event identification given the relations learned from data is explored. I continue the journey using the insights from the previous work to build an information-theoretic framework relying on geostatistical concepts to extract information about spatial patterns (chapter 3). Beyond extracting the spatial dependence characteristics of the data, I propose reproducing them with the histogram via entropy reduction (HER) framework for probabilistic interpolation of unsampled points. An investigation is conducted to explore the features of HER using synthetically generated continuous data with varying sample densities and data properties. Additionally, HER is contrasted to ordinary kriging (OK) in a qualitative and quantitative manner. Due to the importance of the analysis of uncertainties in spatial contexts and the great potential of HER for exploring probability maps, in chapter 4 HER is adapted for dealing with categorical data, threshold-exceeding probabilities, and for reproducing the spatial fluctuation of the dataset reality with sequential simulation (HERs). Here, local and spatial uncertainty is addressed in the context of risk of soil contamination by lead using real-world data. Local uncertainty results of HER are thoroughly compared to indicator kriging (IK) and a proof of concept of the simulation framework HERs is presented. Finally, in chapter 5, I discuss and synthesize the key findings from the use of information to build solutions tailored to different problems at hand and identify the key challenges and opportunities for future investigations.

First application, chapter 2 – Temporal domain under an information perspective: a nonparametric framework for rainfall-runoff event identification.

Second application, chapter 3 – An information theoretic view of spatial data: a nonparametric framework for geostatistics.

Third application, chapter 4 – Categorical geostatistics and simulation: a nonparametric framework for soil contamination analysis.

1.2 CHAPTER II: TEMPORAL DOMAIN UNDER AN INFORMATION PERSPECTIVE

Given enough data to build data-driven models, their potential lies in the way they learn and exploit relations between data unconstrained by functional or parametric assumptions and choices. Here, the relations between target (variable to be predicted) and predictors (variables used to prediction) of time series are explored. The proposed framework is a form of supervised learning in the sense that known/labeled targets are used for training the model. Since it is built to be nonparametric, the framework can handle any kind of relation between the predictor(s) and the target unconstrained by functional assumptions. Each choice of a particular predictor is equivalent to formulating a model hypothesis. Models with different degrees of complexity (number of predictors) are tested to decide whether a data-relation should be applied to a problem/system at hand and to determine the

Chapter 2 addresses the following questions:

- How much information can each variable bring to the model?

contribution of each extra dimension (i.e., each extra predictor) in the model. This decision is based on the information gain (or its uncertainty reduction counterpart) brought to the model by each predictor variable; and measuring it is a key question addressed here by entropy and conditional entropy.

Linked to the size of the dataset, a second question addressed in chapter 2 is whether the dataset is sufficiently large to allow a robust inference of the data-relations, or, in other words, how much uncertainty about the target is due to the only partial *representativeness* of the dataset. This effect is measured via cross entropy and Kullback–Leibler divergence, and especially applies when working with many predictors. The exponential growth of the space of possible hypotheses as the number of dimensions grows is also known as *curse of dimensionality* (Goodman et al., 2008). The curse of dimensionality is closely related to the problem of *overfitting* since an overly complex model (with too many dimensions) tends to require more data to learn patterns in data instead of modeling their noise. Here, the information content of the dataset together with its representativeness analysis is used to quantify the minimal data requirement for a given model. Equally important, representativeness analysis is used as a support tool to decide, for a given amount of data, which number of predictors is optimal in the sense of avoiding both overfitting (by choosing too many predictors) and ignoring the available information (by choosing too few predictors).

- How much uncertainty about the target is due to the limited representativeness of the dataset?

- What is the minimum amount of data required to avoid overfitting?

Another important aspect investigated are the characteristics of recursive and non-recursive data-relations. In recursive relations, the target variable also appears as a predictor with a temporal shift. This is comparable to autoregressive models, where the dependent variable (target) is used as an independent variable (predictor). Recursive relations are advantageous whenever there is large information in the order of data (as for time series) and when extrapolation is required, however they might be less robust than non-recursive relations due to potentially strong feedbacks.

- How to account for time ordering?

Given the importance of events in hydrological problems, several methods have been proposed to replace the cumbersome task of manual event detection, such as Blume et al. (2007), Ehret and Zehe (2011), Koskelo et al. (2012), Mei and Anagnostou (2015), Merz and Blöschl (2009), Merz et al. (2006b), and Seibert et al. (2016). Interestingly, while for a trained hydrologist it is usually straightforward to identify events in a time series, it is hard to identify them automatically based on a set of rigid criteria. This happens due to their relative importance, which can vary over time, and strongly depend on user requirements, hydroclimate, and catchment properties. For the purpose of analyzing data-relations and quantifying uncertainty of models and data, the identification of events in a hydrograph offers challenges in the

lines of temporal context and uncertainties. Event detection, therefore, is an interesting testbed problem for exploring the potential of building models as empirical, discrete, multivariate probability distributions.

As a case study, the framework is applied to identify rainfall-runoff events in discharge time series from the Dornbirner Ach catchment in Austria. It mainly exploits information of discharge and precipitation series by learning relations between them and the occurrence of events from user-supplied classifications. The method measures the predictive power and robustness of the available data and provides optimal (minimum conditional entropy) probabilistic predictions of event occurrence for hydrological analysis and operational practice. Applying the model reduced the uncertainty in event classification by 77.8%. Finally, the results are validated through a holdout method and then compared to a physically-based approach, showing similar behavior for both the physically-based and data-driven models. Beyond probabilistic predictions, the framework learns and exploits relations between data, unconstrained by functional or strong parametric assumptions. One of the strengths of the data-based approach is that it potentially accepts any data to serve as predictors, and although the proposed framework is used to reproduce a hydrologist's way of identifying rainfall-runoff events, this is just one of many potential applications.

1.3 CHAPTER III: AN INFORMATION VIEW OF GEOSTATISTICS

Chapter 3 addresses the following questions:

- How to build a geostatistical framework free of parameterizations and assumptions to honestly deal with data uncertainty?

- How much spatial information is in the dataset, and how can it be used?

Spatial interpolation has a long history of application in Earth science when dealing with sparse spatial data coverage of measurement data. The traditional approach of modeling the uncertainty with respect to geostatistical interpolation consists in computing a kriging estimate and its attached error variance, and explicitly assuming a Gaussian distribution for assessing the confidence interval (Goovaerts, 1997, p. 261; Kitanidis, 1997, p. 68; Bourennane et al., 2007). By doing so, the errors are considered to be independent of the data values and depend only on the data configuration, a condition called *homoscedasticity*. Unfortunately, such restriction is rarely fulfilled for environmental attributes and soil variables (Goovaerts, 1997, p. 261; Bourennane et al., 2007; Kazianka and Pilz, 2010; Hristopulos and Baxevani, 2020), and therefore, it raises the first questions of how to build a geostatistical framework free of strong parameterizations, normality, and homoscedastic assumptions for an honest accounting of data uncertainty, how much spatial information there is in the dataset, and how it can be used.

In the context of data-driven modeling, another key question addressed here is to find a way of dealing with empirical (data-driven) and probabilistic modeling to add generality to the modeling process in order to handle different sources of information while avoiding strong parameterizations and normality assumptions. The problem of combining multiple conditional probability distributions into a single one is treated here using *aggregation methods* (Allard et al., 2012). The principle is to aggregate distributions extracted from the difference between pairs of observations into a global probability distribution. However, beyond the flexibility brought by the data-driven modeling, the variety of ways available to aggregate probabilities distributions added another facet to the problem. This raised the last question addressed in the chapter, which is related to a possible physical interpretation of the aggregation methods. Finally, three distinct aggregation methods are analyzed, aiming to estimate conditional distributions (target point conditioned to the sampled values) by introducing or inferring (dis-)continuity properties into or from the field.

- How to add generality to the modeling process and deal with probabilistic modeling?

- Is it possible to interpret the distinct aggregation methods in terms of the physical characteristics of the field?

As in the previous chapter, to address the issues of having a flexible-general model, free of assumptions, which properly handles uncertainty, I develop a method for geostatistical analysis and prediction directly based on empirical distributions and information theory. The purpose is to bypass the steps of variogram fitting done in traditional kriging methods and, at the same time, to avoid the risk of adding information not available in the data. As an additional outcome, the method minimizes predictive uncertainty expressed by relative entropy and estimates conditional distributions since it accounts for both spatial configuration and data values – in other words, it provides a proper framework for uncertainty estimation. More specifically, I propose a geostatistical, probabilistic, data-driven interpolator which combines measures of information theory with probability aggregation methods for (i) quantifying the available information in the dataset, (ii) extracting the structure of the data spatial dependence, (iii) minimizing the uncertainty of the predictions, (iv) introducing or inferring (dis-)continuity properties of the field, (v) relaxing normality assumptions, (vi) avoiding the addition of information not available in the data with functions, and (vii) handling uncertainty appropriately by means of conditional distributions. The proposed approach is called histogram via entropy reduction (HER).

With HER, it is possible to describe spatial dependence patterns and obtain conditional probabilistic predictions. I investigate and demonstrate its efficacy in ascertaining the underlying field with varying sample densities and data properties using a synthetically generated datasets from known Gaussian processes. HER shows a comparable performance to the popular benchmark model of

OK, with the additional advantage of higher generality. This framework provides spatial predictions with a minimum of assumptions involved and optimizes the use of the available data in terms of uncertainty. The novel method brings a new perspective of spatial interpolation and uncertainty analysis to geostatistics and statistical learning, under the lens of information theory.

1.4 CHAPTER IV: CATEGORICAL GEOSTATISTICS AND SIMULATION

Chapter 4 addresses the following questions:

- Can HER be adapted to handle categorical data?

- Can HER simulate the spatial fluctuation of data to avoid the smoothing effects?

- What is the ability of HER for modeling non-Gaussian data and assessing local uncertainty in a real-world dataset?

The results presented in chapter 3 are promising both for overcoming parameterization with functions and uncertainty trade-offs present in many traditional interpolators, and for assessing the uncertainty about the unknown through conditional distributions. The estimation results presented by HER are locally accurate and are appropriate for visualizing trends. However, the method suffers from the smoothing effect and is therefore inappropriate for simulating extreme values, similarly to OK, as discussed by Rossi and Deutsch (2014, p. 167). *Local uncertainty*, on the other hand, allows us to assess the uncertainty at any specific unsampled location but not the uncertainty when many locations must be considered simultaneously (*spatial uncertainty*; Goovaerts, 2001). Therefore, here I go one step further and explore the adaptability of HER to handle categorical data and to simulate the spatial fluctuation of the dataset reality. Additionally, in continuation to the previous chapter, the ability of HER in handling non-Gaussian data from a real application (non-synthetic data) and assessing their local uncertainty is also investigated.

Soil variables offer interesting properties for exploring and testing HER since they rarely meet assumptions of normality or present errors independent from the actual data values, and frequently display skewed distributions (Bourennane et al., 2007). For this reason, the established soil dataset of the Swiss Jura region (Atteia et al., 1994; Webster et al., 1994) is selected for addressing the previous issues of adaptability and model testing. In this fashion, the nonparametric framework of HER is tailored to assess local uncertainty for the delineation of contaminated areas and to handle categorial data in the context of estimating threshold-exceeding probabilities to map the risk of soil contamination by lead. Beyond exploring the potential adaptability of the method, the study investigates the method properties in contrast to IK and an OK model available in literature. Additionally, HER is extended through sequential simulation (HERs) for generating equiprobable realizations of lead concentrations and assessing spatial uncertainty (uncertainty jointly over several locations).

In the application, HER and IK exhibit comparable accuracy and precision in the performance analysis, albeit their local un-

certainties present different distribution shapes and magnitudes. HER has shown to be a unique framework for dealing with uncertainty estimation in a fine resolution without the need of (i) modeling multiple variograms, (ii) correcting order-relation violations, (iii) interpolating probabilities (or extrapolating tails) to obtain conditional cumulative distribution functions, or (iv) presenting stronger hypotheses of data distribution. In terms of information, it avoids strong loss of information arising from data binarization and the risk of adding information not contained in data caused by parameterization. The chapter presents a new facet of the HER method for modeling uncertainty in a soil contamination and remediation application, which brings together concepts of information theory and probability aggregation methods for measuring the information content of the data and optimizing its use. Finally, the intrinsic interdisciplinarity of the proposed framework has once more proven to entail a higher flexibility to the modeling in terms of purpose, degrees of freedom, and incorporation of properties in the context of spatial statistics.

Part II

IDENTIFYING RAINFALL-RUNOFF EVENTS IN DISCHARGE TIME SERIES: A DATA-DRIVEN METHOD BASED ON INFORMATION THEORY

This study is published in the scientific journal Hydrology and Earth System Science (HESS) and is a reprint of:

Thiesen, Stephanie; Darscheid, Paul; Ehret, Uwe (2019): Identifying rainfall-runoff events in discharge time series – a data-driven method based on Information Theory, Hydrology and Earth System Sciences, 23(2), 1015-1034. doi: 10.5194/hess-23-1015-2019

2

IDENTIFYING RAINFALL-RUNOFF EVENTS IN DISCHARGE TIME SERIES: A DATA-DRIVEN METHOD BASED ON INFORMATION THEORY

ABSTRACT

In this study, we propose a data-driven approach for automatically identifying rainfall-runoff events in discharge time series. The core of the concept is to construct and apply discrete multivariate probability distributions to obtain probabilistic predictions of each time step that is part of an event. The approach permits any data to serve as predictors, and it is nonparametric in the sense that it can handle any kind of relation between the predictor(s) and the target. Each choice of a particular predictor dataset is equivalent to formulating a model hypothesis. Among competing models, the best is found by comparing their predictive power in a training dataset with user-classified events. For evaluation, we use measures from information theory such as Shannon entropy and conditional entropy to select the best predictors and models and, additionally, measure the risk of overfitting via cross entropy and Kullback–Leibler divergence. As all these measures are expressed in "bit", we can combine them to identify models with the best tradeoff between predictive power and robustness given the available data.

We applied the method to data from the Dornbirner Ach catchment in Austria, distinguishing three different model types: models relying on discharge data, models using both discharge and precipitation data, and recursive models, i.e., models using their own predictions of a previous time step as an additional predictor. In the case study, the additional use of precipitation reduced predictive uncertainty only by a small amount, likely because the information provided by precipitation is already contained in the discharge data. More generally, we found that the robustness of a model quickly dropped with the increase in the number of predictors used (an effect well known as the curse of dimensionality) such that, in the end, the best model was a recursive one applying four predictors (three standard and one recursive): discharge from two distinct time steps, the relative magnitude of discharge compared with all discharge values in a surrounding 65 h time window and event predictions from the previous time step. Applying the model reduced the uncertainty in event classification by 77.8 %, decreasing conditional entropy from 0.516 to 0.114 bits. To assess the quality of the proposed method, its results were binarized and validated through a holdout method and then compared to a physically based approach. The comparison showed similar behavior of both models (both with accuracy near 90 %), and the cross-validation reinforced the quality of the proposed model. Given enough data to build data-driven models, their potential lies in the way they learn and exploit relations between data unconstrained by functional or parametric assumptions and choices. And, beyond that, the use of these models to reproduce a hydrologist's way of identifying rainfall-runoff events is just one of many potential applications.

2.1 INTRODUCTION

Discharge time series are essential for various activities in hydrology and water resources management. In the words of Chow et al. (1988), "[. . .] the hydrograph is an integral expression of the physiographic and climatic characteristics that govern the relations between rainfall and runoff of a particular drainage basin." Discharge time series are a fundamental component of hydrological learning and prediction, since they (i) are relatively easy to obtain, being available in high quality and from widespread and long-existing observation networks; (ii) carry robust and integral information about the catchment state; and (iii) are an important target quantity for hydrological prediction and decision-making.

Beyond their value in providing long-term averages aiding water balance considerations, the information they contain about limited periods of elevated discharge can be exploited for baseflow separation; water power planning; sizing of reservoirs and retention ponds; design of hydraulic structures such as bridges, dams or urban storm drainage systems; risk assessment of floods; and soil erosion. These periods, essentially characterized by rising (start), peak and recession (ending) points (Mei and Anagnostou, 2015), will hereafter simply be referred to as "events". They can have many causes (rainfall, snowmelt, upstream reservoir operation, etc.) and equally as many characteristic durations, magnitudes and shapes. Interestingly, while for a trained hydrologist with a particular purpose in mind, it is usually straightforward to identify such events in a time series, it is hard to identify them automatically based on a set of rigid criteria. One reason for this is that the set of criteria for discerning events from non-events typically comprises both global and local aspects, i.e., some aspects relate to properties of the entire time series and some to properties in time windows. And to make things worse, the relative importance of these criteria can vary over time, and they strongly depend on user requirements, hydroclimate and catchment properties.

So why not stick to manual event detection? Its obvious drawbacks are that it is cumbersome, subject to handling errors and hard to reproduce, especially when working with long-term data. As a consequence, many methods for objective and automatized event detection have been suggested. The baseflow separation, and consequently the event identification (since the separation allows the identification of the start and end time of the events), has a long history of development. Theoretical and empirical methods for determining baseflow are discussed since 1893, as presented in Hoyt (1936). One of the oldest techniques according to Chow et al. (1988) dates back to the early 1930s, with the normal depletion curve from Horton (1933). As stated by Hall (1968), fairly complete discussions of baseflow equations, mathematical derivations and applications were already present in the 1960s. In the last 2 decades, more recent techniques embracing a multitude of approaches (graphical-, theoretical-, mathematical-, empirical-, physical- and data-based) aim to automate the separation.

Ehret and Zehe (2011) and Seibert et al. (2016) applied a simple discharge threshold approach with partly unsatisfactory results; Merz et al. (2006b) introduced an iterative approach for event identification based on the comparison of direct runoff and a threshold. Merz and Blöschl (2009) expanded the concept to analyze runoff coefficients and applied it to a large set of catchments. Blume et al. (2007) developed the "constant k" method for baseflow separation, employing a gradient-based search for the end of event discharge. Koskelo et al. (2012) presented the physically based "sliding average

with rain record" – SARR – method for baseflow separation in small watersheds based on precipitation and quick-flow response. Mei and Anagnostou (2015) suggested a physically based approach for combined event detection and baseflow separation, which provides event start, peak and end times.

While all of these methods have the advantage of being objective and automatable, they suffer from limited generality. The reason is that each of them contains some kind of conceptualized, fixed relation between input and output. Even though this relation can be customized to a particular application by adapting parameters, it remains to a certain degree invariant. In particular, each method requires an invariant set of input data, and sometimes it is constrained to a specific scale, which limits its application to specific cases and to where these data are available.

With the rapidly increasing availability of observation data, computer storage and processing power, data-based models have become increasingly popular as an addition or alternative to established modeling approaches in hydrology and hydraulics (Solomatine and Ostfeld, 2008). According to Solomatine and Ostfeld (2008) and Solomatine et al. (2009), they have the advantage of not requiring detailed consideration of physical processes (or any kind of a priori known relation between model input and output); instead, they infer these relations from data, which however requires that there are enough data to learn from. Of course, including a priori known relations among data into models is an advantage as long as we can assure that they really apply. However, when facing undetermined problems, i.e., for cases where system configuration, initial and boundary conditions are not well known, applying these relations may be over-constraining, which may lead to biased and/or overconfident predictions. Predictions based on probabilistic models that learn relations among data directly from the data, with few or no prior assumptions about the nature of these relations, are less bias-prone (because there are no prior assumptions potentially obstructing convergence towards observed mean behavior) and are less likely to be overconfident compared to established models (because applying deterministic models is still standard hydrological practice, and they are overconfident in all but the very few cases of perfect models). This applies if there are at least sufficient data to learn from, appropriate binning choices are made (see the related discussion in Sect. 2.2.2) and the application remains within the domain of the data that was used for learning.

In the context of data-based modeling in hydrology, concepts and measures from information theory are becoming increasingly popular for describing and inferring relations among data (Liu et al., 2016), quantifying uncertainty and evaluating model performance (Chapman, 1986; Liu et al., 2016), estimating information flows (Darscheid, 2017; Weijs, 2011), analyzing spatio-temporal variability in precipitation data (Brunsell, 2010; Mishra et al., 2009), describing catchment flow (Pechlivanidis et al., 2016), and measuring the quantity and quality of information in hydrological models (Nearing and Gupta, 2017).

In this study, we describe and test a data-driven approach for event detection formulated in terms of information theory, showing that its potential goes beyond event classification, since it enables the identification of the drivers of the classification, the choice of the most suitable model for an available dataset, the quantification of minimal data requirements, the automatic reproduction classifications for database generation and the handling of any kind of relation between the data. The method is presented in Sect. 2.2. In Sect. 2.3, we describe two test applications with data from

the Dornbirner Ach catchment in Austria. We present the results in Sect. 2.4 and draw conclusions in Sect. 2.5.

2.2 METHOD DESCRIPTION

The core of the information theory method (ITM) is straightforward and generally applicable; its main steps are shown in Fig. 2.1 and will be explained in the following.

2.2.1 *Model hypothesis step*

The process starts by selecting the target (what we want to predict) and the predictor data (that potentially contain information about the target). Choosing the predictors constitutes the first and most important model hypothesis, and there are almost no restrictions to this choice. They can be any kind of observational or other data, transformed by the user or not; they can be part of the target dataset themselves, e.g., time lagged or space shifted; and they can even be the output of another model. The second choice and model hypothesis is the mapping between items in the target and the predictor dataset, i.e., the relation hypothesis. It is important for the later construction of conditional histograms that a 1:1 mapping exists between target and predictor data, i.e., one particular value of the target is related to one particular value of predictor (in contrast to 1:n or n:m relationships). Often, the mapping relation is established by equality in time.

2.2.2 *Model building step*

The next step is the first part of model building. It consists of choosing the value range and binning strategy for target and predictor data. These choices are important, as they will frame the estimated multivariate probability mass functions (PMFs) constituting the model and directly influence the statistics we compute from them for evaluation. Generally, these choices are subjective and reflect user-specific requirements and should be made while taking into consideration data precision and distribution, the size of the available datasets, and required resolution of the output. According to Gong et al. (2014), when constructing probability density functions (PDFs) from data via the simple bin-counting method, "[...] too small a bin width may lead to a histogram that is too rough an approximation of the underlying distribution, while an overly large bin width may result in a histogram that is overly smooth compared to the true PDF." Gong et al. (2014) also discussed the selection of an optimal bin width by balancing bias and variance of the estimated PDF. Pechlivanidis et al. (2016) investigated the effect of bin resolution on the calculation of Shannon entropy and recommended that bin width should not be less than the precision of the data. Also, while equidistant bins have the advantage of being simple and computationally efficient (Ruddell and Kumar, 2009), hybrid alternatives can overcome weaknesses of conventional binning methods to achieve a better representation of the full range of data (Pechlivanidis et al., 2016).

With the binning strategy fixed, the last part of the model building is to construct a multivariate PMF from all predictors and related target data. The PMF dimension equals the number of predictors plus one (the target), and the way probability

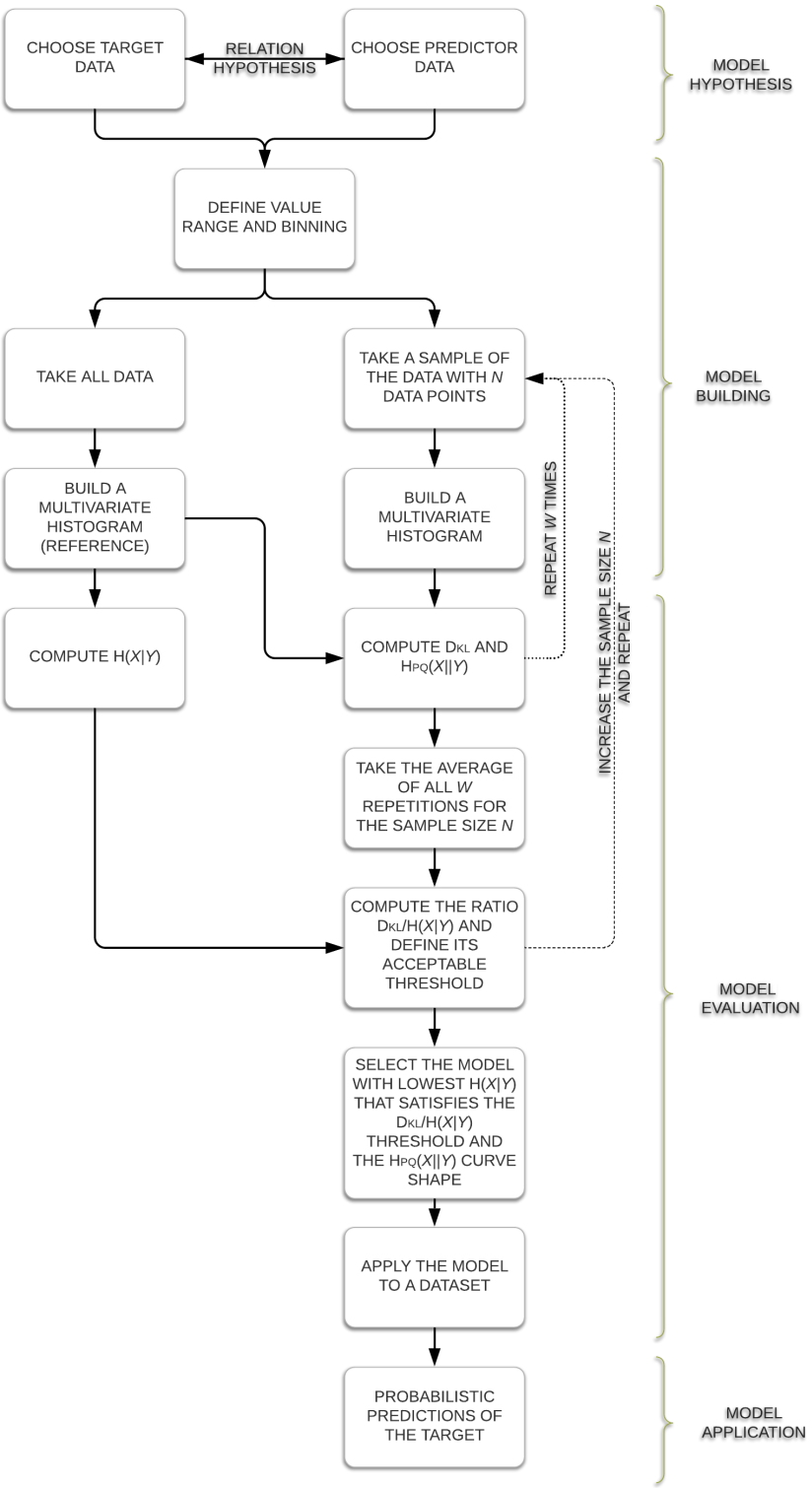

Figure 2.1: Main steps of the ITM.

mass is distributed within it is a direct representation of the nature and strength of the relationship between the predictors and the target as contained in the data. Application of this kind of model for a given set of predictor values is straightforward; we simply extract the related conditional PMF (or PDF) of the target, which, under the assumption of system stationarity, is a probabilistic prediction of the target value.

If the system is non-stationary, e.g., when system properties change with time, the inconsistency between the learning and the prediction situation will result in additional predictive uncertainty. The problems associated with predictions of non-stationary systems apply to all modeling approaches. If a stable trend can be identified, a possible countermeasure is to learn and predict detrended data and then reimpose the trend in a post-processing step.

2.2.3 *Model evaluation step*

Information theory – Measures

In order to evaluate the usefulness of a model, we apply concepts from information theory to select the best predictors (the drivers of the classification) and validate the model. With this in mind, this section provides a brief description of the information theory concepts and measures applied in this study. The section is based on Cover and Thomas (2006), which we recommend for a more detailed introduction to the concepts of information theory. Complementarily, for specific applications to investigate hydrological data series, we refer the reader to Darscheid (2017).

Entropy can be seen as a measure of the uncertainty of a random variable; it is a measure of the amount of information required on average to describe a random variable (Cover and Thomas, 2006). Let X be a discrete random variable with alphabet χ and probability mass function $p(x)$, $x \in \chi$. Then, the Shannon entropy $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x). \tag{2.1}$$

If the logarithm is taken to base two, an intuitive interpretation of entropy is the following: given prior knowledge of a distribution, how many binary (yes or no) questions need to be asked on average until a value randomly drawn from this distribution is identified? We can describe the conditional entropy as the Shannon entropy of a random variable conditional on the (prior) knowledge of another random variable. The conditional entropy $H(X|Y)$ of a pair of discrete random variables (X, Y) is defined as

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y). \tag{2.2}$$

The reduction in uncertainty due to another random variable is called mutual information $I(X, Y)$, which is equal to $H(X) - H(X|Y)$. In the study, both measures, Shannon entropy and conditional entropy, are used to quantify the uncertainty of the models (univariate and multivariate probability distributions, respectively). The first is calculated as a reference and measures the uncertainty of the target dataset. The latter

is applied to the probability distributions of the target conditional on predictor(s), and it corroborates to select the more informative predictors, i.e., the ones which lead to the most significant reduction of uncertainty of the target.

$$D_{\mathrm{KL}}(p||q) = \sum_{x \in X} p\,(x)\ \log_2\ \frac{p(x)}{q(x)}. \tag{2.3}$$

The Kullback–Leibler divergence is also a measure of the inefficiency of assuming that the distribution is q when the true distribution is p (Cover and Thomas, 2006). The Shannon entropy $H(p)$ of the true distribution p plus the Kullback–Leibler divergence $D_{\mathrm{KL}}(p||q)$ of p with respect to q is called cross entropy $H_{\mathrm{pq}}(X||Y)$. In the study, we use these related measures to validate the models and to avoid overfitting by measuring the additional uncertainty of a model if it is not based on the full dataset p but is only based on a sample q thereof.

Note that the uncertainty measured by Eqs. 2.1 to 2.1 depends only on event probabilities, not on their values. This is convenient, as it allows joint treatment of many different sources and types of data in a single framework.

Information theory – Model evaluation

As a benchmark, we can start with the case where no predictor is available, but only the unconditional probability distribution of the target is known. As seen in Eq. 2.1, the associated predictive uncertainty can be measured by the Shannon entropy $H(X)$ of the distribution (where X indicates the target). If we introduce a predictor and know its value in a particular situation a priori, predictive uncertainty is the entropy of the conditional probability function of the target given the particular predictor value. Conditional entropy $H(X|Y)$, where Y indicates the predictor(s), is then simply the probability-weighted sum of entropies of all conditional PMFs. Conditional entropy, like mutual information, is a generic measure of statistical dependence between variables (Sharma and Mehrotra, 2014), which we can use to compare competing model hypotheses and select the best among them.

Obviously, advantages of setting up data-driven models in the described way are that it involves very few assumptions and that it is straightforward when formulating a large number of alternative model hypotheses. However, there is an important aspect we need to consider: from the information inequality, we know that conditional entropy is always less than or equal to the Shannon entropy of the target (Cover and Thomas, 2006). In other words, information never hurts, and consequently adding more predictors will always either improve or at the least not worsen results. In the extreme, given enough predictors and applying a very refined binning scheme, a model can potentially yield perfect predictions if applied to the learning dataset. However, besides the higher computational effort, in this situation, the curse of dimensionality (Bellman, 1957) occurs, which "covers various effects and difficulties arising from the increasing number of dimensions in a mathematical space for which only a limited number of data points are available" (Darscheid, 2017). This means that with each predictor added to the model, the dimension of the conditional target–predictor PMF will increase by 1, but its volume will increase exponentially. For example, if the target PMF is covered by two bins and each predictor by 100, then a single, double and triple predictor model will consist of 200, 20 000 and

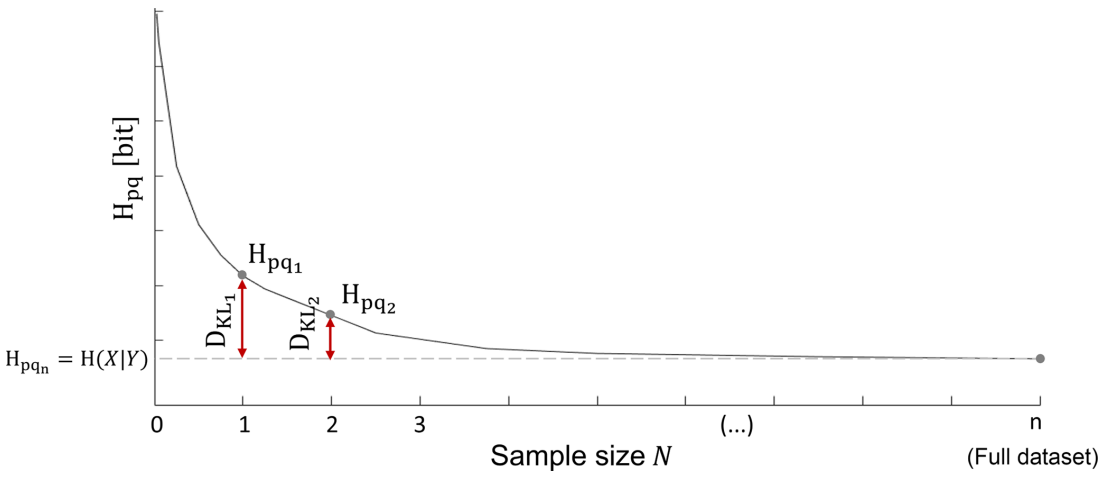

Figure 2.2: Investigating the effect of sample size through cross entropy and Kullback–Leibler divergence.

2 000 000 bins, respectively. Clearly, we will need a much larger dataset to populate the PMF mentioned last than the first. This also means that increasing the number of predictors for a fixed number of available data increases the risk of creating an overfitted or non-robust model in the sense that it will become more and more sensitive to the absence or presence of each particular data point. Models overfitted to a particular dataset are less likely to produce good results when applied to other datasets than robust models, which capture the essentials of the data relation without getting lost in detail.

We consider this effect with a resampling approach: from the available dataset, we take samples of various sizes and construct the model from each sample (see repetition statement regarding N in Fig. 2.1). Obviously, since the model was built from just a sample, it will not reflect the target– predictor relation as well as a model constructed from the entire dataset. It has been shown (Cover and Thomas, 2006; Darscheid, 2017) that the total uncertainty of such an imperfect model is the sum of two components: the conditional entropy $H(X|Y)$ of the "perfect" model constructed from all data and the Kullback–Leibler divergence D_{KL} between the sample-based and the perfect model. In this sense, D_{KL} quantifies the additional uncertainty due to the use of an imperfect model. For a given model (selection of target and predictors), the first summand is independent of the sample size, as it is calculated from the full dataset, but the second summand varies: the smaller the sample, the higher D_{KL}. Another important aspect of D_{KL} is that for a fixed amount of data, it strongly increases with the dimension of the related PMF, in other words, it is a measure of the impact of the curse of dimensionality. In information terms, the sum of conditional entropy and Kullback–Leibler divergence is referred to as cross entropy $H_{\mathrm{pq}}(X||Y)$. A typical example of cross entropy as a function of sample size is, for a single model, shown in Fig. 2.2.

The curve represents the mean of several repetitions, which were randomly taken with replacement among these repetitions. Note that, comparable to the Monte Carlo cross-validation, the analysis presented in Fig. 2.2 summarizes a large number of training and testing splits performed repeatedly, and, in addition, were also performed in different split proportions (subsets of various sizes). The difference here is that, in contrast to a standard split where datasets for training and testing are mutually

exclusive, we build the model in the training set and apply it in the full dataset, where one part of the data has not been seen yet and another part has. In other words, we use the training subsets for building the model (a supervised learning approach), and the resulting model is then applied to and evaluated on the full dataset. If, on the one hand, the use of the full dataset for the application includes data of the training set, on the other hand, the procedure favors the comparison of the results always with the same model. Thus, the stated procedure allows a robust and holistic analysis, in the sense that it works with the mean of W repetition for each subset and compares different sizes of training subset with a unique reference, the model built from the full dataset.

Particularly, Fig. 2.2 shows that for small sample sizes, D_{KL} is the main contributor to total uncertainty, but when the sample approaches the size of the full dataset, it disappears, and total uncertainty equals conditional entropy. From the shape of the curve in Fig. 2.2 we can also infer whether the available data are sufficient to support the model; when D_{KL} approaches zero (cross entropy approaches its minimum), this indicates that the model can be robustly estimated from the data, or, in other words, the sample size is enough to represent the full dataset. In an objective manner, we can also do a complementary analysis by calculating the ratio $D_{\mathrm{KL}}/H(X|Y)$, which is a measure of the relative contribution of D_{KL} to total uncertainty. We can then compare this ratio to a defined tolerance limit (e.g., 5%) to find the minimally required sample size.

Another application for Fig. 2.2 is to use these kinds of plots to select the best among competing models with different numbers of predictors. Typically, for small sample sizes, simple models will outperform multi-predictor models, as the latter will be hit harder by the curse of dimensionality; but with increasing data availability, this effect will vanish, and models incorporating more sources of information will be rewarded.

In order to reduce the effect of chance when taking random samples, we repeat the described resampling and evaluation procedure many times for each sample size (see repetition statement W in Fig. 2.1) and take the average of the resulting D_{KL}'s and H_{pq}'s. Based on these averaged results, we can identify the best model for a set of available data.

The proposed cross entropy curve contains a joint visualization of model analysis and model evaluation and, at the same time, provides the opportunity to compare models with different numbers of predictors, being a support tool to decide, for a given amount of data, which number of predictors is optimal in the sense of avoiding both ignoring the available information (by choosing too few predictors) and overfitting (by choosing too many predictors). And since it incorporates a sort of cross-validation in its construction, one of the advantages of this approach is that it avoids splitting the available data into a training and a testing set. Instead, it makes use of all available data for learning and provides measures of model performance across a range of sample sizes.

2.2.4 *Model application step*

Once a model has been selected, the ITM application is straightforward; from the multivariate PMF that represents the model, we simply extract the conditional PMF of the target for a given set of predictor values. The model returns a probabilistic

representation of the target value. If the model was trained on all available data, and is applied within the domain of these data, the predictions will be unbiased and will be neither overconfident nor underconfident. If instead a model using deterministic functions is trained and applied in the same manner, the resulting single-value predictions may also be unbiased, but due to their single-value nature they will surely be overconfident.

For application in a new time series, if its conditions are outside of the range of the empirical PMF or if they are within the range but have never been observed in the training dataset, the predictive distribution of the target (event yes or no) will be empty and the model will not provide a prediction. Several methods exist to guarantee a model answer, however they come with the cost of reduced precision. The solutions range from (i) coarse graining, where the PMF can be rebuilt with fewer, wider bins and an extension of the range until the model provides an answer to the predictive setting, as have been proposed by Darbellay and Vajda (1999), Knuth (2013), and Pechlivanidis et al. (2016), to (ii) gap filling, where the binning is maintained and the empty bins are filled with non-zero values based on a reasonable assumption. Gap-filling approaches comprise adding one counter to each zero-probability bin of the sample histogram, adding a small probability to the sample PDF, smoothing methods such as kernel density smoothing (Blower and Kelsall, 2002; Simonoff, 1996) or Bayesian approaches based on the Dirichlet and multinomial distribution or a maximum-entropy method recently suggested by Darscheid et al. (2018), the latter being applied in the present study

2.3 DESIGN OF A TEST APPLICATION

In this section, we describe the hydroclimatic properties of the data and the two performed applications. For demonstration purposes, the first test application was developed according to the Sect. 2.2 in order to explain which additional predictors we derived from the raw data and their related binning and to present our strategy for the model setup, classification and evaluation. For benchmarking purposes, the second application compares the proposed data-driven approach (ITM) with the physically based approach proposed by Mei and Anagnostou (2015), the characteristic point method (CPM), and applies the holdout method (splitting the dataset into training and testing set) for the cross-validation analysis.

2.3.1 *Data and site properties*

We used quality-controlled hourly discharge and precipitation observations from a 9-year period (31 October 1996 – 1 November 2005, 78 912 time steps). Discharge data are from the gauge Hoher Steg, which is located at the outlet of the 113 km² Alpine catchment of the Dornbirner Ach in northwestern Austria (GMT+1). Precipitation data are from the station Ebnit located within the catchment.

For the available period, we manually identified hydrological events by visual inspection of the discharge time series. To guide this process, we used a broad event definition, which can be summarized as follows: "an event is a coherent period of elevated discharge compared to the discharge immediately before and after and/or a coherent period of high discharge compared to the data of the entire time series." We

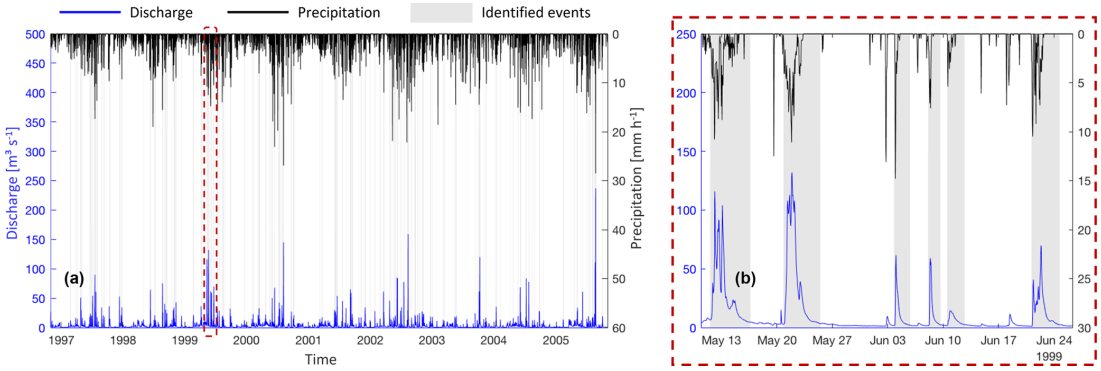

Figure 2.3: Input data of discharge, precipitation, and user-based event classification. Overview of the time series **(a)** and detailed view **(b)**.

suggest that this is a typical definition if the goal is to identify events for hydrological process studies such as analysis of rainfall-runoff coefficients, baseflow separation or recession analysis. Based on this definition, we classified each time step of the time series as either being part of an event (value 1) or not (value 0). Altogether, we identified 177 individual events covering 9 092 time steps, which is 11.5% of the time series. For the available 9-year period, the maximum precipitation is 28.5 mm h^{-1}, and the maximum and minimum discharge values are 237.0 and 0.037 $m^3 s^{-1}$, respectively. A preliminary analysis revealed that all times with discharge exceeding 15.2 $m^3 s^{-1}$ were classified as an event, and all times with discharge below 0.287 $m^3 s^{-1}$ were always classified as a non-event. Both the input data and the event classification are shown in Fig. 2.3.

2.3.2 *Application I – ITM*

2.3.2.1 *Predictor data and binning*

Since we wanted to build and test a large range of models, we not only applied the raw observations of discharge and precipitation but also derived new datasets. The target and all predictor datasets with the related binning choices are listed in Table 2.1; additionally, the predictors are explained in the text below.

For reasons of comparability, we applied uniform binning (fixed-width interval partitions) to all data used in the study, except for discharge; here we grouped all values exceeding 15.2 $m^3 s^{-1}$ (the threshold beyond which an event occurred for sure) into one bin to increase computational efficiency. For each data type, we selected the bin range to cover the range of observed data and chose the number of bins with the objective of maintaining the overall shape of the distributions with the least number of bins.

Discharge Q $[m^3 s^{-1}]$

This is the discharge as measured at Hoher Steg. In order to predict an event at time step t, we tested discharge at the same time step as a predictor, $Q(t)$, and at time steps before and after t, such as $Q(t-2)$, $Q(t-1)$, $Q(t+1)$, and $Q(t+2)$.

Table 2.1: Target and predictors – characterization and binning strategy.

Target (**X**)	**Symbol**	**Unit**	**Bins** [a] **[start : end]**	**Number of bins**
User-based event t classification at time	e	(-)	[0 : 1]	2
Predictors (**Y**)	**Symbol**	**Unit**	**Bins** [a] **[start : step : end]**	**Number of bins**
Discharge	$Q(t–2)$,$Q(t–1)$, $Q(t)$,$Q(t+1)$ $Q(t+2)$	($m^3\ s^{-1}$)	[0 : 0.5 : 16], [16 : end]	34
Natural logarithm of discharge	$\ln Q(t–2)$, $\ln Q(t–1)$, $\ln Q(t)$,$\ln Q(t+1)$ $\ln Q(t+2)$	ln($m^3\ s^{-1}$)	[-3.5 : 0.2 : 2.9], [2.9 : end]	34
Relative Magnitude of discharge	Q_{RMC}, Q_{RML}, Q_{RMR}	(–)	[0 : 0.1 : 1]	11
Discharge slope	$Q_{slope_{e_{before}}}$, $Q_{slope_{e_{after}}}$	($m^3\ s^{-1}\ h^{-1}$)	[-50 : 5 : 90]	29
Precipitation at time t	P	(mm h^{-1})	[0 : 1 : 30]	31
Model-based event probability	$e_p(t–1)$	(–)	[0 : 0.1 : 1]	11

[a] Bins identified by their central values [leftmost center value : step : rightmost center value].

Natural logarithm of discharge $\ln Q$ $[\ln Q(m^3\ s^{-1})]$

We also used a log transformation of discharge to evaluate whether this non-linear conversion preserved more information in Q when mapped into the binning scheme than the raw values. Note that the same effect could also be achieved by a logarithmic binning strategy, but as mentioned we decided to maintain the same binning scheme for reasons of comparability. As for Q, we also applied the log transformation to time-shifted data.

Relative magnitude of discharge Q_{RM} $[-]$

This is a local identifier of discharge magnitude at time t in relation to its neighbors within a time window. For each time step, we normalized discharge into the range $[0, 1]$ using Eq. 2.4, where Q_{max} is the largest value of Q within the window and Q_{min} is the smallest:

$$Q_{RM} = \frac{Q(t) - Q_{min}}{Q_{max} - Q_{min}}. \tag{2.4}$$

A value of $Q_{RM} = 0$ indicates that $Q(t)$ is the smallest discharge within the analyzed window, and a value of $Q_{RM} = 1$ indicates that it is the largest. We calculated these values for many window sizes and for windows with the time step under consideration in the center (Q_{RMC}), at the right end (Q_{RMR}) and at the left end (Q_{RML}) of the window. The best results were obtained for a time-centered window of 65 h.

For further details see Sect. 2.3.2.2.

Slope of discharge Q_{slope} $[m^3\,s^{-1}\,h^{-1}]$

This is the local inclination of the hydrograph. This predictor was created to take into consideration the rate and direction of discharge changes. We calculated both the slope from the previous to the current time step applying Eq. 2.5 and the slope from the current to the next time step applying Eq. 2.6, where positive values always indicate rising discharge:

$$Q_{\mathrm{slope}_{\mathrm{before}}} = \frac{Q(t) - Q(t-1)}{t - (t-1)}, \tag{2.5}$$

$$Q_{\mathrm{slope}_{\mathrm{before}}} = \frac{Q(t+1) - Q(t)}{(t+1) - t}. \tag{2.6}$$

Precipitation P $[mm\,h^{-1}]$

This is the precipitation as measured at Ebnit.

Model-based event probability e_p $[-]$

In general, information about a target of interest can be encoded in related data such as the predictors introduced above, but it can also be encoded in the ordering of data. This is the case if the processes that are shaping the target exhibit some kind of temporal memory or spatial coherence. For example, the chance of a particular time step to be classified as being part of an event increases if the discharge is on the rise, and it declines if the discharge declines. We can incorporate this information by adding to the predictors discharge from increasingly distant time steps, but this comes at the price of a rapidly increasing impact of the curse of dimensionality. To mitigate this effect, we can use sequential or recursive modeling approaches; in a first step, we build a model using a set of predictors and apply it to predict the target. In a next step, we use this prediction as a new, model-derived predictor, combine it with other predictors in a second model, use it to make a second prediction of the target and so forth. Each time we map information from the multi-dimensional set of predictors onto the one-dimensional model output, we compress data and reduce dimensionality while hoping to preserve most of the information contained in the predictors. Of course, if we apply such a recursive scheme and want to avoid iterations, we need to avoid circular references, i.e., the output of the first model must not depend on the output of the second. In our application, we assured this by using the output from the first model at time step $t-1$ as a predictor in the second model to make a prediction at time step t. Comparable to a Markov model, this kind of predictor helps the model to better stick to a classification after a transition from event to non-event or vice versa.

2.3.2.2 *Selecting the optimal window size for the* Q_{RM} *predictor*

To select the most informative window size when using relative magnitude of discharge as a predictor, we calculated conditional entropy of the target given discharge

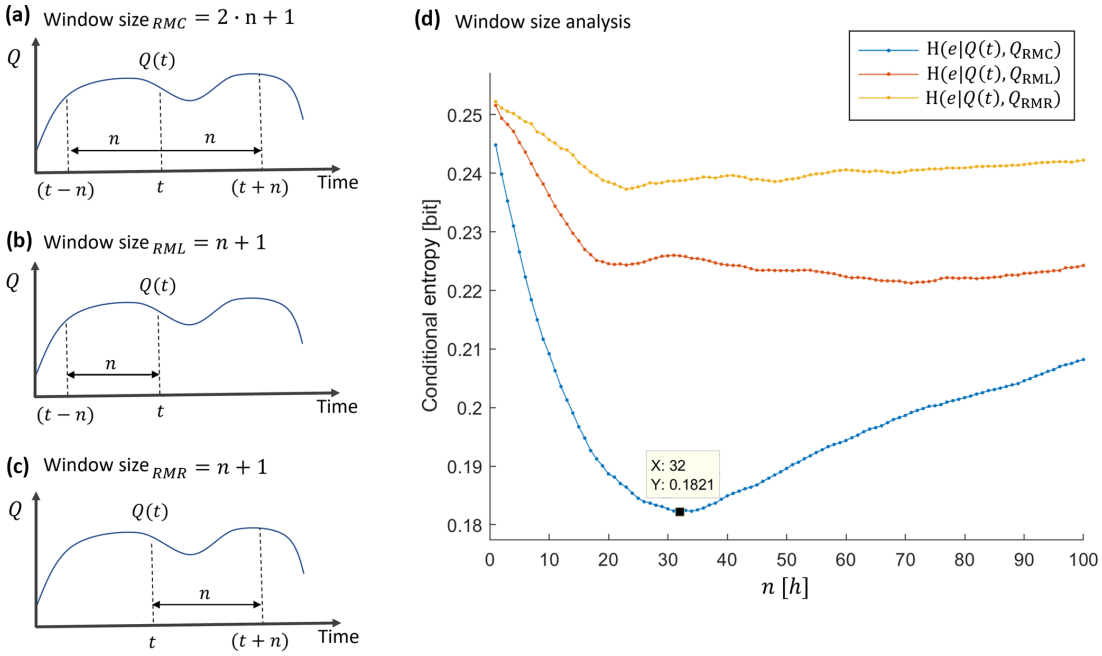

Figure 2.4: Window size definitions for window types. **(a)** Q_{RMC}, **(b)** Q_{RML}, and **(c)** Q_{RMR} window definitions, and **(d)** window size analysis.

and the Q_{RMC}, Q_{RML} and Q_{RMR} predictors for a range of window sizes on the full dataset. The definition of the window sizes for the different window types and the conditional entropies are shown in Fig. 2.4.

The best (lowest) value of conditional entropy was obtained for a time-centered window (Q_{RMC}) with $2 * 32 + 1 = 65$ h of total width. We used this value for all further analyses.

2.3.2.3 *Model classification, selection and evaluation*

Model classification

All the models we set up and tested in this study can be assigned to one of three distinct groups. The groups distinguish both typical situations of data availability and the use of recursive and non-recursive modeling approaches. Models in the Q-based group apply exclusively discharge-based predictor(s). For models in the P-based group, we assumed that in addition to discharge, precipitation data are also available. This distinction was made, because in the literature two main groups of event detection methods exist: one relying solely on discharge data the other using precipitation data additionally. Finally, models in the model-based group all apply a two-step recursive approach as discussed in Sect. 2.3.2.1. In this case, the first model is always from the Q- or P-based group. Later, event predictions at time step $t-1$ of the first model application are then, together with additional predictors from the Q- or P-based group, used as a predictor in the second model.

Model selection

In order to streamline the model evaluation process, we applied an approach

of supervised model selection and gradually increasing model complexity, we started by setting up and testing all possible one-predictor models in the Q- and P-based group. From these, we selected the best-performing model and combined it with each remaining predictor into a set of two-predictor models. The best-performing two-predictor model was then expanded to a set of three-predictor models using each remaining predictor and so forth. For the model-based group, the strategy was to take the best-performing models from both the Q- and the P-based group as the first model and then combine it with an additional predictor. In the end, we stopped at four-predictor models, since beyond it, the uncertainty contribution due to limited sample size became dominant.

Model evaluation

Among models with the same number of predictors, we compared model performance via the conditional entropy (target given the predictors), calculated from the full dataset. However, when comparing models with different numbers of predictors, the influence of the curse of dimensionality needs to be taken into account. To this end, we calculated sample-based cross entropy and Kullback–Leibler divergence as described in Sect. 2.2.3 for samples of size of 50 up to the size of the full dataset, using the following sizes: 50, 100, 500, 1000, 1500, 2000, 2500, 5000, 7500, 10 000, 15 000, 20 000, 30 000, 40 000, 50 000, 60 000, 70 000 and 78 912. To eliminate effects of chance, we repeated the resampling 500 times for each sample size and took their averages. In Appendix A, the resampling strategy and the choice of repetitions are discussed in more detail.

2.3.3 *Application II – ITM and CPM comparison*

The second application aims to compare the performances of the ITM and another automatic event identification method from a more familiar perspective. The predictions were performed in a separate dataset, and, as a measure of diagnostic, concepts from the receiver operating characteristic (ROC) curve quantified the hits and misses of the predictions of both models according to a time series of user-classified events (considered the true value). More about the ROC analysis can be found in Fawcett (2005).

For the comparison, the characteristic point method (CPM) was chosen, because, in contrast with the data-driven ITM, it is a physically based approach for event identification, which is applicable to and recommended for the characteristics of the available dataset (hourly timescale data on catchment precipitation and discharge) and open source. The essence of the method is to characterize flow events with three points (start, peak(s) and end of the event) and then associate the event to a corresponding rainfall event (Mei and Anagnostou, 2015). For the event identification, a baseflow separation is previously needed and proposed by coupling the revised constant k method (Blume et al., 2007) and the recursive digital filter proposed by Eckhardt (2005). More about the CPM can be attained in Mei and Anagnostou (2015).

Since the outcome of the CPM is dichotomous, classified as either event or non-event, the probabilistic outcome of the ITM must be converted into a binary solution. The binarization was reached in the study by choosing an optimum threshold of the

probabilistic prediction ($p_{threshold}$), where all time steps with probabilities equal to or greater than it were classified as being part of an event. The objective function of the optimization was based on the ROC curve and sought to minimize the distance to the top-left corner of the ROC curve, i.e., the Euclidean distance between the true positive rate (R_{TP}, proportion of events correctly identified in relation to the total of true events) and false positive rate (R_{FP}, proportion of false events in relation to the total of true non-events) to the perfect model (where $R_{TP_{\mathrm{perfect}}} = 1$ and $R_{FP_{\mathrm{perfect}}} = 0$), as expressed in Eq. 2.7 [1]:

$$\min \quad \sqrt{(1 - R_{\mathrm{TP}})^2 + (0 - R_{\mathrm{FP}})^2}. \tag{2.7}$$

Even though the physically based CPM method theoretically does not require a calibration step, for avoiding misleading comparison, the parameter Rnc (rate of no change, used to quantify null-change ratio in recession coefficient k) was optimized by Eq. 2.7. Thus, R_{TP} and $R_{\mathrm{FP}} \in [0, 1]$ are calculated as a function of the optimized parameter $p_{threshold}$ (for the ITM) and R_{nc} (for the CPM).

Due to the pthreshold and Rnc optimization and to enable the cross-validation of the models in a new dataset, the available data were divided into training and testing sets. And, since the ITM model requires a minimum dataset size to guarantee the model robustness, the holdout split was based on the data requirement of the selected ITM model obtained according to application I, Sect. 2.3.2. Therefore, the training dataset was used to build the ITM model and to calibrate the $p_{threshold}$ (needed for the binarization) and R_{nc}.

After that, the calibrated models (ITM and CPM) were applied to a new dataset (testing dataset), and measures of quality based on the ROC curve were computed in order to evaluate and compare their performance, such as (i) the true positive rate (R_{TP}), which represents the percentage of event classification hits (counting of events correctly classified by the model, P_{T}, divided by the amount of the true events in the testing dataset, P); (ii) the false positive rate (R_{FP}), which represents the percentage of false events identified by the model (counting of events misclassified by the model, P_{F}, divided by the amount of the true non-events in the testing dataset, N); (iii) the accuracy, which reflects the total proportion of events (P_{T}) and non-events (or true negative, N_{T}) that were correctly predicted by the model; and (iv) the distance to the perfect model given by the Eq. 2.7, which represents the norm between the results obtained by the method and a perfect prediction.

2.4 RESULTS AND DISCUSSION

2.4.1 *Results for application I*

2.4.1.1 *Model performance for the full dataset*

Here we present and discuss the model results when constructed and applied to the complete dataset. As we stick to the complete dataset, Kullback-Leibler divergence will always be zero, and model performance can be fully expressed by conditional entropy (see Sect. 2.3.2.3; Model Evaluation), with the (unconditional) Shannon entropy of

1 A detailed discussion about the cut-off values of the ROC curve can be found in Habibzadeh et al. (2016).

Table 2.2: Conditional entropy and relative uncertainty reduction of one-predictor models.

no.	Predictive model $(\mathbf{X}\|\mathbf{Y})$	$\mathbf{H}(\mathbf{X}\|\mathbf{Y})$ [bit]	$\mathbf{H}(\mathbf{X}\|\mathbf{Y})\,/\mathbf{H}(\mathbf{X})$[a]
Q-based group			
1	$e \mid Q(t-2)$	0.269	52.1%
2	$e \mid Q(t-1)$	0.264	51.3%
3	$e \mid Q(t)$	0.260	50.3%
4	$e \mid Q(t+1)$	0.255	49.4%
5	$e \mid Q(t+2)$	0.250	48.6%
6	$e \mid \ln Q(t-2)$	0.269	52.2%
7	$e \mid \ln Q(t-1)$	0.265	51.3%
8	$e \mid \ln Q(t)$	0.260	50.4%
9	$e \mid \ln Q(t+1)$	0.255	49.4%
10	$e \mid \ln Q(t+2)$	0.251	48.6%
11	$e \mid Q_{\mathrm{RMC}}$	0.505	97.9%
12	$e \mid Q_{\mathrm{slope}_{\mathrm{before}}}$	0.473	91.8%
13	$e \mid Q_{\mathrm{slope}_{\mathrm{after}}}$	0.473	91.8%
P-based group			
14	$e \mid P$	0.472	91.6%

[a] $H(\mathrm{X}) = H(e) = 0.516\ bits$.

the target data $H(e) = 0.516\ bits$ as an upper limit, which we use as a reference to calculate the relative uncertainty reduction for each model. In Table 2.2, conditional entropies and their relative uncertainty reductions are shown for each Q- and P-based one-predictor model.

One-predictor models based on Q and $\ln Q$ reduced uncertainty to about 50% (models no. 1–10 in Table 2.2, fourth column), with a slight advantage of Q over $\ln Q$. Interestingly, both show their best results for the time offset $t+2$, i.e., future discharge is a better predictor of event detection than discharge at the current time step. As we were not sure whether this also applies to two-predictor models, we decided to test both the $t+2$ and t predictors of Q and $\ln Q$ in the next step. Compared to Q and $\ln Q$, relative magnitude of discharge Q_{RMC} and discharge slope Q_{slope} performed poorly, and so did P, the only model in the P-based group. This is most likely because for a certain time step, being part of an event is not as dependent on precipitation at this particular time step but is rather dependent on the accumulated rainfall in a period preceding it. Despite its poor performance, we decided to use it in higher-order models to see whether it becomes more informative in combination with other predictors.

Based on these considerations and the model selection strategy described in Sect. 2.3.2.3, we built and evaluated all possible two-predictor models. The models and results are shown in Table 2.3.

As could be expected from the information inequality, adding a predictor improved the results, and for some models (no. 16 and no. 20), the t predictors outperformed their $t+2$ counterparts (no. 17 and no. 21, respectively). Once more, Q predictors

Table 2.3: Conditional entropy and relative uncertainty reduction of two-predictor models.

no.	Predictive model (X|Y)	H (X|Y) [bit]	H (X|Y) /H(X)[a]
Q-based group			
15	$e \mid Q(t+2),\ Q(t)$	0.226	43.9%
16	$e \mid Q(t),\ Q_{\mathrm{RMC}}$	0.182	35.3%
17	$e \mid Q(t+2),\ Q_{\mathrm{RMC}}$	0.191	37.1 %
18	$e \mid Q(t),\ Q_{\mathrm{slope}_{\mathrm{after}}}$	0.254	49.3 %
19	$e \mid \ln Q(t+2),\ \ln Q(t)$	0.233	45.1%
20	$e \mid \ln Q(t),\ Q_{\mathrm{RMC}}$	0.185	35.8%
21	$e \mid \ln Q(t+2),\ Q_{\mathrm{RMC}}$	0.194	37.5%
22	$e \mid \ln Q(t),\ Q_{\mathrm{slope}_{\mathrm{after}}}$	0.254	49.3%
P-based group			
23	$e \mid Q(t),\ P$	0.248	48.2%
24	$e \mid Q(t+2),\ P$	0.247	48.0%
25	$e \mid \ln Q(t),\ P$	0.249	48.2%
26	$e \mid \ln Q(t+2),\ P$	0.249	48.2%

[a] $H(X) = H(e) = 0.516\ bits$.

Table 2.4: Conditional entropy and relative uncertainty reduction of three-predictor models.

no.	Predictive model (X|Y)	H (X|Y) [bit]	H (X|Y) /H(X)[a]
Q-based group			
27	$e \mid Q(t),\ Q_{\mathrm{RMC}},\ Q(t+2)$	0.144	28.0%
P-based group			
28	$e \mid Q(t),\ P,\ Q_{\mathrm{RMC}}$	0.167	32.5%

[a] $H(X) = H(e) = 0.516\ bits$.

performed slightly better than $\ln Q$ such that for all higher-order models, we only used $Q(t)$ and ignored $Q(t+2)$, $\ln Q(t)$ and $\ln Q(t+2)$.

In the P-based group, adding any predictor greatly improved results by about 50%, but not a single P-based model outperformed even the worst of the Q-based group.

Finally, from both the Q- and P-based group, we selected the best model using t predictors (no. 16 and no. 23, respectively) and extended them to three-predictor models with the remaining predictors. The models and results are shown in Table 2.4.

Again, for both models, the added predictor improved results considerably, and we used both of them to build a recursive four-predictor model as described in Sect. 2.3.2.3. The new predictor, $e_p(t-1)$ is simply the probabilistic prediction of a model (no. 27 or no. 28, in this case) for time step $t-1$ of being part of an event, with a value range of $[0,1]$. This means that $e_{p_{27}}(t-1)$ carries the memory from the previous predictions of model no. 27 (and $e_{p_{28}}(t-1)$ from model no. 28, accordingly), and the new four-predictor models no. 29 and no. 30 as shown in Table 2.5 are simply copies of these models, extended by a memory term: $e_p(t-1)$.

Table 2.5: Conditional entropy and relative uncertainty reduction of recursive four-predictor models.

no.	Predictive model $(\mathbf{X}\|\mathbf{Y})$	$\mathbf{H}(\mathbf{X}\|\mathbf{Y})$ [bit]	$\mathbf{H}(\mathbf{X}\|\mathbf{Y})/\mathbf{H}(\mathbf{X})$[a]
Model-based group			
29	$e \mid Q(t),\ Q_{\mathrm{RMC}},\ Q(t+2),\ e_{p_{27}}(t-1)$	0.114	22.2%
30	$e \mid Q(t),\ P,\ Q_{\mathrm{RMC}},\ e_{p_{28}}(t-1)$	0.142	27.6%

[a] $H(X) = H(e) = 0.516$ *bits*.

Table 2.6: Models selected for sample-based tests.

Model group	1 predictor	2 predictors	3 predictors	4 predictors
Q-based group[a]	$Q(t)$ no. 3	$Q(t)$, Q_{RMC} no. 16	$Q(t)$, Q_{RMC}, $Q(t{+}2)$ no. 27	–
P-based group[b]	–	$Q(t)$, P no. 23	$Q(t)$, P, Q_{RMC} no. 28	–
Model-based with Q-based predictors[a]	–	–	–	$Q(t)$, Q_{RMC}, $Q(t{+}2)$, $e_{p_{27}}(t{-}1)$ no.29
Model-based with P-based predictors[b]	–	–	–	$Q(t)$, P, Q_{RMC}, $e_{p_{28}}(t{-}1)$ no. 30

[a] Models which apply exclusively discharge-based predictor(s).
[b] Models which apply discharge- and precipitation-based predictor(s).

Again, model performance improved, and model no. 29 was the best among all tested models, though so far the effect of sample size was not considered, which might have a strong impact on the model rankings. This is investigated in the next section.

2.4.1.2 *Model performance for samples*

The sample-based model analysis is computationally expensive, so we restricted these tests to a subset of the models from the previous section. Our selection criteria were to (i) include at least one model from each predictor group, (ii) include at least one model from each dimension of predictors, and (iii) choose the best-performing model. Altogether we selected the seven models shown in Table 2.6. Please note that despite our selection criteria, we ignored the one-predictor model using precipitation due to its poor performance.

For these models, we computed the cross entropies between the full dataset and each sample size N for W repetitions, and in the end, for each sample size N, we took the average of the W repetitions. The results are shown in Fig. 2.5. For comparison, the cross entropies between the target dataset and samples thereof are also included and labeled as model no. 0.

In Fig. 2.5, the cross entropies at the right end of the the x axis, where the sample contains the entire dataset, equal the conditional entropies, as the effect of sample size is zero. However, with decreasing sample size, cross entropy grows in a non-linear

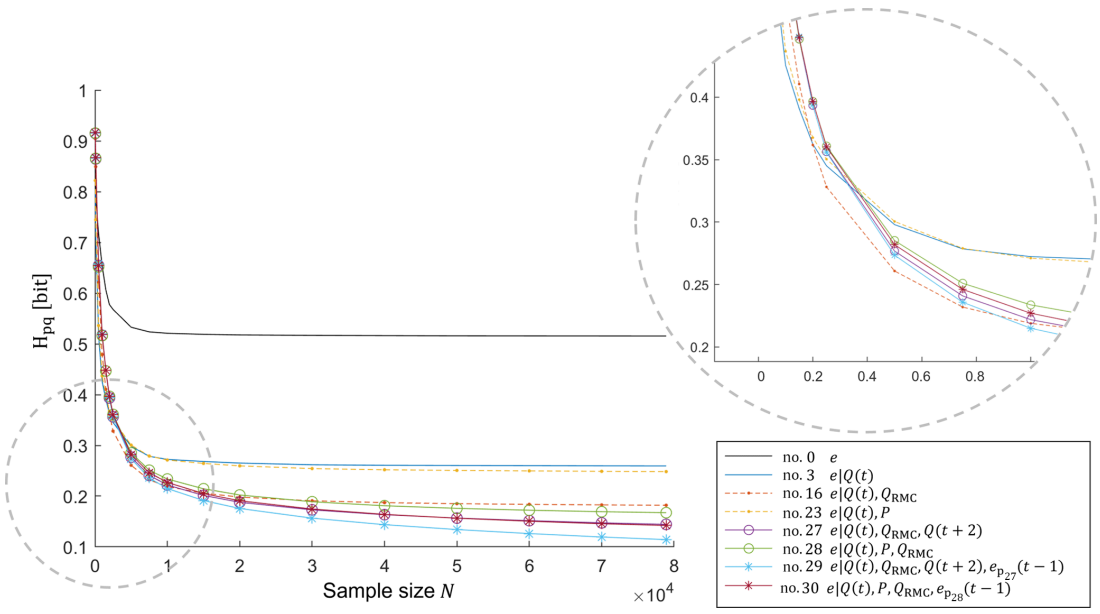

Figure 2.5: Cross entropy for models in Table 2.6 as a function of sample size.

fashion as D_{KL} starts to grow. If we walk through the space of sample sizes in the opposite direction, i.e., from left to right, we can see that as the samples grow, the rate of change of cross entropy decreases, the reason being that the rate of change of D_{KL} decreases, which means that the model learns less and less from new data points. Thus, by visually exploring these "learning curves" of the models we can make two important statements related to the amount of data required to inform a particular model: we can state how large a training dataset should be to sufficiently inform a model, and we can compare this size to the size of the actually available dataset. If the first is much smaller than the latter, we gain confidence that we have a well-informed, robust model. If not, we know that it may be beneficial to gather more data, and if this is not possible, we should treat model predictions with caution.

As mentioned in Sect. 2.2.3, besides Fig. 2.5 informing the amount of data needed to have a robust model (implying that sample size is enough to represent the full dataset), it allows the comparison of competing models with different dimensions and selection of the optimal number of predictors (taking advantage of the available information and avoiding overfitting). In this sense, in the P-based group and for sample sizes smaller 5000, the two-predictor model no. 23 performs best, but for larger samples sizes, the four-predictor model no. 30 takes the lead. Likewise, in the Q-based group and for sample sizes smaller than 2500, the single-predictor model no. 3 is the best but is outperformed by the two-predictor model no. 16 from 2500 until 10 000, which in turn is outperformed by the four-predictor model no. 29 from 10 000 to the end. Across all groups, models no. 3, no. 16 and no. 29 form the lower envelope curve in Fig. 2.5, which means that one of them is always the best model choice, depending on the sample size.

Interestingly, the best-performing model for large sample sizes (no. 29) includes predictors which reflect the definition criteria that guided manual event detection (Sect. 2.3.1): $Q(t)$ and $Q(t+2)$ contain information about the absolute magnitude of discharge, Q_{RMC} expresses the magnitude of discharge relative to its vicinity, and $e_{p_{27}}(t-1)$ relates it to the requirement of events to be coherent.

Table 2.7: Application I – curse of dimensionality and data size validation for models in Table 2.6.

no.	Predictive model	$\mathbf{H(X)}$ [bit]	$\mathbf{H(X)/H(X)}$[a]	Sample size where $\mathbf{D_{KL}/H(X) \leq 5\%}$ % of the full dataset[b]	Sample size [a]	Number of bins
0	e	0.516	100%	$\geq$ 4398 (5.6%)	0.5	2
no.	**Predictive model**	$\mathbf{H(X\|Y)}$ [bit]	$\mathbf{H(X\|Y)/H(X)}$[a]	**Sample size where** $\mathbf{D_{KL}/H(X\|Y) \leq 5\%}$ **% of the full dataset**[b]	**Sample size [a]**	**Number of bins**
3	$e \mid Q(t)$	0.260	50.4%	$\geq$ 9952 (12.6%)	1.1	68
16	$e \mid Q(t)$, Q_{RMC}	0.182	35.3%	$\geq$ 29 460 (37.3%)	3.4	748
23	$e \mid Q(t)$, P	0.248	48.2%	$\geq$ 18 880 (23.9%)	2.2	2108
27	$e \mid Q(t)$, Q_{RMC}, $Q(t{+}2)$	0.144	28.0%	$\geq$ 60 178 (76.3%)	6.9	25 432
28	$e \mid Q(t)$, P, Q_{RMC}	0.167	32.5%	$\geq$ 50 377 (63.8%)	5.8	23 188
29	$e \mid Q(t)$, Q_{RMC}, $Q(t{+}2)$, $e_{p_{27}}(t{-}1)$	0.114	22.2%	$\geq$ 69 102 (87.6%)	7.9	279 752
30	$e \mid Q(t)$, P, Q_{RMC}, $e_{p_{28}}(t{-}1)$	0.142	27.6%	$\geq$ 62 667 (79.4%)	7.2	255 068

[a] $H(X) = H(e) = 0.516$ *bits*.
[b] Size of the full dataset: 78 912 data points (9 years).

We also investigated the contribution of sample size effects to total uncertainty by analyzing the ratio of D_{KL} and $H(X|Y)$ as described in Sect. 2.2.3. As expected, for all models the contribution of sample size effects to total uncertainty decreases with increasing sample size, but the absolute values and the rate of change strongly differ. For the one-predictor model no. 3, the D_{KL} contribution is small already for small sample sizes (circa 65% for a sample size equal to 50), and it quickly drops to almost zero with increasing sample size. For multi-predictor models such as no. 29, the D_{KL} and $H(X|Y)$ contribution to uncertainty exceeds that of $H(X|Y)$ by a factor of 7 for small samples (circa 700% for sample size equal 50), and it decreases only slowly with increasing sample size.

In Table 2.7 (fifth column), we show the minimum sample size to keep the D_{KL} contribution below a threshold of 5% for each model.

As expected, the models with few predictors require only small samples to meet the 5% requirement (starting from a subset of 12.6% of the full dataset for the one-predictor model to 37.3% for the two-predictor model), but for multi-predictor models such as models no. 29 and no. 30, more than 60 000 data points are required (87.6% and 79.4% of the full dataset, respectively). This happens because the greater the number of predictors, the greater the number of bins in the model. This means that we need a much larger dataset to populate the PMF with the largest number of bins; for example, model no. 29 has 279 752 bins and requests 7.9 years of data. Considering that the amount of data available in the study is limited, this also means that increasing the number of predictors and/or bins also increases the risk of creating an overfitted or non-robust model. Thus, the ratio $D_{\mathrm{KL}}/H(X|Y)$ and visual inspection of the curve in Fig. 2.5 orientate the user when to stop adding new predictors to avoid overfitting. In this fashion, Table 2.7 shows that each of the models tested meets the 5% requirement, claiming up to 87.6% of the available dataset (69 102 out of 78 912 data points for

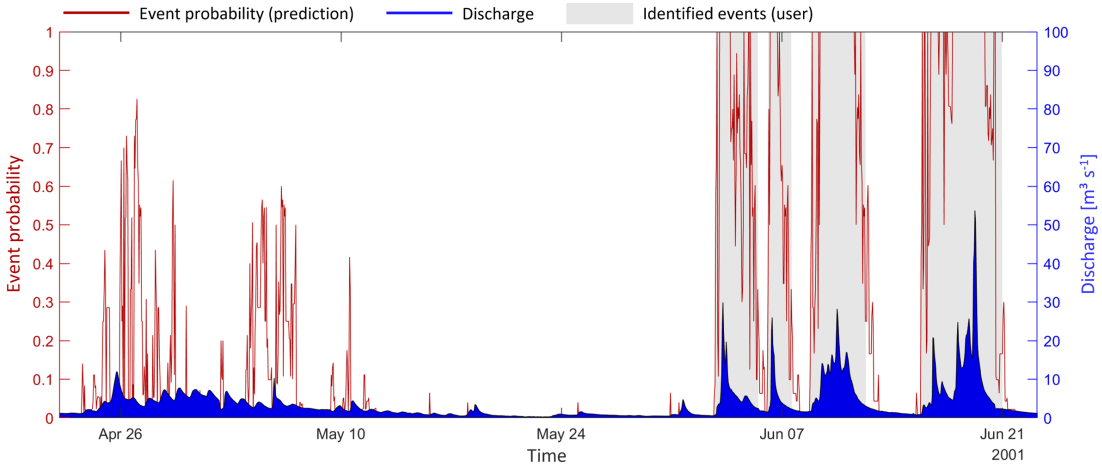

Figure 2.6: Application I – probabilistic prediction of four-predictor model no. 29 (Table 2.5) for a subset of the training data.

model no. 29), which indicates that all of them are robustly supported by the data. In this case, we can confidently choose the best-performing model among them (no. 29, with uncertainty equal to 0.114 bits) for further use. Interestingly, with this analysis, it was also possible to identify the drivers of the user classification, which, in the case of model no. 29, were the predictors $Q(t)$, Q_{RMC}, $Q(t+2)$ and $e_p(t-1)$.

2.4.1.3 *Model application*

In the previous sections, we developed, compared and validated a range of models to reproduce subjective, manual identification of events in a discharge time series. Given the available data, the best model was a four-predictor recursive model built with the full dataset and $Q(t)$, Q_{RMC}, $Q(t+2)$ and $e_p(t-1)$ as predictors (no. 29; Table 2.7). This model reduced the initial predictive uncertainty by 77.8%, decreasing conditional entropy from 0.516 to 0.114 bits. This sounds reasonable, but what do the model predictions actually look like? As an illustration, we applied the model to a subset of the training data, from 22 April to 22 June 2001. For this period, the observed discharge, the manual event classification by the user and the model-based prediction of event probability are shown in Fig. 2.6.

In the period from 1 to 21 June, four distinct rainfall-runoff events occurred which were also classified as such by the user. During these events, the model-based predictions for event probability remained consistently high, except for some times at the beginning and end of events or in times of low flow during an event. Obviously, the model here agrees with the user classification, and if we wished to obtain a binary classification from the model, we could get it by introducing an appropriate probability threshold (as further described in Sect. 2.4.2).

Things look different, though, in the period of 26 April to 10 May, when snowmelt induced diurnal discharge patterns. During this time, the model identified several periods with reasonable (above 50%) event probability, but the user classified the entire period as a non-event. Arguably, this is a difficult case for both manual and automated classification, as the overall discharge is elevated, but it is not elevated by much, and diurnal events can be distinguished but are not pronounced. In such

Table 2.8: Cross-validation dataset – characteristics of the user event classification set.

Dataset	Time steps classified as positive events (P)	Time steps classified as non-events (N)	Percentage of events (P/T)	Percentage of events (N/T)	Total (T)
Training	8150	60 952	11.8%	88.2%	69 102
Testing	942	8868	9.6%	90.4%	9810
Sum	9092	69 820	11.5%	9.9%	78 912

cases, both the user-based and the model-based classifications are uncertain and may disagree.

To identify snowmelt events or potentially improve the information contained in the precipitation set, other predictors could have been used in the analysis (such as aggregated precipitation, snow depth, air temperature, nitrate concentrations, moving average of discharge, etc.), or the target could have been classified according to it type (rainfall, snowmelt, upstream reservoir operation, etc.), instead of having a dichotomous outcome, i.e., event and non-event. The choice of target and potential predictors occurs according to user interest and data availability.

Another point that may be of interest to the user is the improvement of the consistency of the event duration. This can be reached by selection of predictors or through a post-processing step. As previously discussed in Sect. 2.3.2.1, by applying a recursive predictor $e_p(t-1)$, a memory effect is incorporated into the model, bringing some inertia for the transition from event to non-event or vice versa. If it is in the user's interest, the memory effect could be further enhanced by adding more recursive predictors, such as $e_p(t-2)$, $e_p(t-3)$ and so on. An alternative option for clearing very short discontinuous time steps or very short events would be to increase event coherence in a post-processing step with an autoregressive model, with model parameters found by maximizing agreement with the observed events.

Finally, in contrast to the evaluation approach presented, where the subsets are compared to the full dataset (subset data plus data not seen during training), the next section will present the evaluation of the ITM and CPM applied for mutually exclusive training and testing sets.

2.4.2 *Results for application II*

Sect. 2.4.1 showed that, for the full dataset, the best model was the recursive one with $Q(t), Q_{RMC}$, $Q(t+2)$ and $e_p(t-1)$ as the drivers of the user classification (model no. 29, Table 2.7), which could be robustly built with a sample size of 69 102. Thus, to assure its robustness for the second application, since we are creating a new PMF based only on the training set, the split of the data (discharge, precipitation and user event classification) divided the 78 912 time steps into two periods composed of (i) 87.6% of the full dataset (69 102 time steps) forming the training dataset (from 31 October 1996 at 01:00 to 18 September 2004 at 06:00 GMT+1) and (ii) the remaining 12.4% (9810 time steps) forming the testing dataset (from 18 September 2004 at 07:00 to 1 November 2005 at 00:00 GMT+1). The characteristics of the user event classification dataset, used as the true classification for accounting the hits and misses of the ITM and CPM, is presented in Table 2.8.

Table 2.9: Application II – ITM and CPM performance.

Event detection method	**True positive** ($\mathbf{P}_T$)	$\mathbf{R}_{TP}$ ($\mathbf{P}_T/\mathbf{P}$ [a])	**False positive** ($\mathbf{P}_F$)	$\mathbf{R}_{FP}$ ($\mathbf{P}_F/\mathbf{N}$ [a])	**Accuracy %** ($\mathbf{P}_T$+$\mathbf{N}_T$ [b])/($\mathbf{P}$ [a]+$\mathbf{N}$ [a])	**Eq. 2.7 distance** [c]
ITM	918	97.5%	1113	12.6%	88.4%	0.13
CPM	796	84.5%	877	9.9%	89.6%	0.18

[a] $P = 942, N = 8868$ (Table 2.8). [b] $N_T = N - P_F$. [c] Distance to the perfect model of the ROC curve.

For model training, input data from both models, the ITM and CPM, were smoothed. First, a 24 h moving average was applied to the discharge of the CPM (this was recommended by the first author of the method, Yiwen Mei, during personal communications in 2018), and to avoid misleading comparison, it was then applied to the probabilities of the ITM right before the binarization. The smoothing improved the results of both models and worked as a post-processing filter which removed some noise (events with a very short duration) and attenuated effects from snowmelt. Note that this is a feature of our training dataset, and it is therefore not necessarily applicable to other similar problems and neither is a required step.

Following the data smoothing, we proceeded with the optimization of the following parameters: the threshold for the probability output of the ITM and rate of no change for the CPM (Sect. 2.3.3). The results of the two models also improved with the optimization performed. The optimum parameters obtained were $p_{\text{threshold}} = 0.26$ and $R_{nc} = -6.6$. For these values, the final distances in the training dataset given by Eq. 2.7 were 0.05 and 0.23 for the ITM and CPM, respectively.

After the model training, the calibrated models were applied to the testing dataset to predict binary events. The event predictions were then compared to the true classification (Table 2.8, testing row), and their hits and misses were calculated in order to evaluate and compare their performance. The results are compiled in Table 2.9.

The quality parameters presented in Table 2.9 show that the ITM true positive rate equals 97.5%, i.e., it is 13.0% higher than the CPM R_{TP}). In contrast, the CPM false negative rate is equal to 9.9%, while the ITM R_{FP} is equal to 12.6% (2.7% higher). These results indicate that the ITM is more likely to predict events than the CPM but at the cost of increasing the false positive rate. Combining these two rates into a single success criterion according to Eq. 2.7 showed that the ITM is slightly superior to the CPM (Table 2.9, last column).

Considering only the hits of the models, both methods performed similarly, reaching almost 90% accuracy, with the CPM being slightly better than the ITM. However, it should be emphasized that although the accuracy of the model gives a good notion of the model hits, it was not used as a criterion for success because it is a myopic criterion for the false event classifications. False positives are essential in the context of event prediction, since most of the data are non-events (88.2% of the training dataset; Table 2.8), and a blind classification of all time steps as being non-event, for example, would overcome the accuracy obtained by both models (90.4% of the testing dataset; Table 2.8), even though it is not a useful model.

As an illustration, in the context of the binary analysis, the observed discharge, the true event classification (manually made by an expert), the ITM-predicted events and

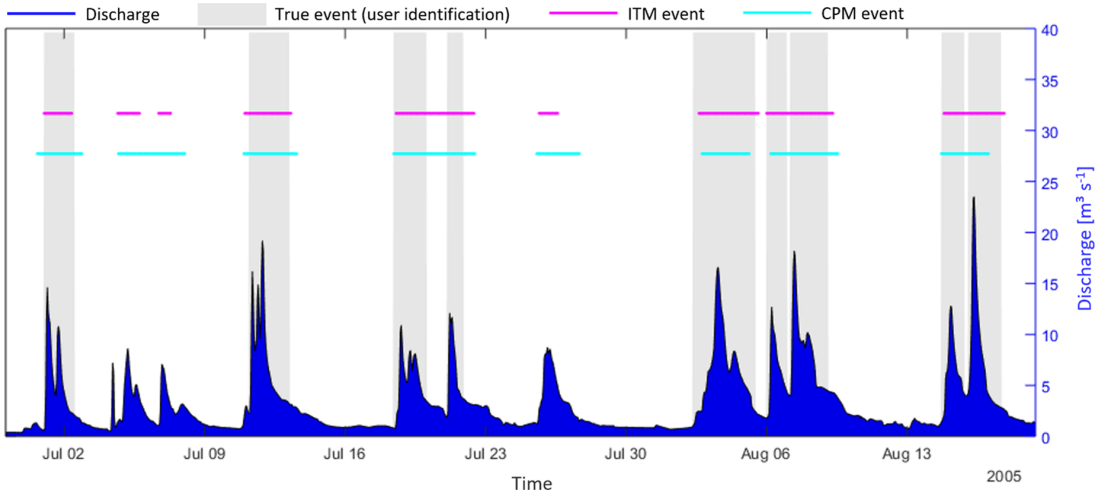

Figure 2.7: Application II – binary prediction of ITM and CPM for a subset of the testing dataset.

CPM-predicted events are shown in Fig. 2.7 for a subset of the testing data, from 29 June to 19 August 2005.

For the analyzed subset, nine distinct rainfall-runoff events occurred and were identified as such by the ITM and CPM. However, different from the true identification, both models grouped some of these events (20 July, 7 and 16 August) with events with longer duration. False events were also observed in both models, where three false events were identified by the ITM (5, 7 and 26 July), and two (but contemplating the same period as the ITM) were identified by the CPM. It should be noted that they are false in relation to the user classification; however, we can not exclude the possibility of false classification by the visual inspection process. A further criticism is that the holdout cross-validation involves a single run, which is not as robust as multiple runs. Nevertheless, the way that the split was proposed recognizes the logical order of obtaining the data. Thus, despite the subjectivity of event selection by a user and the application of a simplified method of cross-validation, it is possible to conclude that, overall, the ITM and CPM behaved similarly and provided reasonable predictions, as seen numerically in Table 2.9 and qualitatively through Fig. 2.7.

An interesting conclusion is that the ITM was able to overcome the CPM while requiring only discharge data and a training dataset of classified events (also based on the discharge set), whereas the CPM demanded precipitation, catchment area and discharge as inputs. It is important to note that the CPM can be modified to be used without precipitation data; however in our case it resulted in a considerably higher false positive rate, since the rainfall event-related filters cannot be applied. In contrast, since the CPM is a physically based approach, it does not require a training dataset with identified events (although the optimization in the calibration step has representatively improved its results), and there are no limitations in terms of dataset size, which eliminates the robustness analysis, being then a method more easily implemented for binary classification. The binarization of the ITM predictions and parameter optimization in the CPM are not included in the original methods, however, they were essential adaptations to allow a fair comparison of the models. Finally, the suitability or not of the existing event detection techniques depends mainly on the user's interest and the data available for application.

2.5 SUMMARY AND CONCLUSIONS

Typically, it is easy to manually identify rainfall-runoff events due to the high discriminative and integrative power of the brain–eye system. However, this is (i) cumbersome for long time series; (ii) subject to handling errors; and (iii) hard to reproduce, since it dependents on acuity and knowledge of the event identifier. To mitigate these issues, this study has proposed an information theory approach to learn from data and to choose the best predictors, via uncertainty reduction, for creating predictive models that automatically identify rainfall-runoff events in discharge time series.

The method was established in four main steps: the model hypothesis, building, evaluation and application. Each association of predictor(s) to the target is equivalent to formulating a model hypothesis. For the model building, nonparametric models constructed discrete distributions via bin-counting, requiring at least a discharge time series and a training dataset containing a yes or no event identification as target. In the evaluation step, we used Shannon entropy and conditional entropy to select the more informative predictors and Kullback–Leibler divergence and cross entropy to analyze the model in terms of overfitting and curse of dimensionality. Finally, the best model was applied to its original dataset to compare the predictability of the events. For the purpose of benchmarking, a holdout cross-validation and a comparison of the proposed data-driven method with an alternative physically based approach were performed.

The approach was applied to discharge and precipitation data from the Dornbirner Ach catchment in Austria. In this case study, 30 models based on 16 predictors were built and tested. Among these, seven predictive models with a number of predictors varying from one to four were selected. Interestingly, across these models, the three best-performing ones were obtained using only discharge-based predictors. The overall best model was a recursive one applying four predictors: discharge from two different time steps, the relative magnitude of discharge compared to all discharge values in a surrounding 65 h time window and event predictions from the previous time step. When applying the best model, the uncertainty of event classification was reduced by 77.8%, decreasing conditional entropy from 0.516 to 0.114 bits. Since the conditional entropy reduction of the models with precipitation was not higher than the ones exclusively based on discharge information, it was possible to infer that (i) the information coming from precipitation was likely already contained in the discharge data series and (ii) the event classification is not so much dependent on precipitation at a particular time step but rather on the accumulated rainfall in the period preceding it. Furthermore, precipitation data are often not available for analysis, which makes the model exclusively based on discharge data even more attractive.

Further analysis using cross entropy and Kullback–Leibler divergence showed that the robustness of a model quickly dropped with the number of predictors used (an effect known as the curse of dimensionality) and that the relation between number of predictors and sample size was crucial to avoid overfitting. Thus, the model choice is a tradeoff between predictive power and robustness, given the available data. For our case, the minimum amount of data to build a robust model varied from 9952 data points (one-predictor model with 0.260 bits of uncertainty) to 69 102 data points (four-predictor model with 0.114 bits of uncertainty). Complementarily, the quality of the model was verified in a more traditional way, by a cross-validation analysis

(where the model was built in a training dataset and validated in a testing dataset), and a comparative investigation between our data-driven approach and a physically based model. As a result, in general, both models presented reasonable predictions and reached similar quality parameters, with almost 90% of accuracy. In the end, the comparative analysis and cross-validation reinforced the quality of the method, previously validated in terms of robustness using measures from information theory.

In the end, the data-driven approach based on information theory is a consolidation of descriptive and experimental investigations, since it allows one to describe the drivers of the model through predictors and investigates the similarity of the model hypothesis with respect to the true classification. In summary, it presents advantages such as the following: (i) it is a general method that involves a minimum of additional assumptions or parameterizations; (ii) due to its nonparametric approach, it preserves the full information of the data as much as possible, which might get lost when expressing the data relations with functional relationships; (iii) it obtains data relations from the data itself; (iv) it is flexible in terms of data requirement and model building; (v) it allows one to measure the amount of uncertainty reduction via predictors; (vi) it is a direct way to account for uncertainty; (vii) it permits explicitly comparing information from various sources in a single currency, the bit; (viii) it allows one to quantify minimal data requirements; (ix) it enables one to investigate the curse of dimensionality; (x) it is a way of understanding the drivers (predictors) of the model (also useful in machine learning, for example); (xi) it one permits to choose the most suitable model for an available dataset; and (xii) the predictions are probabilistic, which compared to a binary classification, additionally provides a measure of the confidence of the classification.

Although the procedure was employed to identify events from a discharge time series, which for our case were mainly triggered by rainfall and snowmelt, the method can be applied to reproduce user classification of any kind of event (rainfall, snowmelt, upstream reservoir operation, etc.) and even identify them separately. Moreover, one of the strengths of the data-based approach is that it potentially accepts any data to serve as predictors, and it can handle any kind of relation between the predictor(s) and the target. Thus, the proposed approach can be conveniently adapted to another practical application.

Part III

HISTOGRAM VIA ENTROPY REDUCTION (HER): AN INFORMATION-THEORETIC ALTERNATIVE FOR GEOSTATISTICS

This study is published in the scientific journal Hydrology and Earth System Science (HESS) and is a reprint of:

Thiesen, Stephanie; Vieira, Diego M.; Mälicke, Mirko; Loritz, Ralf; Wellmann, J. Florian; Ehret, Uwe (2020): Histogram via entropy reduction (HER) – an information-theoretic alternative for geostatistics, Hydrology and Earth System Sciences, 24(9), 4523-4540. doi: 10.5194/hess-24-4523-2020

3

HISTOGRAM VIA ENTROPY REDUCTION (HER): AN INFORMATION-THEORETIC ALTERNATIVE FOR GEOSTATISTICS

ABSTRACT

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

3.1 INTRODUCTION

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

Stochastic geostatistical approaches, such as ordinary kriging (OK), have been widely studied and applied in various disciplines since their introduction to geology and mining by Krige (1951), bringing significant results in the context of environmental sciences. However, like other parametric regression methods, it relies on prior assumptions about theoretical functions and, therefore, includes the risk of suboptimal performance due to suboptimal user choices (Yakowitz and Szidarovszky, 1985). OK uses fitted functions to offer uncertainty estimates, while deterministic estimators (NN and IDW) avoid function parameterizations at the cost of neglecting uncertainty analysis. In this sense, researchers are confronted with the trade-off between avoiding parameterization 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 as "data-driven" methods) have become increasingly popular as a substitute for or complement to established modeling approaches. In the context of data-based modeling in the environmental sciences, concepts and measures from information theory are being used for describing and inferring relations among data (Liu et al., 2016; Mälicke et al., 2020; Thiesen et al., 2019), quantifying uncertainty and evaluating model performance (Chapman, 1986; Liu et al., 2016; Thiesen et al., 2019), estimating information flow (Darscheid, 2017; Weijs, 2011), and measuring similarity, quantity, and quality of information in hydrological models (Loritz et al., 2018, 2019; Nearing and Gupta, 2017). In the spatial context, information-theoretic measures were used to obtain longitudinal profiles of rivers (Leopold and Langbein, 1962), to solve problems of spatial aggregation and quantify information gain, loss, and redundancy (Batty, 1974; Singh, 2013), to analyze spatiotemporal variability (Brunsell, 2010; Mishra et al., 2009), to address risk of landslides (Roodposhti et al., 2016), and to assess spatial dissimilarity (Naimi, 2015), complexity (Pham, 2010), uncertainty (Wellmann, 2013), and heterogeneity (Bianchi and Pedretti, 2018).

Most of the popular data-driven methods have been developed in the computational intelligence community and, since they are not built for solving particular problems, applying these methods remains a challenge for the researchers outside this field (Solomatine and Ostfeld, 2008). The main issues for researchers in hydroinformatics for applying data-driven methods lie in testing various combinations of methods for particular problems, combining them with optimization techniques, developing robust modeling procedures able to work with noisy data, and providing the adequate model uncertainty estimates (Solomatine and Ostfeld, 2008). To overcome these challenges and the mentioned parameterization–uncertainty trade-off in the context

of spatial interpolation, this paper is concerned with formulating and testing a novel method based on principles of geostatistics, information theory, and probability aggregation methods to describe spatial patterns and to obtain stochastic predictions. In order to avoid fitting of spatial correlation functions and assumptions about the underlying distribution of the data, it relies on empirical probability distributions to (i) extract the spatial dependence structure of the field, (ii) minimize entropy of predictions, and (iii) produce stochastic estimation of unsampled points. Thus, the proposed histogram via entropy reduction (HER) approach allows nonparametric and stochastic predictions, avoiding the shortcomings of fitting deterministic curves and, therefore, the risk of adding information not contained in the data, but still relying on geostatistical concepts. HER is seen as a solution in between geostatistics (knowledge driven) and statistical learning (data driven) in the sense that it allows automated learning from data bounded by a geostatistical framework.

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

The paper is organized as follows. The method is presented in Sect. 3.2. In Sect. 3.3, we describe the data properties, performance parameters, validation design, and benchmark models. In Sect. 3.4, we explore the properties of three different aggregation methods, present the results of HER for different samples sizes and data types, compare the results to benchmark models, and, in the end, discuss the achieved outcomes and model contributions. Finally, we draw conclusions in Sect. 3.5.

3.2 METHOD DESCRIPTION

Histogram via entropy reduction method (HER) has three main steps, namely (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. The first and third steps are shown in Fig. 3.1.

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

3.2.1 *Information theory*

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

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x). \tag{3.1}$$

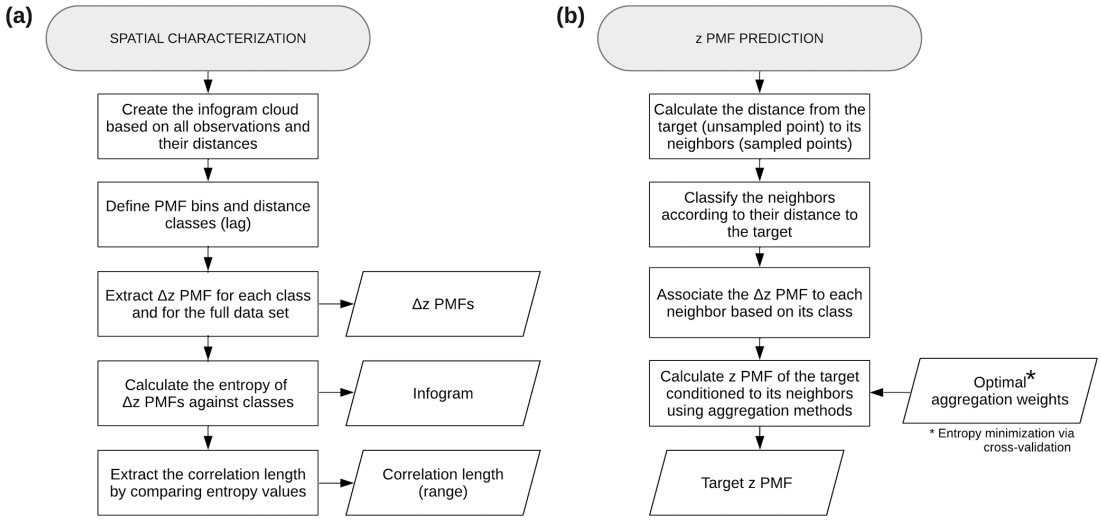

Figure 3.1: HER method. Flowcharts illustrating **(a)** spatial characterization and **(b)** z probability mass function (PMF) prediction.

We use the logarithm to base two so that the entropy is expressed in bits. Each bit corresponds to an answer to one optimal yes–no question asked with the intention of reconstructing the data. It varies from zero to $\log_2 n$, where n represents the number of bins of the discrete distribution. In the study, Shannon entropy is used to extract the infogram and correlation length of the dataset (explored in Sect. 3.2.2).

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

$$D_{\mathrm{KL}}(p||q) = \sum_{x \in X} p\,(x)\ \log_2 \frac{p(x)}{q(x)}. \tag{3.2}$$

Also referred to as relative entropy, $D_{\mathrm{KL}}(p||q)$ can be understood as a measure of information loss of assuming that the distribution is q when in reality it is p (Weijs et al., 2010). It is nonnegative and is zero strictly if $p = q$. In HER context, Kullback–Leibler divergence is optimized to select the weights for aggregating distributions (detailed in Sect. 3.2.3). The measure is also used as a scoring rule for performance verification of probabilistic predictions (Gneiting and Raftery, 2007; Weijs et al., 2010).

Note that the measures presented by Eqs. 3.1 and 3.2 are defined as functionals of probability distributions and do not depend 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.

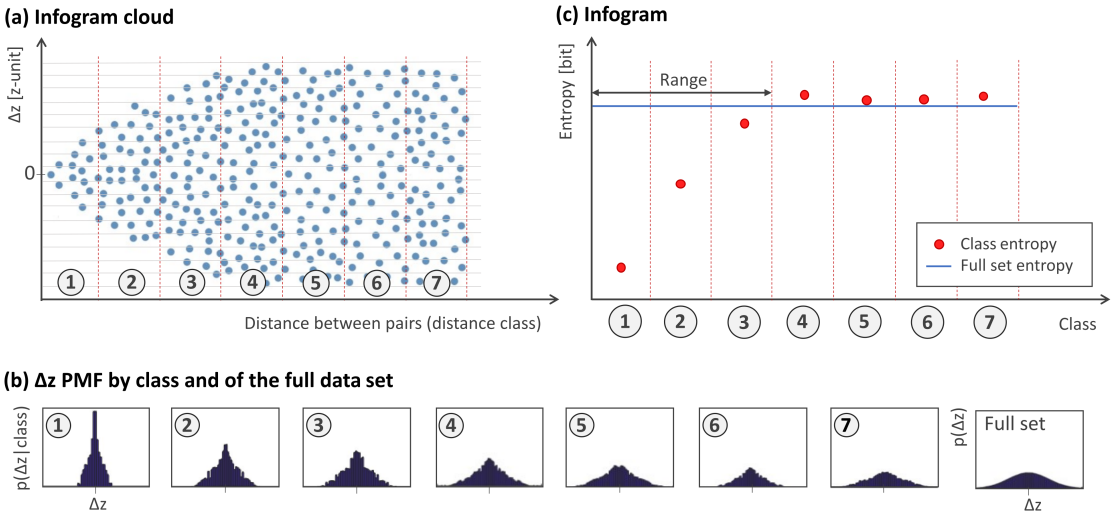

Figure 3.2: Spatial characterization. Illustration of **(a)** infogram cloud, **(b)** Δz probability mass functions (PMFs) by class, and **(c)** infogram.

3.2.2 *Spatial characterization*

The spatial characterization (Fig. 3.1a) 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. To capture the spatial variability and related uncertainties, concepts of geostatistics and information theory are integrated into the method. As shown in Fig. 3.1a, the spatial characterization phase aims to, first, obtain Δz probability mass functions (PMFs), where z is the variable under study; second, the behavior of entropy as a function of lag distance (which the authors denominate as "infogram"); and, finally, the correlation length (range). These outputs are outlined in Fig. 3.2 and attained in the following steps:

(i) Infogram cloud (Fig. 3.2a): calculate the difference in the z values (Δz) between pairs of observations; associate each Δz to the Euclidean separation distance of its respective point pair. Define the lag distance (demarcated by red dashed lines), here called distance classes or, simply, classes. Divide the range of Δz values into a set of bins (demarcated by horizontal gray lines);

(ii) Δz PMFs (Fig. 3.2b): construct, for each distance class, 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. 3.2c): calculate the entropy of each Δz PMF and of the unconditional PMF. Compute the range of the data; this is the distance at which the conditional entropy exceeds the unconditional entropy. Beyond this point, the neighbors start becoming uninformative, and it is pointless to use information outside of this neighborhood.

The infogram cloud is the preparation needed for constructing the infogram. It contains a complete cloud of point pairs. The infogram plays a role similar to that of

the variogram; through the lens of information theory, we can characterize the spatial dependence of the dataset, calculate the spatial (dis-)similarities, and compute its correlation length (range). It describes the statistical dispersion of pairs of observations for the distance class separating these observations. Quantitatively, it is a way of measuring the uncertainty about Δz given the class. Graphically, the infogram shape is the fingerprint of the spatial dependence, where the larger the entropy of one class, the more uncertain (disperse) its distribution. It reaches a threshold (range) where the data no longer show significant spatial correlation. We associate neighbors beyond the range to the Δz PMF of the full dataset. By doing so, we restrict ourselves to the more informative classes and reduce the number of classes to be mapped, thus improving the results and the speed of calculation. Note that, in the illustrative case of Fig. 3.2, 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. 3.5 in Sect. 3.4.1).

Naimi (2015) introduced a similar concept to the infogram called an 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 to quantify the degree of spatial order and rank different structures. Both works, and the present study, are carried out with a variogram-like shape and entropy-based measures and are looking for data (dis-)similarity, yet with different purposes and metrics. The proposed infogram terminology seeks to provide an easy-to-follow association with the quantification of information available in the data.

Converting the frequency distributions of Δz into PMFs requires a cautious choice of bin width, since this decision will frame the distributions used as the model and directly influence the statistics we compute for evaluation (D_{KL}). Many methods for choosing an appropriate binning strategy have been suggested (Gong et al., 2014; Knuth, 2013; Pechlivanidis et al., 2016; Thiesen et al., 2019). 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 the 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 of introducing minimal prior information (Knuth, 2013). The bin size was defined, based on Thiesen et al. (2019), by comparing the cross entropy $H_{\mathrm{pq}} = H(p) + D_{\mathrm{KL}}(p||q))$ between the full learning set and subsamples for various bin widths. The selected one shows a stabilization of the cross entropy for small sample sizes, meaning that the bin size is reasonable for small and large sample sizes and analyzed distribution shapes. For favoring comparability, the bins are kept the same for all applications and performance calculations.

Additionally, to avoid distributions with empty bins, which might make the PMF combination (discussed in Sect. 3.2.3.1) unfeasible, 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 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, to fit 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.

3.2.3 *Minimization of estimation entropy*

To infer the conditional distribution of the target z_0 (unsampled point) given its neighbors z_i (where $i = 1, \ldots, n$ are the indices of the sampled points), we use the Δz PMFs obtained at the spatial characterization step (Sect. 3.2.2). To do so, each neighbor z_i is associated to a class and, hence, to a Δz distribution according to their distance to the target z_0. This implies the assumption that the empirical Δz PMFs apply everywhere in the field, irrespective of specific location, and only depend on the distance between points. Each Δz PMF is then shifted by the z_i value of the observation it is associated to, yielding the z PMF of the target given the neighbor i, which is denoted by $p(z_0|z_i)$. Assume, for instance, three observations, z_1, z_2, and z_3, for which we want to predict the probability distribution of the target z_0. In this case, what we infer at this stage is the conditional probability distributions, $p(z_0|z_1)$, $p(z_0|z_2)$, and $p(z_0|z_3)$.

Now, since we are in fact interested in the probability distribution of the target conditioned to multiple observations, namely $p(z_0|z_1, z_2, z_3)$, how can we optimally combine the information gained from individual observations to predict this target probability? In the next sections, we address this issue by using aggregation methods. After introducing potential ways to combine PMFs (Sect. 3.2.3.1), we propose an optimization problem, via entropy minimization, to define the weight parameters needed for the aggregation (Sect. 3.2.3.2).

3.2.3.1 *Combining distributions*

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

The main objective of this process is to aggregate probability distributions coming from different sources into a global probability distribution. For this purpose, the computation of the full conditional probability $p(z_0|z_1, \ldots, z_n)$ – where z_0 is the event we are interested in (target), and z_i with $i = 1, \ldots, n$ is a set of data events (or

neighbors) – is obtained by the use of an aggregation operator, P_G, called pooling operator, with the following:

$$p\left(z_0|z_1,\ldots,z_n\right) \approx P_G\left(p(z_0|z_1),\ldots,p(z_0|z_n)\right). \tag{3.3}$$

From now on, we will adopt a similar notation to that of Allard et al. (2012), using the more concise expressions $P_i(z_0)$ to denote $p(z_0|z_i)$ and $P_G(z_0)$ for the global probability, $P_G\left(P_1(z_0),\ldots,P_n(z_0)\right)$.

The most intuitive way to aggregate the probabilities $p_1,\ldots,p_n$ is by linear pooling, which is defined as follows:

$$P_{G_{OR}}(z_0) = \sum_{i=1}^{n} w_{OR_i}\, P_i(z_0), \tag{3.4}$$

where n is the number of neighbors, and w_{OR_i} are positive weights verifying $\sum_{i=1}^{n} w_{OR_i} = 1$. Eq. 3.4 describes mixture models in which each probability p_i represents a different population. If we set equal weights w_{OR_i} to every probability P_i the method reduces to an arithmetic average, coinciding with the disjunction of probabilities proposed by Tarantola (2005) and Tarantola and Valette (1982), as illustrated in Fig. 3.3b. Since it is a way of averaging distributions, the resulting distribution $P_{G_{OR}}$ is often multimodal. Additive methods, such as linear pooling, are related to union of events and to the logical operator OR.

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

$$\ln P_{G_{AND}}\ (z_0) = \ln\zeta + \sum_{i=1}^{n} w_{AND_i} \ln P_i\left(z_0\right), \tag{3.5}$$

or equivalently $P_{G_{AND}}(z_0) \propto \prod_{i=1}^{n} P_i\left(z_0\right)^{w_{AND_i}}$, where ζ is a normalizing constant, n is the number of neighbors, and w_{AND_i} are positive weights. One particular case consists of setting $w_{AND_i} = 1$ for every i. This refers to the conjunction of probabilities proposed by Tarantola (2005) and Tarantola and Valette (1982), as shown in Fig. 3.3c. In contrast to linear pooling, log-linear pooling is typically unimodal and less dispersed.

Aggregation methods are not limited to the 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. 3.5) to fields with continuous properties. This idea is supported by Journel (2002), who remarked that a logarithmic expression evokes the simple kriging expression (used for continuous variables). For example, if we have two points z_1 and z_2 with different values and want to estimate the target z_0 at a location between them in a continuous field, we would expect that the estimate z_0 would be somewhere between z_1 and z_2, which can be achieved by an AND combination. In a more intuitive way, if we notice that, for kriging, the shape of the predicted distribution is assumed to be fixed (Gaussian, for example), multiplying two distributions with different means would result in a Gaussian distribution as well, less dispersed than the original ones, as also seen

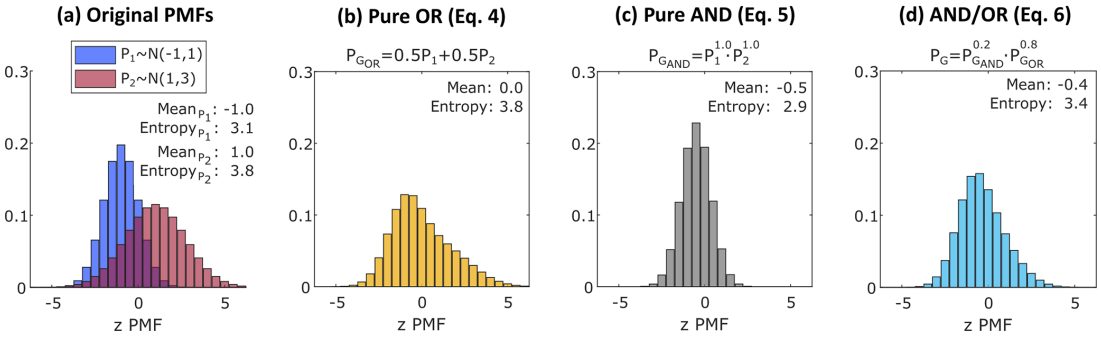

Figure 3.3: Examples of the different pooling operators. Illustration of **(a)** normal PMFs $\mathcal{N}(\mu, \sigma^2)$ to be combined; **(b)** linear aggregation of (a), Eq. 3.4; **(c)** log-linear aggregation of (a), Eq. 3.5; and **(d)** log-linear aggregation of (b) and (c), Eq. 3.6.

for the log-linear pooling. It is worth mentioning that some methods for modeling spatially dependent data, such as copulas (Bárdossy, 2006; Kazianka and Pilz, 2010) and effective distribution models (Hristopulos and Baxevani, 2020), also use log-linear pooling to construct 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 the idea that phenomena that show abrupt changes should be modeled as mixture of populations. In this case, if we have two points z_1 and z_2 belonging to different categories, a target z_0 between them will either belong to the category of z_1 or z_2, which can be achieved by the mixture distribution given by the OR pooling. In other words, the OR aggregation is a way of combining information from different sides of the truth; thus, a conservative way of considering the available information from all sources.

Note that, for both linear and log-linear pooling, weights equal to zero will lead to uniform distributions, therefore bypassing the PMFs 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 at the end of this section.

The selection of the most suitable aggregation method depends on the specific problem (Allard et al., 2012), and it will influence the PMF prediction and, therefore, the uncertainty structure of the field. Thus, depending on the knowledge about the field, a user can either add information to the model by applying an a priori chosen aggregation method or infer these properties from the field. Since, in practice, there is often a lack of information to accurately describe the interactions between the sources of information (Allard et al., 2012), inference is the approach we tested in the comparison analysis (Sect. 3.4.2). For that, we propose estimating the distribution P_G of a target, by combining $P_{G_{\text{AND}}}$ and $P_{G_{\text{OR}}}$, as follows:

$$P_G(z_0) \propto P_{G_{\text{AND}}}(z_0)^{\alpha} \; P_{G_{\text{OR}}}(z_0)^{\beta}, \tag{3.6}$$

where α and β are positive weights varying from zero to one, which will be found by optimization. Eq. 3.6 is the choice made by the authors as a way of balancing both

natures of the PMF aggregation. The idea is to find the appropriate proportion of α (continuous) and β (discontinuous) properties of the field by minimizing the estimated relative entropy. Note that, when the weight α or β is set to zero, the final distribution results, respectively, in a pure OR, Eq. 3.4, or pure AND aggregation, Eq. 3.5, as special cases. The equation is based on the log-linear aggregation, as opposed to linear aggregation, since the latter is often multimodal, which is an undesirable property for geoscience applications (Allard et al., 2012). Alternatively, Eqs. 3.4 and 3.5 or a linear pooling of $P_{G_{\text{AND}}}(z_0)$ and $P_{G_{\text{OR}}}(z_0)$ could be used. We explore the properties of the linear and log-linear pooling in Sect. 3.4.1.

The practical differences between the pooling operators used in this paper are illustrated in Fig. 3.3, where Fig. 3.3a introduces two PMFs to be combined, and Figs. 3.3b to 3.3d show the resulting PMFs for Eqs. 3.4 to 3.6, respectively. In Fig. 3.3b, we use equal weights for both PMFs, and the resulting distribution is the arithmetic average of the bin probabilities. In Fig. 3.3c, 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. Figure 3.3d shows a log-linear aggregation of the two previous distributions (Fig. 3.3b,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.

3.2.3.2 *Weighting PMFs*

Scoring rules assess the quality of probabilistic estimations (Gneiting and Raftery, 2007) and, therefore, can be used to estimate the parameters of a pooling operator (Allard et al., 2012). We selected Kullback–Leibler divergence (D_{KL}, Eq. 3.2) as the loss function to optimize α and β, Eq. 3.6, and the w_{OR_k} and w_{AND_k} weights (Eqs. 3.4 and 3.5, 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-theoretic point of view by Roulston and Smith (2002), is a strictly proper scoring rule since it provides summary metrics that address calibration and sharpness simultaneously by rewarding narrow prediction intervals and penalizing intervals missed by the observation (Gneiting and Raftery, 2007).

By means of a 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 so 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. 3.4 and 3.5, we assign one weight to each distance class k. This means that, given a target z_0, the neighbors grouped in the same distance class will be assigned the same weight. For a more continuous weighting of the neighbors, as an extra step we linearly interpolate the weights according to the Euclidean distance and the weight of

the next class. Another option could be narrowing down the class width, in which case more data are needed to estimate the respective PMFs.

Firstly, we obtained, in parallel, the weights of Eqs. 3.4 and 3.5 by convex optimization, and later α and β by grid search with both weight values ranging from 0 to 1 (steps of 0.05 were used in the application case). In order to facilitate the convergence of the convex optimization, the following constraints were employed: (i) for linear pooling, set $w_{\mathrm{OR}_1} = 1$, to avoid non-unique solutions; (ii) force weights to decrease monotonically (i.e., $w_{k+1} \leq w_k$); (iii) define a lower bound to avoid numerical instabilities (e.g., $w_k \geq 10^{-6}$); iv) define an upper bound ($w_k \leq 1$). Finally, after the optimization, normalize the weights to verify $\sum_k w_{\mathrm{OR}_k} = 1$ for linear pooling (for log-linear pooling, the resulting PMFs are normalized).

In order to increase computational efficiency, and due to the minor contribution of neighbors in classes far away from the target, the authors only used the 12 neighbors closest to the target when optimizing α and β and when predicting the target. Note that this procedure is not applicable for the optimization of the w_{OR_k} and w_{AND_k} weights, since we are looking for one weight w_{k} for each class k, and therefore, we cannot risk neglecting those classes for which we have an interest in their weights. For the optimization phase discussed here, and for the prediction phase (in next section), the limitation of the number of neighbors together with the removal of classes beyond the range are efficient means of reducing the computational effort involved in both phases.

3.2.4 *Prediction*

With the results of the spatial characterization step (classes, Δz PMFs, and range, as described in Sect. 3.2.2), the definition of the aggregation method and its parameters (Sect. 3.2.3.1 and 3.2.3.2, respectively), and the set of known observations, we have the model available to predict distributions.

Thus, to estimate a specific unsampled point (target), first, we calculate the Euclidean distance from the target to its neighbors (sampled observations). Based on this distance, we obtain the class of each neighbor and associate to each its corresponding Δz PMF. As mentioned in Sect. 3.2.2, neighbors beyond the range are associated to the Δz PMF of the full dataset. To obtain the z PMF of target z_0 given each neighbor z_i, we simply shift the Δz PMF of each neighbor by its z_i value. Finally, by applying the defined aggregation method, we combine the individual z PMFs of the target given each neighbor to obtain the PMF of the target conditional on all neighbors. Fig. 3.1b presents the z PMF prediction steps for a single target.

3.3 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.

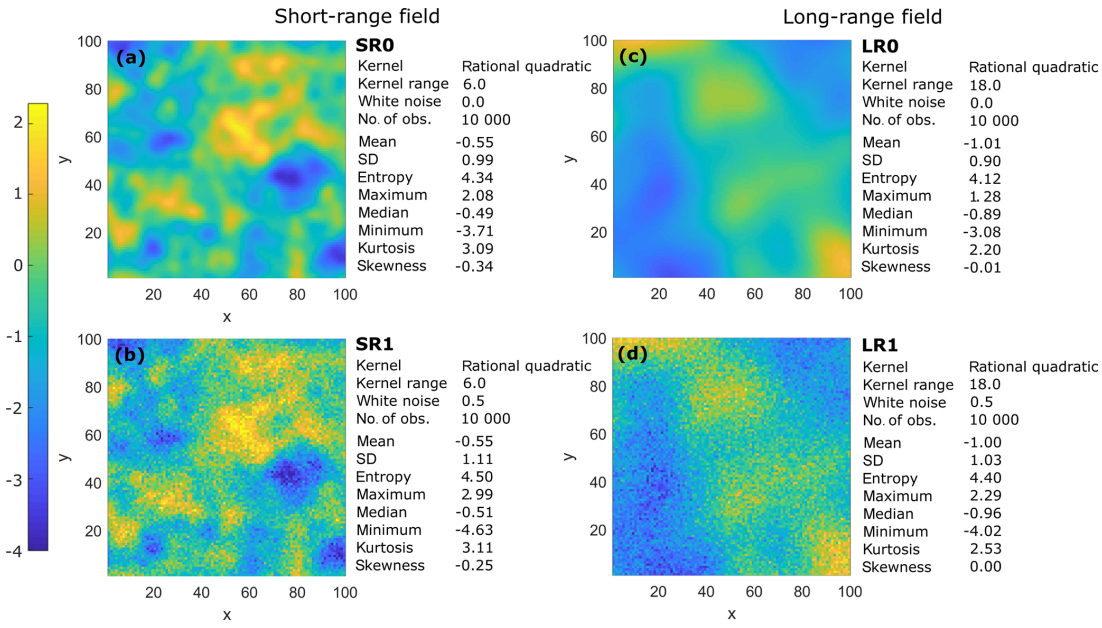

Figure 3.4: Synthetic fields and summary statistics. **(a)** Short-range field without noise (SR0), **(b)** short-range field with noise (SR1), **(c)** long-range field without noise (LR0), and **(d)** long-range field with noise (LR1).

3.3.1 *Data properties*

To test the proposed method in a controlled environment, four synthetic 2D spatial datasets with grid size 100×100 were generated from known Gaussian processes. A Gaussian process is a stochastic method that is specified by its mean and a covariance function or kernel (Rasmussen and Williams, 2006). The data points are determined by a given realization of a prior, which is randomly generated from the chosen kernel function and the associated parameters. In this work, we used a rational quadratic kernel (Pedregosa et al., 2011) as the covariance function, with two different correlation length parameters for the kernel, namely 6 and 18 units, to produce two datasets with fundamentally different spatial dependence. For both short- and long-range fields, white noise was introduced by a Gaussian distribution, with a mean of zero and standard deviation equal to 0.5. The implementation was taken from the Python library, namely scikit-learn (Pedregosa et al., 2011). The generated sets comprise (i) a short-range field without noise (SR0), (ii) a short-range field with noise (SR1), (iii) a long-range field without noise (LR0), and (iv) a long-range field with noise (LR1). Fig. 3.4 presents the field characteristics and their summary statistics. The summary statistics of each field type are included in Appendix B.1.

3.3.2 *Performance criteria*

To evaluate the predictive power of the models, a quality assessment was carried out with three criteria, namely mean absolute error (E_{MA}) and Nash–Sutcliffe efficiency (E_{NS}), for the deterministic cases, and mean of the Kullback–Leibler divergence (D_{KL}), for the probabilistic cases. E_{MA} was selected because it gives the same weight to all

errors, while E_{NS} penalizes variance as it gives more weight to errors with larger absolute values. E_{NS} also shows a normalized metric (limited to one), which favors general comparison. All three metrics are shown in Eqs. 3.7, 3.8 and 3.2, 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 follows:

$$E_{MA} = \frac{1}{n} \sum_{i=1}^{n} |\widehat{z}_i - z_i| \,, \tag{3.7}$$

$$E_{NS} = 1 - \frac{\sum_{i=1}^{n} (\widehat{z}_i - z_i)^2}{\sum_{i=1}^{n} (z_i - \overline{z})^2}, \tag{3.8}$$

where $\widehat{z}_i$ and z_i are, respectively, the predicted and observed values at the ith location, $\overline{z}$ is the mean of the observations, and n is the number of tested locations. For the probabilistic methods, $\widehat{z}_i$ 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 in which the observed value is located; i.e., in Eq. 3.2, we set p equal to one for the corresponding bin and compared it to the probability value of the same bin in the predicted distribution. This procedure is just applicable to probabilistic models, and it enables one to measure how confident the model is in predicting the correct observation. In order to calculate this metric for ordinary kriging, we must convert the predicted probability density functions (PDFs) to PMFs, employing the same bins used in HER.

3.3.3 *Calibration and test design*

To benchmark and investigate the effect of sample size, we applied holdout validation as follows. Firstly, we randomly shuffled the data, and then divided it into 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) were used as the test set. We calibrated the models on learning subsets with increasing sizes of 200, 400, 600, 800, 1000, 1500, and 2000 observations. We used the validation set for fine adjustments and plausibility checks. To avoid multiple calibration runs, the resampling was designed in a way that the learning subsets increased in size by adding new data to the previous subset; i.e., the observations of small sample sizes were always contained in the larger sets. To facilitate model comparison, the validation and test datasets were fixed for all performance analyses, independently of the analyzed learning set. This procedure also avoided variability of results coming from multiple random draws since, by construction, we improved the learning with growing sample size, and we always assessed the results in the same set. The test set was kept unseen until the final application of the methods, as a "lock-box approach" (Chicco, 2017), and its results were used to evaluate the model performance presented in Sect. 3.4. See Appendix B.1 for the summary statistics of the learning, validation, and test subsets.

3.3.4 *Benchmark interpolators*

In addition to presenting a complete application of HER (Sect. 3.4.1), a comparative analysis among the best-known and used methods for spatial interpolation in the earth sciences (Li and Heap, 2011; Myers, 1993) is performed (Sect. 3.4.2). Covering deterministic, probabilistic, and geostatistical methods, three interpolators were chosen for the comparison, namely nearest neighbor (NN), inverse distance weighting (IDW), and ordinary kriging (OK).

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

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

NN and IDS do not require calibration. To calibrate HER aggregation weights, we applied LOOCV, as described in Sect. 3.2.3.2, to optimize the performance of the left-out sample in the learning set. As the loss function, the minimization of the mean D_{KL} was applied. After learning the model, we used the validation set for plausibility check of the calibrated model and, eventually, adjustment of parameters. Note that no function fitting is needed to apply HER.

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

The selection of lag size has important effects on the HER infogram and, as discussed in Oliver and Webster (2014), on the empirical variogram of OK. However, since the goal of the benchmarking analysis was to find a fair way to compare the methods, we fixed the lag distances of OK and HER at 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.

3.4 RESULTS AND DISCUSSION

In this section, three analyses are presented. Firstly, we explore the results of HER using three different aggregation methods on one specific synthetic dataset (Sect. 3.4.1). In Sect. 3.4.2, we summarize the results of the synthetic datasets LR0, LR1, SR0, and SR1 for all calibration sets and numerically compare HER performance with traditional interpolators. For all applications, the performance was calculated on the same test set. For brevity, the model outputs were omitted in the comparison analysis, and only the performance metrics for each dataset and interpolator are shown. Finally, Sect. 3.4.3 provides a theoretical discussion on the probabilistic methods (OK and HER), contrasting their different properties and assumptions.

3.4.1 *HER application*

This section presents three variants of HER, applied to the LR1 field with a calibration subset of 600 observations (LR1-600). This dataset was selected since, due to its optimized weights, α and β (which reach almost the maximum value of one suggested for Eq. 3.6), it favors contrasting the uncertainty results of HER when applying the three distinct aggregation methods proposed in Eqs. 3.4–3.6.

As a first step, the spatial characterization of the selected field is obtained and shown in Fig. 3.5. For brevity, only the odd classes are shown in Fig. 3.5b. In the same figure, the Euclidean distance (in grid units) relative to the class is indicated after the class name in interval notation (left-open, right-closed interval). For both z PMFs and Δz PMFs, a bin width of 0.2 (10% of the distance class width) was selected and kept the same for all applications and performance calculations. As mentioned in Sect. 3.3.4, we fixed the lag distances to equal intervals of two distance units. Based on the infogram cloud (Fig. 3.5a), the Δz PMFs for all classes were obtained. Subsequently, the range was identified as the point beyond which the class entropy exceeded the entropy of the full dataset (seen as the intersect of the blue and red-dotted lines in Fig. 3.5c). This occurs at class 23, corresponding to a Euclidean distance of 44 grid units. In Fig. 3.5c, it is also possible to notice a steep reduction in entropy (red curve) for furthest classes due to the reduced number of pairs composing the Δz PMFs. A similar behavior is also typically found in experimental variograms (not shown).

The number of pairs forming each Δz PMF and the optimum weights obtained for Eqs. 3.4 and 3.5 are presented in Fig. 3.6. Fig. 3.6a shows the number of pairs which

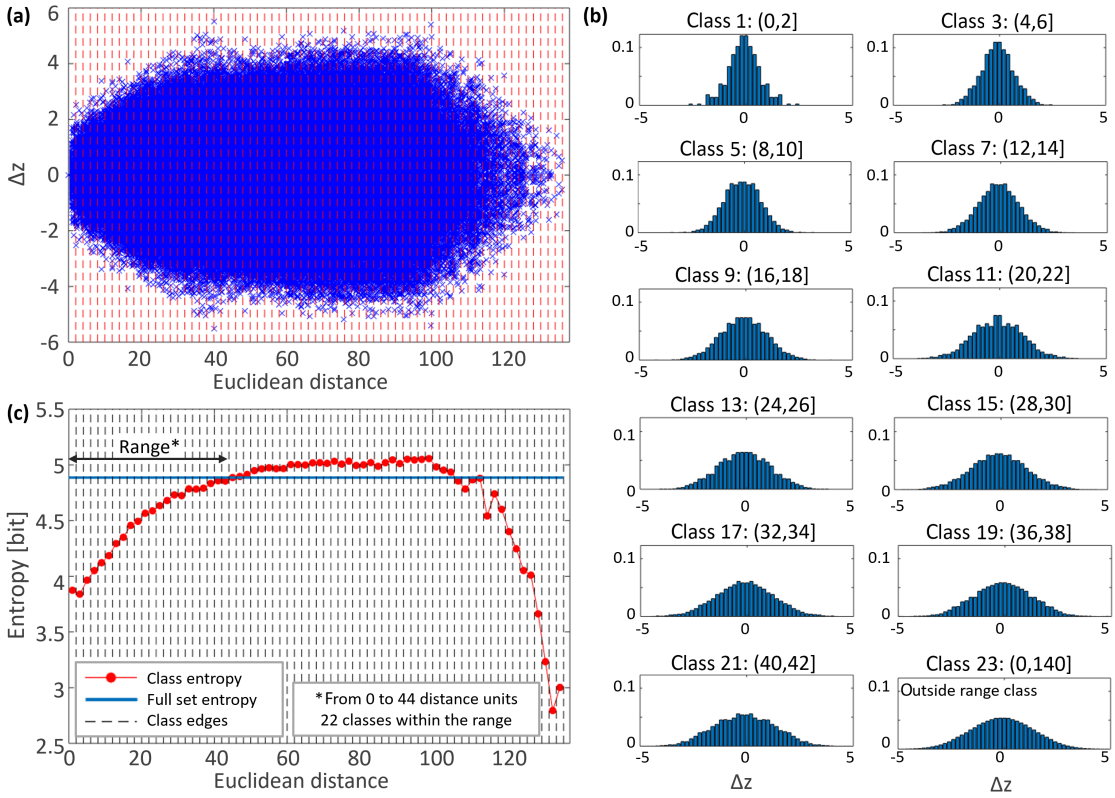

Figure 3.5: Spatial characterization of LR1-600 showing the **(a)** infogram cloud, **(b)** Δz PMFs by class, and **(c)** infogram.

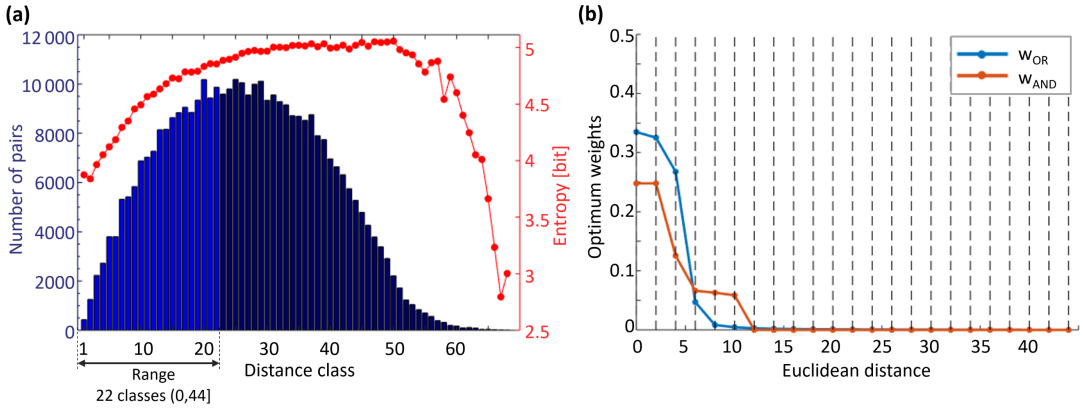

Figure 3.6: LR1-600, with **(a)** class cardinality and **(b)** optimum weights – Eqs. 3.4 and 3.5.

compose the Δz PMF by class, where the first class has just under 500 pairs and the last class inside the range (light blue) has almost 10 000 pairs. About 40% of the pairs (142 512 out of 359 400 pairs) are inside the range. We obtained the weight of each class by convex optimization, as described in Sect. 3.2.3.2. The dots in Fig. 3.6b represent the optimized weights of each class. As expected, the weights reflect the decreasing spatial dependence of variable z with distance. Regardless of the aggregation method, LR1-600 models are highly influenced by neighbors up to a distance of 10 grid units (distance class 5). To estimate the z PMFs of target points, the following three different methods were tested:

(i) Model 1: AND/OR combination, proposed by Eq. 3.6, where LR1-600 weights resulted in $\alpha = 1$ and $\beta = 0.95$;

(ii) Model 2: pure AND combination, given by Eq. 3.5;

(iii) Model 3: pure OR combination, given by Eq. 3.4.

The model results are summarized in Table 3.1 and illustrated in Fig. 3.7, where the first column of the panel refers to the AND/OR combination, the second column to the pure AND combination, and the third column to the pure OR combination. To assist in visually checking the heterogeneity of z, the calibration set representation is scaled by its z value, with the size of the cross increasing with z. For the target identification, we used its grid coordinates (x,y).

Fig. 3.7a shows the E-type estimate[1] of z (expected z obtained from the predicted z PMF) for the three analyzed models. Neither qualitatively (Fig. 3.7a) nor quantitatively (Table 3.1) is it possible to distinguish the three models based on their E-type estimate or its summary statistics. Deterministic performance metrics (E_{MA} and E_{NS}; Table 3.1) are also similar among the three models. However, in probabilistic terms, the representation given by the entropy map (Fig. 3.7b; which shows the Shannon entropy of the predicted z PMFs), the statistics of predicted z PMFs, and the D_{KL} performance (Table 3.1) reveal differences.

By its 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, for example, the case of targets A (10,42) and B (25,63). Target B (25,63) is located in between two sampled points in a heterogeneous region (small and large z values, both in the first distance class) and presents distributions with a bimodal shape and higher uncertainty (Fig. 3.7c), especially for model 3 (4.68 bits). For the more assertive models (1 and 2), the distributions of target B (25,63) have lower uncertainty (3.42 and 3.52 bits, respectively). They shows some peaks, due to small bumps in the PMF neighbors (not shown), which are boosted by the w_{AND_k} exponents in Eq. 3.5. In contrast, target A (10,42), which is located in a more homogeneous region, with the closest neighbors in the second distance class, shows a sharper z PMF in comparison to target B (25,63) for models 1 and 3 and a Gaussian-like shape for all models.

1 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 of neighboring values.

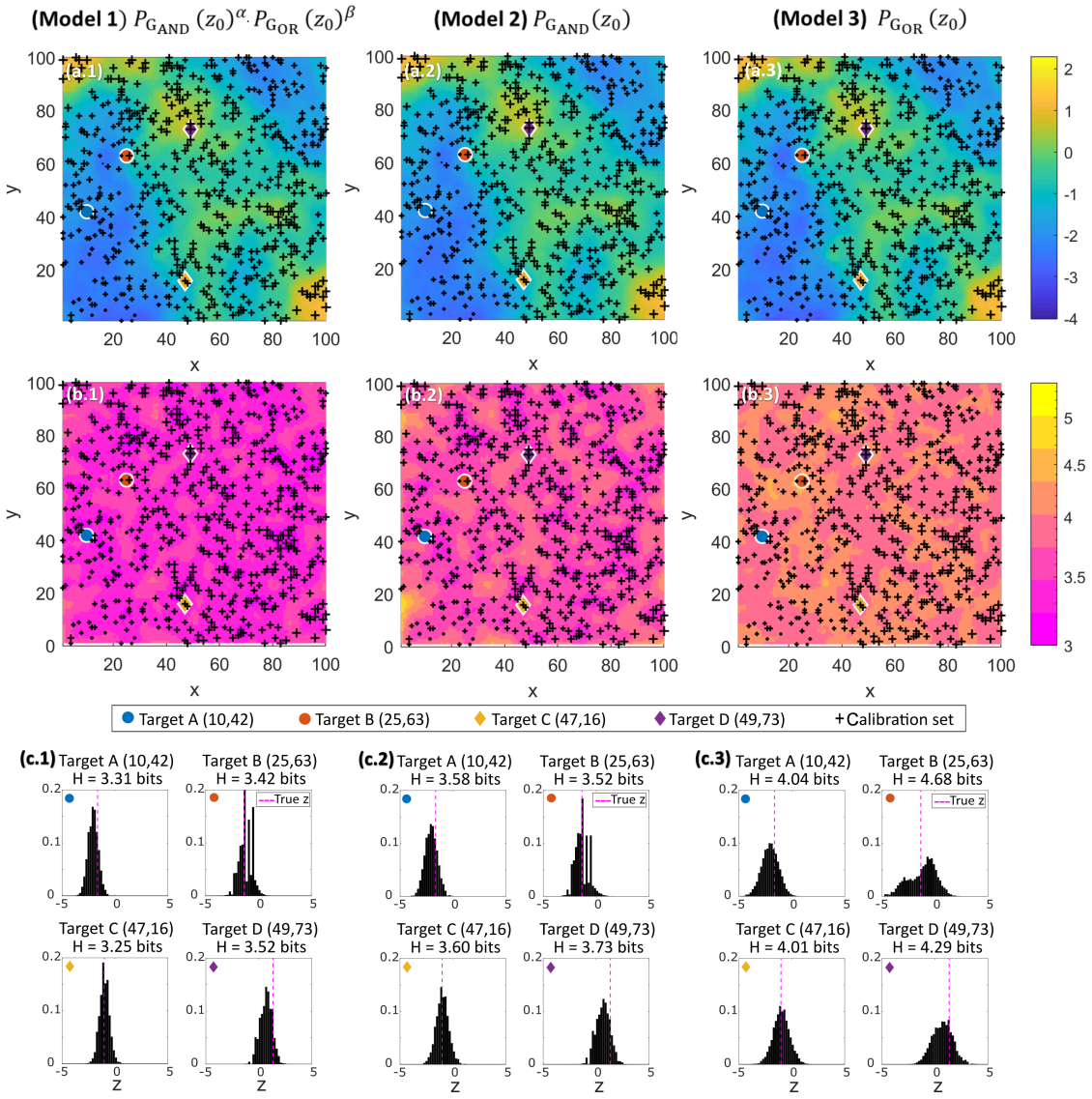

Figure 3.7: LR1-600 results showing the **(a)** E-type estimate of z, **(b)** entropy map (bit), and **(c)** z PMF prediction for selected points. The first, second, and third columns of the panel refer to the results of model 1 (AND/OR), model 2 (AND), and model 3 (OR), respectively.

Table 3.1: Summary statistics and model performance of LR1-600.

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

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

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 Δz (not shown) in comparison to target D (49,73) due to the homogeneity of z values in the region.

Although the z PMFs (Fig. 3.7c) from models 1 and 2 present comparable shapes, the uncertainty structure (color and shape displayed in Fig. 3.7b) of the overall field differs. Since model 1 is derived from the aggregation of models 2 and 3, as presented in Eq. 3.6, 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 3.1). On the other hand, due to the averaging of PMFs, model 3 is the more conservative model, verified by the highest overall uncertainty (median entropy of 4.17 bits). Model 3 also predicts a smaller minimum and higher maximum of the E-type estimate; in addition, for the selected targets, it provides the widest confidence interval.

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

3.4.2 *Comparison analysis*

In this section, the test set was used to calculate the performance of all methods (NN, IDS, OK, and HER) as a function of sample size and dataset type (SR0, SR1, LR0, and LR1). HER was applied using the AND/OR model proposed by Eq. 3.6. See Appendix B.2 for the calibrated parameters of all models discussed in this section.

Fig. 3.8 summarizes the values of mean absolute error (E_{MA}), Nash–Sutcliffe efficiency (E_{NS}), and mean Kullback–Leibler divergence (D_{KL}) for all interpolation methods, sampling sizes, and dataset types. The SR fields are located in the left column and the LR in the right. Datasets without noise are represented by continuous lines, and datasets with noise are represented by dashed lines. E_{MA} is presented in Fig. 3.8a,b for the SR and LR fields, respectively. All models have the same order of magnitude of E_{MA} for the noisy datasets (SR1 and LR1; dashed lines), with the performance of the NN model being the poorest, and OK being slightly better than IDS and HER. For the datasets without noise (SR0 and LR0; continuous lines), OK performed better than the other models, with a decreasing difference given sample size. In terms of E_{NS}, all models have comparable results for LR (Fig. 3.8d), except NN in the LR1 field. A larger contrast in the model performances can be seen for the SR field (Fig. 3.8c), where, for SR1, NN performed the worst and OK the best. For SR0, especially for small sample sizes, OK performed better and NN poorly, while IDS and HER had similar results, with a slightly better performance for HER.

The probabilistic models of OK and HER were comparable in terms of D_{KL}, with OK being slightly better than HER, especially for small sample sizes (Fig. 3.8e,f). An exception is made for OK in LR0. Since the D_{KL} scoring rule penalizes extremely

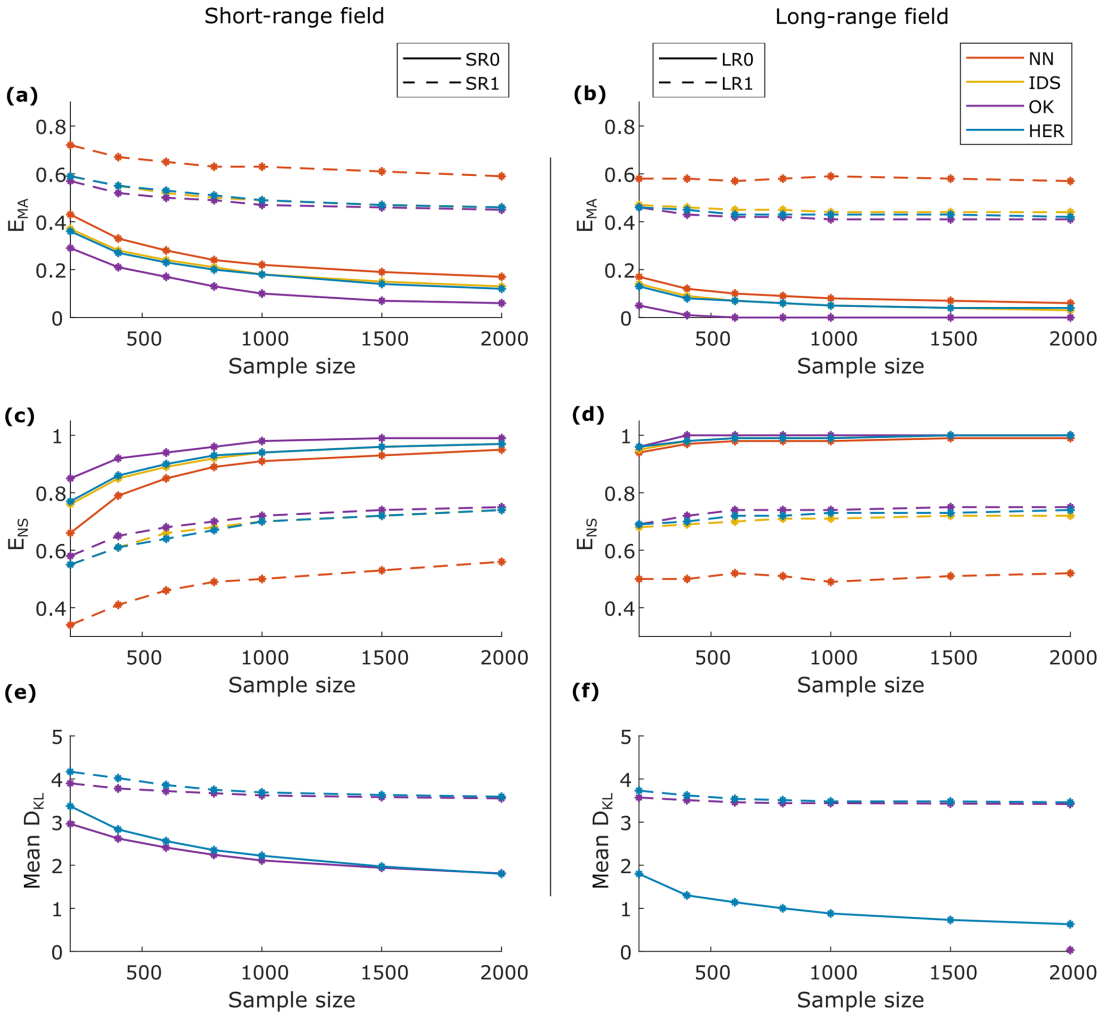

Figure 3.8: Performance comparison of NN, IDS, OK, and HER. **(a,b)** Mean absolute error, **(c,d)** Nash–Sutcliffe efficiency, and **(e,f)** Kullback–Leibler divergence scoring rule for the SR datasets in the left panels (a, c, and e) and the LR datasets in the right panels (b, d, and f). Continuous line refers to datasets without noise and dashed lines to datasets with noise.

confident but erroneous predictions, D_{KL} of OK tended to infinity for LR0 and, therefore, it is not shown in Fig. 3.8f.

For all models, the performance metrics for LR showed better results when compared to SR (compare the left and right columns in Fig. 3.8). The performance improvement given the sample size is similar for all models, which can be seen by the similar slopes of the curves. In general, we noticed a prominent improvement in the performance in SR fields up to a sample size of 1000 observations. On the other hand, in LR fields, the learning process already stabilizes at around 400 observations. In addition to the model performance presented in this section, the summary statistics of the predictions and the correlation of the true value and the residue of predictions can be found in Appendix B.3.

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

3.4.3 *Discussion*

3.4.3.1 *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. 3.4.1). Independent of the choice of the aggregation method, the deterministic results (E-type estimate of z) of all the models were remarkably similar. In contrast, we could see different uncertainty structures of the estimates for all three cases analyzed, ranging from a more confident method to a more conservative one. The uncertainty structures also reflected the expected behavior of larger errors in locations surrounded by data that are very different in value, as mentioned in Goovaerts (1997, p. 180, 261). In this sense, HER has proved effective in considering both the spatial configuration of data and the data values regardless of which aggregation method is selected.

As previously introduced in Sect. 3.2.3.1, the choice of pooling method can happen beforehand in order to introduce physical knowledge to the system, or several can be tested to learn about the response of the field to the selected model. Aside from their different mathematical properties, the motivation behind the selection of the two aggregation methods (linear and log-linear) was the incorporation of continuous or discontinuous field properties. The interpretation is supported by Goovaerts (1997, p. 420), Journel (2002) and Krishnan (2008), where the former connects a logarithmic expression (AND) to continuous variables, while the latter two associate linear pooling (OR) to abrupt changes in the field and categorical variables.

As verified in Sect. 3.4.1, the OR (=averaging) combination of distributions to estimate target PMFs was the most conservative (with the largest uncertainty) method among all those tested. For this method of PMF merging, all distributions are considered feasible, and each point adds new possibilities to the result, whereas the AND combination of PMFs was a bolder approach, intersecting distributions to extract their agreements. Here, we are narrowing down the range of possible values so that the final distribution satisfies all observations at the same time. Complementarily, considering the lack of information to accurately describe the interactions between

the sources of information, we proposed inferring α and β weights (the proportion of AND and OR contributions, respectively) using Eq. 3.6. It resulted in a reasonable trade-off between the pure AND and the pure OR model and was hence used for benchmarking HER against traditional interpolation models in Sect. 3.4.2.

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

3.4.3.2 *Benchmarking and applicability*

Although the primary objective of this study is to investigate the characteristics of HER, Sect. 3.4.2 compares it to three established interpolation methods. In general, HER performed comparably to OK, which was the best-performing method among the analyzed ones. The probabilistic performance comparison was only possible between HER and OK where both methods also produced comparable results. Note that the datasets were generated using Gaussian process (GP) so that they perfectly fulfilled all recommended requisites of OK (field mean independent of location; normally distributed data), thus favoring its performance. Additionally, OK was also favored when converting their predicted PDFs to PMFs, since the defined bin width was often orders of magnitude larger than the standard deviation estimated by OK. However, the procedure was a necessary step for the comparison, since HER does not fit continuous functions for their predicted PMFs.

Although environmental processes hardly fulfill Gaussian assumptions (Hristopulos and Baxevani, 2020; Kazianka and Pilz, 2010), GP allows the generation of a controlled dataset in which we could examine the method performances in fields with different characteristics. 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 nonparametric, HER handles different data properties without the need to transform the available data to fulfill model assumptions. And since HER uses binned transformations 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.

3.4.3.3 *Model generality*

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

HER does not require the fitting of a theoretical function; its spatial dependence structure (Δz PMFs; infogram) is derived directly from the available data, while, according to Putter and Young (2001), OK predictions are only optimal if the weights are calculated from the correct underlying covariance structure, which, in practice, is not the case since the covariance is unknown and estimated from the data. Thus, the choice of the theoretical variogram for OK can strongly influence the predicted z, depending on the data. In this sense, for E-type estimates, HER is more robust against user decisions than OK. Moreover, HER is flexible in the way that it aggregates the probability distributions, not being a linear estimator like OK. In terms of the number of observations, and being a nonparametric 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 with respect to computational effort. Nevertheless, relying on fitted functions can mask the lack of observations since it still produces attractive, but not necessarily reliable, maps (Oliver and Webster, 2014).

OK and HER have different levels of generality. OK weights depend on how the fitted variogram varies in space (Kitanidis, 1997, p. 78), whereas 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 only the spatial geometry from the data (Goovaerts, 1997, p. 180), while HER 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 their spatial correlation structure, a characteristic which was first possible with the advent of indicator kriging (Journel, 1983). Conversely, when no nugget effect is expected, HER can lead to undesired uncertainty when predicting the value at or near sampled locations. This can be overcome by defining a small distance class for the first class, changing the binning to obtain a point–mass distribution as a 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.

3.4.3.4 *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 one), which, on the one hand, is a useful characteristic for reducing redundancy and predicting values outside the range of the data (Goovaerts, 1997, p. 176) but, on the other hand, can lead to unacceptable results, such as negative metal concentrations (Goovaerts, 1997, p. 174–177) and negative kriging variances (Manchuk and Deutsch, 2007). HER weights are limited to the range of $[0, 1]$. Since the used dataset was evenly spaced, a possible issue of redundant information in the case of clustered samples was not considered in this paper. The influence of data clusters could be reduced by splitting the search neighborhood into equal-angle sectors and retaining within each sector a specified number of nearest data (Goovaerts, 1997, p. 178) or discarding measurements that contain 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, 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. As with 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 (Appendix B.3; Fig. B.3). However, HER revealed a reduced smoothing (residue correlation closer to zero) compared to OK for SR0, SR1, and LR1. In particular, for points beyond the range, both methods predict by averaging the available observations. While OK calculates the same weight for all observations beyond the range and proceeds with their linear combination, HER associates Δz PMF of the full dataset to all observations beyond the range and aggregates them using the same weight (last-class weight).

3.5 SUMMARY AND CONCLUSION

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

Throughout the paper, three aggregation methods (OR, AND, and AND/OR) were analyzed in terms of uncertainty and resulted in predictions ranging from conservative to more confident ones. HER's performance was also compared to

popular interpolators (nearest neighbor, inverse distance weighting, and ordinary kriging). All methods were tested under the same conditions. HER and ordinary kriging (OK) were the most accurate methods for different sample sizes and field types. HER has featured the following properties: (i) it is nonparametric in the sense that predictions are directly based on empirical distribution, thus bypassing function fitting and, therefore, avoiding the risk of adding information not available in the data; (ii) it allows one to incorporate different uncertainty properties according to the dataset and user interest by selecting the aggregation method; (iii) it enables the calculation of confidence intervals and probability distributions; (iv) it is nonlinear, and the predicted conditional distribution depends on both the spatial configuration of the data and the field values; (v) it has the flexibility of adjusting the number of parameters to be optimized according to the amount of data available; (vi) it is adaptable for handling binary or even categorical data, since HER uses binned transformations of the data; and (vii) it can be extended to conditional stochastic simulations by directly performing sequential simulations on the predicted conditional distribution.

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

Part IV

ASSESSING LOCAL AND SPATIAL UNCERTAINTY WITH NONPARAMETRIC GEOSTATISTICS

This study is published in the scientific journal Stochastic Environmental Research and Risk Assessment (SERRA) and is a reprint of:

*Thiesen, Stephanie; Ehret, Uwe (2021): Assessing local and spatial uncertainty with nonparametric geostatistics. Stochastic Environmental Research and Risk Assessment. doi:*10.1007/s00477-021-02038-5

4

ASSESSING LOCAL AND SPATIAL UNCERTAINTY WITH NONPARAMETRIC GEOSTATISTICS

ABSTRACT

Uncertainty quantification is an important topic for many environmental studies, such as identifying zones where potentially toxic materials exist in the soil. In this work, the nonparametric geostatistical framework of histogram via entropy reduction (HER) is adapted to address local and spatial uncertainty in the context of risk of soil contamination. HER works with empirical probability distributions, coupling information theory and probability aggregation methods to estimate conditional distributions, which gives it the flexibility to be tailored for different data and application purposes. To explore how HER can be used for estimating threshold-exceeding probabilities, it is applied to map the risk of soil contamination by lead in the well-known dataset of the region of Swiss Jura. Its results are compared to indicator kriging (IK) and to an ordinary kriging (OK) model available in the literature. For the analyzed dataset, IK and HER predictions achieve the best performance and exhibit comparable accuracy and precision. Compared to IK, advantages of HER for uncertainty estimation in a fine resolution are that it does not require modeling of multiple indicator variograms, correcting order-relation violations, or defining interpolation/extrapolation of distributions. Finally, to avoid the well-known smoothing effect when using point estimations (as is the case with both kriging and HER), and to provide maps that reflect the spatial fluctuation of the observed reality, we demonstrate how HER can be used in combination with sequential simulation to assess spatial uncertainty (uncertainty jointly over several locations).

4.1 INTRODUCTION

Modeling the uncertainty about the unknown is of crucial importance for evaluating the risk involved in any decision-making process. The traditional approach of modeling the uncertainty with respect to geostatistical interpolation consists of computing a kriging estimate and its attached error variance, and explicitly assuming a Gaussian distribution for assessing the confidence interval (Goovaerts, 1997, p. 261; Kitanidis, 1997, p. 68; Bourennane et al., 2007). The major restrictions of this approach are (i) that the distribution of the estimation error is assumed to be normal, and (ii) that the variance of the errors is assumed to be independent of the data values, and only dependent on the data configuration (Kitanidis, 1997, p. 68; Goovaerts, 1997, p. 261). These Gaussian and homoscedastic assumptions are unfortunately rarely fulfilled for environmental attributes and soil variables. Instead, they often display skewed distributions (Bourennane et al., 2007; Goovaerts, 1997, p. 261).

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

Recently, for avoiding the risk of adding information not present in data, Thiesen et al. (2020b) proposed combining information theory with probability aggregation methods in a geostatistical framework as a novel nonparametric method for stochastic estimation at unsampled locations. Histogram via entropy reduction (HER) was primarily proposed to bypass fitting spatial correlation functions and assumptions about the underlying distribution of the data. In addition, it is a proper framework for uncertainty estimation since it accounts for both spatial configuration and data values and offers higher generality than ordinary kriging (OK). HER uses binned transformation of the data and optimization of the information content, which gives some flexibility to adapt the method to handle different kinds of data and problems. Furthermore, it allows incorporating different uncertainty properties by selecting the aggregation method. For the present paper, these primary findings paved the way for

the further development of the spatial interpolation framework of HER to assess both (i) the local uncertainty when dealing with categorical data and threshold-exceeding probabilities, and (ii) the spatial uncertainty by reproducing the spatial fluctuation of the dataset with sequential simulation.

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

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

The paper is organized as follows. HER method and its adaptations are presented in Sect. 4.2. In Sect. 4.3, we describe the dataset, performance criteria, and benchmark models; apply OK, IK, and HER to a real dataset; and compare their estimation and local uncertainty results. Finally, a proof of concept of HERs is presented. In Sect. 4.4 we discuss results, and in the closing Sect. 4.5, we summarize the key findings and draw conclusions.

4.2 METHOD DESCRIPTION

In the following sections, we give a brief presentation of information theoretic measures employed in the HER method (Sect. 4.2.1) and introduce its three main steps (Sect. 4.2.2). Specifically in Sect. 4.2.2.3, we propose an adaptation of the minimization problem tailored to estimating local threshold-exceeding probabilities. Finally, we expand HER for spatial uncertainty analysis in Sect. 4.2.3.

4.2.1 *Information theoretic measures employed in HER*

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

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

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x). \tag{4.1}$$

The logarithm to base two denotes entropy in unit of bits, which is associated to the number of binary questions needed to reconstruct a random variable. This means that, e.g., the entropy of a fair coin toss is 1 bit or, in other words, the answer of one yes-no question (e.g., is it tails?) is enough to identify the toss output. Therefore, the above expression measures the average uncertainty of a probability distribution. HER uses Shannon entropy to evaluate the spatial dependence of the dataset and its correlation length.

Kullback-Leibler divergence (or relative entropy) compares similarities between two probability distributions p and q:

$$D_{\mathrm{KL}}(p||q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)}. \tag{4.2}$$

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

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

4.2.2 *HER for local uncertainty*

The brief introduction to HER presented in the following is based on Thiesen et al. (2020b), further details can be found there. HER is a distribution-free interpolator enclosed in a geostatistical framework. It was formulated to describe spatial patterns and solve spatial interpolation problems. In HER, we incorporate concepts from information theory and probability aggregation methods for globally minimizing uncertainty and predicting conditional probability distributions (CPD) directly based on empirical

discrete distributions (also referred to as probability mass functions, PMFs). HER comprises three main steps: (i) characterization of spatial dependence, (ii) selection of an aggregation method and associated optimal weights, and (iii) prediction of the target CPD. These steps are explained in the following sections.

4.2.2.1 *Characterization of spatial dependence*

Let us consider the situation illustrated in Fig. 4.1c, where z is the attribute under study and we are interested in inferring the z PMF of the target 0 ($p(z_0)$ is the estimated probability mass function of z at the unsampled location u_0) given its neighbors 1, 2, and 3 (z_1, z_2, and z_3 are available observations sampled at locations u_1, u_2, and u_3). In order to characterize the spatial dependence, we extract the distribution associated to each neighbor and the correlation length (range) in the following actions. First, for each lag distance interval k – also called distance class or simply class – with bounds d_{k-1} and d_k, we calculate the difference of the z-values between all pairs of observations within the interval ($\Delta Z_k = \{z_i - z_j \mid i \neq j, \quad d_{k-1} < |u_i - u_j| \leq d_k\}$) and generate the corresponding Δz PMF ($p_{\Delta Z_k}(\Delta z)$, Fig. 4.1a)[1]. The entropy values of each Δz PMF (one for each distance class k) is visualized as a 2D plot called infogram ($H(\Delta Z_k)$, Fig. 4.1b). The infogram describes the statistical dispersion of pairs of observations for the distance separating these observations (Thiesen et al., 2020b). Quantitatively, it is a way of measuring the uncertainty about Δz given the separation distance of the data, meaning that observations start becoming less informative as the distance increases. Note that in the same figure, the range can be identified as the distance where the entropy of the classes exceeds the full dataset entropy $H(\Delta Z)$, calculated over the difference of z-values between all pairs of observations in the dataset ($\Delta Z = \{z_i - z_j \mid i \neq j\}$). This range definition is based on the principle that the observations beyond this distance start becoming uninformative, and it is pointless to use information outside of this neighborhood[2]. Finally, we associate to each neighbor the Δz PMF of the corresponding class k, according to its absolute lag distance from the target, then shift this distribution by its z-value $p(z_0|z_i) = p_{\Delta Z_k}(z_0 - z_i)$, as outlined in Fig. 4.1c. In the end of this first step, we have inferred the conditional PMFs $p(z_0|z_1)$, $p(z_0|z_2)$, and $p(z_0|z_3)$. A practical example using HER is shown in Fig. C.1 with more details.

4.2.2.2 *Probability aggregation*

For the second step of the method, the individual conditional distributions obtained in the previous step are combined by using probability aggregation methods. The aggregation method is based on work by Allard et al. (2012), which we recommend as a summary of existing aggregation methods. The probability aggregation yields a

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

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

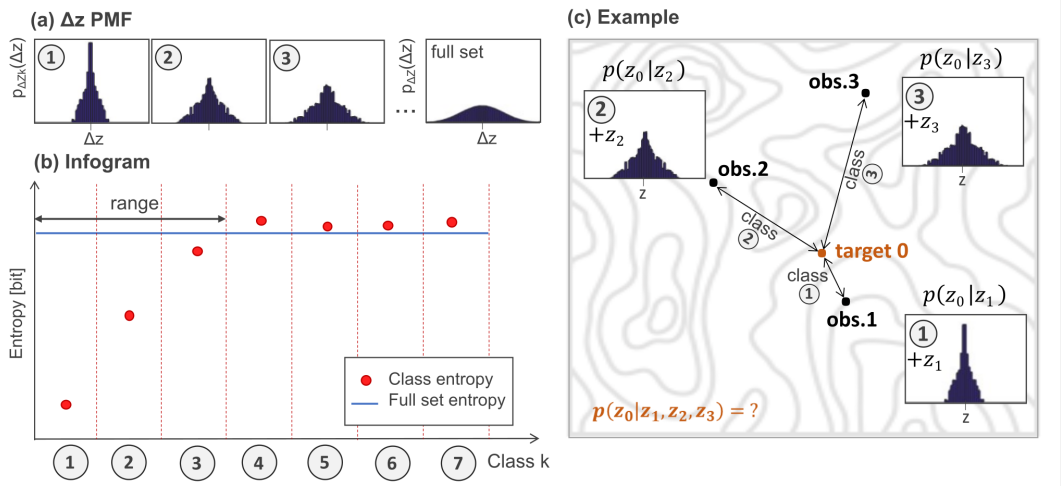

Figure 4.1: Schematic of the HER method. **(a)** Δz PMFs $p_{\Delta Z_k}(\Delta z)$ of the difference in the z-values (Δz) between all pairs of observations within distance class k and Δz PMF $p_{\Delta Z}(\Delta z)$ of the full dataset; **(b)** infogram, obtained by calculating the entropy $H(\Delta Z_k)$ of PMFs in (a) and plotting them against their respective distance class, with the range determined by the entropy of the full dataset $H(\Delta Z)$; and **(c)** practical example where the target value to be estimated is z_0 and the available observations are z_1, z_2, and z_3.

single, global distribution for the target 0, so that the joint probability $p\left(z_0|z_1,...,z_n\right) \approx P_G\left(p\left(z_0|z_1\right),\ldots,p\left(z_0|z_n\right)\right)$, with z_0 being the estimation of the target value (at an unsampled location) and z_i values at neighboring locations, where $i = 1,...,n$ are the indices of the sampled observations and z is the variable under study. For brevity, from now on we use $P_i(z_0)$ to denote $p\left(z_0|\,z_i\right)$ and $P_G\left(z_0\right)$ for the global probability $P_G\left(P_1\left(z_0\right),\ldots,P_n(z_0)\right)$.

Two basic aggregation methods were discussed by Thiesen et al. (2020b), namely linear pooling and log-linear pooling. Linear pooling (Eq. 4.3) is a way of averaging distributions. It is related to the union of events and associated with the logical operator OR. Multiplication of probabilities, or log-linear pooling in Eq. 4.4, in turn, is associated with the logical operator AND, and related to the intersection of events. Due to their distinct characteristics, Thiesen et al. (2020b) associated the linear aggregation to discontinuous field properties, and the log-linear to continuous ones. The authors exemplified that, if we have two points A and B with different z-values (z_A, z_B) and want to estimate the z-value of a the target point X located between both in a continuous field, we would expect that z_X would be somewhere between the z-values of A and B, which can be achieved by an AND combination. On the other hand, in the case of categorical data (or abrupt changes; Goovaerts, 1997, p. 420), considering A and B belonging to different categories, a target X located between both will either belong to the category of A or B, which can be achieved by an OR combination.

The third pooling operator (Eq. 4.5), which combines $P_{G_{\text{AND}}}$ and $P_{G_{\text{OR}}}$, was proposed and explored in Thiesen et al. (2020b). It optimally expresses continuous and discontinuous properties of a field (controlled by parameters α and β, respectively) by

minimizing the relative entropy (D_{KL}) of the estimation and the true data. Since the final distribution of this pooling contains a pure OR, Eq. 4.3, and pure AND, Eq. 4.4, aggregation as special cases, it was recommended by the authors for cases where the field properties are not known a priori.

$$P_{G_{\mathrm{OR}}}(z_0) = \sum_{i=1}^{n} w_{\mathrm{OR}_i} \, P_i(z_0), \tag{4.3}$$

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

$$\ln P_{G_{\mathrm{AND}}}(z_0) = \ln \zeta + \sum_{i=1}^{n} w_{\mathrm{AND}_i} \ln P_i(z_0), \tag{4.4}$$

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

$$P_G(z_0) \propto P_{G_{\mathrm{AND}}}(z_0)^{\alpha} \, P_{G_{\mathrm{OR}}}(z_0)^{\beta}, \tag{4.5}$$

where α and β are positive weights varying from 0 to 1.

4.2.2.3 *Entropy minimization*

After selecting the appropriate aggregation method, we address the optimization problem for estimating the weights of the pooling operators. In Thiesen et al. (2020b), the authors were interested in comparing HER results with OK estimates. Therefore, by means of leave-one-out cross-validation, they chose a global set of weights such that the disagreement of the "true" observation (left-out measurement) and the estimated probability of the bin containing the true observation was minimized. For doing so, the optimization problem was tailored to find the set of weights (one for each distance class) which minimizes the expected relative entropy (D_{KL}) of all targets. Note that when dealing with single-value observations (or categorical data), this is equivalent to subtracting the probability of the bin containing the true value from one. The D_{KL} evaluation of a single prediction is outlined in Fig. 4.2a. In the present study, we propose an adaptation of this loss function (Fig. 4.2a) to

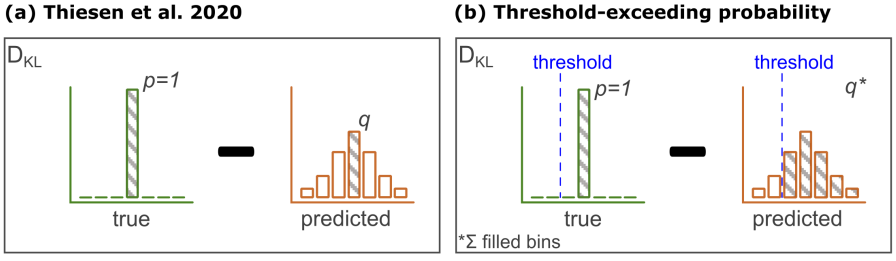

Figure 4.2: Optimization problem. **(a)** Maximizing the probability of the "true" observation (Thiesen et al., 2020b) and **(b)** maximizing the estimation of threshold-exceeding probability.

focus on the estimation of threshold-exceeding probabilities (Fig. 4.2b). Here, instead

of optimizing the probability of a single bin containing the true observation, we minimize the probability disagreement (relative entropy, D_{KL}) of the binarized left-out measurement (above or below z_c threshold) and the cumulative probability of the estimated distribution (also binary, above or below z_c threshold). With this adaptation, the optimization problem focusses on selecting weights which maximize the probability of the target matching the true classification. The authors' goals were to reduce the risk that an unsampled site is declared "safe" when in reality the soil is "toxic" and vice versa, and to open the possibility of working with categorical data. The method adaptation proposed in Fig. 4.2b will be used throughout the paper and will simply be referred to as HER.

For both optimization problems (Fig. 4.2a,b), one optimum weight is obtained for each distance class k and used in Eqs. 4.3 and 4.4, referred to as w_{OR_k} and w_{AND_k}, respectively (here generalized as w_k). After that, α and β from Eq. 4.5 are optimized by grid search, with candidate values ranging from 0 to 1 (steps of 0.05 were used in the application case).

Particularly for the present study, another adaptation was done to avoid undesired non-zero uncertainty when predicting z-values at sampled locations: within the first distance class, we asymptotically increase the weight towards infinity as the distance approaches zero, by scaling with the inverse of the distance. For all other distance classes, similarly to Thiesen et al. (2020b), we linearly interpolate the weights according to the Euclidean distance and the weight of the next class. A practical example of the proposed interpolation is illustrated in Fig. C.2.

4.2.2.4 *PMF prediction*

As seen before, to estimate the z-value of the target 0 (i.e., the unknown observation z_0), first we classify its neighbors z_i (sampled observations) according to their distance to the target. Each neighbor is then associated to its corresponding Δz PMF and shifted by its z_i value. Finally, by applying the selected aggregation method and its optimum weights, we combine the individual z PMFs of the neighbors to obtain the z distribution of the target conditioned on all neighbors (z PMF). By construction, the assessed PMF is nonparametric since no prior assumption is made regarding the shape of the distribution of possible values.

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

4.2.3 *HER for spatial uncertainty*

So far, we proposed modeling distributions to obtain estimates of values and related uncertainties at specific locations (local uncertainty) using the HER method. However, these single-point PMFs do not allow to simultaneously assess the uncertainty about

attribute values at several locations (Goovaerts, 1997, p. 262). Simply multiplying CPDs of several locations to obtain their joint probability would assume independence between the data, a case of little interest (Goovaerts, 1997, p. 372). Therefore, we address multiple-point – or spatial – uncertainty by combining HER with sequential simulation (HERs). Stochastic simulation was introduced in the early 1970's to correct for the smoothing effect of kriging and to provide maps that reflect the spatial fluctuation of the observed reality (Deutsch and Journel, 1998, p. 18; Journel, 1974). Geostatistical simulation generates a model of uncertainty that is represented by multiple sets of possible values distributed in space, one set of possible outcomes is referred to as a realization (Leuangthong et al., 2004). Different yet equiprobable realizations, all conditioned on the same dataset and reflecting the same dispersion characteristics, can be produced to be used for numerical and visual appreciation of spatial uncertainty (Deutsch and Journel, 1998, p. 19; Journel, 2003; Journel and Huijbregts, 1978). Such equiprobable realizations are known as stochastic images and share the same sample statistics and conditioning data (Gómez-Hernández and Cassiraga, 1994).

Sequential simulations with HER are generated by first establishing a random path along all nodes in the grid network. Then, for each node, and in the order of the random path we (i) derive the PMF of the node using HER as explained in Sect. 4.2.2, (ii) randomly draw a single value from this PMF, and (iii) assign the value to the grid as an additional observation. With this procedure, we sequentially include the simulated values to the original dataset and used them to condition predictions at the remaining locations. The simulated value (step ii) is derived from a Monte Carlo simulation (Metropolis and Ulam, 1949), where we randomly draw a p-value uniformly distributed between 0 and 1 and obtain the z value from the estimated PMF. Equiprobability is ensured by triggering each realization by one random seed drawn from a uniform distribution (Deutsch and Journel, 1998, p. 19; Goovaerts, 1999).

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

4.3 APPLICATION TO REAL DATA

4.3.1 *Jura dataset*

We evaluate HER (Sect. 4.2.2) and HERs (Sect. 4.2.3) by applying them to the well-known Jura dataset, which is often used as benchmarking in the geostatistical literature, e.g., Allard et al. (2011), Atteia et al. (1994), Bandarian et al. (2018), Bel et al.

(2009), Dabo-Niang et al. (2016), Goovaerts (1997), Goovaerts et al. (1997), Loquin and Dubois (2010), and Webster et al. (1994). The data were collected by the Swiss Federal Institute of Technology at Lausanne from a 14.5 km² area in the Swiss Jura region. A comprehensive description of the sampling, field, and laboratory procedures is available in Atteia et al. (1994) and Webster et al. (1994), and a detailed exploratory data analysis can be found in Goovaerts (1997).

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

Fig. 4.3 illustrates the Pb concentrations at the locations of the calibration set, the locations of the validation set, and histogram and cumulative distribution of the calibration set. Table 4.1 presents the summary statistics of Pb for all datasets. The Swiss federal ordinance defined the regulatory threshold used as the tolerable maximum for healthy soil (Fix and Hodges Jr., 1987): locations with lead concentrations above the critical threshold (z_c) of 50 $mg\ kg^{-1}$ (or $z_c = 1.699$ in its logarithm transform) are considered contaminated. For the available dataset, this limit is exceeded at 42.1% of the calibration set locations, see Fig. 4.3c. The dotted line in Fig. 4.3a indicates the transect SW-NE to be discussed in Sect. 4.3.4.1, which was based on the cross section shown in Goovaerts (1997).

Table 4.1: Summary statistics of $\log_{10}$(Pb) datasets.

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

[a] Evenly spaced bins, with intervals of 0.015 (more in Sect. 4.3.3).
Regulatory threshold: $z_c = 1.699$.

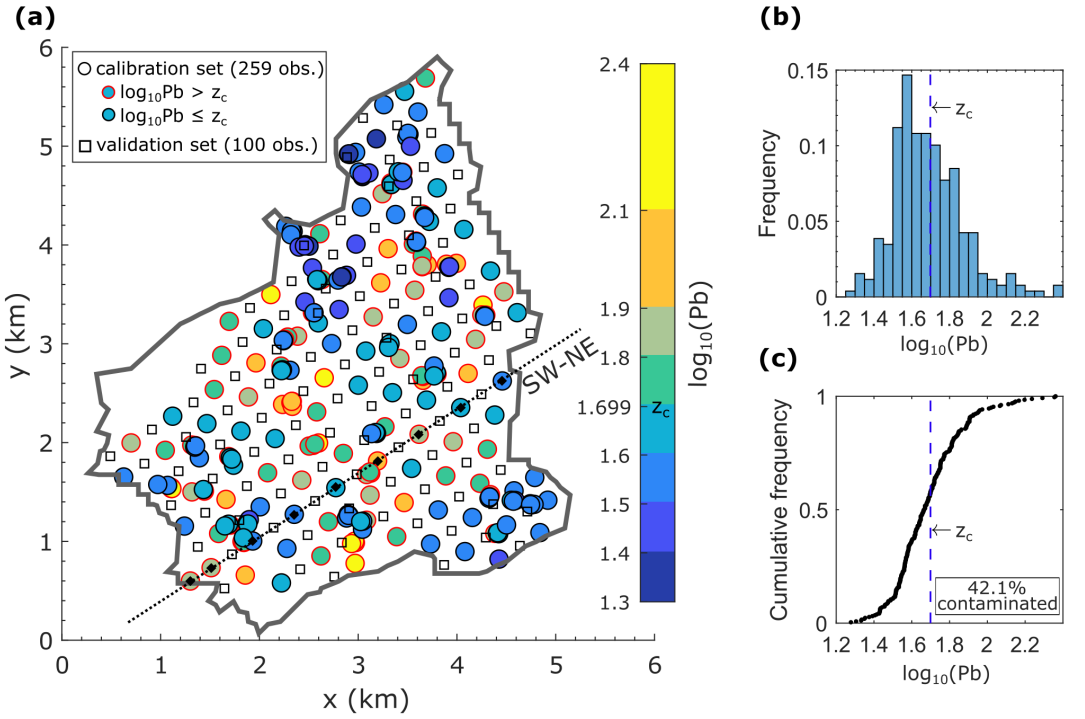

Figure 4.3: Calibration set. **(a)** Concentration values, **(b)** histogram, and **(c)** cumulative distribution.

4.3.2 *Performance criteria*

The quantitative evaluation of the predictive power of the models was carried out with two criteria for the deterministic results, namely, mean absolute error (E_{MA}) and Nash-Sutcliffe efficiency (E_{NS}), and another two for the probabilistic outcomes, i.e., Kullback-Leibler divergence (D_{KL}) and goodness statistic (G). These metrics are presented in Eqs. 4.6, 4.7, 4.2, and 4.9, respectively.

The deterministic performance metrics are defined as:

$$E_{\mathrm{MA}} = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{z}_{\mathrm{i}} - z_{\mathrm{i}}\right|, \tag{4.6}$$

$$E_{\mathrm{NS}} = 1 - \frac{\sum_{i=1}^{n}\left(\widehat{z}_{\mathrm{i}} - z_{\mathrm{i}}\right)^2}{\sum_{i=1}^{n}\left(z_{\mathrm{i}} - \overline{z}\right)^2}, \tag{4.7}$$

where $\widehat{z}_i$ and z_i are, respectively, the expected value of the predictions and observed values at the i-th location, $\overline{z}$ is the mean of the measurements, and n is the number of tested locations. E_{MA} was selected because it gives the same weight to all errors, while E_{NS} penalizes variance as it gives more weight to errors with larger absolute values. With its limitation to a maximum value of 1, E_{NS} facilitates general comparison.

For verifying the quality of predicted probability distributions, their accuracy and precision will be calculated for the validation set (where a "true" measurement

is available). While precision is a measure of the narrowness of the distribution, accuracy measures if the true value is contained in some fixed symmetric probability p-probability intervals (PI), e.g., interquartile range (Deutsch, 1997). For evaluating accuracy and precision together, we assess the Kullback-Leibler divergence (D_{KL}, Eq. 4.2 between the binary probability distribution (above-below threshold) and the true measurement (as shown in Fig. 4.2b) and take the mean over all validation points. D_{KL} is more than a measure of accuracy, since it does not need the definition of a probability cutoff to classify the binary distribution as hit or misclassification, and it is dependent on the probability values predicted. A maximum agreement $D_{\mathrm{KL}} = 0$ is obtained when all binary PMFs are very precise (probability of 1) and accurate (correct prediction) in predicting the true (above or below threshold), and it goes to infinity when a maximum disagreement is met.

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

$$\overline{\xi}\left(p\right) = \frac{1}{n}\sum_{i=1}^{n} \xi\left(u_i; p\right) \; \forall \; p \in [0,1]\,, \tag{4.8}$$

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

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

$$G = 1 - \frac{1}{L}\sum_{l=1}^{L} w_l \left|\overline{\xi}\left(p_l\right) - p_l\right|, \tag{4.9}$$

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

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

$$\overline{W}(p) = \frac{1}{n\overline{\xi}(p)}\sum_{i=1}^{n} \xi\left(u_i; p\right)\left[F^{-1}\left(u_i; \frac{1+p}{2}\right) - F^{-1}\left(u_i; \frac{1-p}{2}\right)\right]. \tag{4.10}$$

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

Overall, the validity of the model can be asserted when the mean error is close to 0, Nash-Sutcliffe efficiency is close to 1, mean of Kullback-Leibler divergence is close to 0, and accuracy (given by the goodness statistic) close to 1. Visually, a goodness statistic equal to 1 corresponds to an "accuracy plot" with maximum agreement between $\overline{\xi}(p)$ and p-PI. Note that the precision is only visually verified throughout the "PI-width plot", where the narrower the width of the PI (y-axis) the better. In Sect. 4.3.4.2, we discuss with real examples how these two plots (Fig. 4.10) interact.

4.3.3 *Benchmark models and setup of HER*

This section presents how HER was set up for the described dataset (Sect. 4.3.1) and briefly describes the two benchmark models, namely ordinary kriging (OK) and indicator kriging (IK). The authors suggest consulting Deutsch and Journel (1998), Goovaerts (1997), and Kitanidis (1997) for a more detailed explanation of the OK and IK methods. For brevity, details of the implemented models were included in Appendix C.1.

In OK, the unsampled values are estimated by a linear combination of the available data, which are weighted according to a spatial variability function (variogram) fitted to the data. It was selected for the comparison analysis due to the availability of a complete model for (the logarithm base of) lead concentrations of the Jura dataset in the literature. Therefore, OK parameters and results were taken directly from Atteia et al. (1994). The fitted variogram parameters are specified in Appendix C.1 (Table C.1). It is noteworthy that Atteia et al. (1994) estimated the model parameters by training on the full dataset (calibration plus validation set) while for all other models used in this paper, parameters are estimated by training exclusively on the calibration dataset and the performance is obtained in the validation set only. Since the uncertainty of OK models ignores the observation values, retaining only the spatial geometry from the data (Goovaerts, 1997, p. 180), we used the explicit assumption of normally distributed estimation errors in this study, which is a common practice for modeling local uncertainty in linear geostatistics (Goovaerts, 1998; Kitanidis, 1997, p. 68). Finally, to keep the results comparable, we discretized the predicted probability density functions employing the same discretization (bins) as used in HER. This binning scheme is presented and discussed in the next paragraph.

Similar to HER, the objective of IK is to directly estimate the distribution of z at an unsampled location without assuming a predefined uncertainty shape. For that, considering a defined cutoff value, an indicator transform (above-below cutoff) of the available data is combined with kriging weights to assess the probability of the z unsampled locations being above or below this threshold. When dealing with continuous variables, many cutoffs can be defined so that putting together their probabilities results in a full cumulative distribution. Since we are dealing with

continuous lead concentrations, for a fair comparison between HER and IK, the IK cutoffs were defined to coincide with the bins of HER. Therefore, in total, 69 cutoff values were specified, varying from 1.290 to 2.295 in steps of 0.015 (plus the critical limit z_c for the logarithm of lead concentration of 1.699). We defined the extremes of the distributions predicted by IK as the minimum and maximum Pb concentration of the calibration set (1.278 and 2.361, Table 4.1) as proposed by Deutsch and Journel (1998, p. 238) and Goovaerts (2009). Furthermore, the lag spacing used for the IK variogram was also the same as that used for the HER infogram, namely 70 meters (0.07 km). The parameter file used to model IK is shown in Appendix C.1 (Fig. C.3). Although choosing such a large number of thresholds is not common practice, it facilitates local uncertainty comparison (entropy maps and CPDs).

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

For comparison purposes, we fixed the lag distances of IK and HER at equal intervals of 70 meters (0.07 km) and the predicted log_{10}(Pb) distributions of OK, IK, and HER were equally discretized with evenly spaced intervals of 0.015. We selected this bin width for HER according to Thiesen et al. (2019), in which the size of 0.015 (equivalent to a concentration difference of 1.7 ppm around z_c)[3] showed a stabilization of the cross-entropy ($H_{pq} = H(p) + D_{KL}(p||q)$) when comparing the full calibration set and subsamples for various bin widths. Furthermore, to increase computational efficiency, and due to the minor contribution of faraway neighbors, we used only the 30 neighbors closest to the target. With the lag (or class), bin width, and number of neighbors defined, it was possible to assess the spatial characterization and, consequently, to proceed with the weight optimization (both available in Appendix C.1, Fig. C.1 and Fig. C.2). As shown in Fig. C.1, the calculated range contains 20 distance classes, reaching 1.4 km (roughly a third of the length of the x-domain). Considering the optimization problem proposed in Sect. 4.2.2.3, the optimum weights (w_{OR} and w_{AND}) obtained for Eqs. 4.3 and 4.4 are illustrated in Appendix C.1 (Fig. C.2b). Both contributions considerably decrease until the sixth class (circa 0.4 km), beyond which they stabilize and decrease almost linearly until reaching the range (1.4 km, class 20). The optimum contributions obtained for AND and OR aggregation in Eq. 4.5 are $\alpha = 0.65$ and $\beta = 0$, therefore exclusively intersecting distributions. The spatial characterization, aggregation method, optimal weights, and the set of known observations define the HER model for predicting local distributions.

3 Note that 1.7 ppm is approximately half of the standard deviation of various-sources errors estimated in Atteia et al. (1994) for the Pb dataset.

Table 4.2: Summary of the method procedures and associated performance metrics.

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

[a] All distributions are discretized by the same binning scheme.

The general procedures to obtain target estimates, distributions, and the binary probability for the contamination classification are summarized for each method in Table 4.2. The performance metrics related to each output are also shown.

4.3.4 *Results from local estimation with HER, IK, and OK*

Considering the similarities between HER and IK (both nonparametric methods with data dependent distributions), Sect. 4.3.4.1 focuses on presenting the local predictions of these two methods. OK maps are offered in Appendix C.2. In Sect. 4.3.4.2, the performance of all three interpolators is compared and discussed.

4.3.4.1 *Model application*

This section presents maps and distributions produced by IK and HER, using exclusively the Jura calibration set in their logarithm transform. Hereafter, we omit its logarithm form and refer to the data and results simply as lead (Pb) concentrations. For comparison purposes, an identical color range was used for maps presenting

the same information. Additionally, the color bars of Figs. 4.4 and 4.5 discriminate, respectively, the z_c threshold of lead concentration (1.699) and the entropy of the calibration set (5.348 bits, Table 4.1). All maps were developed using a grid with size of 0.05 km by 0.05 km.

In Fig. 4.4, we show the expected values (E-type) of lead concentrations. In general, a similar trend (given by the color shapes) for HER and IK can be seen, with similar low and high pollutant concentration areas. HER is slightly bolder in predicting extremely low (below 1.5) and high (above 2.1) concentrations, presenting larger areas in dark blue and yellow. The estimate map of OK is available in Fig. C.5a (Appendix C.2).

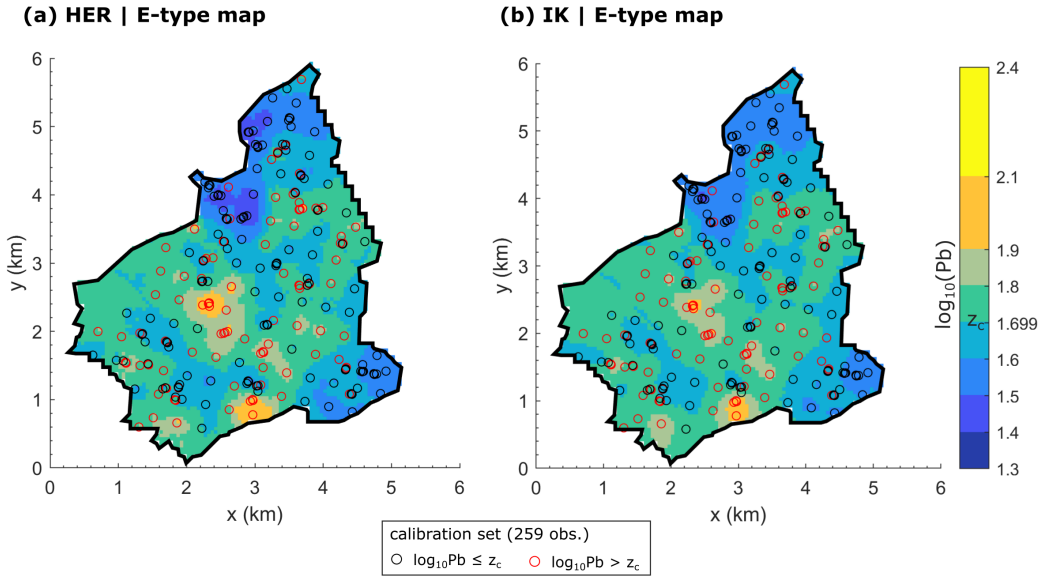

Figure 4.4: E-type map. **(a)** HER method, and **(b)** IK method.

Despite the similar trend of E-type values, the local uncertainty (Fig. 4.5) consistently differs between HER and IK. While IK predictions show generally lower uncertainty (all values are below the calibration set entropy of 5.348 bits), HER shows a broader range of entropy values. As expected, HER modeled a higher uncertainty to the west of the study area (Fig. 4.5a), where no nearby measurements are available, and lower uncertainty in the regions with a higher density of observations. Conversely, IK presents higher entropy in these denser areas.

The generally lower entropy of the IK map can be attributed, in this case, to the resolution of the local PMF, which is given by the numbers of cutoffs used for modeling. Although supporting the comparison analysis, the use of a finer resolution resulted in local distributions with empty bins (visible in Fig. 4.8), thus reducing the uncertainty of the distribution in terms of entropy. The entropy map and predicted distributions of an IK model with coarse resolution (IK_{10}) are available in Appendix C.2 (Fig. C.4 and Fig. C.6, respectively). Although different in magnitude, the same behavior of higher uncertainty in denser areas can be seen in IK_{10} (Appendix C.2, Fig. C.4). The entropy map of OK is available in Fig. C.5b (Appendix C.2).

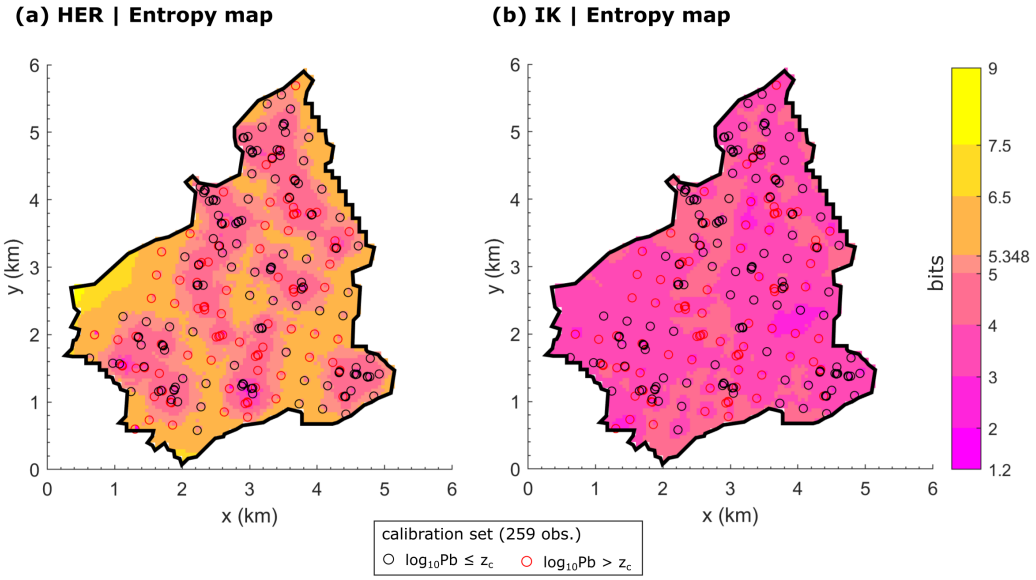

Figure 4.5: Entropy map. Local uncertainty in terms of entropy. **(a)** HER method, and **(b)** IK method.

Using the maximum acceptable concentration of lead (z_c), probability maps for exceeding this critical threshold were produced (Fig. 4.6). These maps were built by cumulating probabilities above z_c. Both methods, HER and IK, show high probability of contamination (in black) in zones of higher Pb concentrations and low probability of contamination (in light gray) in areas of lower concentration. HER shows larger areas in black and light gray than IK, being therefore a bit bolder in its predictions. Note that IK maps in Figs. 4.6b and 4.7b do not suffer any negative impact due to a large number of cutoffs, since only one cutoff (z_c) was used. The probability map of OK is available in Fig. C.5c (Appendix C.2).

According to Goovaerts (1997, p. 362), contaminated areas can be delineated by setting a location as "contaminated" if the probability of exceeding the tolerable maximum ($z_c = 1.699$) is larger than the marginal probability of contamination (0.421, estimated in Sect. 4.3.1), and "safe" otherwise. The proportion of wrongly classified points generally reaches its minimum close to the marginal probability of contamination (Goovaerts, 1997, p. 366). In the present application, all lead models (OK, IK, and HER) presented the minimum misclassification occurring close to the probability of 0.5 instead of the marginal probability of 0.421 (further discussed in Appendix C.2, Fig. C.8). However, considering that there are several ways to account for uncertainty in the decision-making process, and therefore greatly different results may be reached depending on the classification criteria (Goovaerts, 1997, p. 347, 362), comparing their differences is not within the scope of this work.

Thus, based on the probability map for z_c (Fig. 4.6) and the marginal probability of contamination (0.421), we binarize the probabilities to classify the results in "contaminated" and "safe" areas. HER and IK results are shown in Fig. 4.7, and OK in Fig. C.5d (Appendix C.2).

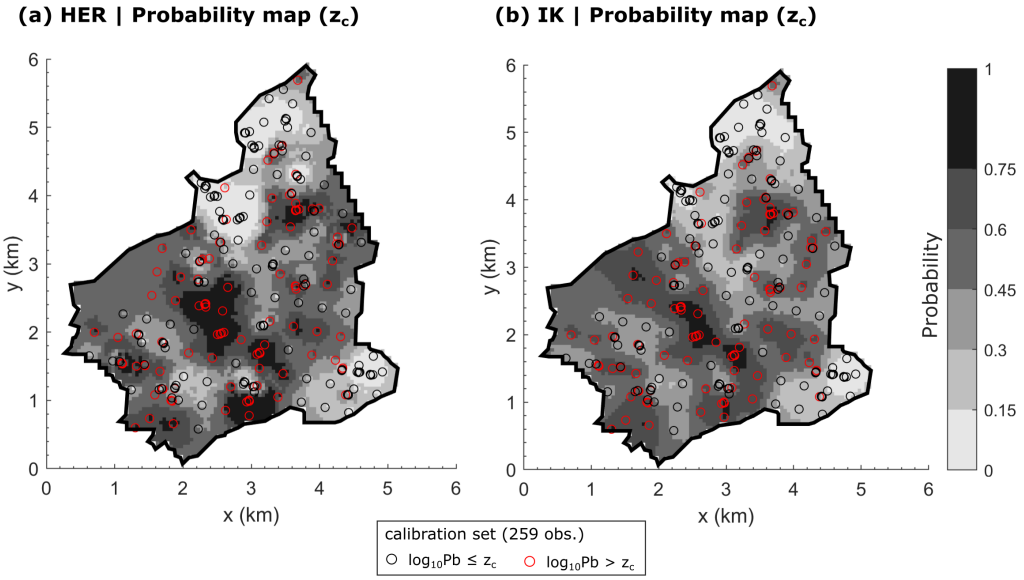

Figure 4.6: Probability map. Probability of exceeding the critical threshold ($z_c = 1.699$). **(a)** HER method, and **(b)** IK method.

The classification maps of HER and IK are relatively similar, however areas declared safe by IK are slightly more connected (Fig. 4.7b). In contrast, contaminated areas are more connected in the HER map (Fig. 4.7a). The respective OK maps can be found in Appendix C.2 (Fig. C.5), revealing a very local influence of each calibration point. For a more detailed theoretical comparison between HER and OK, please refer to Thiesen et al. (2020b).

Finally, six locations were selected to be explored in more detail. Four of them are from the validation set, and therefore represent a ground truth (targets A to D, Fig. 4.8), and two of them were selected from the grid by their distance to neighbors and their homogeneity (targets E and F, Fig. 4.8). These points, neighbors, and results are presented in Fig. 4.8. The locations were chosen with the goal to encompass targets with low (targets A and B) and high (targets C and D) concentration as ground truth, and a more homogeneous (targets A, C, and E) and a more heterogeneous (targets B, D, and F) neighborhood.

In general, all IK distributions (Fig. 4.8) contain empty bins between sampled values, while by construction, HER offers a higher resolution in the sense that the estimated CPD is more continuous. As a trade-off for these empty bins, in IK_{10} (Appendix C.2, Fig. C.6), fewer IK cutoffs were used, and the resolution was artificially increased by linearly interpolating the probability values within each cutoff. Nevertheless, IK and HER show relatively similar shapes and spread for targets A and E, locations with more homogeneous neighbors. Although their uncertainty differs, the expected values are also comparable, being equal for target E. Despite the homogeneity of their neighborhood, the expected values of targets A and C are not equal to their true value. One reason for this is that just a few (or no) nearby calibration points have a concentration as low (target A) or as high (target C) as their true value. The

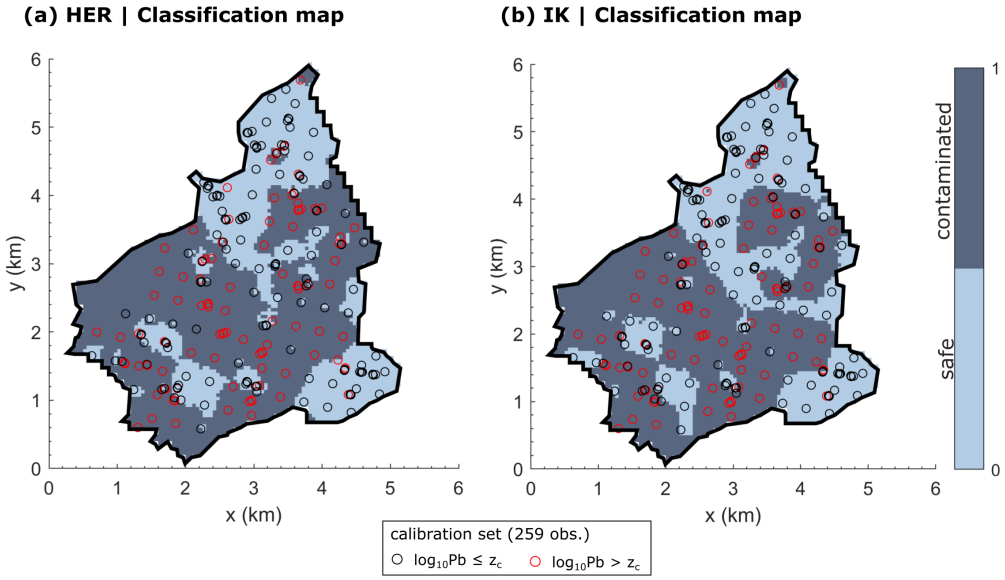

Figure 4.7: Classification map. Classification of locations as contaminated by lead on the basis that the probability of exceeding the critical threshold (z_c= 1.699) is larger than the marginal probability of contamination (0.421). **(a)** HER method, and **(b)** IK method.

same applies to target D, although it is in a heterogeneous neighborhood. At last, target F, which is located far from the calibration set, presents a higher entropy when predicted with HER, and a more certain distribution for IK. The local distributions of these targets and the IK_{10} model are available in Appendix C.2 (Fig. C.6). Neither IK nor IK_{10} achieved the finer resolution of HER.

Finally, Fig. 4.9 depicts the mean and two confidence intervals (CI) of the SW-NW cross section exclusively for the HER model. The SW-NW cross section location and its neighborhood are shown in Fig. 4.3a. The CI image also contains nine points from the calibration set (black circles), and seven points from the validation set (red squares), all of them located close to the cross section.

Some of the calibration points exactly match the SW-NE cross section. They can be identified in Fig. 4.9 as locations where the uncertainty goes to zero (from left to right, 1st, 4th, and 9th black circles). For points not exactly on the cross section, their influence in reducing the uncertainty due to their proximity to the transect is visible. In particular, the 3rd and 4th calibration points (black circles, Fig. 4.9) are in contrasting situations. The 3rd one is in a region with homogeneous calibration points close by – which result in a narrower uncertainty band –, while the 4th one presents an abrupt uncertainty reduction since it is located exactly in the transect, but its surrounding is rather heterogenous – which explains the wider CI in its surrounding.

Validation points of high Pb concentrations (2nd and 3rd red squares, Fig. 4.9) are outside the 95% CI. This happens due to relatively homogeneous neighbors in the first six distance classes (within a radius of circa 0.4 km), where none presents such high Pb concentration. On the other hand, for the more homogeneous regions (4th, 6th, and 7th red squares), E-type predictions are close to the true values. Note that

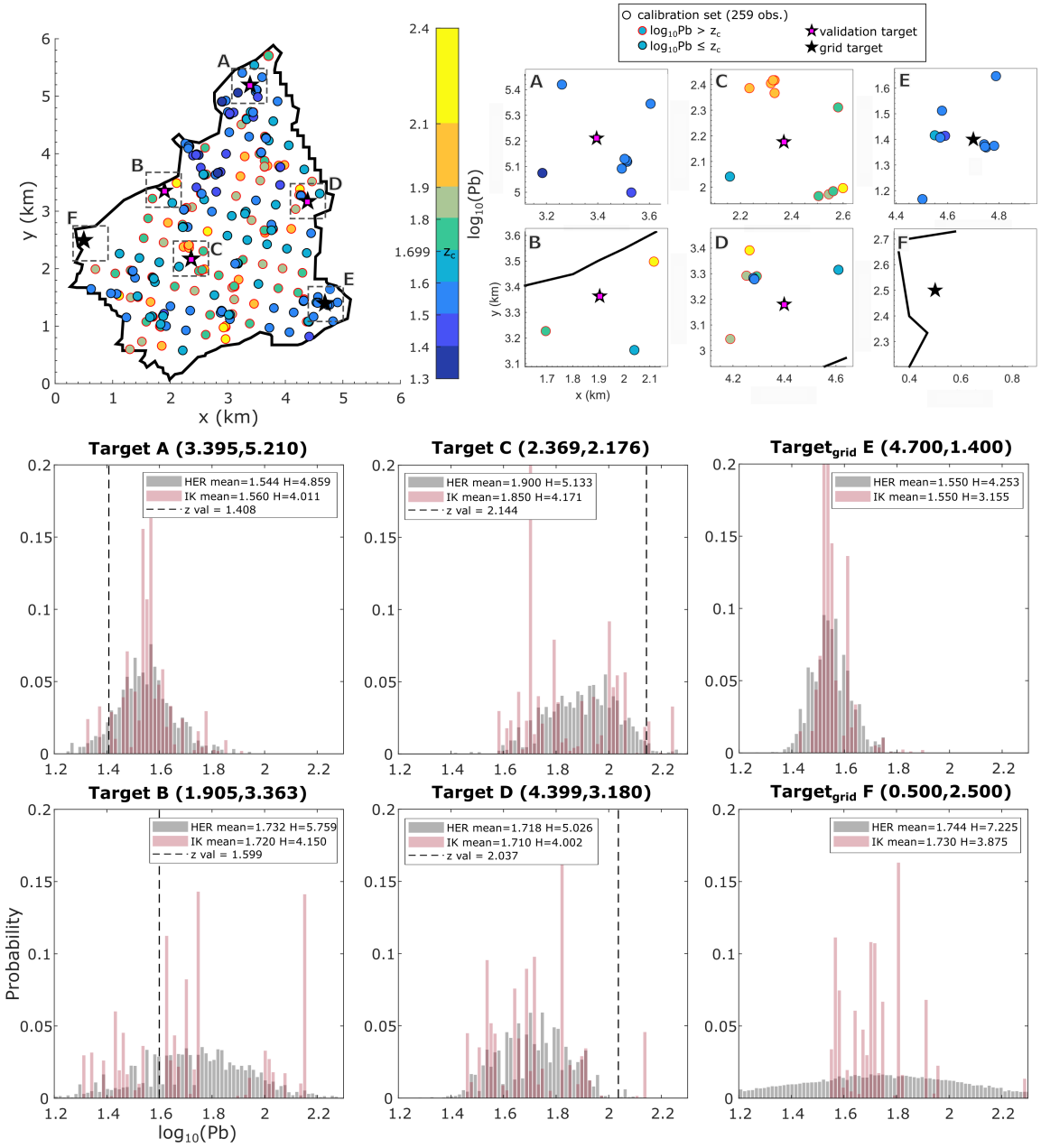

Figure 4.8: Local distribution of targets of the validation set (targets A to D) and grid (targets E and F) for HER (gray) and IK (red). Targets are identified by their coordinates (x, y). The location of each target is shown in a buffer of 600 m by 600 m.

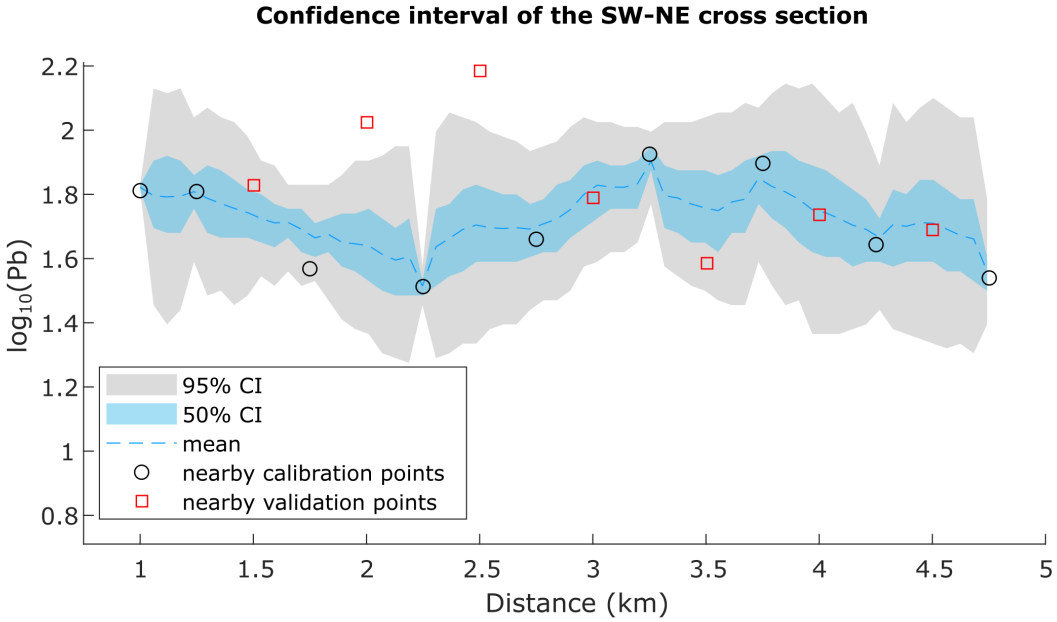

Figure 4.9: HER confidence interval (CI) of the SW-NE cross section (shown in Fig. 4.3a).

despite their continuous vicinity (with an increasing or decreasing tendency), these three validation points present different uncertainty band sizes. It is wider for 6^{th} and 7^{th} since they are located in a more heterogeneous region.

4.3.4.2 *Performance comparison*

In this section, the validation set is used to calculate the performance metrics of OK, IK, and HER. Table 4.3 summarizes their mean absolute error (E_{MA}), Nash-Sutcliffe efficiency (E_{NS}), Kullback-Leibler divergence (D_{KL}), and goodness statistic (G). Accuracy and precision are shown in Fig. 4.10.

Considering the deterministic metrics (based on the expected value), all models have a comparable E_{MA}. OK presents larger E_{NS} errors than IK and HER (Table 4.3). IK and HER have similar efficiency E_{NS}. On the other hand, when we cumulate the predicted distributions for the validation set in two bins (above and below threshold z_c) and compare its results to the true observation (as in Fig. 4.2b), HER presents the smallest divergence D_{KL} (mean over all validations points) between predicted and true probability, and OK the largest.

With respect to the Goodness statistic, OK and HER obtained the best G (Table 4.3). This reflects their accuracy in estimating distributions. Accuracy results are also shown in Fig. 4.10a. The nonparametric models (IK and HER) present points below the 45° line, which indicates the inaccuracy of these probabilistic models for large p-PI (mainly $p > 0.70$). The lower G of IK can be attributed to the goodness statistic, Eq. 4.9, penalizing inaccurate predictions, which shows points further away from the bisector line (around 0.80-PI, Fig. 4.10a) in comparison to OK and HER. Since a high G can be obtained by distributions with large spread, we used Fig. 4.10b to

Table 4.3: Cross-validation results for OK, IK, and HER method.

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

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

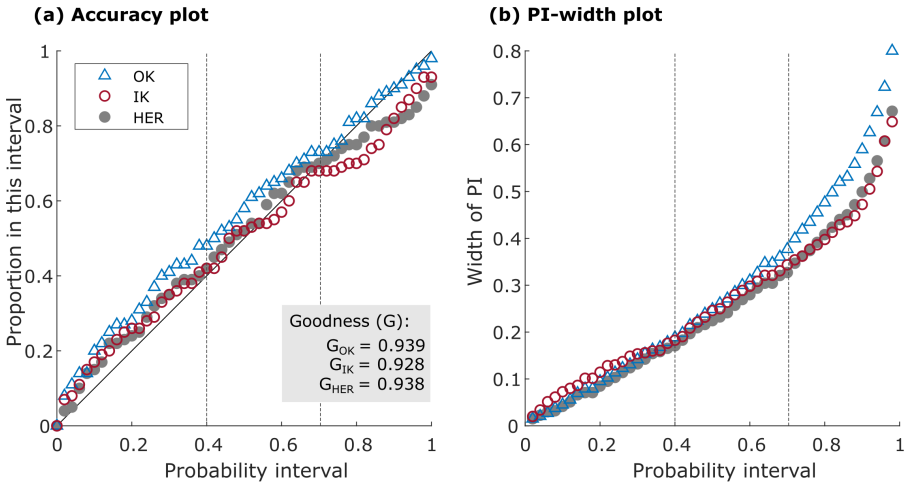

Figure 4.10: OK, IK, and HER performance. **(a)** Proportion of the true lead values falling within the probability intervals (p-PI) of increasing sizes, and **(b)** width of these intervals versus p-PI. The goodness statistic (G) quantify the similarity between the expected and observed proportions in the accuracy plots.

evaluate the precision of the models. The PI-width plot shows the estimated $\overline{W}(p)$ versus expected fractions p.

Considering that the smaller the PI-width (y-axis), the narrower (more precise) the distribution, Fig. 4.10b indicates that HER and OK predict more precise distributions approximately for $p < 0.40$, HER for $0.40 < p < 0.70$, and IK for $p > 0.70$. Besides being the model with narrower predicted distributions until $p < 0.70$ (Fig. 4.10b), HER points in Fig. 4.10a are above the bisector line being, therefore, considered accurate. On the other hand, for intervals of $p > 0.70$, HER and IK are considered more precise than OK (Fig. 4.10b), but at the cost of increasing their inaccuracy (Fig. 4.10a), i.e., their narrowness in the predicted distributions may cause the proportion of true values falling into these intervals to be smaller than for the OK model.

The accuracy and PI-width plots of the coarse model IK_{10} with linear interpolation of cutoffs are available in Appendix C.2 (Fig. C.7). Even though IK and IK_{10} present similar E_{MA}, E_{NS}, and D_{KL} (Appendix C.2, Table C.2), IK_{10} linear extrapolation of the distribution tails contributes to its increase in uncertainty (PI-widths as large as OK for large intervals, Fig. C.7b), therefore increasing accuracy ($G = 0.960$, Fig. C.7a).

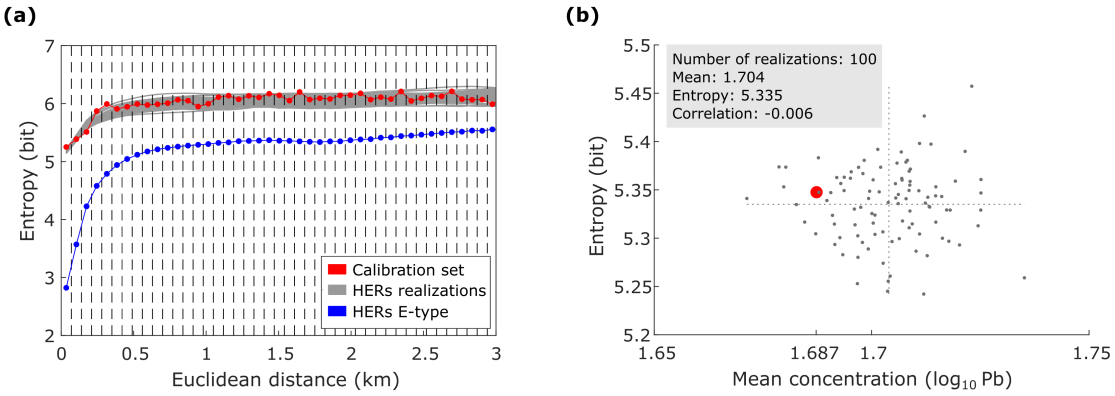

Figure 4.11: Ergodic fluctuations of 100 realizations generated with HERs. **(a)** Infogram and **(b)** scatterplot of the mean and entropy values.

4.3.5 *Results from spatial simulation with HERs*

Smooth interpolated maps, such as the ones produced by IK and HER, although locally accurate on average and appropriate for visualizing trends (Rossi and Deutsch, 2014, p. 167), fail to reproduce clusters of large concentrations, and consequently, should not be used for applications sensitive to the presence of extreme values and their patterns of continuity (Goovaerts, 1997, p. 370). Therefore, in this section, we show the results from applying HER in combination with sequential simulation (HERs, detailed in Sect. 4.2.3) for generating multiple realizations of the Pb concentration that match the calibration statistics and conditioning data. By construction, all these realizations honor the calibration values at their locations and should reflect the statistics deemed consequential for the problem at hand (Goovaerts, 1997, p. 370).

HERs was calibrated such that the statistical fluctuations of the realizations were reasonable and unbiased (Leuangthong et al., 2005). The statistical fluctuations due to a finite domain size are referred to as ergodic fluctuations, which mainly happen due to the size of the domain relative to the correlation length. We can expect these statistical fluctuations for anything less than an infinite domain (Leuangthong et al., 2005). In HER and HERs case, the correlation length reaches 1.4 km, i.e., circa one third of the x-domain length. Additionally, Rossi and Deutsch (2014, p. 168) argue that between 20 and 50 simulations are generally sufficient to characterize the range of possible values for the simulated values. We used 100 realizations to match the number of simulations done by Goovaerts (1997) for the Jura dataset. The fluctuation analysis of one hundred realizations is presented in Fig. 4.11, where we show their discrepancies in relation to the calibration infogram and marginal distribution. The challenges faced during the model calibration and details about the entropy calculation due to finite sample can be found in Appendix C.1.

As desired, the fluctuations of the infogram of the 100 realizations (gray curves in Fig. 4.11a) are unbiased in relation to the calibration infogram (red curve), spreading above and below it. This means that the spatial variability of the calibration set is reproduced by the realizations (although with some fluctuation). Departures between

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

Fig. 4.11b depicts that the entropy of the realizations (gray dots) is above and below the entropy of the calibration set (red dot), and that the mean entropy of the realizations (5.335 bits, represented by the gray dashed line) is close to the entropy of the calibration (red dot, 5.348 bits), indicating a reasonable reproduction of the uncertainty in the observed data. On the other hand, the mean of the realizations (1.704) is approximately 1% higher than the mean of the calibration set (1.687) and less than 0.25% higher than the mean of the E-type of IK (1.704) and HER (1.700). In this sense, the difference between the mean values of the simulation and the calibration dataset could reflect a bias due to spatial clustering of the observations, instead of a bias in the realizations with respect to the true mean of the population (Goovaerts, 1997, p. 370). Although it was not done here, when the simulated PMF is deemed too different from the target PMF an adjustment of the simulated PMFs is possible (Goovaerts, 1997, p. 427). According to Deutsch and Journel (1998, p. 134), any realization can be postprocessed to reproduce the sample histogram; hence the sample mean and variance. To do so, Journel and Xu (1994) proposed a posterior identification of the histogram, which allows improving reproduction of the target PMF while still honoring the conditioning data and without significant modification of the spatial correlation patterns in the original realization. For the sake of brevity, the improved reproduction of PMFs is beyond the scope of this paper. We should bear in mind that verifying the quality of the reproduction does not provide an indication on the goodness of the set of realizations as a whole, because unlike models of local uncertainty (that have true observations to be compared), there is no reference spatial distribution of values to be used in models of spatial uncertainty (Goovaerts, 2001).

For illustration, two arbitrary stochastic images constructed with HERs and the calibration dataset are pictured in Fig. 4.12.

One can notice that the generated stochastic images (Fig. 4.12) do not smooth out details of the spatial variation of the Pb concentration as in estimation maps (Fig. 4.4). And compared to interpolation techniques like OK, IK, and HER, the variability of the simulated maps is higher due to the incorporation of variability between unsampled points. A comparison between the E-type and simulation variability in space is available in Fig. 4.11a.

In general, both images present low concentration zones (blue) to the North and Southeast of the study area, which are derived from the low uncertainty and the tendency of low concentration previously verified in the regions (Fig. 4.5a and Fig. 4.4a, respectively). Similarly, the zone with high concentration and low uncertainty (around

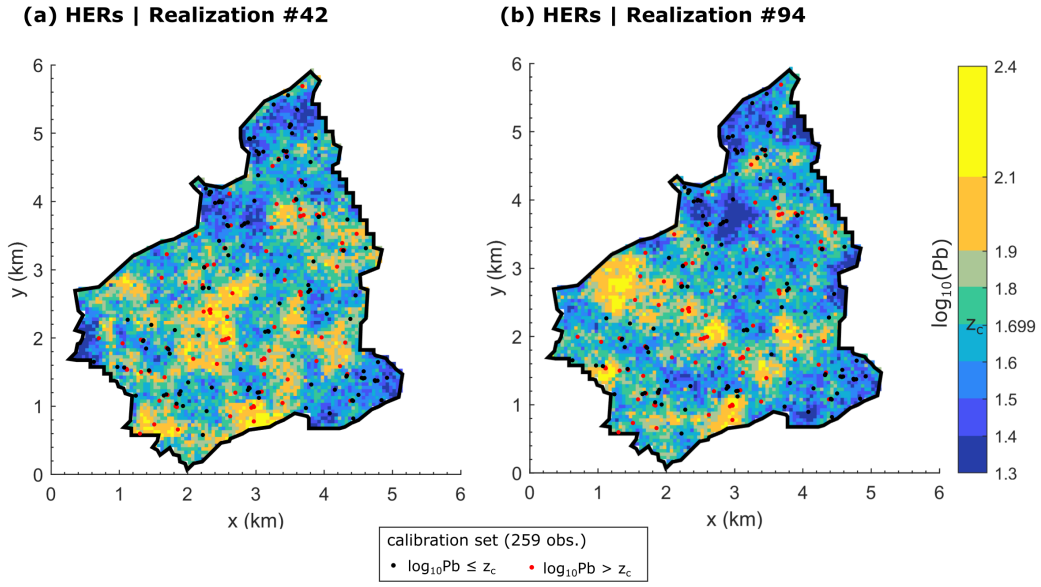

Figure 4.12: Realizations generated with HERs. **(a)** Realization #42 and **(b)** realization #94. Simulation grid size of 0.05 km x 0.05 km.

x=2.5 and y=2.5, Fig. 4.4a and Fig. 4.5a) presents, in both realizations, high Pb concentrations. On the other hand, regions with higher uncertainty (due to the heterogeneity of the sample data or because they are far away from sample data) present a more variable concentration when comparing both images.

4.4 DISCUSSION

In general, IK and HER are conceptually different in their modeling. HER relies on empirical probability distributions to describe the spatial dependence of the study area and uses aggregation methods to combine distributions. IK estimates a number of probabilities for a series of cutoffs, for each of which an indicator variogram is modeled to describe the spatial continuity of the study area, and the estimated probabilities are then interpolated to obtain the full distribution. Furthermore, a global set of weights for the classes is obtained with HER, while IK performs multiple local optimizations, one for each target and cutoff. Both methods share similarities: they are nonparametric in the sense that no prior assumption about the shape of the distribution being estimated is made, their results are data dependent, and they can be applied to continuous or categorical variables. Such characteristics do not apply to OK, therefore, we focused our analysis on IK and HER. A detailed conceptual discussion comparing OK and HER is available in Thiesen et al. (2020b). Although HER is considered nonparametric, two assumptions are implicit in defining the weights used for the PMF aggregation: one in linearly interpolating the optimum weights obtained for each class, and the other in defining the optimization problem (both topics are discussed in Sect. 4.2.2.3). An analogous interpretation of these assumptions can be applied to IK, where the weights are obtained by minimizing the variance and applied

to the linear combination of the observations. The latter step is comparable to the choice of the aggregation method in HER.

IK and HER are distance models between any two pair of points, with different forms of inference. While in IK the spatial variability of the attribute values can be fully characterized by a single covariance function, which differs for each cutoff (Goovaerts, 1997, p. 33), HER relies directly on the dataset to extract one distribution for each distance class (as seen in Fig. C.1). The stationarity assumption behind the inference is a model decision (and not a characteristic of the physical phenomenon) and can be deemed inappropriate if its consequences do not allow one to reach the goal of the study (Goovaerts, 1997, p. 438). The inference of the spatial characterization together with the aggregation procedure allows the spread of the local distributions in HER as well as the simulated values of HERs to naturally reach values beyond the calibration set (both above the maximum and below the minimum). For IK, this is only possible if the user imposes extremes beyond the calibration set. Likewise, the extremes of HER distributions can be restricted by the user according to their interest.

Interestingly, despite their conceptual differences, in this study HER and IK show comparable performance in both deterministic and probabilistic terms (Table 4.3 and Fig. 4.10). One exception is the Kullback-Leibler divergence (D_{KL}), for which HER was able to classify "contaminated" and "safe" areas with higher precision and accuracy. Such accomplishment may be explained by the fact that the HER optimization problem was built around this metric (Sect 4.2.2.3), although this does not guarantee the best performance in the validation set. Regardless of the performance comparison presented, we should be mindful that there is no unique, best, or true model for modeling uncertainty (Journel, 2003). Consequently, there can be several alternatives that depend on the user decision to model the uncertainty which can be more suitable to the problem at hand.

When applying IK, two major issues arise, namely, inconsistent (negative) probabilities when estimating distributions and the choice of interpolation/extrapolation models to increase the resolution of the estimated distribution (Goovaerts, 1997, p. 441, 319, 326; Goovaerts, 2009). The first is known as order relation deviations and is typically treated by a posteriori correction of the estimated probabilities, which imposes nonnegative slopes to the cumulative distribution (Goovaerts, 2009). For the latter, there are different ways of achieving a finer resolution of the distribution. Increasing the number of cutoffs leads to cumbersome inference and modeling of multiple indicator variograms (one for each cutoff), which consequently increases the likelihood of order relation deviations due to the empty cutoff classes (Goovaerts, 1997, p. 326; Rossi and Deutsch, 2014, p. 160). As an alternative to that, multiple interpolation and extrapolation models are available in the literature. In such cases, where interpolation/extrapolation models are used, besides the arbitrariness of the model selection (Goovaerts, 2009), distribution statistics such as the mean or variance may overly depend on the modeling of the upper and lower tails of the distribution (Goovaerts, 1997, p. 337). Therefore, due to the trade-off between increasing the number of thresholds and using models to derive continuous distributions, both alternatives were discussed in this paper (IK and IK_{10}). Regardless of the chosen approach, the risk of suboptimal choices by the user remains. Conversely, HER avoids imposing these corrections to the distributions and multiple variogram fitting, but its

parameter choices (such as distance class size, bin width, number of neighbors, and aggregation type) are also subjective. Yet, for both methods HER and IK, parameter decisions can be based on performance metrics via leave-one-out cross-validation, for example.

Both IK and HER estimated remarkably similar values of Pb concentration (E-type map, Fig. 4.4). On the other side, the maps associated with the probabilistic results (entropy map in Fig. 4.5, probability of exceeding the critical threshold in Fig. 4.6, and classification map in Fig. 4.7) are distinct, with increasing uncertainty of HER in data sparse regions. We noticed that when dealing with sparse data, there is not enough data to fill each cutoff in IK, which, due to the resulting empty bins, decreases the uncertainty (entropy). The opposite happens in denser regions, where more data is available and the chances of more bins being filled is higher, increasing therefore the entropy for heterogeneous regions. As discussed in Sect. 4.3.4.1 (Fig. 4.8), both methods reflected the expected behavior of larger errors in locations surrounded by data that are very different in value (as expected and argued by Goovaerts, 1997, p. 180). However, in terms of PMF resolution, the greater computational and inference cost of HER in comparison to IK is balanced by a finer resolution of the distributions, which could be neither achieved by the IK nor the IK_{10} model. The lack of resolution in IK is particularly severe when using indicator-related algorithms with only a few cutoff values such as the nine deciles of the sample (Deutsch and Journel, 1998, p. 134). In this case, the loss of information available in continuous data is more accentuated in IK than in HER, due to the indicator transform of the data (Fernández-Casal et al., 2018) and few cutoffs. In contrast, the resolution of HER distributions is given by the selected bin width and, consequently, an indicator transform would only be needed as a post-processing step (such as for a probability analysis of exceeding a critical threshold or a classification map).

In terms of simulation, HERs has proven to be difficult to calibrate. Many parameters were tested until the entropy (variability) of the realizations converged to the entropy of the calibration dataset. In the sensitivity analysis performed (not shown), the authors verified a strong impact of the number of aggregated distributions (thus, number of neighbors) when intersecting distributions. The stronger the contribution of the AND combination (which is the case here), and the higher the homogeneity of the data, the more sensitive the spatial variability of HERs is to the number of neighbors. Therefore, in general, too many equal (homogeneous) PMFs would result in a very narrow output (deflation of the spatial variability), whereas too few could inflate it. Although a first analysis of the simulation procedure and results of HERs was introduced in this paper with promising results, further investigations considering the influence of different data properties, implementation of strategies (such as search neighborhood and multiple-grid simulation available in Goovaerts (1997, p. 378 p. 379), and the addition of transfer functions are needed.

Finally, we should bear in mind that uncertainty arises from our lack of knowledge about the phenomenon under study and, therefore, it is not an intrinsic property of the phenomenon (Goovaerts, 1997, p. 441). Uncertainty is data-dependent and, most importantly, model-dependent, and, consequently, can be controlled by the expert according to their wishes (Journel, 2003). No model, hence, no uncertainty measure, can ever be objective: the point is to accept this limitation and to document clearly

all its aspects (Goovaerts, 1997, p. 441; Journel, 2003). Thus, despite the uncertainty differences between IK and HER and our attempt to quantify their performances, IK and HER presented legitimate results, which exhibited similar accuracy and precision performances.

4.5 SUMMARY AND CONCLUSION

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

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

HER-based sequential simulation (called HERs) allowed generating realizations that reproduced the spatial variability of the sample set. The quality of the realizations was verified in terms of their statistical fluctuation in relation to the sample set. However, no further analyses of the results (such as benchmarking comparison or adding transfer functions) were carried out, due to the typical absence of a spatial distribution of values to be used as a reference (Goovaerts, 2001).

HER and its adaptation HERs allow nonparametric estimation and stochastic predictions, avoiding the shortcomings of fitting any kind of deterministic curves and, therefore, the risk of adding information that is not contained in the data (or losing available information), but still relying on two-point geostatistical concepts. In relation to IK, HER has shown to be a unique tool for estimating nonparametric conditional distributions with the advantage of (i) not presenting problems of order-relation deviations, (ii) being free of function assumptions for interpolating probabilities or extrapolating tails of distributions, (iii) not requiring the definition of various cutoffs and, consequently, their respective indicator variogram modeling, (iv) displaying a finer resolution of the predicted distribution, (v) avoiding strong loss of information due to data binarization, and (vi) bringing more flexibility to uncertainty prediction through the different aggregation methods and optimization strategies. Finally, due to the growing use of stochastic simulation algorithms for uncertainty assessment in soil science and the potential improvement of results given the consideration of soft variables (secondary data), the authors believe that additional investigations of HERs and model adaptations of HER are topics worth of further research.

Part V

CONCLUSION

This closing chapter compiles the key findings and results of this thesis, and discusses their general relevance for Earth system modeling and geostatistics. In addition, supported by the limitations faced in the study, I raise opportunities for future research.

5 CONCLUSION

5.1 SUMMARY AND CONTRIBUTIONS

Motivated by the challenges of complexity and underdeterminism of Earth system science problems, this thesis develops and tests a nonparametric, probabilistic framework to express and apply geoscientific knowledge. In particular, it focuses on uncertainty analysis using information theory (IT) in spatial and temporal contexts. The key challenges for building such a modeling framework are to find efficient ways to work with large datasets, to consider the effects of learning from limited data, to find generalized ways to combine various sources of information, and to deal with uncertainty. The results indicate that IT is a proper choice for uncertainty and information quantification, which, together with probabilistic modeling directly derived from data, adds generality to the modeling process and helps to learn relations between data unconstrained by functional or strong parametric assumptions. All three applications proposed here explore a new framework firmly rooted in probability and information theory, which together allow nonparametric learning and prediction in temporal and spatial domains. I start with the analysis of time series in the context of rainfall-runoff events in chapter 2, followed by spatial interpolation (chapters 3 and 4), and simulation for soil contamination analysis (chapter 4). The main contributions of the work are both theoretical and application-oriented, which drives the research to the following key conclusions:

Probability and information theory are applied to nonparametric learning and to address uncertainty.

The three testbed applications explore learning and prediction from temporal and spatial data in terms of information.

Properties of the framework proposed in this thesis:

- Usually, raw data carry some uncertainty caused by equipment errors, calibration, or different kinds of methodological assumptions and expert's judgment (Savelyeva et al., 2010). In the proposed framework, predictive relationships are directly derived from data and expressed as discrete probability distributions. The advantage of this is that it helps us to make good use of the available data since, as much as possible, it avoids the introduction of undesirable side information or erasing existing information coming, e.g., from suboptimal expert's parametric choices, data transformation, or lossy compressions (Neuper and Ehret, 2019). Furthermore, applying a probabilistic relation in the way it is proposed has the benefit of providing joint statements about the target and the related estimation uncertainty unconstrained by functional assumptions.

- avoids the introduction of undesirable side information or erasing existing information by using discrete probability distributions;

- The guiding theme of this work is the application of IT to Earth science problems. IT allows to extract information about the patterns and relations from data or to compress data while preserving the information they contain. Particularly in this thesis, IT allows to directly quantify uncertainty and information content of datasets, measure data-relations and uncertainty reduction of models, investigate the representativeness and predictive power of variables and models, compute minimal data requirements to avoid overfitting, characterize the spatial dependence fingerprint of a variable in a field, calculate model performance, and optimize model structures – all of this using the same universal currency of bit. The use of a single unit is beneficial, as it allows explicitly comparing and joint treatment of many different sources of information in a single framework.

- enables to directly quantify uncertainty and information content of datasets, and to analyze patterns and data-relations in a single unit, bit;

- Specifically in chapter 2, the proposed data-driven approach based on IT is seen as a consolidation of descriptive and experimental investigations since it allows one to describe the drivers of the model by quantifying the information contribution of the predictors and to investigate the similarity of the model hypothesis with respect to the ground truth. Beyond being a way of understanding the drivers of the system (also useful for, e.g., feature selection in the machine learning context), the framework enables to consider the effects of time ordering, learning from limited data and from models with increasing complexity and, consequently, choosing the most suitable model for the available dataset.

- describes the drivers of a system;

- allows the selection of the most informative model according to the dataset;

- These primary findings pave the way for proposing an information-theoretic framework for spatial interpolation (chapter 3) called HER. Here, IT is anchored to principles of geostatistics, allowing to characterize the spatial dependence of a variable by quantifying the information content of the data conditioned on the lag distance to extract its correlation length and to minimize the disagreement between observed data and predictions. By the same token, the probabilistic embedding (data-driven and nonparametric distributions) allows an honest accounting of the related uncertainties, bypassing function fitting of the spatial dependence structure and, therefore, avoiding the risk of adding information not available in data. It also brings more flexibility to the model since it is feasible to adjust the number of lags to be optimized according to the amount of data available. Additionally, the use of aggregation methods for combining distributions (Allard et al., 2012) brings a new facet to spatial interpolation, allowing one to incorporate

- relaxes assumptions and minimizes uncertainties;

- incorporates different uncertainty properties with aggregation methods;

different uncertainty properties according to the dataset and expert interest.

- In the subsequent chapter 4, the spatial interpolation framework of HER is further developed to assess local uncertainty when dealing with categorical data and threshold-exceeding probabilities, and to reproduce the spatial fluctuation of the dataset reality with sequential simulation and, thus, assessing spatial uncertainty with HERs. Here, it is verified that, different from traditional approaches, HER does not present problems of order-relation deviations, is free of function assumptions for interpolating probabilities or extrapolations tails of distributions, allows a fine resolution of the predicted distribution, avoids a strong loss of information caused by data binarization, does not require the definition of various cutoffs and, consequently, their respective indicator variogram modeling.

- allows simulating data properties and fluctuations;

- permits to deal with categorical or continuous data without issues presented in traditional approaches;

Altogether, the three applications have proven to be successful across a range of applications: from event detection (chapter 2) to spatial interpolation and simulation of toxic elements in soil (chapters 3 and 4). They explore data-based modeling firmly rooted in probability and information theory. This integration, on the one hand, entails the generality and flexibility needed to handle any kind of data-relations and limitations in data volume while, on the other hand, provides a tool for interpretation in terms of information content or its counterpart of uncertainty.

- addresses any kind of data-relations; and

- is flexible to be adapted to different problems at hand.

5.2 OUTLOOK AND RECOMMENDATIONS

The benefits of working data-driven and being unconstrained by strong assumptions however come at a price. Although mitigated by the increasing availability of data volume and computer power (Bell et al., 2009), learning robust data-based relations requires a considerable amount of data, and applying them for predictions is computationally more expensive than using deterministic functions (Neuper and Ehret, 2019). To enhance the computational effort, with further developments, it might be advantageous to work with kernels (or fitting probability density functions, pdf) to replace empirical distributions with mathematical functions. Especially for the HER interpolation method, kernels and pdfs have the potential benefit of providing a transition in the spatial characterization model, which is currently defined for each distance lag individually. Although this has the potential of improving the calculations and model results, the choice of kernel brings the risk of adding side information and new assumptions to the framework, which is the reason why nonparametric distributions are used so far. Other possibilities would be to reduce redun-

The work limitations of this thesis encompass:

- the issue of computational performance and dataset size;

dant computations by dividing the study domain in subdomains according to their similarity (as in Ehret et al., 2020).

Eventually, the use of mathematical functions might also help to deal with the caveats in defining the discretization of the data in bins to build distributions. The perils related to the bin width selection are challenging and directly influence the entropy quantification (Gong et al., 2014; Pechlivanidis et al., 2016). Choosing a too fine discretization (too many bins) increases the risk of overfitting and, therefore, higher uncertainty, while too few bins can lead to oversmoothing distributions (Larson, 2010) and, hence, lower uncertainty. In this thesis, the effect of bin resolution is addressed when calculating the amount of data needed to obtain a robust learning curve in order to avoid the problem of overfitting. Additionally, the binning scheme is kept constant throughout each analysis, so that the uncertainty increase (or decrease) is always relative to a fixed model. Although challenging, the binning transformations of data brings an intuitive interpretation of uncertainty, data resolution, and information content of the related distribution, allowing, for example, associating the bins to physical or uncertainty knowledge of the variable under study. It further facilitates the adaptation of frameworks built to deal with continuous data to also handle categorical data while maintaining the underlying logic of the framework.

- the issue of binning transformations of data;

Specifically related to the spatial interpolator, it is not possible to mathematically compare the proposed HER to kriging equations to argue that kriging results are a particular case of many possibilities of HER. This happens because both methods, although based on fundamentals of geostatistics, are fundamentally different in their implementation. While in HER a global set of weights are kept fixed for the lags, kriging performs multiple local optimizations (one per target), and, therefore, the weight of the observations varies according to each target. Additionally, the inclusion of aggregation methods has a two-fold impact. On the one hand, it adds complexity to the framework, raising questions on how to select the aggregation method to be used, how to define and solve the optimization problem, or how to interpret the different possibilities of aggregation. On the other hand, it brings flexibility to adapt the method according to the type of available data and to work with different uncertainty properties.

- the issue of complexity;

The modeling approach presented in this thesis is limited to dealing with problems that assume stationarity in space or time. In this fashion, they will fail if addressed to questions of change, i.e., if the past from which we have learned does not represent the future we want to predict or if the spatial dependence is not the same along the field. The stationarity assumption behind the inference is a model decision (and not a characteristic of the physical phenomenon) and can be deemed inappropriate if its

consequences do not allow one to reach the goal of the study (Goovaerts, 1997, p. 438). In this sense, further research is required to tackle nonstationarity questions. Similarly to Ehret et al. (2020), who proposed an adaptive clustering to address space-time data by grouping the model domain into similar subdomains, an adaptive model (of spatial, temporal, or spatio-temporal data) could be applied according to the similarity of subdomains. In this case, information-theoretical concepts could be properly used to measure similarity and grouping the subdomains. In turn, the multiple subdomains, each one with its own model, comes at the price of rapidly increasing data demand with system size and number of subdomains, demands that consequently amplify the computational effort.

- the issue of stationarity; and

Along with the increasing interest in data-driven methods, an interesting avenue of research is to incorporate aspects of physical theory into data-driven models. This desire for coupling physics-based and data-driven approaches is engaging for a number of reasons. While data-driven methods and statistical learning contribute to both accounting and extracting patterns from data, the use of models which are based in, or constrained by, physical properties allows us to both learn about the underlying processes of the systems we are modeling and to extrapolate the modeling to situations that have never been seen before. As a step in this direction, in chapters 2 the physically-based approach of CPM (characteristic point method; Mei and Anagnostou, 2015) is improved using data to calibrate the CPM model. In a more subtle manner, the proposed frameworks already incorporate some physical knowledge implicitly by the choice of predictors in chapter 2 and by the assumption that near things are more related than distant things (a condition known as the first law of geography; Tobler, 1970) in chapters 3 and 4.

- the issue of physics-based and data-driven models.

The studies presented here are limited to dealing with time series and spatial interpolation separately and focus on a particular problem. Therefore, additional investigations are required to analyze the method in the face of spatio-temporal domains and to quantify the degree of consistency (similarity) between data of two systems in order to transfer relations learned from a particular system to another (a process referred to as regionalization). Most especially in the spatial context, improvements in the theoretical and modeling areas are important to further address the use of information of additional variables (also known as covariables), handling redundant data, integrating continuous and categorical data in the same framework, analyzing the contributions of the searching neighborhoods, and the influence of the aggregation method in the inflation/deflation of the variability of simulated locations.

Topics worth of further research:
- spatio-temporal problems;
- regionalization; and
- improving predictions with covariables.

5.3 CONCLUDING REMARKS

This thesis looks afresh at typical Earth science problems through the lens of information theory.

The proposed framework entails the generality needed for modeling in terms of purpose, degrees of freedom, and data availability.

As argued by Singh (2018), due to the computing prowess and sophisticated instrumentation available these days, integration of hydrology, and therefore, Earth science, with allied areas is rapidly increasing and will so continue. Here, at the same time that this work proposes an interdisciplinary approach to analyze different Earth science problems, this can be contrasted to the fact that the model construction and evaluation are firmly rooted in a single property – information. Notwithstanding the discussed caveats and limitations, several important findings emerge from the three studies displayed in this thesis. In general, the knowledge about relations between data is represented by discrete, multivariate, probability distributions derived directly from observed data. The common goal of the proposed applications is to avoid conceptualization and compression of data-relations by nonparameterization of distributions, helping to preserve the information content of the data and, at the same time, allowing an honest account of the related uncertainties. Overall, the developed framework shows that the integration of probability and information theory allows a generalized way to build solutions tailored to the problem at hand. Altogether, the three applications have proven to be successful across a wide range of domains, showing great modeling flexibility in terms of purpose, degrees of freedom, and data availability. All things considered, it is my expectation that the research work presented in this thesis has contributed to look afresh at typical geoscientific problems through the lens of information theory.

Part VI

APPENDIX

A

APPENDIX TO CHAPTER II

A.1 RESAMPLING STRATEGY AND NUMBER OF REPETITIONS

In the study, samples of size N from the dataset were obtained through bootstrapping, i.e., they were taken randomly but continuously in time, with replacement among the W repetitions. For each sample size, we repeated draws W times and took the average cross entropy and D_{KL} to eliminate effects of chance (see repetition statements N and W in Fig. 2.1). Thus, in order to find the value of W which balances statistical accuracy and computational efforts, we did a dispersion analysis through calculating the Shannon entropy (as a measure of dispersion) of the cross entropy distribution of the (unconditional) target model (model no. 0 in Table 2.7). Sixty one bins ranging from 0 to 6 in steps of 0.1 bits were used; this contemplates the range of all possible cross entropy values among the tested pairs of N and W. Fig. A.1 presents the Shannon entropy applied as a dispersion parameter to analyze the effect of the number of repetitions W for different sample sizes N.

Considering the graph in Fig. A.1, in general, the behavior of the Shannon entropy among the repetitions is similar for each sample size analyzed, indicating that the dispersion of the results according to the number of repetitions does not vary too much, i.e., the bins are similarly filled. However, it is possible to see that, as the sample size increases, the Shannon entropy for the different number of repetitions approaches that for the 50 000 repetitions. For sample sizes up to 7500, the bars from 50, 100 and 300 repetitions present some peaks and troughs, indicating some dispersion in filling the bins. Thus, in this case study, the minimum of 500 repetitions was assumed as a reasonable number of repetitions for computing the mean of the cross entropy in the sample size investigation. This number of repetitions was also validated considering the smoothness and logical behavior of the curves obtained during the data size validation and curse of dimensionality analyses (Fig. 2.5 in Sect. 2.4.1.2).

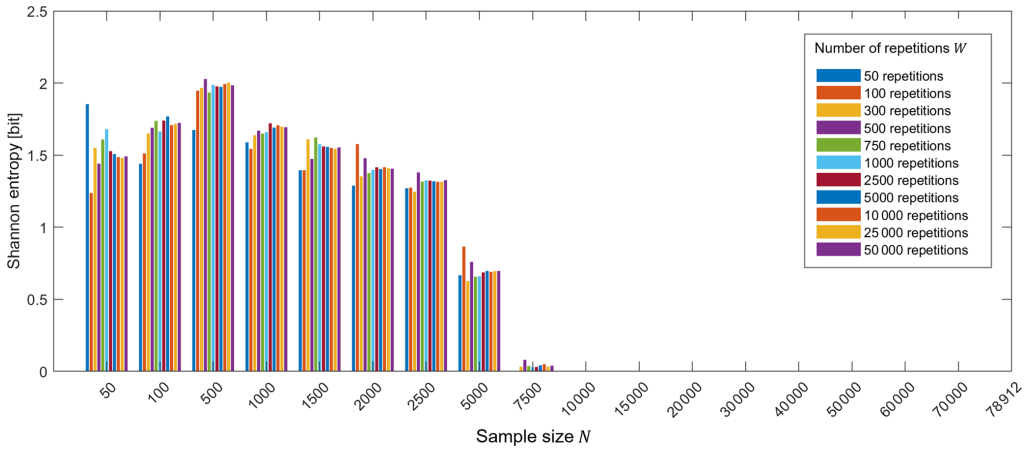

Figure A.1: Dispersion analysis of the cross entropy. The effect of the number of repetitions in the target model (no. 0 in Table 2.7).

B

APPENDIX TO CHAPTER III

B.1 SUMMARY STATISTICS OF THE RESAMPLED DATASETS

Table B.1: Summary statistics of the resampled datasets – Short-range (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 B.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.

B.2 PARAMETER TUNING

This appendix consolidates the final parameters used in the models presented in Sect. 3.4.2. Particularly for HER, Fig. B.1 presents the final weights optimized for Eqs. 3.4 and 3.5. It was limited to 18 grid units (nine distance classes), due to the small contribution of the faraway classes. Similarly, Fig. B.2 shows α and β weights of Eq. 3.6. Finally, Table B.3 and Table B.4 summarize the calibrated parameters obtained for each model (varying method, sample size and dataset type).

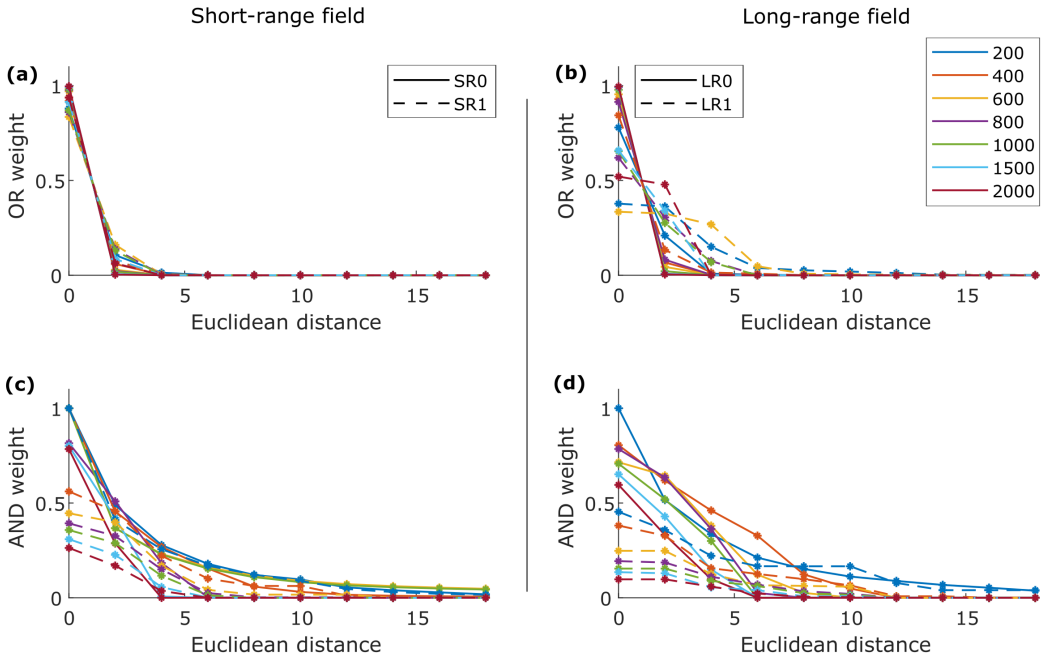

Figure B.1: HER optimized weights by distance class: **(a,b)** w_{OR}, Eq. 3.4, and **(c,d)** w_{AND}, Eq. 3.5. SR datasets on the left panel and LR datasets on the right panel. Continuous line refers to datasets without noise and dashed lines to datasets with noise.

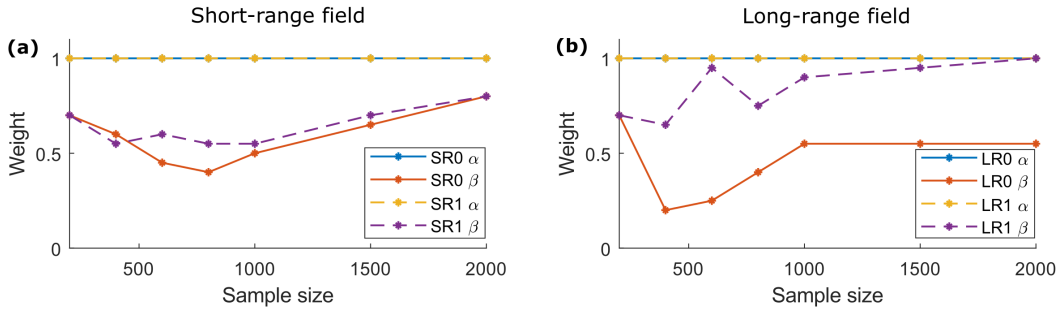

Figure B.2: HER α and β weights by sample size, Eq. 3.6: **(a)** SR datasets on the left panel and **(b)** LR datasets on the right panel. Continuous line refers to datasets without noise and dashed lines to datasets with noise

Table B.3: Method calibration by sample size – parameters of the models for the short-range dataset (SR0 and SR1).

	Sample size	**200**	**400**	**600**	**800**	**1000**	**1500**	**2000**
Method	**Parameter**				**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**				**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

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

Table B.4: Method calibration by sample size – parameters of the models for the long-range dataset (LR0 and LR1).

	Sample size	**200**	**400**	**600**	**800**	**1000**	**1500**	**2000**
Method	**Parameter**				**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	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**				**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

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

B.3 SUMMARY STATISTICS OF THE MODEL PREDICTIONS

This appendix 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 B.5 to Table B.8. Finally, Fig. B.3 illustrates their residue correlation (obtained by calculating the Pearson correlation coefficient between the true values and the residue of the predictions).

Fig. B.3 illustrates 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.

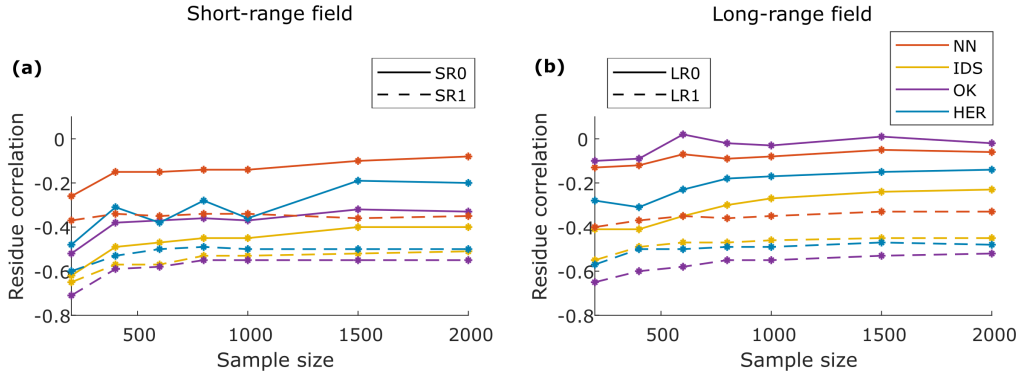

Figure B.3: Performance comparison of NN, IDS, OK and HER: **(a)** residue correlation for SR datasets and **(b)** residue correlation for LR datasets. Continuous line refers to datasets without noise and dashed lines to datasets with noise.

Table B.5: Summary statistics of the prediction on test set by model – short-range dataset without noise (SR0).

Method	Statistics	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

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

Table B.6: Summary statistics of the prediction on test set by model – short-range dataset with noise (SR1).

Method	Statistics	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

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

Table B.7: Summary statistics of the prediction on test set by model – long-range dataset without noise (LR0).

Method	Statistics	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

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

Table B.8: Summary statistics of the prediction on test set by model – long-range dataset with noise (LR1).

Method	Statistics	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

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

C

APPENDIX TO CHAPTER IV

C.1 MODEL PARAMETERS

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

OK

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

Table C.1: Parameters of OK fitted variograms as proposed by Atteia et al. (1994).

log_{10}(Pb)	Nugget	Sill	Range (km)
spherical model 1	0.0096	0.0228	0.287
spherical model 2	0.0131	–	2.605

HER

This section presents the spatial characterization of the lead dataset using HER (Fig. C.1) and the optimum weights obtained to be used in aggregation methods (Fig. C.2).

Fig. C.1a presents the raw infogram from where the class PMFs (Fig. C.1b) and, consecutively, the infogram (Fig. C.1c) were obtained. In Fig. C.1b, the Euclidean distance (in km) relative to the class is indicated after the class name in interval notation (left-open, right-closed interval) and, for brevity, only the odd classes are shown. The visual increasing of the spread of the Δz PMFs given the distance class (Fig. C.1b) is numerically verified also in the infogram (red curve, Fig. C.1c), which presents increasing entropy (therefore, decreasing spatial dependence or increasing spatial disorder) with distance. As shown in Fig. C.1c, the calculated range included 20 classes, reaching 1.4 km (circa three times smaller than the x-domain length of about 4 km). The range was identified as the point beyond which the class entropy exceeded the entropy of the full dataset (seen as the intersect of the blue and red-dotted lines).

The number of pairs forming each Δz PMF and the optimum weights (w_{OR} and w_{AND}) obtained for Eqs. 4.3 and 4.4, respectively, are illustrated in Fig. C.2. About 30% of the pairs (20 294 out of 66 822 pairs) are inside the range, where the first class has just under 500 pairs and the last class inside the range (light blue) has above 1500 pairs. Decreasing contribution of the weight with the distance is seen in Fig. C.2b, with strong influence of the first six classes (until about 0.4 km). Furthermore, the optimum

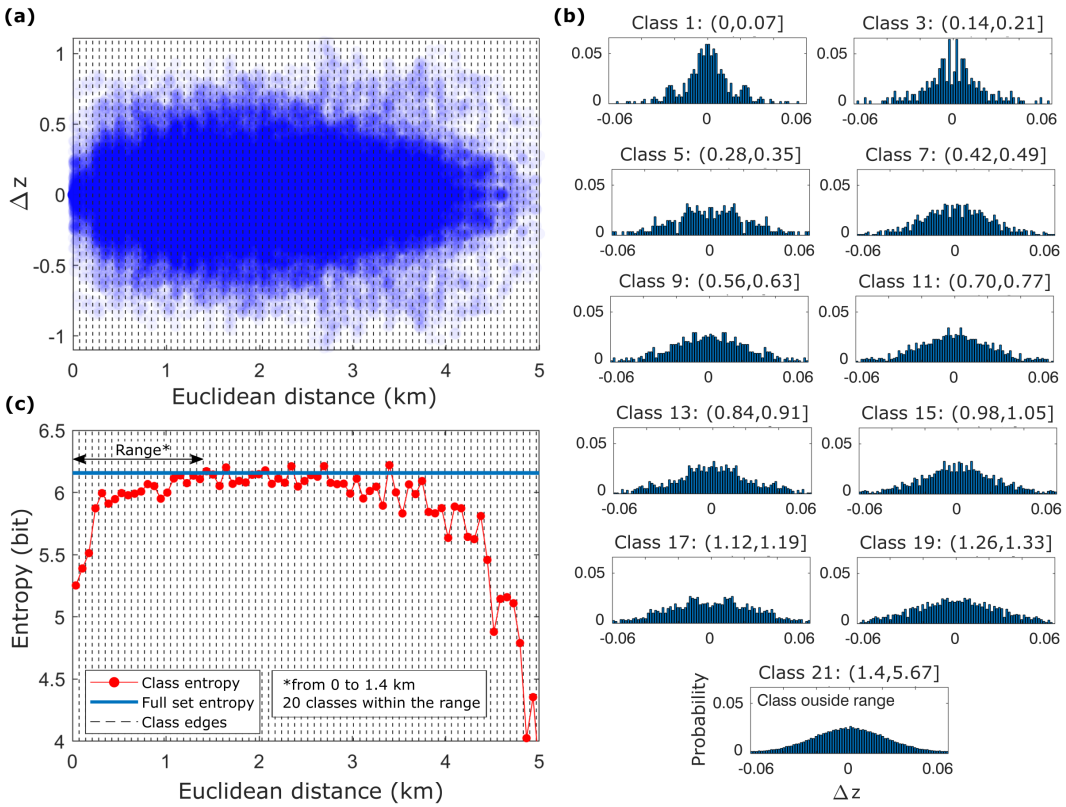

Figure C.1: Spatial characterization of the lead dataset using HER. **(a)** Infogram cloud, **(b)** Δz PMFs by class, and (c) infogram.

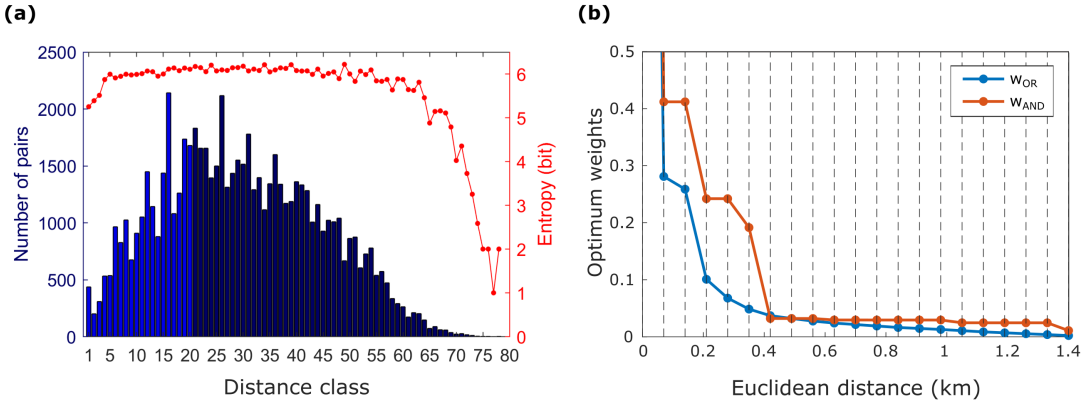

Figure C.2: HER model characteristics of the lead dataset. **(a)** Class cardinality and **(b)** optimum weights – Eqs. 4.3 and 4.4.

contribution of AND and OR aggregation, Eq. 4.5, for this model was $\alpha = 0.65$ and $\beta = 0$.

IK and IK$_{10}$

This section presents the parameters used in AUTO-IK program (developed by Goovaerts, 2009) to calibrate the indicator kriging model (called IK) for the paper

dataset. The parameter file employed is available Fig. C.3. The program AUTO-IK described in Goovaerts (2009) is available on his personal website sites.google.com/site/goovaertspierre/pierregoovaertswebsite/download/indicator-kriging.

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

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

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

Figure C.3: Parameter file used for geostatistical analysis of lead required by AUTO-IK.exe. Indicator semivariograms for thresholds corresponding to 68 equally spaced cutoffs plus z_c threshold, are computed using 30 lags of 0.07 km. The models are fitted automatically and used to perform full ordinary indicator kriging using up to the 32 closest observations located within a radius of 2 km.

Based on this IK model, the authors also generate a model using 10 cutoffs, of which nine are equally spaced p-quantiles of the sample histogram and one is the z_c threshold, i.e., [1.488, 1.543, 1.576, 1.619, 1.667, 1.699 (z_c), 1.709, 1.752, 1.816, 1.907]. The decision was based on (Goovaerts, 1997, p. 285), who recommends using z_c as a cutoff to avoid the later interpolation of its probability and argues that cutoff values beyond the first and ninth decile of the calibration set may be inappropriate, since they depend on the spatial distribution of a few pairs of points. In general, (Rossi and Deutsch, 2014, p. 160) also recommend between 8 and 15 cutoff values. Thus, due to its 10 cutoff values, this model is called IK_{10}.

HERs

For the sequential simulation model (HERs), we verified the quality of the reproduction of the realizations similarly to the work of Goovaerts (1997) and Leuangthong et al. (2005). The final optimum weights were practically the same as HER model, with the identical infogram and PMF of the classes of HER (as in Fig. C.1), same cardinality and similar w_{OR} and w_{AND} (as in Fig. C.2), $a = 0.55$ (intersecting PMFs), and

$\beta = 0$ (averaging PMFs). The small changes on the optimum weights (automatically obtained) happened since the number of neighbors used for HERs was set to seven (instead of 30 used for HER).

Although HER and HERs models resulted both in a pure intersection of PMFs (since we have just α contribution), the influence in the number of neighbors plays an important role when intersecting distributions and, therefore, we reduced it to seven in HERs. As explored in (Thiesen et al., 2020b), the higher the number of (similar) distributions to be intersected, the smaller the uncertainty of the resultant distribution. Consequently, due to the sequential procedure of HERs – in which for each iteration we artificially add an extra sample to the data to condition the next prediction – the number of distributions to be intersected greatly increase in relation to the validation set. Thus, to balance this decrease in the entropy (uncertainty), the authors have chosen to reduce the number of neighbors. This implementation decision (number of neighbors) was done by simply checking the infogram of each realization, until it was unbiased in relation to the sample set (Fig. 4.11a). This is how we also validate the model regarding ergodic fluctuations.

It is important to note that estimating entropy via a finite sample have the tendency to be underestimated (Darscheid, 2017). Therefore, considering the great discrepancy in the amount of data between the calibration set (259 observations) and realizations (grid with more than 10 000 targets), we introduced a bias in the realizations so that they could be compared to the calibration set (Fig. 4.11b). This was conducted by drawing 259 points from each realization (with no replacement), calculating their entropy, repeating it 1 000 times, and taking the mean of these repetitions. Although the bias of the calibration set could be estimated (as proposed by Darscheid, 2017; Steck and Jaakkola, 2004), a bias correction of the entropy of the calibration set is not straightforward since the obtained value is just a reference to bound the maximum bias and not its exact value. Conversely, adding a bias to the realizations allowed the comparison of the entropy of the calibration set and of the realizations.

Additionally, the authors verified the existence of connectivity of extremely high and small concentration values using indicator variograms for the deciles of 0.2 and 0.8 and different realizations (not shown). The results pointed out no destructuration effect (also known as maximum entropy property, (Goovaerts, 1997, p. 272, 393), e.g., for the realizations #42 and #94 (Fig. 4.12), due to the similarity of the indicator variogram of the calibration set and simulated realizations for the different deciles. Therefore, HERs present itself as an appropriate method for cases where extreme values are spatially correlated.

C.2 EXTRA RESULTS

This section consolidates extra results for the local uncertainty of OK, IK, IK_{10} and HER models. Fig. C.4 displays the entropy map of IK_{10}. It is noteworthy that the E-type, probability, and classification maps were not included for IK_{10} due to their similarity to the ones produced to the refined IK model.

Fig. C.5 displays the local results for the OK model, including estimation, entropy, probability and classification maps. Similar to Goovaerts (1997, p. 362), the estimation map of OK (Fig. C.5a), which is optimal for least-square criterion, tends to overestimate the Pb concentration, leading to most locations being classified as contaminated (Fig. C.5d). While the OK estimates (Fig. C.5a) and E-type estimates presented in

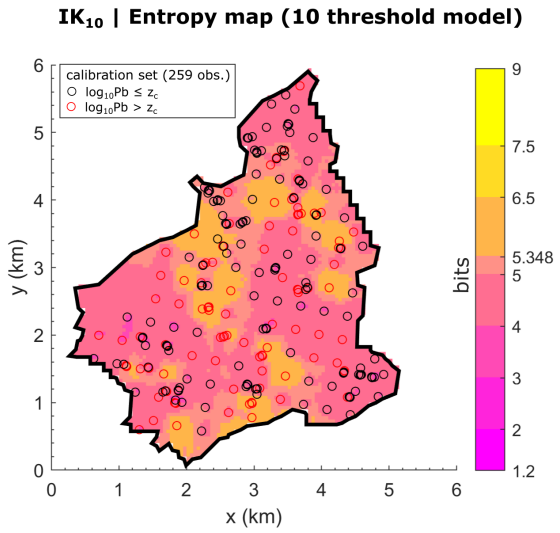

Figure C.4: Entropy map. Local uncertainty in terms of entropy for IK_{10}.

Table C.2: Cross-validation results for OK, IK, IK_{10}, and HER method.

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

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

the paper (Fig. 4.4) are similar, their uncertainty (Figs. C.5b and 4.5) are completely different. The map of OK entropy indicates greater uncertainty where data are sparse, whereas the uncertainty is smallest near data locations. Such effect is expected since OK ignores the observation values, retaining only the spatial geometry from the data (Goovaerts, 1997, p. 180).

The local distributions of IK, IK_{10}, and HER models are displayed in Fig. C.6. In this image, we can relate the bin-filling effect of the linear interpolation and extrapolation of the distribution assumed in IK_{10} with IK.

Table C.2 (performance results) and Fig. C.7 (accuracy and PI-width plots) contain information already presented in the paper, with the inclusion of IK_{10}.

The misclassification given different probability cutoffs is shown in Fig. C.8. Different than expected, all lead models (OK, IK, and HER) presented the minimum misclassification occurring close to the probability of 0.5 instead of the marginal probability of 0.421 (estimated in Sect. 4.3.1). This could be explained by the fact that the marginal probability was calculated on the calibration set and we are analyzing the models on the validation set, or by the fact that no declustering of the calibration data was done before calculating the marginal probability. Although, for all models, misclassification is not minimal at the marginal probability of 0.421, they have a similar monotonic tendency of decreasing its values until the minimum (at about 0.5).

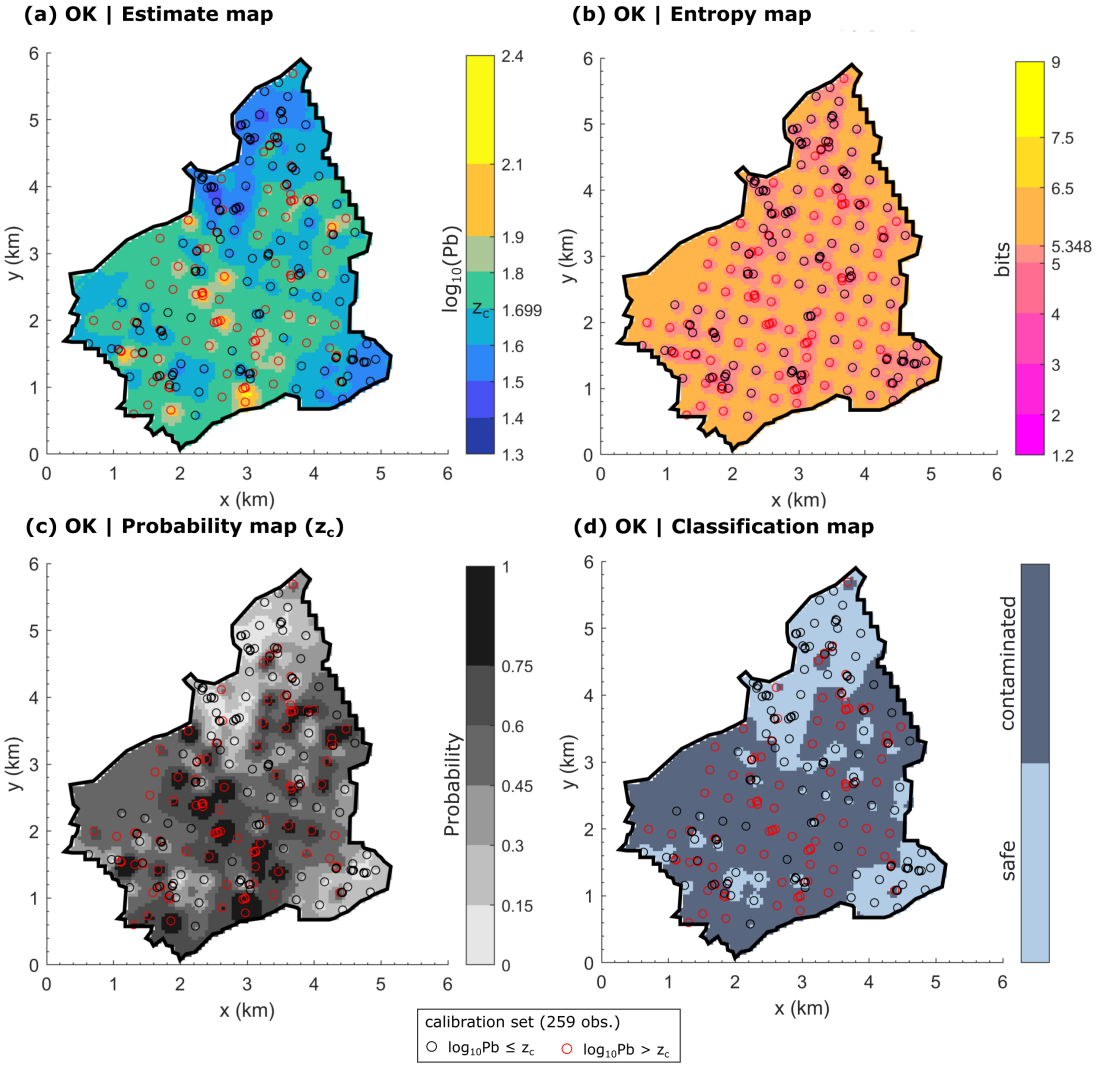

Figure C.5: OK maps for log_{10}(Pb) dataset. **(a)** Estimates, **(b)** local uncertainty in terms of information, **(c)** probability of exceeding the critical threshold (z_c= 1.699), and **(d)** classification of locations as contaminated by lead on the basis that the probability of exceeding the critical threshold z_c is larger than the marginal probability of contamination (0.421).

IK_{10} presented similar misclassification in comparison to IK, which was not plotted to avoid interference with the visualization.

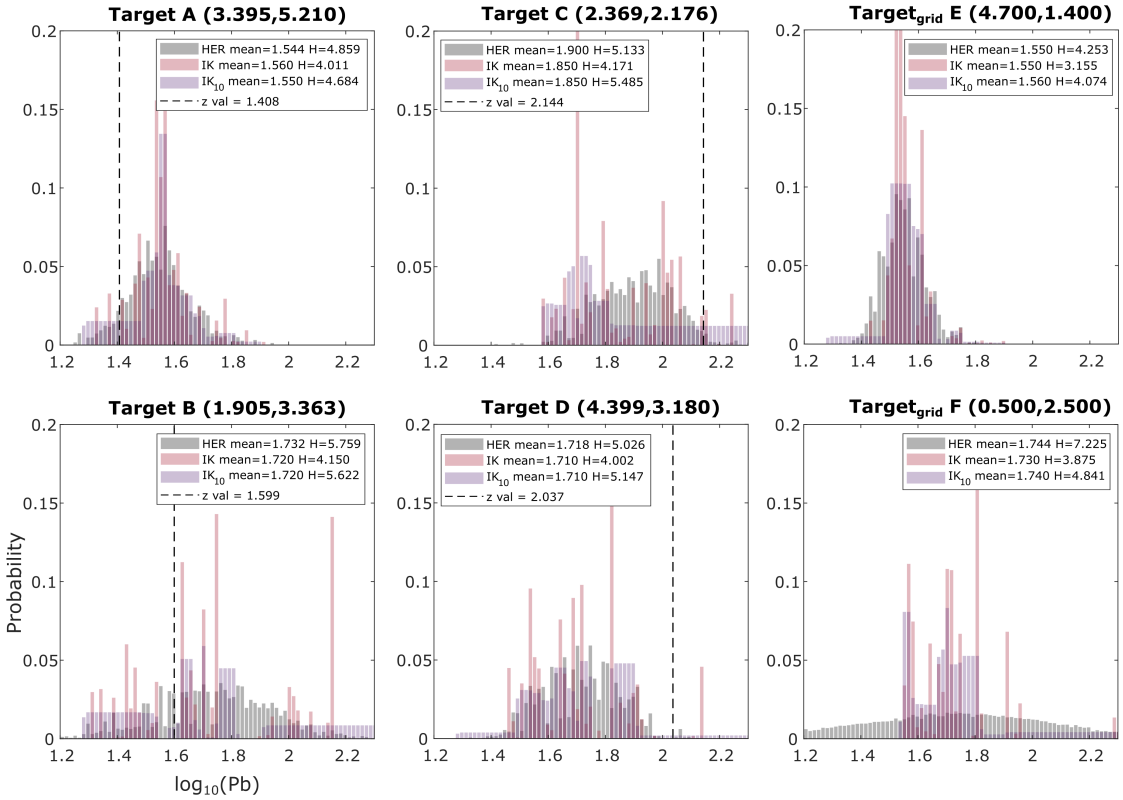

Figure C.6: Local distribution of targets of the validation set (targets A to D) and grid (targets E and F) for HER (gray), IK (red), and IK_{10} (purple).

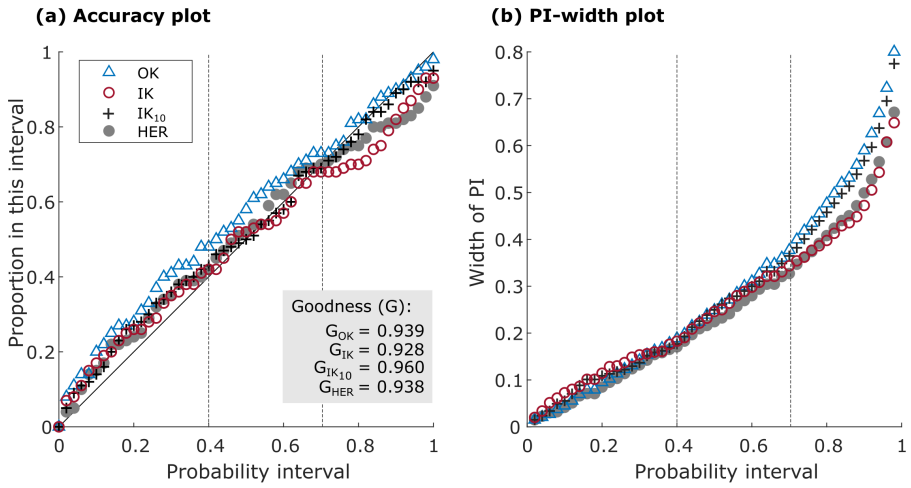

Figure C.7: OK, IK, IK_{10}, and HER performance. **(a)** Proportion of the true lead values falling within the probability intervals (p-PI) of increasing sizes and **(b)** the width of these intervals versus p-PI.

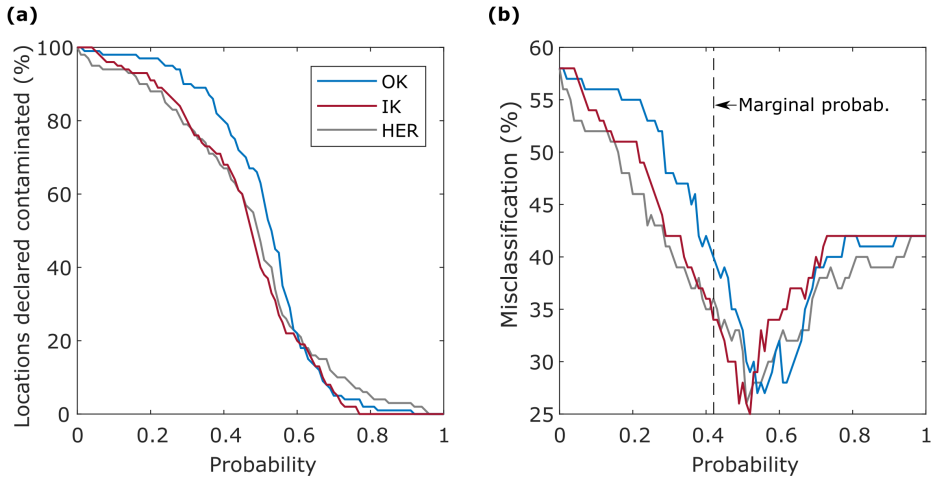

Figure C.8: Proportion of validation locations **(a)** that are declared contaminated with respect to lead concentration and **(b)** that are wrongly classified for OK, IK, and HER models.

BIBLIOGRAPHY

Allard, D., D. D'Or, and R. Froidevaux (2011). "An efficient maximum entropy approach for categorical variable prediction." In: *European Journal of Soil Science* 62.3, pp. 381–393. DOI: 10.1111/j.1365-2389.2011.01362.x.

Allard, D., A. Comunian, and P. Renard (2012). "Probability aggregation methods in geoscience." In: *Mathematical Geosciences* 44.5, pp. 545–581. DOI: 10.1007/s11004-012-9396-3.

Atteia, O., J.-P. Dubois, and R. Webster (1994). "Geostatistical analysis of soil contamination in the Swiss Jura." In: *Environmental Pollution* 86.3, pp. 315–327. DOI: 10.1016/0269-7491(94)90172-4.

Bandarian, E. M., U. A. Mueller, J. Fereira, and S. Richardson (2018). "Transformation methods for multivariate geostatistical simulation – Minimum/Maximum autocorrelation factors and alternating columns diagonal centres." In: *Advances in Applied Strategic Mine Planning*, pp. 371–394. DOI: 10.1007/978-3-319-69320-0_24.

Bárdossy, A. (2006). "Copula-based geostatistical models for groundwater quality parameters." In: *Water Resources Research* 42.11, pp. 1–12. DOI: 10.1029/2005WR004754.

Batty, M. (1974). "Spatial entropy." In: *Geographical Analysis* 6.1, pp. 1–31. DOI: https://doi.org/10.1111/j.1538-4632.1974.tb01014.x.

Bel, L., D. Allard, J. M. Laurent, R. Cheddadi, and A. Bar-Hen (2009). "CART algorithm for spatial data: application to environmental and ecological data." In: *Computational Statistics and Data Analysis* 53.8, pp. 3082–3093. DOI: 10.1016/j.csda.2008.09.012.

Bell, G., T. Hey, and A. Szalay (2009). "Computer science: beyond the data deluge." In: *Science* 323.5919, pp. 1297–1298. DOI: 10.1126/science.1170411.

Bellman, R (1957). *Dynamic programming*. Princeton.

Bianchi, M. and D. Pedretti (2018). "An entrogram-based approach to describe spatial heterogeneity with applications to solute transport in porous media." In: *Water Resources Research* 54.7, pp. 4432–4448. DOI: 10.1029/2018WR022827.

Blöschl, G. and M. Sivapalan (1995). "Scale issues in hydrological modelling: A review." In: *Hydrological Processes* 9.3-4, pp. 251–290. DOI: 10.1002/hyp.3360090305.

Blower, G. and J. E. Kelsall (2002). "Nonlinear kernel density estimation for binned data: Convergence in entropy." In: *Bernoulli* 8.4, pp. 423–449.

Blume, T., E. Zehe, and A. Bronstert (2007). "Rainfall-runoff response, event-based runoff coefficients and hydrograph separation." In: *Hydrological Sciences Journal* 52.5, pp. 843–862. DOI: 10.1623/hysj.52.5.843.

Bourennane, H., D. King, A. Couturier, B. Nicoullaud, B. Mary, and G. Richard (2007). "Uncertainty assessment of soil water content spatial patterns using geostatistical simulations: An empirical comparison of a simulation accounting for single attribute and a simulation accounting for secondary information." In: *Ecological Modelling* 205.3-4, pp. 323–335. DOI: 10.1016/j.ecolmodel.2007.02.034.

Branch, M. A., T. F. Coleman, and Y. Li (1999). "A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization prob-

lems." In: *SIAM Journal on Scientific Computing* 21.1, pp. 1–23. DOI: 10.1137/S1064827595289108.

Brunsell, N. A. (2010). "A multiscale information theory approach to assess spatial-temporal variability of daily precipitation." In: *Journal of Hydrology* 385.1-4, pp. 165–172. DOI: 10.1016/j.jhydrol.2010.02.016.

Chapman, T. G. (1986). "Entropy as a measure of hydrologic data uncertainty and model performance." In: *Journal of Hydrology* 85.1-2, pp. 111–126. DOI: 10.1016/0022-1694(86)90079-X.

Chicco, D. (2017). "Ten quick tips for machine learning in computational biology." In: *BioData Mining* 10.1, pp. 1–17. DOI: 10.1186/s13040-017-0155-3.

Chow, V. T., D. R. Maidment, and L. W. Mays (1988). *Applied hydrology*. New York: McGraw-Hill.

Cover, T. M. and J. A. Thomas (2006). *Elements of information theory*. 2nd ed. New Jersey: John Wiley & Sons, pp. 1–748. DOI: 10.1002/047174882X. arXiv: ISBN0-471-06259-6.

Dabo-Niang, S., C. Ternynck, and A. F. Yao (2016). "Nonparametric prediction of spatial multivariate data." In: *Journal of Nonparametric Statistics* 28.2, pp. 428–458. DOI: 10.1080/10485252.2016.1164313.

Darbellay, G. A. and I. Vajda (1999). "Estimation of the information by an adaptive partitioning of the observation space." In: *IEEE Transactions on Information Theory* 45.1, pp. 1315–1321. DOI: 10.1109/18.761290.

Darscheid, P. (2017). "Quantitative analysis of information flow in hydrological modelling using Shannon information measures." Master thesis. Karlsruhe Institute of Technology, p. 73.

Darscheid, P., A. Guthke, and U. Ehret (2018). "A maximum-entropy method to estimate discrete distributions from samples ensuring nonzero probabilities." In: *Entropy* 20.8, p. 601. DOI: 10.3390/e20080601.

Deutsch, C. V. (1997). "Direct assessment of local accuracy and precision." In: *Geostatistics Wollongong '96*. Ed. by E. Baafi and N. Schofield. Vol. 1. Kluwer Academic Publishing, pp. 115–125.

Deutsch, C. V. and A. G. Journel (1998). *GSLIB: Geostatistical software library and user's guide*. New York: Oxford University Press, p. 369.

Eckhardt, K. (2005). "How to construct recursive digital filters for baseflow separation." In: *Hydrological Processes* 19.2, pp. 507–515. DOI: 10.1002/hyp.5675.

Ehret, U. and E. Zehe (2011). "Series distance – an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events." In: *Hydrology and Earth System Sciences* 15.3, pp. 877–896. DOI: 10.5194/hess-15-877-2011.

Ehret, U., P. Darscheid, G. Nearing, and H. Gupta (2018). "An information perspective on hydrological learning and prediction." In: *EGU General Assembly 2018*. Vol. 20, 8–13 Apr 2018, EGU2018-13836-1. Vienna/Austria.

Ehret, U., R. V. Pruijssen, M. Bortoli, R. Loritz, E. Azmi, and E. Zehe (2020). "Adaptive clustering: reducing the computational costs of distributed (hydrological) modelling by exploiting time-variable similarity among model elements." In: *Hydrol. Earth Syst. Sci.*, pp. 4389–4411.

Fawcett, T. (2005). "An introduction to ROC analysis Tom." In: *Irbm* 35.6, pp. 299–309. DOI: 10.1016/j.patrec.2005.10.010. arXiv: /dx.doi.org/10.1016/j.patrec.200 [http:].

Fernández-Casal, R., S. Castillo-Páez, and M. Francisco-Fernández (2018). "Nonparametric geostatistical risk mapping." In: *Stochastic Environmental Research and Risk Assessment* 32.3, pp. 675–684. DOI: 10.1007/s00477-017-1407-y.

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

Fix, E and J. L. Hodges Jr. (1987). *Ordinance relating to pollutants in soil (VSBo of June 9, 1986)*. Tech. rep. Bern, Switzerland: Swiss Federal Office of Environment, Forest and Landscape.

Gneiting, T. and A. E. Raftery (2007). "Strictly proper scoring rules, prediction, and estimation." In: 102.477, pp. 359–378. DOI: 10.1198/016214506000001437.

Gómez-Hernández, J. J. and E. F. Cassiraga (1994). "Theory and practice of sequential simulation." In: *Geostatistical Simulations*. Ed. by M. Armstrong and P. A. Dowd. Kluwer Academic Publishers, pp. 111–124. DOI: doi:10.1007/978-94-015-8267-4_10.

Gómez-Hernández, J. J. and X. H. Wen (1998). "To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology." In: *Advances in Water Resources* 21.1, pp. 47–61. DOI: 10.1016/S0309-1708(96)00031-0.

Gong, W., D. Yang, H. V. Gupta, and G. Nearing (2014). "Estimating information entropy for hydrological data: one dimensional case." In: *Water Resources Research* 1, pp. 5003–5018. DOI: 10.1002/2014WR015874.Received.

Good, I. J. (1952). "Rational decisions." In: *Journal of the Royal Statistical Society* 14.1, pp. 107–114.

Goodman, N. D., T. D. Ullman, and J. B. Tenenbaum (2008). "Learning a theory of causality." In: *Psychological Review* 118, pp. 110–119.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford Uni. New York, p. 483.

Goovaerts, P. (1998). "Geostatistics in soil science: State-of-the-art and perspectives." In: *Geoderma* 89.1-2, pp. 1–45. DOI: 10.1016/S0016-7061(98)00078-0.

Goovaerts, P. (1999). "Impact of the simulation algorithm, magnitude of ergodic fluctuations and number of realizations on the spaces of uncertainty of flow properties." In: *Stochastic Environmental Research and Risk Assessment* 13.3, pp. 161–182. DOI: 10.1007/s004770050037.

Goovaerts, P. (2001). "Geostatistical modelling of uncertainty in soil science." In: *Geoderma* 103.1-2, pp. 3–26. DOI: 10.1016/S0016-7061(01)00067-2.

Goovaerts, P. (2009). "AUTO-IK: a 2D indicator kriging program for the automated non- parametric modeling of local uncertainty in earth sciences." In: *Comput Geosci.* 23.1, pp. 1–7. DOI: 10.1038/jid.2014.371.

Goovaerts, P., R. Webster, and J. P. Dubois (1997). "Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics." In: *Environmental and Ecological Statistics* 4.1, pp. 31–48.

Habibzadeh, F., P. Habibzadeh, and M. Yadollahie (2016). "On determining the most appropriate test cut-off value: The case of tests with continuous results." In: *Biochemia Medica* 26.3, pp. 297–307. DOI: 10.11613/BM.2016.034.

Hall, F. R. (1968). "Base-flow recessions – a review." In: *Water Resour. Res.* 4.973-983.

Horton, R. E. (1933). "The role of infiltration in the hydrologic cycle." In: *Trans. Am. Geophys. Union* 14, pp. 446–460.

Hoyt, W. G. (1936). *Studies of relations of rainfall and run-off in the United States*. Tech. rep. Washington: Geol. Surv. of U.S., Water-Supply Paper 772, p. 301. DOI: 10.3133/wsp772.

Hristopulos, D. T. and A. Baxevani (2020). "Effective probability distribution approximation for the reconstruction of missing data." In: *Stochastic Environmental Research and Risk Assessment* 34.2, pp. 235–249. DOI: 10.1007/s00477-020-01765-5.

Journel, A. G. (1974). "Geostatistics for conditional simulation of ore bodies." In: *Economic Geology* 69.5, pp. 673–687. DOI: 10.2113/gsecongeo.69.5.673.

Journel, A. G. (1983). "Nonparametric estimation of spatial distributions." In: *Journal of the International Association for Mathematical Geology* 15.3, pp. 445–468. DOI: 10.1007/BF01031292.

Journel, A. G. (1989). *Fundamentals of Geostatistics in Five Lessons*. Ed. by M. L. Crawford and E. Padovani. Vol. 8. Washington, D.C.: American Geophysical Union. DOI: 10.1029/sc008.

Journel, A. G. (2002). "Combining knowledge from diverse sources: an alternative to traditional data independence hypotheses." In: *Mathematical Geology* 34.5, pp. 573–596. DOI: 10.1023/A:1016047012594.

Journel, A. G. (2003). "Multiple-point geostatistics: a state of the art." In: *Stanford Center for Reservoir Forecasting*, pp. 1–52.

Journel, A. G. and C. J. Huijbregts (1978). *Mining geostatistics*. London, UK: Academic Press, p. 610.

Journel, A. G. and W. Xu (1994). "Posterior identification of histograms conditional to local data." In: *Mathematical Geology* 26.3. DOI: https://doi.org/10.1007/BF02089228.

Kazianka, H. and J. Pilz (2010). "Spatial interpolation using copula-based geostatistical models." In: *geoENV VII – Geostatistics for Environmental Applications*, pp. 307–319. DOI: 10.1007/978-90-481-2322-3.

Kitanidis, P. K. (1997). *Introduction to geostatistics: applications in hydrogeology*. Cambridge, United Kingdom: Cambridge University Press, p. 249.

Klemeš, V. (1983). "Conceptualization and scale in hydrology." In: *Journal of Hydrology* 65.1-3, pp. 1–23. DOI: 10.1016/0022-1694(83)90208-1.

Knuth, K. H. (2013). "Optimal data-based binning for histograms." In: p. 30. arXiv: 0605197v2 [physics].

Koskelo, A. I., T. R. Fisher, R. M. Utz, and T. E. Jordan (2012). "A new precipitation-based method of baseflow separation and event identification for small watersheds (<50km2)." In: *Journal of Hydrology* 450-451, pp. 267–278. DOI: 10.1016/j.jhydrol.2012.04.055.

Krige, D. G. (1951). "A statistical approach to some mine valuation and allied problems on the Witwatersrand." Master's thesis. University of Witwatersrand, p. 136.

Krishnan, S. (2008). "The tau model for data redundancy and information combination in earth sciences: theory and application." In: *Mathematical Geosciences* 40.6, pp. 705–727. DOI: 10.1007/s11004-008-9165-5.

Larson, J. W. (2010). "Can we define climate using information theory?" In: *IOP Conference Series: Earth and Environmental Science* 11, p. 012028. DOI: 10.1088/1755-1315/11/1/012028.

Leopold, L. B. and W. B. Langbein (1962). *The concept of entropy in landscape evolution*. Washington.

Leuangthong, O., J. A. McLennan, and C. V. Deutsch (2005). "Acceptable ergodic fluctuations and simulation of skewed distributions." In: *Application of Computers and Operations Research in the Mineral Industry - Proc. of the 32nd Int. Symposium on the Application of Computers and Operations Research in the Mineral Industry, APCOM 2005* c, pp. 211–218. DOI: 10.1201/9781439833407.ch27.

Leuangthong, O., J. A. McLennan, and C. V. Deutsch (2004). "Minimum acceptance criteria for geostatistical realizations." In: *Natural Resources Research* 13.3, pp. 131–141. DOI: 10.1023/B:NARR.0000046916.91703.bb.

Li, J. and A. D. Heap (2008). "A review of spatial interpolation methods for environmental scientists." In: *Geoscience Australia* 2008/23, p. 137.

Li, J. and A. D. Heap (2011). "A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors." In: *Ecological Informatics* 6.3-4, pp. 228–241. DOI: 10.1016/j.ecoinf.2010.12.003.

Li, J. and A. D. Heap (2014). "Spatial interpolation methods applied in the environmental sciences: A review." In: *Environmental Modelling and Software* 53, pp. 173–189. DOI: 10.1016/j.envsoft.2013.12.008.

Liu, D. et al. (2016). "Entropy of hydrological systems under small samples: uncertainty and variability." In: *Journal of Hydrology* 532, pp. 163–176. DOI: 10.1016/j.jhydrol.2015.11.019.

Loquin, K. and D. Dubois (2010). "Kriging and epistemic uncertainty: A critical discussion." In: *Studies in Fuzziness and Soft Computing* 256, pp. 269–305. DOI: 10.1007/978-3-642-14755-5_11.

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

Loritz, R., A. Kleidon, C. Jackisch, M. Westhoff, U. Ehret, H. Gupta, and E. Zehe (2019). "A topographic index explaining hydrological similarity by accounting for the joint controls of runoff formation." In: *Hydrology and Earth System Sciences Discussions* March, pp. 1–22. DOI: 10.5194/hess-2019-68.

Mälicke, M and H. D. Schneider (2019). *Scikit-GStat 0.2.6: A scipy flavored geostatistical analysis toolbox written in Python*. DOI: 10.5281/zenodo.3531816.

Mälicke, M., S. Hassler, T. Blume, M. Weiler, and E. Zehe (2020). "Soil moisture: variable in space but redundant in time." In: *Hydrology and Earth System Sciences Discussions*, pp. 1–28. DOI: 10.5194/hess-2019-574.

Manchuk, J. G. and C. V. Deutsch (2007). "Robust solution of normal (kriging) equations." In: p. 10.

Mei, Y. and E. N. Anagnostou (2015). "A hydrograph separation method based on information from rainfall and runoff records." In: *Journal of Hydrology* 523, pp. 636–649. DOI: 10.1016/j.jhydrol.2015.01.083.

Merz, R. and G. Blöschl (2009). "A regional analysis of event runoff coefficients with respect to climate and catchment characteristics in Austria." In: *Water Resources Research* 45.1, pp. 1–19. DOI: 10.1029/2008WR007163.

Merz, R., G. Blöschl, and J. Parajka (2006a). "Regionalization methods in rainfall-runoff modelling using large catchment samples." In: *IAHS-AISH Publication* 307, pp. 117–125.

Merz, R., G. Blöschl, and J. Parajka (2006b). "Spatio-temporal variability of event runoff coefficients." In: *Journal of Hydrology* 331.3-4, pp. 591–604. DOI: 10.1016/j.jhydrol.2006.06.008.

Metropolis, N. and S. Ulam (1949). "The Monte Carlo method." In: *Journal of the American Statistical Association* 44.247, pp. 335–341.

Mishra, A. K., M. Özger, and V. P. Singh (2009). "An entropy-based investigation into the variability of precipitation." In: *Journal of Hydrology* 370.1-4, pp. 139–154. DOI: 10.1016/j.jhydrol.2009.03.006.

Myers, D. E. (1993). "Spatial interpolation: an overview." In: *Geoderma* 62.1-3, pp. 17–28. DOI: 10.1016/0016-7061(94)90025-6.

Naimi, B. (2015). "On uncertainty in species distribution modelling." PhD thesis. University of Twente. DOI: 10.3990/1.9789036538404.

Nearing, G. S. and H. V. Gupta (2017). "Information vs. uncertainty as the foundation for a science of environmental modeling." In: *eprint arXiv:1704.07512*, pp. 1–23. arXiv: 1704.07512.

Neuper, M. and U. Ehret (2019). "Quantitative precipitation estimation with weather radar using a data- and information-based approach." In: *Hydrology and Earth System Sciences* 23.9, pp. 3711–3733. DOI: 10.5194/hess-23-3711-2019.

Oliver, M. A. and R. Webster (2014). "A tutorial guide to geostatistics: Computing and modelling variograms and kriging." In: *Catena* 113, pp. 56–69. DOI: 10.1016/j.catena.2013.09.006.

Ortiz, J. C., O. Leuangthong, and C. V. Deutsch (2004). "A multiGaussian approach to assess block grade uncertainty." In: *Center for Computational Geostatistics Annual Report Papers*, pp. 1–12.

Pechlivanidis, I. G., B. Jackson, H. Mcmillan, and H. V. Gupta (2016). "Robust informational entropy-based descriptors of flow in catchment hydrology." In: *Hydrological Sciences Journal* 61.1, pp. 1–18. DOI: 10.1080/02626667.2014.983516.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in Python." In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Perdigão, R. A., U. Ehret, K. H. Knuth, and J. Wang (2020). "Debates: Does Information Theory Provide a New Paradigm for Earth Science? Emerging Concepts and Pathways of Information Physics." In: *Water Resources Research* 56.2, pp. 1–13. DOI: 10.1029/2019WR025270.

Pham, T. D. (2010). "GeoEntropy: A measure of complexity and similarity." In: *Pattern Recognition* 43.3, pp. 887–896. DOI: 10.1016/j.patcog.2009.08.015.

Putter, H and G. A. Young (2001). "On the effect of covariance function estimation on the accuracy of kriging predictors." In: *Bernoulli* 7.3, pp. 421–438.

Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian processes for machine learning*. The MIT Press, p. 248.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat (2019). "Deep learning and process understanding for data-driven Earth system science." In: *Nature* 566.7743, pp. 195–204. DOI: 10.1038/s41586-019-0912-1.

Roodposhti, M. S., J. Aryal, H. Shahabi, and T. Safarrad (2016). "Fuzzy Shannon entropy: a hybrid GIS-based landslide susceptibility mapping method." In: *Entropy* 18.10. DOI: 10.3390/e18100343.

Rossi, M. E. and C. V. Deutsch (2014). *Mineral resource estimation*. London: Springer. DOI: https://doi.org/10.1007/978-1-4020-5717-5.

Roulston, M. S. and L. A. Smith (2002). "Evaluating probabilistic forecasts using information theory." In: *Monthly Weather Review* 130.6, pp. 1653–1660. DOI: 10.1175/1520-0493(2002)130⟨1653:EPFUIT⟩2.0.CO;2.

Ruddell, B. L. and P. Kumar (2009). "Ecohydrologic process networks: 1. Identification." In: *Water Resources Research* 45.3, pp. 1–23. DOI: 10.1029/2008WR007279.

Savelyeva, E., S. Utkin, S. Kazakov, and V. Demyanov (2010). "Modeling spatial uncertainty for locally uncertain data." In: *geoENVVII – Geostatistics for Environmental Applications*. July. Springer, pp. 295–306. DOI: 10.1007/978-90-481-2322-3_26.

Scott, D. W. (1979). "Scott bin width." In: *Biometrika* 66.3, pp. 605–610. DOI: 10.1093/biomet/66.3.605.

Seibert, S. P., U. Ehret, and E. Zehe (2016). "Disentangling timing and amplitude errors in streamflow simulations." In: *Hydrology and Earth System Sciences* 20.9, pp. 3745–3763. DOI: 10.5194/hess-20-3745-2016.

Shannon, C. E. (1948). "A mathematical theory of communication." In: *The Bell System Technical Journal* 27, pp. 379–423.

Sharma, A. and R. Mehrotra (2014). "An information theoretic alternative to model a natural system using observational information alone." In: *Water Resources Research* 50.1, pp. 650–660. DOI: 10.1002/2013WR013845.

Shepard, D. (1968). "A two-dimensional interpolation function for irregularly-spaced data." In: *Proceedings of the 1968 23rd ACM National Conference*, pp. 517–524. DOI: 10.1145/800186.810616.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. Springer-Verlag New York, pp. 1–656. DOI: 10.1007/978-1-4612-4026-6.

Singh, V. P. (2013). *Entropy theory and its application in environmental and water engineering*. first edit. West Sussex: John Wiley & Sons, Ltd, pp. 1–642. DOI: 10.1002/9781118428306.

Singh, V. P. (2018). "Hydrologic modeling: progress and future directions." In: *Geoscience Letters* 5.1. DOI: 10.1186/s40562-018-0113-z.

Solomatine, D, L. M. See, and R. J. Abrahart (2009). "Data-driven modelling: concepts, approaches and experiences." In: *Practical hydroinformatics*. Ed. by Springer. Berlin, Heidelberg, pp. 17–31.

Solomatine, D. P. and A. Ostfeld (2008). "Data-driven modelling: some past experiences and new approaches." In: *J. Hydroinform.* 10.1, pp. 3–22. DOI: 10.2166/hydro.2008.015.

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

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. Philadelphia: Seam, p. 342.

Tarantola, A. and B. Valette (1982). "Inverse problems = quest for information." In: *Journal of Geophysics* 50, pp. 159–170.

Thiesen, S., P. Darscheid, and U. Ehret (2018). *Rainfall-runoff events identification using Information Theory (ITM method)*. DOI: 10.5281/zenodo.1404637.

Thiesen, S., P. Darscheid, and U. Ehret (2019). "Identifying rainfall-runoff events in discharge time series: a data-driven method based on Information Theory." In: *Hydrol. Earth Syst. Sci.* 23.2, pp. 1015–1034. DOI: https://doi.org/10.5194/hess-23-1015-2019.

Thiesen, S. and U. Ehret (2021). "Assessing local and spatial uncertainty with non-parametric geostatistics." In: *Stochastic Environmental Research and Risk Assessment*, p. 27. DOI: 10.1007/s00477-021-02038-5.

Thiesen, S., D. M. Vieira, and U. Ehret (2020a). *Spatial interpolation with histogram via entropy reduction (HER method)*. DOI: 10.5281/zenodo.3614718.

Thiesen, S., D. M. Vieira, and U. Ehret (2021). *HER adaptation for categorical data, probability maps and spatial simulation*. DOI: 10.5281/zenodo.4501328.

Thiesen, S., D. M. Vieira, M. Mälicke, R. Loritz, J. F. Wellmann, and U. Ehret (2020b). "Histogram via entropy reduction (HER): an information-theoretic alternative for geostatistics." In: *Hydrol. Earth Syst. Sci.* 24.9, pp. 4523–4540. DOI: https://doi.org/10.5194/hess-24-4523-2020.

Tobler, W. R. (1970). "A computer movie simulating urban growth in the Detroit region." In: *Economic Geography* 46, p. 234. DOI: 10.2307/143141.

Webster, R., O. Atteia, and J. P. Dubois (1994). "Coregionalization of trace metals in the soil in the Swiss Jura." In: *European Journal of Soil Science* 45.2, pp. 205–218. DOI: 10.1111/j.1365-2389.1994.tb00502.x.

Weijs, S. V., R. van Nooijen, and N. van de Giesen (2010). "Kullback–Leibler divergence as a forecast skill score with classic reliability– resolution– uncertainty decomposition." In: *Monthly Weather Review* 138.9, pp. 3387–3399. DOI: 10.1175/2010mwr3229.1.

Weijs, S. V. (2011). "Information theory for risk-based water system operation." PhD thesis. Technische Universiteit Delft, p. 210.

Wellmann, J. F. (2013). "Information theory for correlation analysis and estimation of uncertainty reduction in maps and models." In: *Entropy* 15, pp. 1464–1485. DOI: 10.3390/e15041464.

Yakowitz, S. J. and F. Szidarovszky (1985). "A comparison of kriging with nonparametric regression methods." In: *Journal of Multivariate Analysis* 16, pp. 21–53.

AUTHOR CONTRIBUTION AND CODE AVAILABILITY

Chapter 2 Thiesen, Stephanie; Darscheid, Paul; Ehret, Uwe (2019): Identifying rainfall-runoff events in discharge time series – a data-driven method based on Information Theory, Hydrology and Earth System Sciences, 23(2), 1015-1034. doi: 10.5194/hess-23-1015-2019

UE and PD developed the model program (calculation of information theory measures, multivariate histograms operations, event detection) and developed a method to avoid infinitely large values of D_{KL} (Darscheid et al., 2018). ST performed the simulations, cross-validation, parameter optimization, comparative analysis to a second model and the justification of the number of repetitions of the resampling stage. 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.

The Event Detection program, containing the functions to develop multivariate histograms and calculate information theory measures, is published alongside this manuscript via GitHub github.com/KIT-HYD/EventDetection (Thiesen et al., 2018). The repository also includes scripts to exemplify the use of the functions and the dataset of identified event, discharge and precipitation time series from the Dornbirn-erach catchment in Austria used in the case study.

Chapter 3: Thiesen, Stephanie; Vieira, Diego M.; Mälicke, Mirko; Loritz, Ralf; Wellmann, J. Florian; Ehret, Uwe (2020): Histogram via entropy reduction (HER) – an information-theoretic alternative for geostatistics, Hydrology and Earth System Sciences, 24(9), 4523-4540. doi: 10.5194/hess-24-4523-2020

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

The source code for an implementation of HER, containing spatial characterization, convex optimization and distribution prediction is published alongside this

manuscript via GitHub at github.com/KIT-HYD/HER (Thiesen et al., 2020a). The repository also includes scripts to exemplify the use of the functions and the dataset used in the case study. The synthetic field generator using Gaussian process is available in scikit-learn (Pedregosa et al., 2011), while the code producing the fields can be found at github.com/mmaelicke/random_fields.

Chapter 4: Thiesen, Stephanie; Ehret, Uwe (2021): Assessing local and spatial uncertainty with nonparametric geostatistics. Stochastic Environmental Research and Risk Assessment. doi:10.1007/s00477-021-02038-5

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

The source code of the adapted version of HER and its sequential simulation HERs, containing spatial characterization, convex optimization and distribution prediction, is published alongside this manuscript via GitHub at github.com/KIT-HYD/HERs (Thiesen et al., 2021). The repository also includes scripts to exemplify the use of the functions and the dataset used in the case study. The Jura dataset and AUTO-IK (Goovaerts, 2009) script were obtained directly on Goovaert's personal website, namely sites.google.com/site/goovaertspierre/pierregoovaertswebsite/download/, options 'Jura Data' and 'Automatic Indicator Kriging Program (AUTO-IK)'.

OWN PUBLICATIONS

PEER-REVIEWED INTERNATIONAL PUBLICATIONS

Thiesen, Stephanie; Ehret, Uwe (2021): *Assessing local and spatial uncertainty with nonparametric geostatistics. Stochastic Environmental Research and Risk Assessment*. doi:10.1007/s00477-021-02038-5

Thiesen, Stephanie; Vieira, Diego M.; Mälicke, Mirko; Loritz, Ralf; Wellmann, J. Florian; Ehret, Uwe (2020): *Histogram via entropy reduction (HER) – an information-theoretic alternative for geostatistics*, Hydrology and Earth System Sciences, 24(9), 4523-4540. doi: 10.5194/hess-24-4523-2020

Thiesen, Stephanie; Darscheid, Paul; Ehret, Uwe (2019): *Identifying rainfall-runoff events in discharge time series – a data-driven method based on Information Theory*, Hydrology and Earth System Sciences, 23(2), 1015-1034. doi: 10.5194/hess-23-1015-2019

COMPUTER PROGRAMS

Thiesen, Stephanie; Vieira, Diego M.; Ehret, Uwe (2021): *HER adaptation for categorical data, probability maps and spatial simulation*, version v1.0, available at: github.com/KIT-HYD/HERs. Zenodo, doi: 10.5281/zenodo.4501328. (Matlab)

Thiesen, Stephanie; Vieira, Diego M.; Ehret, Uwe (2020): *Spatial interpolation with Histogram via entropy reduction (HER)*, version v1.4, available at: github.com/KIT-HYD/HER. Zenodo, doi: 10.5281/zenodo.3614718. (Matlab)

Thiesen, Stephanie; Darscheid, Paul; Ehret, Uwe (2018): *Rainfall-runoff events identification using Information Theory*, version v1.0, available at: github.com/KIT-HYD/EventDetection-ITM-method. Zenodo, doi: 10.5281/zenodo.1404637. (Matlab)

CONFERENCE CONTRIBUTIONS

(Posters)

Thiesen, Stephanie; Ehret, Uwe (2021): Assessing local and spatial uncertainty with HER method, EGU21-568. *In: EGU General Assembly 2021 – EGU 2021*, 19–30 April 2021, online. PICO presentation & Online discussion. doi: 10.5194/egusphere-egu21-568

Thiesen, Stephanie; Ehret, Uwe (2021): Assessing local and spatial uncertainty with non-parametric geostatistics. *In: 13th International Conference on Geostatistics for Environmental Applications Proceedings – geoENV2020*, 18 June 2021, online. Abstract in book of proceedings. Available at 2020.geoenvia.org/wp-content/uploads/sites/6/

2021/04/Confirmed_for_website.pdf

Thiesen, Stephanie; Vieira, Diego M.; Mälicke, Mirko; Loritz, Ralf; Wellmann, J. Florian; Ehret, Uwe (2020): HER – an information theoretic alternative for geostatistics, EGU2020-1355. *In: EGU General Assembly 2020 – EGU 2020*, 4–8 May 2020, online. Poster & Online discussion. doi: 10.5194/egusphere-egu2020-1355

Thiesen, Stephanie; Vieira, Diego M.; Mälicke, Mirko; Loritz, Ralf; Wellmann, J. Florian; Ehret, Uwe (2020): HER – an information theoretic alternative for geostatistics. *In: Global Young Scientists Summit – GYSS 2020*, 14-17 January 2020, Singapore/Singapore. Poster.

Thiesen, Stephanie; Darscheid, Paul; Ehret, Uwe (2019). Identifying rainfall-runoff events in discharge time series: A data-driven method based on information theory. *In: Tag der Hydrologie*, 28-29 March 2019, Karlsruhe/Germany. Poster.

Thiesen, Stephanie; Darscheid, Paul; Ehret, Uwe (2018): Identifying rainfall-runoff events in discharge time series – a data-driven method based on information theory. *In: II Workshop on Information Theory & Earth Sciences – SWITES 2018*, 16–19 May 2028, Santander/Spain. Poster.

(Papers)

Thiesen, Stephanie; Geraldi, Matheus Soares; Kaestner, Camile Luana (2017): Simulador online para dimensionamento otimizado de reservatório de água da chuva associado à economia financeira. [Online simulator for optimized design of rainwater reservoir associated with the financial economy]. *In: Simpósio Brasileiro de Recursos Hídricos, Florianópolis/Brazil*. Paper.

Thiesen, Stephanie; Santos, Juliana Vieira; Higashi, Rafael Augusto dos Reis (2015): Application of GIS tools for geotechnical Mapping - a case study in Brazil. *In: International Conference on Geotechnical Engineering*, Colombo/Sri Lanka. Paper & Oral presentation.

Santos, Juliana Vieira; **Thiesen, Stephanie**; Higashi, Rafael Augusto dos Reis (2015): Geographic Information System: Methodological proposal for the development of foundation maps based on SPT investigation. *In: 15th Pan-American Conference on Soil Mechanics and Geotechnical Engineering*, Buenos Aires/Argentina. Paper.

Thiesen, Stephanie; Vieira, Diego Machado (2013): Análise e otimização por programação linear da distribuição de bicicletas na região universitária de Florianópolis/SC. [Analysis and optimization by linear programming of the bicycle distribution in the university zone of Florianópolis/SC]. *In: XXVII ANPET – Congresso Nacional de Ensino e Pesquisa em Transportes*. ANPET, Belém/Brazil. Paper & Oral presentation.

BOOK CHAPTERS

Barletta, Rodrigo; **Thiesen, Stephanie**; Nicolau, Ana Paula; Bonanata, Rafael; Noronha, Marcos (2019): Análise hidrodinâmica e de transporte de sedimentos para elaboração de alternativas de recuperação da praia da Armação do Pântano do Sul – Florianópolis/SC. [Hydrodynamic and sediment transport pattern analyses for recovering Armação do Pântano do Sul beach in Florianópolis/SC]. Book Chapter [online]. *Sistema de Modelagem Costeira do Brasil: estudos de caso*, pp. 351–376, Ministério do Meio Ambiente/Brazilian Ministry of Environment. ISBN 978-85-328-0835-6. Available at: repositorio.ufsc.br/handle/123456789/194405

Santos, Juliana Vieira; **Thiesen, Stephanie**; Rafael Augusto dos Reis (2018): Geological-geotechnical database from standard penetration test investigations using geographic information systems. Book Chapter [online]. *Management of Information Systems*, InTech Open, 2018. ISBN 978-1-78984-198-5. doi: 10.5772/intechopen.74208

BACHELOR AND MASTER'S THESES

Thiesen, Stephanie (2016): Aplicação de ferramenta SIG para mapeamento geotécnico e cartas de aptidão para fundação a partir de ensaios SPT – um estudo de caso em Blumenau/SC. *[Application of GIS tool for geotechnical mapping and foundation suitability maps based on standard penetration tests – a case study in Blumenau/SC]*. Master's thesis, Department of Civil Engineering, Federal University of Santa Catarina (UFSC), Florianópolis/Brazil.

Thiesen, Stephanie (2012): Ergonomia – análise dos fatores ambientais na empresa RKS Engenharia de Estruturas. *[Ergonomics – analysis of environmental factors in the company RKS structural engineering]*. Bachelor's thesis, College of Business Administration, Santa Catarina State University (UDESC), Florianópolis/Brazil.

Thiesen, Stephanie; Nicolau, Ana Paula (2011): Análise hidrodinâmica e de padrões de transporte de sedimentos e propostas para a recuperação da praia da Armação do Pântano do Sul Florianópolis/SC. *[Hydrodynamic and sediment transport pattern analyses and proposals for recovering Armação do Pântano do Sul beach in Florianópolis/SC]*. Bachelor's thesis, Department of Civil Engineering, Federal University of Santa Catarina (UFSC), Florianópolis/Brazil.

DECLARATION OF AUTHORSHIP

Eidesstattliche Versicherung gemäß §6 Abs. 1 Ziff. 4 der Promotionsordnung des Karlsruher Instituts für Technologie für die Fakultät für Bauingenieur-, Geo- und Umweltwissenschaften:

1. Bei der eingereichten Dissertation zu dem Thema *Information theory for non-parametric learning and probabilistic prediction: applications in Earth science and geostatistics* handelt es sich um meine eigenständig erbrachte Leistung.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die straf-rechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Karlsruhe, 2021

Stephanie Thiesen

COLOPHON

This document was typeset using the typographical look-and-feel classicthesis developed by André Miede and Ivo Pletikosić. The style was inspired by Robert Bringhurst's seminal book on typography *"The Elements of Typographic Style"*. classicthesis is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Final Version as of December 6, 2021 (classicthesis v4.6).