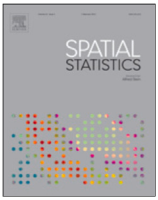
Spatial Statistics 54 (2023) 100737

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



SciKit-GStat Uncertainty: A software extension to cope with uncertain geostatistical estimates

Mirko Mälicke [a,b,*], Alberto Guadagnini [c], Erwin Zehe [a]

[a] Institute for Water and River Basin Management, Karlsruhe Institute of Technology (KIT), Kaiserstr. 12, Karlsruhe, 76131, Baden-Württemberg, Germany

[b] hydrocode GmbH, Mombertstr. 2, Karlsruhe, 76131, Baden-Württemberg, Germany

[c] Dipartimento di Ingegneria Civile Ambientale, Politecnico di Milano, Piazza L. Da Vinci 32, Milano, 20133, Italy

ARTICLE INFO

Article history:
Received 8 November 2022
Received in revised form 6 March 2023
Accepted 6 March 2023
Available online 17 March 2023

Keywords:
Uncertainty
Geostatistics
Python
Variogram estimation

ABSTRACT

This study is focused on an extension of a well established geostatistical software to enable one to effectively and interactively cope with uncertainty in geostatistical applications. The extension includes a rich component library, pre-built interfaces and an online application. We discuss the concept of replacing the empirical variogram with its uncertainty bound. This enables one to acknowledge uncertainties characterizing the underlying geostatistical datasets and typical methodological approaches. This allows for a probabilistic description of the variogram and its parameters at the same time. Our approach enables (1) multiple interpretations of a sample and (2) a multi-model context for geostatistical applications. We focus the sample application on propagating observation uncertainties into manual variogram parametrization and analyze its effects. Using two different datasets, we show how insights on uncertainty can be used to reject variogram models, thus constraining the space of formally equally probable models to tackle the issue of parameter equifinality.

[*] Corresponding author at: Institute for Water and River Basin Management, Karlsruhe Institute of Technology (KIT), Kaiserstr. 12, Karlsruhe, 76131, Baden-Württemberg, Germany.

E-mail addresses: mirko.maelicke@kit.edu (M. Mälicke), alberto.guadagnini@polimi.it (A. Guadagnini), erwin.zehe@kit.edu (E. Zehe).

https://doi.org/10.1016/j.spasta.2023.100737

1. Introduction

Geostatistical analyses are key in several research and industrial areas, including environmental and Earth sciences and engineering application. In this broad context, geostatistics typically considers (statistical) dependences of spatial or spatio-temporal datasets. In viewing a given quantity as a correlated random field, it has been shown to provide critical insights on ways to interpolate, assess, re-scale, and model scenarios of interest in the presence of scarce information. A broad variety of studies is geared towards assessing uncertainty through geostatistical estimation or simulation frameworks (Handcock and Stein, 1993; Journel, 1994; Mowrer, 1997; Zehe et al., 2005; Nowak and Verly, 2005; Delbari et al., 2009; Emery and Peláez, 2011; Todini, 2001; Lloyd and Atkinson, 2001), including some recent hydrological applications focused on preferential pathway analysis (Zehe et al., 2021; Schiavo et al., 2022). Otherwise, only a limited number of studies focus on a rigorous framework of analysis to explicitly include uncertainties associated with the empirical variogram and the way these can impact the estimation of an appropriate interpretive model. In this context, our study aims at providing enhanced insights on this, as the reliability of a geostatistical analysis hinges on an appropriate estimation of the empirical variogram. Thus, our distinctive objective relates to the way one can incorporate uncertainties into the variogram estimation. We then assess the way uncertainty associated with the assessment of the empirical variogram can propagate onto subsequent analysis steps. This allows seamless inclusion of uncertainty into geostatistical interpolations.

To the best of our knowledge, only a limited series of studies address uncertainty in the empirical variogram. Webster and Oliver (1993) define confidence limits for individual spatial models and their parametrizations. Their study considers sub-sampling of a dense datasets and focuses solely on the impact of sample size and the way a threshold can be defined for it through numerical Monte Carlo simulations. Pardo-Igúzquiza and Dowd (2001) describe various approaches to yield approximations of the standard error associated with the variance evaluated across a sample. These authors point out that exact confidence intervals for the empirical variogram are difficult to construct in practice and only a number of approximations can be employed. Some of the methods discussed therein are detailed in Section 3. Their studies relate the uncertainty of empirical variograms to the nature of the semi-variance estimator. Metrics of statistical robustness are then proposed on the basis of the size of the underlying finite sample.

While building on these approaches, here we address the joint effect of several sources of uncertainty on the empirical variogram. We highlight ways these can be tackled and ultimately be included into a variogram modeling context. Some of these sources of uncertainty are aleatory. These include e.g., the inherently limited precision of data in terms of accuracy of an observed quantity as well as of the spatial locations at which observations are taken. Other sources of uncertainty are epistemic and stem from incomplete knowledge about a system functioning and/or processes taking place therein (Hora, 1996; Der Kiureghian and Ditlevsen, 2009; Hüllermeier and Waegeman, 2021). Observations are never perfect in terms of precision and accuracy associated with a given measurement. Furthermore, in some cases one cannot observe directly a target quantity, while only data (corrupted by uncertainty) associated with other related quantities can be monitored. As a common example, one can refer to a rainfall radar, which is not rendering rainfall observation, but reflectivity of hydro-meteors. The latter depends on size and shape of the meteors, their chemical phase and a variety of additional factors (Neuper and Ehret, 2019). All of these sources of uncertainty jointly contribute to what we term *observation uncertainty* in this study.

An exemplary scenario underpinning of our study is associated with the geostatistical Python package SciKit-GStat (Mälicke, 2022b) and corresponds to an image of a pancake taken at a given time during browning. Fig. 1(a) illustrates the actual image and an inset of a target area. Color gradation corresponds to the red channel pixel value, which has a resolution of 8-bit, as common for images. We rely on this pixel value as observation here. This is also consistent with common remote sensing observation techniques, the pancake surface and the image corresponding to the random field under study and to the measurement, respectively. Note that this representation (along with the 8-bit resolution) already implies observation uncertainty. Any given RGB value in the photograph does not reflect the real color of the actual pancake. There are systematic and random

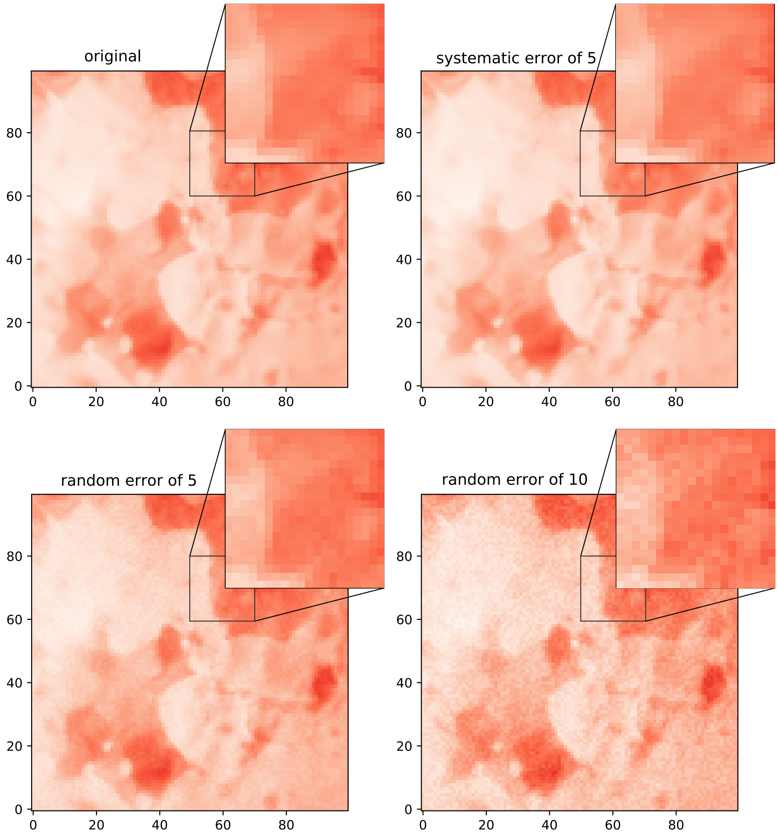


Fig. 1. Image of the pancake, which motivated this work. It shows the image of a pancake with several conceptualized errors applied. **(a)** Red channel of the original image with a 20 × 20 zoom of an area with apparent gradient on short distance. **(b)** Image from (a) with a systematic shift of 5 in the red channel value. **(c)** Image from (a) with a random error of 5 applied to each cell in the red channel. **(d)** Image from (a) with a random error of 15 applied to each cell in the red channel. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

errors influencing the measurement. These include e.g., the moisture of the air between the camera and the pancake or oscillations of the light bulb brightness slightly affecting room illumination. To assist evaluation of observation uncertainties in the context of pancakes and as an example to provide a visual depiction of the effect of uncertain observation, we apply a systematic shift in value (Fig. 1(b)) and a random variation in value (Fig. 1(c), (d)) of a different magnitude in each sub-panel to the original image. Differences in color from Fig. 1(a) (original) to Fig. 1(b) and c are visually very hard to detect. This illustrates that even a considerable variation in value might manifest in a subtle way from a visual standpoint. Fig. 1(c) depicts the magnitude of measurement error, which forms the basis for some of the analyses detailed in Sections 3 and 4.

These kinds of observation uncertainties are somehow less subtle in remote sensing, groundwater hydrology or soil science. Sensor sensitivity studies have shown that observation values are typically subject to much larger ranges of uncertainty (ie. fig. 4 Jackisch et al., 2020; Zehe and Blöschl, 2004; Arthur and Robinson, 2015). In addition to the above mentioned elements, one should note that some research studies can also be affected by un-calibrated sensors and/or, in some instances, on community-sourced sensors (Chapman et al., 2017), which do not comply with the same measurement standards and might also provide only indirect information about the quantity of interest.

Prompted by these elements, we illustrate here the software library SKGstat-Uncertainty that has been developed to specifically address these outstanding issues. The latter is built on existing and established packages for geostatistics in Python. It implements existing and original methods to analyze, assess, quantify, visualize, and propagate uncertainties in variogram estimation. Existing software solutions in Python include SciKit-GStat (Mälicke, 2022b). The latter is a variogram estimation toolbox that is currently characterized by only limited capabilities to handle observation uncertainties. For example, in the current implementation one could add error bars to semi-variance values on the basis of a manual input. Additionally, GSTools (Müller et al., 2021), an advanced geostatistical toolbox in Python, implements uncertainty elements for Kriging only if the user can supply the measurement error as a parameter. In this context, SKGstat-Uncertainty can be identified primarily as an extension to SciKit-GStat and is also compatible with GSTools.

SKGstat-Uncertainty is designed as a general toolbox, that is aimed at performing uncertainty analyses associated with variogram estimation in a way that is accessible to a broad audience. As such, end-users are envisioned to be associated with education, research, and industry sectors. In addition to providing a thorough introduction to the various functions of the toolbox, we exemplify the importance of variogram uncertainty upon considering two exemplary datasets.

Note that our study does not involve automatic fitting of a variogram model, even as the toolbox includes these features (namely the method-of-moments and the Maximum Likelihood approach). For the purpose of our exemplary study, we favor manual fitting of variogram functions to the uncertainty bounds. Doing so enables users to readily inspect various dimensions of uncertainty arising in the context of variogram analysis. By replacing the empirical semi-variance with its confidence limits (see Section 3.3), we explore the uncertainty in the parametrization of a given variogram model. Importantly, we also show that the choice of the theoretical model itself becomes uncertain. In this sense, the heart of the toolbox is a processing module that implements a suite of methods for the quantification of uncertainty associated with empirical semi-variance. Each of these is conducive to an uncertainty bound against which a collection of variogram models and ensuing parametrizations can be assessed. A rich selection of visualization routines enables the user to inspect various aspects of uncertainty. This offers a considerable added value with respect to parameterizing a black-box workflow to obtain a result, which might possibly be considered as the *right* or *most probable* one.

We perform the uncertainty analysis for **(a)** the pancake dataset depicted in Fig. 1 and **(b)** a hydrogeological dataset. The latter comprises a set of well-established and broadly used air permeability data collected across a Berea sandstone rock on a regular, dense grid (Tidwell and Wilson, 1997, 1999, 2002) and is detailed in Section 3.1.2.

After an introducing the software package and the sample application for manual variogram parametrization, we explore the following two research hypotheses:

- hypothesis H1: Empirical variograms (or semi-variances) are uncertain due to inherent observation and estimation uncertainty.
- hypothesis H2: Uncertain empirical semi-variances imply that an interpreting variogram model and the embedded parameters are uncertain; this, in turn, yields uncertain geostatistical interpolation results.

Testing both hypotheses relies on the presented toolbox.

Our study is structured as follows: Section 2 describes the toolbox from a technical perspective; Section 3 includes all methodologies used for the presented analysis; Section 4 illustrates the results and our findings, which are then discussed in Section 5. Conclusions are presented in 6.

2. Software implementation

Our software is a toolbox that is designed to extend the functionality of two well-established geostatistical Python packages, i.e., SciKit-GStat (Mälicke, 2022b) and GSTools (Müller et al., 2021). Key extensions include the implementation of geostatistical analysis tools and functions with options for uncertainty analysis and propagation. The toolbox implements building blocks to form applications governed through a dedicated graphical user interface. While the main focus is set on variography and Kriging, the toolbox is general and can be readily extended to include additional features.

The toolbox SKGstat-Uncertainty is written in Python and is published as open source (Mälicke, 2022a). It is a collection of functions, which can be run through the Python framework `streamlit`. This opens a web-browser based interface to operate the underlying Python code and its settings. As such, advanced programming skills are not required to load data, set up geostatistical libraries, pre- and post-process data, set model parameters, run analyses, and visualize the ensuing results.

Applications built with our toolbox can be scaled. With minimal overhead, it can be run locally on any client computer, a feature that enables one to readily interact with locally hosted data. Alternatively, public streamlit applications can be hosted on a cloud infrastructure of streamlit with limited resources, freely available. It is further noted that deploying a streamlit application on custom infrastructure is straightforward and in line with common web-based deployment strategies. This enables one to use the software at any scale in educational and professional scenarios, in a freely accessible mode, or as the foundation of a paid model. Finally, the toolbox is distributed as a Docker image with fixed software versions and architecture. Docker is a common solution to ensure reproducible software deployment independent of the host architecture and operating system. This enables one to repeat analyses ensuring consistent results.

The software toolbox is structured across several units. First, the *Data Models* describes the structure of the data used by the application. Exemplary, one model describes the attributes and structure of uploaded samples, while another one describes the attributes, which represent an empirical variogram. Data models also include relations between data model instances (usually called entities). Each model is implemented as a Python class and can easily be exported to the open standard format JSON.[1] Thus, students, scientists, or engineers and practitioners can easily export data and results from the application and use these for further analyses in any other framework of their choice. SKGstat-Uncertainty uses an SQLite database to save application data and intermediate results, as a default option. Connecting the toolbox to other database systems is also possible, as it uses the widely spread Python module `SQLAlchemy` (Bayer, 2012), which can connect to (almost) any relational database management system. The demo application stores the data in a remote PostgreSQL database.

Another unit termed *processor* implements algorithms for model evaluation, sampling, uncertainty propagation, and analysis. These algorithms are detailed in Section 3. An *Application Programming Interface (API)* unit collects functions for all common data management tasks, including filtering, creating, editing, or deleting information. While the API is used by the application, it is also usable as a standalone Python module and can be run as a command line interface directly from the operating system. The core unit is termed *components*. It includes the main functions, which are used by the streamlit framework to build the application. These functions run and operate the analysis as specified by the developer.

The *chapters* unit is a collection of standalone streamlit applications. These can be composed together into a final application, or can be run individually. Each of the chapters covers a given topic. Most chapters build on others, e.g., the chapter about Kriging algorithms can only be used after variograms are estimated for a target dataset. The software toolbox currently implements the following chapters:

- **Data management** — This chapter can upload, list and edit existing datasets. New samples can be created by re-sampling existing datasets.
- **Learn geostatistics** — This chapter provides an interactive and guided step-by-step introduction to geostatistics, which might be appropriate for an undergraduate or early stage graduate student. The details are not covered in this work, given their introductory nature and target audience.

[1] Human readable JSON format specification. URL:https://www.json.org/json-en.html, last accessed: 25.10.2022.

- **Variogram estimation** – The chapter implements an interactive interface to estimate sample variograms and propagate various kinds of uncertainty into the empirical variogram. This yields a uncertainty bound-based empirical variogram.
- **Model parametrization** – The chapter implements an interactive interface to identify an arbitrary amount of models and associated model parametrization within the uncertainty bounds of each variogram.
- **Kriging** – The chapter implements four different Kriging algorithms (simple, ordinary, universal, and external drift Kriging) leveraging on the identified variogram model functions to project data onto unobserved locations.
- **Geostatistical simulations** – The chapter implements an interface to perform geostatistical simulations for each theoretical variogram model function. For simplicity, the simulation feature of the tool is not included in this study.
- **Analysis tool** – The chapter enables one to visualize estimation (i.e., Kriging) or simulation results with a variety of pre-defined visualization options (see, e.g., Section 4).

A scientific demo application (termed *uncertain geostatistics*) is implemented to assist the user and can be reached publicly at https://geostat.hydrocode.de.[2] It does not add any significant functionality in terms of geostatistics or uncertainty analysis. The demo application runs an additional PostgreSQL database instead of the default sqlite database. Besides the chapters of SKGstat-Uncertainty described above, three more chapters were added to the application. The *help page* chapter loads documentation from the underlying Python packages SciKit-GStat and GSTools for reference. A *tutorials page* lists a number of short video tutorials about the other chapters. Additionally, a landing page including a login was added. Authenticated users are granted full access to additional data samples, which are not available under an open data license. Without authentication, data are still available when using the application. Otherwise, re-sampling and downloading non-open data (e.g., the Berea sandstone dataset illustrated in Section 3.1.2) are disabled. Authenticated access to the scientific sample application is managed by a third party[3], access to the Berea sandstone dataset can be obtained from the original publication (Tidwell and Wilson, 1997).

3. Data and methods

3.1. Data

3.1.1. Pancake dataset

A detailed description of the pancake dataset is offered by Mälicke (2022b). In line with this study, we consider the red channel of the RGB image in our analyses. For the purpose of our analysis, we re-scale the original red channel image described by Mälicke (2022b) (and associated with a 500 × 500 resolution) to a 100 × 100 resolution using a mean filter. Note that this step corresponds to smoothing the original image, hence decreasing the sample spatial variance. Otherwise, **(a)** it does not affect the workflow underpinning the application of our approach to tackling sample variogram uncertainty and **(b)** it enables us to obtain a sample that is approximately the same size as the one associated with the air permeability information evaluated across the block of Berea sandstone described in Section 3.1.2. We then apply our workflow considering a reduced size data sample constructed upon re-sampling the 100 × 100 resolution image according to a uniform 10 × 10 grid without any offset from the border, to avoid extrapolations in Kriging analyses.

[2] The whole geostatisitcal ecosystem around SciKit-GStat, SKGstat-Uncertainty and demo applications can be reached at https://geostat.hydrocode.de. The standalone demo application is deployed at https://uncertain.geostat.hydrocode.de.

[3] As of this writing, the demo application and the Python package are properties of hydrocode GmbH (https://hydrocode.de). The Python package is open source, while the demo application is free of charge.

Table 1
Overview of all lag class binning methods implemented in SciKit-GStat.
Source: From Mälicke (2022b).

Function	Identifier	Description	Implementation
Equidistant lags	`‘even’`	N lags of same width; Almost always used.	Mälicke et al. (2021)
Uniform lags	`‘uniform’`	N lags of same sample size; Estimates are based on the same sample size & no empty bins	Mälicke et al. (2021)
Sturge’s rule	`‘sturges’`	Equidistant lags derived from Sturge’s rule; use for small normal distributed distance matrices	Virtanen et al. (2020)
Scott’s rule	`‘scott’`	Equidistant lags derived from Scott’s rule; use for large datasets	Virtanen et al. (2020)
Freedman–Diaconis estimator	`’fd’`	Equidistant lags; use for small datasets with outliers in the distance matrix	Virtanen et al. (2020)
Square-root	`‘sqrt’`	Equidistant lags; Very fast function, but usually not recommended	Virtanen et al. (2020)
Doane’s rule	`‘doane’`	Equidistant lags; based on data skewness, use for small non-normal distance matrices	Virtanen et al. (2020)
K-Means	`‘kmeans’`	Non-equidistant lags; clustered distance matrix is used as binning; slow but statistically robust	Pedregosa et al. (2011)
Hierarchical clusters	`‘ward’`	Non-equidistant lags; clustered distance matrix is used as binning; Based on Ward’s criterion for minimizing cluster variance. Computational intensive	Pedregosa et al. (2011)

3.1.2. Berea sandstone

The second dataset we consider is well established and representative of a Darcy-scale collection of air-permeability data (Tidwell and Wilson, 1997, 1999, 2002). The latter are sampled on the six faces of a 81 × 74 × 63 cm^3 block of Berea sandstone, across an area of 30 × 30 cm^2. The sampling grid comprises 36 × 36 regularly spaced nodes (horizontal resolution $\Delta = 0.85$ cm). Data collection relies on four air minipermeameters, each with a given tip-seal (inner and out radius of the minipermeameter are $r \quad i = \{0.15, 0.31, 0.63, 1.27\}$ and $r_2 \quad i = \{1, 2, 3, 4\}$, respectively). For the purpose of our analyses, we focus on the set of data associated with the smallest tip-seal radius.

Recent geostatistical analyses of these data include the works of Riva et al. (2013) and Dell’Oca et al. (2020).

Given the size of the minipermeameter tip, the original Berea sandstone dataset can be considered exhaustive and is used as the (hydrogeological) field equivalent of the pancake image. A sub-sample of the air permeability data to be used in our uncertainty analyses is then obtained upon considering the information available on a uniform 8 × 8 grid, approximately corresponding to 10% of the field. This enables us to perform the same types of analyses for the two selected datasets and consistently compare results across these.

3.2. Empirical variogram estimation

Empirical variograms are estimated using the Python package SciKit-GStat (Mälicke, 2022b). The package offers various options to this end. The scientific demo application integrates nine out of the ten binning algorithms implemented in SciKit-GStat (Table 1). Depending on the binning algorithm, the user may select the number of lag classes for the evaluation of the variogram and the associated confidence limits. The largest separating distance at which point pairs are formed can be set directly or selected from predefined values such as, e.g., the median separating distance.

Table 2
Overview of all semi-variance estimator functions implemented in SciKit-GStat (Mälicke, 2022b).
Source: Modified after Mälicke (2022b).

Estimator	Identifier	Description	Reference
Mathéron	`'matheron'`	Default, most popular estimator	Matheron (1963)
Cressie–Hawkins	`'cressie'`	Power transformation based - robust to outliers	Cressie and Hawkins (1980)
Dowd	`'dowd'`	Median based, fast estimator for non-normal distributed residuals	Dowd (1984)
Genton	`'genton'`	Percentile-based estimator - powerful for skewed residuals, but very computationally intensive	Genton (1998)
Shannon entropy	`'entropy'`	Information theory metric focusing on information content of residuals	Shannon (1948)

The semi-variance of the resulting population of increments corresponds to the sample variogram for a given separation distance (or lag) and can be estimated through one of the five implemented estimators (Table 2). In case of a positively skewed dataset and in the presence of outliers, we recommend the use of robust semi-variance estimators (see, e.g., Table 2).

The empirical variogram for the pancake dataset (Fig. 2 d) is estimated upon relying on Matheron semi-variance (Matheron, 1963) according to 14 evenly spaced bins. The largest separating distance between a point pair was set to 100 grid units, thus coinciding with the length of the side of the domain across which data are sampled. Visual inspection of the results shows that the empirical variogram is characterized by a nugget/sill ratio of about 0.25. This is deemed as a remarkable amount of the total observed variability that could not be explained by the observed degree of spatial dependence (or correlation) of the target quantity.

The empirical variogram for the data associated with the Berea sandstone sample is depicted in Fig. 2(c). The KMeans based binning algorithm (see Table 1) is employed to form 10 lag classes up to the largest considered lag of 24 cm. Similar to the pancake dataset, this corresponds to the length of the side of the domain across which data are sampled. The semi-variance is evaluated using the Matheron estimator, consistent with the pancake dataset. These results (see Fig. 2) suggest that the nugget/sill ratio of the empirical variogram might be smaller for the Berea than for the pancake dataset.

3.3. Uncertainty bounds of the empirical variogram

The key element of the application is the possibility to propagate observation uncertainties onto the estimation of the empirical variogram. These are then ultimately employed to characterize the empirical variogram through bounds of uncertainty. We note that we specifically tailor our approach to empirical variograms and consider the underlying random field to be either second-order stationary or to satisfy the intrinsic hypothesis.

The first option available relies on the quantification of a confidence interval through the standard deviation of the empirical density of the (zero-mean) residuals of the squared increments of the target quantity corresponding to a given lag. The approach is straightforward and can be used, e.g., when no other information on observation uncertainty is available. The characteristic width, δ, associated with the confidence interval is evaluated as:

$$\delta = z\frac{\sigma}{\sqrt{N}} \tag{1}$$

where σ is the standard deviation of sample squared increments, N is sample size, and z is the z-score of the desired confidence level of the standard Normal distribution function Z. As the uncertainty bound is evaluated on the basis of the confidence interval of the mean point pair residual, the central limit theorem is expected to hold. The latter may be violated, though, when using a high number of lag classes in combination with a small sample size for some of these.

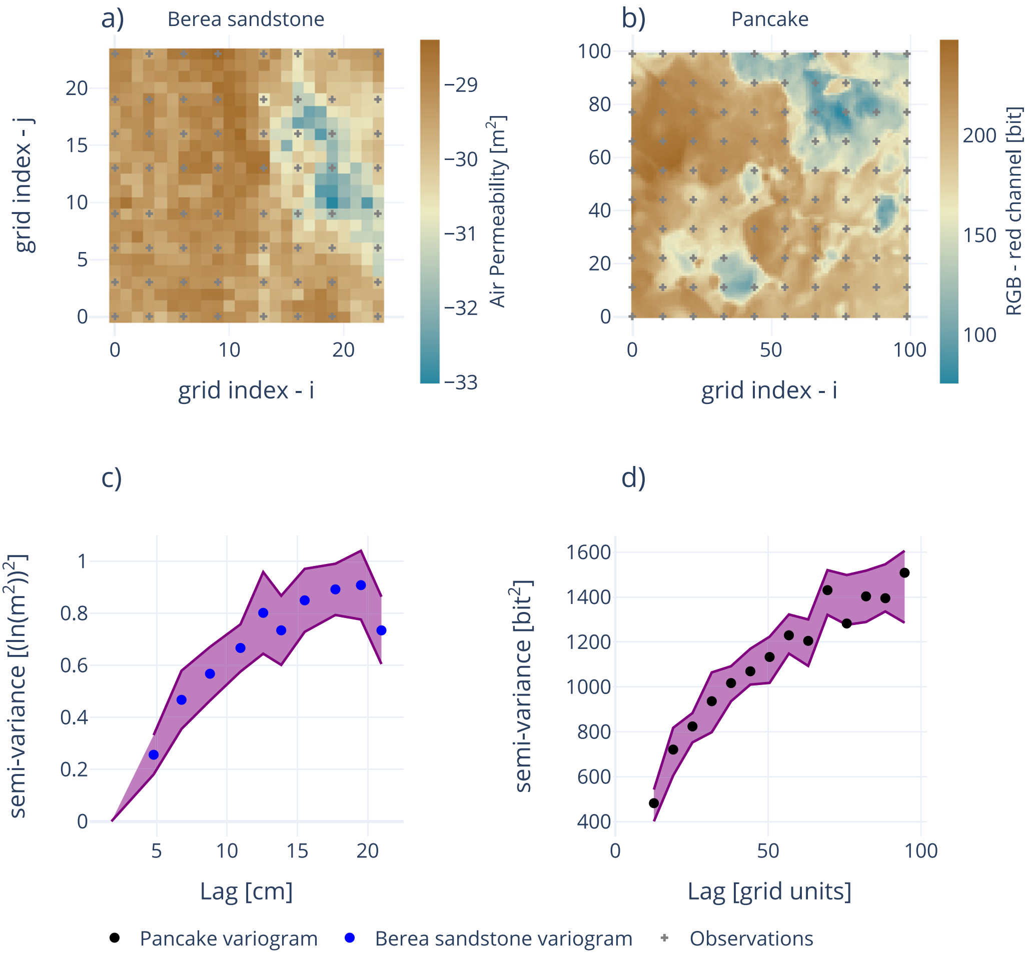


Fig. 2. Data Overview: **(a)** Permeability data associated with one of the faces (denoted as face 1) of the Berea sandstone sample obtained through the minipermeameter characterized by a 0.15 cm inner radius of the tip. Data are originally published and described in Tidwell and Wilson (1997, 1999). **(b)** Spatial distribution of the data associated with the pancake setting (see also Fig. 1(a)), color gradation being adjusted to match the corresponding visualization related to the Berea sandstone sample. Symbols in (a) and (b) correspond to the data employed in our exemplary analyses. **(c, d)** Empirical variogram obtained considering the sampled data depicted in (a) and (b) for the Berea sandstone grid sample (**c**, blue circles) and the pancake dataset (**d**, black circles). The purple area corresponds to the uncertainty bounds estimated for the variograms. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We thus encourage the user to carefully inspect the histogram of point pairs associated with all lag classes. Note that in the following we consider typical 95% confidence intervals for the Berea sandstone sample variograms. This approach is employed for the Berea sandstone scenario, as no further information on actual observation uncertainties is available.

The second approach is based on the evaluation of semi-variance values for each lag class in the context of a k-fold statistical robustness test. The application implements options to subdivide each class associated with a given lag into 3, 5, 7, or 10 folds and evaluate the semi-variance k times for $k - 1$ folds comprised in the bin. Upon relying on 100 iterations, values of squared increments are allocated randomly to the folds and the uncertainty bounds are evaluated for the

$i \times k$ estimated semi-variance values. The number of iterations can be adjusted by the user. The key assumption underlying this approach is that the robustness of semi-variance values calculated for a large number of smaller subsets strongly reflects the true uncertainty associated with the semi-variance. The main advantage of the methodology is that it does not require any particular assumptions about the residual distribution because it simply evaluates the actual semi-variance given the reduced size dataset. Otherwise, a weak element of the procedure is that it is quite sensitive to the settings of the robustness test (especially to sample size). If the number of pairs within each lag class is not sufficiently large, the k-subsets might be too small to infer robust statistics. Otherwise, when considering large samples, the computational demand for this iterative process needs to be carefully considered and might hamper its efficiency. The approach is well suited to tackle scenarios where the user cannot quantify observation uncertainties and the amount of data enables one to avoid resorting to the simple approach encapsulated in Eq. (1).

The third approach implemented is set within a numerical Monte Carlo simulation context. It is here demonstrated considering the (re-sampled) field of observations resulting from the original data. The array of observations is replaced by a randomly generated array, given a specified aleatory uncertainty measure. Here, we consider three kinds of uncertainty metrics that can be propagated onto the variogram.

A first metric is based on considering measurement error to be represented by a uniform distribution with a given mean (corresponding to the observed value) and a minimum/maximum value specified by the user, which we will denote as *measurement error bounds*. This enables one to assign the same weight to all of the values included in the support of the distribution.

A second metric relies on the standard error of the mean (SEM) of the observations. The latter needs to be specified by the user as an input parameter to the procedure. By doing so one considers observation errors to be characterized by a Normal distribution with a given mean (corresponding to the observed value) and standard deviation, σ, given by:

$$\sigma = SEM * \sqrt{N} \tag{2}$$

where N corresponds to the sample size.

A third option considers specifying directly the standard deviation of the aforementioned Normal distribution.

Resorting to a given observation error metric depends on available metadata, i.e., on additional information eventually complementing the analyzed dataset. For example, some manufacturers of physical sensor devices might supply SEM values, while modeling results might rather be associated with a well defined error bound. It is quite often possible to estimate one of the three aforementioned metrics from expert knowledge. When knowledge on the uncertainty metrics described above is available, the Monte Carlo approach is preferable, as compared to the other options described, which are based on stronger assumptions. If available, SEM is possibly a preferred aleatory uncertainty measure, as it describes observation uncertainties by definition.

The empirical variogram is then represented through the evaluated uncertainty bounds. These embed the concept of uncertainty we propose to employ in the context of geostatistical analyses fully encapsulating uncertainty in the empirical variogram. In line with the spirit of our study, we then obtain a collection of variogram models (and ensuing parametrizations) that are consistent with an interpretation of a variogram based on the concept of uncertainty bounds. As previously stated, the ensuing collection of models (and parameters) can then be employed to propagate variogram uncertainty onto geostatistical analyses (i.e., in the context of estimation and/or simulation scenarios).

3.4. Theoretical model performance metrics

Accounting for uncertainty bounds of the empirical variogram enables one to consider **(a)** multiple parameter sets conditional to a given model and/or **(b)** multiple competing model formulations that are all consistent with the level of uncertainty associated with observations. Thus, model selection is a major epistemic source of uncertainty, directly tied to our research hypothesis H2 (Section 1). A key research question tackled through the tool hinges on the identification of

Table 3
Overview of all theoretical variogram model functions implemented in SciKit-GStat.
Source: Modified after Mälicke (2022b).

Model	Identifier	Description	Implementation
Spherical	`'spherical'`	Short ranged correlation length, popular model in geoscience; for smooth, but steep gradients in fields.	Burgess and Webster (1980)
Exponential	`'exponential'`	Long ranged for smooth fields with less steep gradients.	Journel and Huijbregts (1976)
Gaussian	`'gaussian'`	Mid ranged for sharply changing fields	Journel and Huijbregts (1976)
Cubic	`'cubic'`	Similar to Gaussian models, but with a shorter correlation length.	Montero et al. (2015)
Matérn	`'matern'`	Has an additional smoothness parameter to adapt shapes between Exponential and Gaussian models.	Zimmermann et al. (2008)
Stable	`'stable'`	Has an additional shape (power) parameter to adapt the range.	Montero et al. (2015)

theoretical variogram models that, following a given parametrization, are fully comprised within the identified uncertainty margins.

Model formulations available in the toolbox are listed in Table 3.

The toolbox implements a variety of metrics to assess model performance, as detailed in the following Sections.

3.4.1. Root squared mean error — RMSE

An adjusted version of the root squared mean error ($RMSE$) can be used as a goodness-of-fit metric for a given variogram model parametrization. In this context, for uncertainty bounds of width $\Delta\gamma = u - l$ (u and l being an upper and lower limit) at a given lag and for a target model variogram γ, we set $RMSE := 0$ if $l < \gamma' < u$. Otherwise, $RMSE$ is evaluated as:

$$RMSE = \sqrt{\frac{\sum_{i=0}^{N} min(\gamma_i' - u, \quad l - \gamma_i')^2}{N}} \tag{3}$$

where N is the number of lags at which the empirical variogram (and hence the uncertainty bounds) is estimated from available data. We note that $RMSE$ is used to assess the model solely on the basis of the fit of the theoretical model to the empirical variogram uncertainty bounds. As such, it does not provide information about the ability of a given model (or model parameter set) to correctly estimate or simulate the analyzed quantity at unobserved locations.

3.4.2. Cross-validation through ordinary Kriging

As a second metric that can be employed to evaluate the performance of a given variogram model, we also rely on a classical leave-one-out cross-validation. For $Z(s_N)$ observations, the model is applied considering $N - 1$ observations to estimate $Z(s_N)^*$ at the omitted location via Ordinary Kriging. The ensuing differences between observed and interpolated values are then assessed upon relying on their associated RMSE. A value of $RMSE = 0$ indicates that the model is capable of reproducing the observations. Increasing values of $RMSE$ correspond to an increased mismatch between observation and interpolation-based estimates.

3.4.3. Deviance information criterion — DIC

The application also allows for the evaluation of a given variogram model upon relying on model selection criteria. These are employed to evaluate the relative skill of a candidate model (as compared against other model analyzed) to interpret available observations. We rely here on formal model selection criteria to evaluate (in a relative sense) the ability of each of the models we consider to be consistent with the estimated uncertainty bounds related to the empirical variogram. Among the various model selection criteria proposed in the literature to discriminate amongst models (see e.g. Riva et al., 2011; Höge et al., 2018), we rest here on the Deviance Information Criterion *DIC* (Spiegelhalter et al., 2002, 2014), which is a generalization of the Akaike Information Criterion *AIC* (Akaike, 1973; Hurvich and Tsai, 1989).

The deviance *D* of a given model (parameterized through a set of parameters collected in vector $\vec{\Theta}$) is given by:

$$D(\vec{\Theta}) = -2ln(L) \tag{4}$$

where *L* is the likelihood function of the considered theoretical variogram model. Here, we consider the following definition of a negative log-likelihood function from Lark (2000, Eq. (14)):

$$\begin{aligned} L(r, s|\vec{\hat{m}}, \hat{\sigma}^2, \vec{z}) = \frac{n}{2}ln(2\pi) + \frac{n}{2} - \frac{n}{2}ln(n) + \frac{1}{2}ln|\vec{A}| + \\ \frac{n}{2}ln\left((\vec{z} - \vec{\hat{m}})^T A^{-1}(\vec{z} - \vec{\hat{m}})\right) \end{aligned} \tag{5}$$

where $\vec{z}$ is a vector whose entries correspond to n available observations; r and s are the range and sill of the considered variogram model, respectively; $\vec{\hat{m}}$ is a vector of maximum likelihood estimates of the available data at the observation points (see also Lark, 2000, Eq. (12)); $\hat{\sigma}^2$ is a maximum likelihood estimate of the sample variance (see also Lark, 2000, Eq. (13)); and A is the auto-correlation matrix for the sample and specified (Lark, 2000, Eq. (9)) as follows:

$$\begin{aligned} \vec{A}(i,j) &= 1 & i = j, \\ &= s\{1 - f(\vec{x}_i - \vec{x}_j|r)\} & i \neq j \end{aligned} \quad s = \frac{c}{c_0 + c} \tag{6}$$

Here, i, j are the indices corresponding to the observation locations; $f(\vec{x}_i - \vec{x}_j|r)$ is the spatially structured component of the variogram model conditioned only to the range parameter, r; s is a term associated with the nugget to sill ratio, c and c_0 corresponding to the variogram sill and nugget, respectively. We note that Eq. (5) underlies the assumption of Gaussian distribution for the associated variogram model parameters.

The deviance information criterion penalizes a model with respect to its competing counterparts through the complexity of its parametrization. The latter is quantified via the concept of *effective parameters*, *pD*, defined as:

$$pD = \overline{D(\vec{\Theta})} - D(\bar{\vec{\Theta}}) \tag{7}$$

where $\bar{\vec{\Theta}}$ is the mean of all parameters associated with a given model (i.e., a given functional format of the variogram) and $\overline{D(\vec{\Theta})}$ is the sample mean of deviance evaluated across all models and parameter sets.

Considering the sample probability density of model parameters, *DIC* is then evaluated as:

$$DIC = \overline{D(\vec{\Theta})} + pD \tag{8}$$

Following Spiegelhalter et al. (2002), one could assess *pD* upon relying on the mode or on the median of the distribution of model parameters assessed through model characterization on the basis of the uncertainty bounded empirical variogram. All of these options are implemented in the toolbox. As an additional option to evaluate *pD*, we also consider Gelman et al. (2014, Eq. (7).10):

$$pD = \frac{1}{2}var(D(\vec{\Theta})) \tag{9}$$

This formulation always yields positive values for *pD*, which, in turn, makes the use of *DIC* very intuitive. Thus, we use the latter approach and formulation for this study and as a default option for the toolbox due to its readily intuitive nature.

3.4.4. Structural risk minimization

Another area where one usually needs to balance between model complexity and over-fitting is machine learning. In this context, an appealing framework is provided by the concept of structural risk minimization (Vapnik and Chervonenkis, 1974). While the toolbox implements a variation of the latter, we not pursue it further in this study. The interested reader is referred to Appendix A, where the available option from the toolbox is briefly illustrated.

3.5. Variogram model assessment

The toolbox function for manual variogram fitting enables the user to **(a)** select any of the available theoretical variogram models and **(b)** interactively parameterize these for the desired number of model parameter sets while considering the estimated uncertainty bounds related to the given empirical variogram. The quantitative metrics described in Section 3.4 are evaluated for the collection of all models and associated parameters employed for data interpretation. We recall that the objective here is to sample the set of possible theoretical model functions and their parametrizations. We further note that other techniques conductive to (posterior) distributions of model parameters such as, e.g., acceptance–rejection sampling (e.g. Russian et al., 2017, and references therein) are not yet embedded in the toolbox. Otherwise, the modular nature of the toolbox facilitates the integration of additional simulation tools. Thus, users are foreseen to be able to choose among various approaches (as soon as these are implemented) to obtain a collection (i.e., an ensemble) of candidate theoretical models (and ensuing model parameter sets) in their scenarios of interest.

The collection of model functions and ensuing parameter sets are then filtered to retain the best-performing models. With reference to this issue, our toolbox implements an interactive, feature-rich selection interface. The user may perform model selection analysis upon relying on one of the metrics detailed in Sections 3.4.1 to 3.4.4 or comparing the results associated with the use of all of these. While the demo application is currently confined to a given number of options for model selection, its flexible structure enables one to seamlessly expand on these. The user can either **(a)** retain a fixed amount of parameter sets (e.g 10 best ones), **(b)** retain a fixed amount of parameter sets stratified by model type (like 3 Gaussian, 3 Spherical, and so on) or **(c)** calculate a threshold by defining an acceptable relative deviation from the best parameter set. Only the selected parameter sets are then considered for the estimation of a Kriging uncertainty bound, as described in the following Section.

3.6. Kriging uncertainty bounds

Our toolbox includes four different Kriging algorithms from GSTools (Müller et al., 2021). While the default option is Ordinary Kriging, the user may select to rely on either Simple or Universal Kriging. If auxiliary information is available, external drift Kriging can be used, incorporating such data as drift. For Simple Kriging, the mean of the field needs to be specified by the user. For Universal Kriging, a linear or a quadratic internal drift term is currently available.

To propagate uncertainties to a Kriging application, each of the selected models is used with each of the associated parameter sets to project the data onto a target grid. While the size of such grid can be specified interactively by the user, the toolbox also implements some options to automatically evaluate the coordinate locations for each grid cell.

The following option is of interest for our demo software. In case the user uploads a field (such as, e.g., the pancake scenario we consider) and uses the toolbox to sub-sample it, the toolbox automatically uses the grid of the originally uploaded field, if Kriging is applied to the **sub-sample**. The advantage of this procedure is that one can associate a value from the originally uploaded field to any location of the target grid which is not tied to the sub-sample. This enables the user to objectively assess the overall performance of each kriged field.

The uncertainty propagated onto the Kriging-based estimates corresponds to the range of interpolation estimates associated with each grid location. We note that the number of available estimates matches the number of selected models and model parameter sets. In some cases, it is possible that a given parameter set is conducive to kriged estimates that markedly differ from those of the remaining models and model parameter sets, either across the whole target field or only within a certain region. Thus, an important feature implemented in the tool enables one to examine and compare the contribution of a given parameter set to the overall uncertainty of the results. We do so upon relying on the Shannon entropy (Shannon, 1948) associated with the collection of predictions at each grid cell/node. The Shannon entropy is defined as:

$$H = -\sum_{i=0}^{M} p_i * log_2(p_i) \tag{10}$$

where p_i is the empirical probability of non-exceedance of the ith value of the collection of estimates related to a target location in the domain. The Shannon entropy is well suited to analyze redundancy within a model and model parameter collection (Loritz et al., 2018; Mälicke et al., 2020, e.g.). Non-exceedance probabilities are evaluated upon subdividing the range of the obtained interpolated values across the whole domain into a number M of bins, which is typically set to the number of selected parameter sets. This is tantamount to considering the same binning for obtaining p_i at all grid locations.

In order to compare Shannon entropy across datasets and assess the agreement of estimates between the models and ensuing parameter sets, the Shannon entropy is normalized. A suitable normalization considers the Shannon entropy of a distribution of M uniformly sized bins, H_{max}. For the whole domain, the same H_{max} will be considered. The normalized Entropy $H_n = \frac{H}{H_{max}}$ is a measure of how close the distribution of estimates in each grid cell is to a uniform distribution (corresponding to $H_n = 1$). Thus, it can be used to identify grid locations of high estimate variability within a set of Kriging results. It can also be used to compare results across multiple datasets, with respect to the number of parameter sets selected. We note that a $H = 0$ for a given grid location implies that all estimates reside within the same bin. This does not imply that all estimates are numerically close, because M (i.e., the number of selected parameter sets) might be quite small in some cases. Here, we use the normalized Shannon Entropy to identify regions of the domain where there is high estimate variability, that can then be compared across multiple datasets.

4. Results

4.1. Variograms and related uncertainty

Here, we illustrate all details of the application with reference to the pancake sample. We then present and analyze our findings for the Berea sandstone sample (see Fig. 2, a&b). Visual inspection suggests that the two fields display a similar spatial structure and their variograms exhibit a similar pattern (Fig. 2 c&d). The two variograms differ clearly with respect to the width of their uncertainty bounds. The latter is larger for the Berea sandstone dataset. We note that, taking only the uncertainty bounds into account, almost any theoretical model might fit each of these empirical variograms and a prior selection of a specific model is not justified.

4.2. Theoretical variogram models and associated performance metrics

With respect to the uncertainty bound of the empirical variogram, almost none of the theoretical model (here manually parameterized) can be rejected (see Fig. 3(a)). Fig. 3(b) provides a graphical depiction of the relevant metrics for all of these models. Here, the different model types are listed in the first column and each band represents one set of model parameters, color gradation being indicative of a given model type. The first connection to the second column ranks the models by their fit in terms of *RMSE* (Eq. (3)). For visual reasons, *RMSE* values are ranked and grouped into quartiles, with the 25% best performing model parameter sets at the top of the column. The bands spread out significantly and are not grouped by model type anymore. This stresses the visual impression that no model instances are significantly off when considering empirical variogram uncertainty.

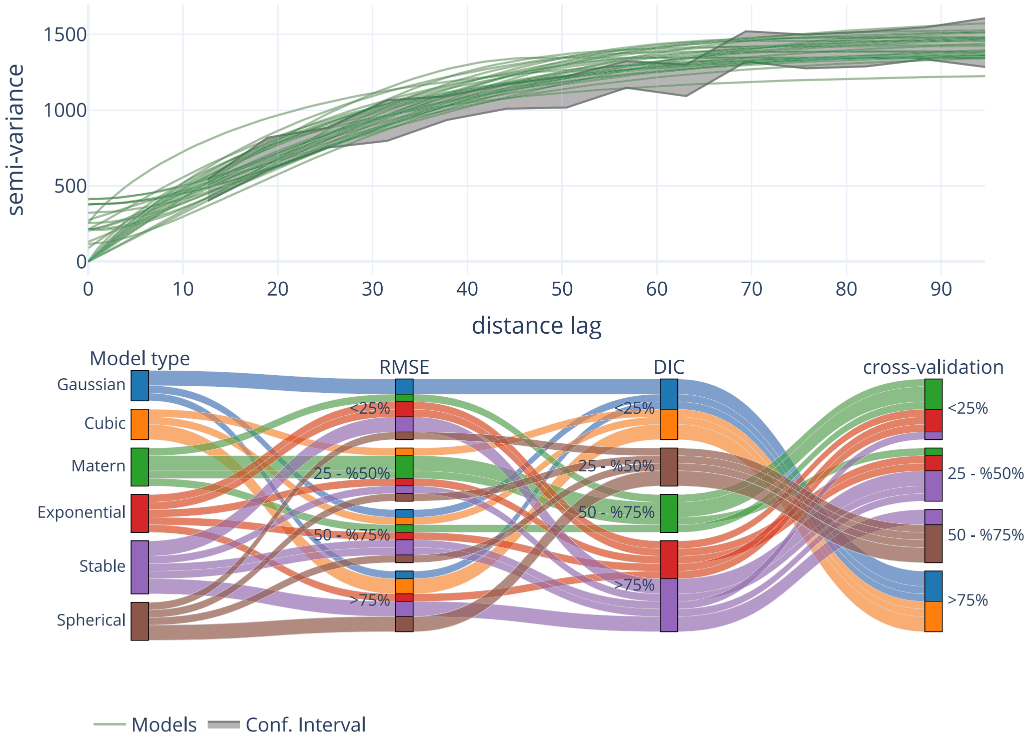


Fig. 3. **(a)** Uncertainty bounds (gray area) associated with the empirical variogram related to the **pancake** dataset, including with all theoretical variogram models fitted (green curves). **(b)** Parallel coordinates plot for the models depicted in (a) showing the considered performance metrics, i.e., *RMSE* (2nd column), *DIC* (3rd column), and cross-validation (4th column). The first column groups the individual models by their type and corresponding color gradation. For each of the measures, the models are ranked into quartiles; as an illustrative example, we consider $<$ *25%* to delineate the collection of the best performing models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The third column in Fig. 3(b) ranks the model parameter sets by the corresponding *DIC* value (Eq. (8)). By design the model parameter sets are grouped by model type, as *DIC* is a performance metric on model type and does not distinguish among the different parametrizations. Similar to other information criteria, *DIC* grounds the suitability of a given model on the likelihood of the model parameters, given the sample distribution. In terms of *DIC*, the Cubic and Gaussian models perform best, while exponential and stable models are characterized by poorer performance. We further note that, due to manual parametrization, the model collection sizes are quite small and *DIC* values might change when additional parameter sets are added to the collection.

The last column in Fig. 3(b) provides a ranking of the model parameter sets grounded on cross-validation results. The predictive power of each (manually fitted) model and model parameter set is assessed by applying Kriging interpolation via a leave-one-out cross-validation for all observation points. Interestingly, all bands cross on the connection of the third and fourth column (see Fig. 3(b)). Thus, model and model parameter ranking change to favor Matérn parametrizations over Gaussian and cubic models. One has to keep in mind that for the purpose of our demonstration Kriging is only applied to the sample considered and model performance might differ for unobserved locations. This is expected to depend on the density and structure of the observation points.

Taking all of the above elements into consideration, one can conclude that the uncertain observations allow for various models and ensuing parametrizations to be considered as suitable

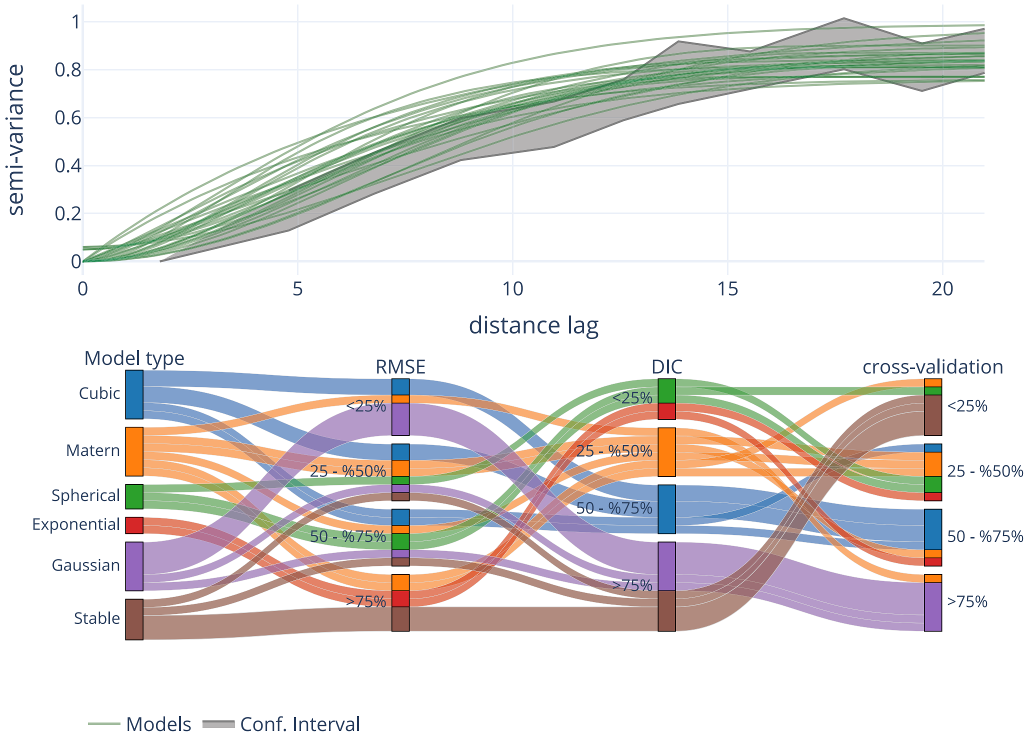


Fig. 4. **(a)** Uncertainty bounds (gray area) associated with the empirical variogram related to the **Berea sandstone** dataset, including all theoretical variogram models fitted (green curves). **(b)** Parallel coordinates plot for the models depicted in (a) showing the considered performance metrics, i.e., *RMSE* (2nd column), *DIC* (3rd column), and cross-validation (4th column). The first column groups the individual models by their type and corresponding color gradation. For each of the measures, the models are ranked into quartiles; as an illustrative example, we consider $<$ *25%* to delineate the collection of the best performing models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in a virtually indistinguishable way. This is largely supported by the *RMSE* results. In practice all parameter sets would be accepted in any least-square based automatic procedure. The *DIC* criterion rests on variogram likelihoods given the sample distribution and does favor specific model types over others. This can loosely be seen as an assessment of how and which models an automated maximum likelihood approach favors. Results from cross-validation, which is conceptualized as a visualization of the training error of the model, are in contrast with those provided by *DIC*. This finding is unexpected and raises some interesting questions about when and how to apply automatic and semi-automatic fitting procedures.

The shape of the uncertainty bound is slightly different for the Berea sandstone sample and is characterized by a less pronounced increase of semi-variances within the first few lag classes. Similar to what can be observed for the pancake dataset, all theoretical variogram functions (green curves in Fig. 4(a)) appear to be equally compatible with the estimated uncertainty bounds. This results in a straightforward parametrization of Gaussian or Gaussian-shaped Matérn models. Otherwise, the exponential and exponentially shaped stable models appear not to be fully compatible with the estimated uncertainty bounds, with special reference to the upper limit of these. This behavior is also reflected by the *RMSE* values in the second column (Fig. 4(b)), which rank the exponential model parametrizations slightly lower than for the pancake dataset.

According to the values of *DIC* (Fig. 4(b)), spherical and exponential models are highly ranked, as opposed to their Gaussian and stable counterparts. Similar to the pancake dataset, results of cross-validation based on Kriging appear to favor some models that were ranked low according to the other metrics employed. While the Gaussian models still perform worse than others, the stable models are ranked significantly higher according to this metric. It is worth noting that all parameter sets of the stable model lie in the best performing quartile in terms of Kriging cross-validation, even those that visually show notable deviations when juxtaposed to the uncertainty bound. The same finding holds for the exponential model. All parametrizations of the latter are ranked in the lowest quartile for RMSE, in the highest quartile for *DIC* and close to median for the cross-validation metric.

In summary, there is no model type that is ranked low consistently by all metrics across both datasets considered. Likelihood- and uncertainty bound-driven metrics do not yield a unique and unequivocal outcome when analyzed jointly and neither of these is entirely supported by Kriging cross-validation across the collection of the corresponding parameters. Possibly, a clear conclusion is that Gaussian models should be avoided, although they appear to fit the uncertainty bound best.

All of these results suggest that any kind of automatic variogram fit should always be complemented by careful inspection of results of the kind we illustrate, on the basis of multiple metrics, each revealing a particular aspect of uncertainty.

4.3. Kriging uncertainty bounds

A collection of about 30 different model parameter sets has been identified for the pancake dataset. A critical element in the analysis of the way variogram uncertainty propagates onto Kriging results is the possibility of ranking model parameter sets according to the performance metrics selected. This is accomplished through the implementation of a filtering step in the tool. The latter allows for various functions to filter the model parameter sets with respect to one of the performance metrics detailed in Section 3.4.

All models parameter sets are then ranked with reference to each of the metrics considered (i.e., *RMSE*, *DIC*, and cross-validation). The filter rejects the 10% worst parameter sets for each metric. For both datasets we find that six instances were rejected, most of these associated with Gaussian models, which is seen to be ranked lowest in more than one metric.

Propagating variogram uncertainties onto the Kriging results generally leads to large corresponding uncertainty bounds. By taking different model parametrizations into consideration, one finds a spread of Kriging interpolation results which is typically of about 25 units, while attaining peaks of about 70 units (Fig. 5(a)), which corresponds to about 30% of the range of values of the available data. The width of the Kriging uncertainty bounds is highly heterogeneous in space. In some areas the uncertainty bounds are not much larger than the observation uncertainty propagated into the procedure, while being markedly larger in other regions. In general, uncertainty band widths correlate with the location on the grid and most model parameter sets seem to disagree in terms of Kriged values close to the domain boundaries.

As expected, the Kriging error variance generally tends to vanish close to observation locations. Fig. 5(b) shows the range of Kriging error variances for all selected model parameter sets. As expected, and consistent with the dense sampling arrangement, no particular spatial differences can be identified. We remark that a value of 0 in Fig. 5(b) implies that all Kriging variance values coincide, all models being in agreement.

While the range of kriged values for a given unobserved location can be large, this can be due, in some cases, to a single parameter set or to a limited number thereof. This element can be investigate through the analysis of the entropy map of model Kriging results. Fig. 5(c) depicts the spatial distribution of the values of the normalized information entropies. Here, a value of 1 or 0 implies large variability across the collection of estimates or that all estimates fall into the same bin, respectively. Values of the normalized information entropy of estimates (Fig. 5(c)) are largely spread evenly across the domain, even as some clusters are noted around a number of observation locations. Values are small in most areas. This finding suggests that only a few model parameter sets are driving the width of the uncertainty bands in Fig. 5(a). The entropy map displays a high level of spatial organization.

The Berea sandstone sample is characterized by similar results (Fig. 6). There is a considerable overlap of larger normalized entropies and wider uncertainty bands. This is especially evident

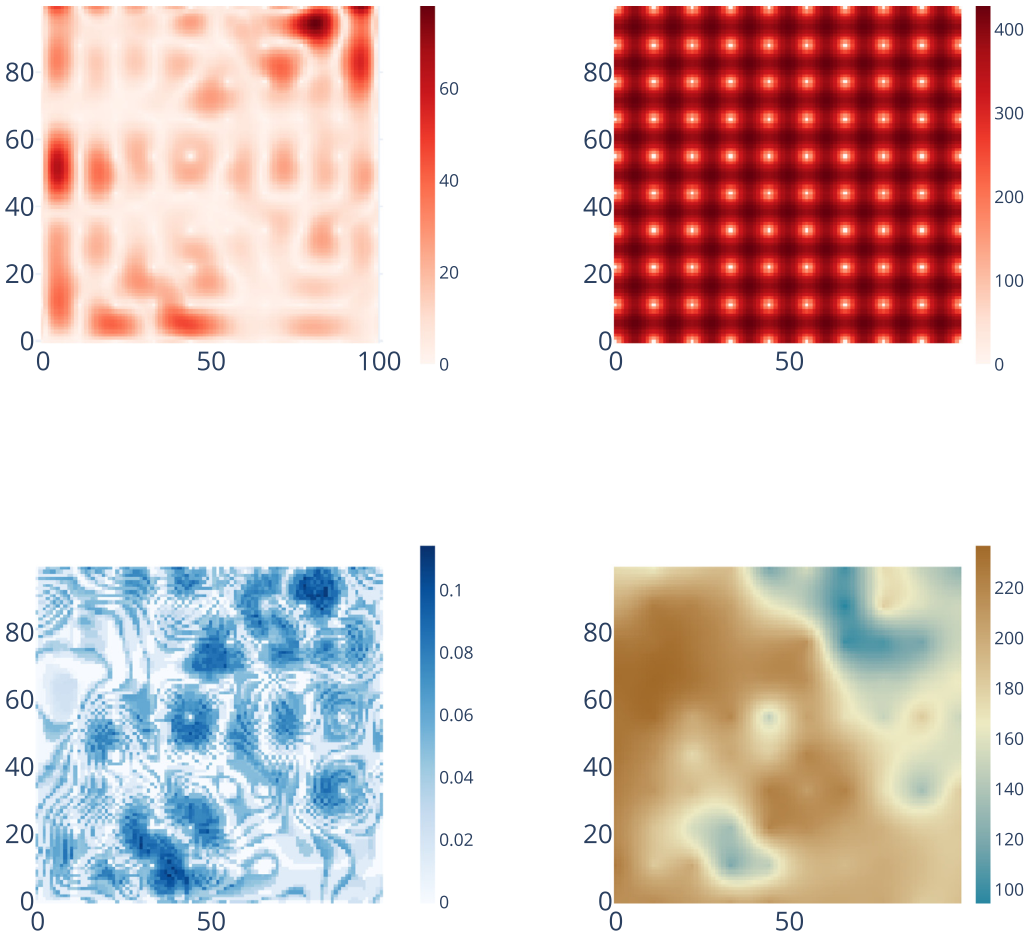


Fig. 5. Kriging estimation results for the **pancake dataset** after re-sampling on a **regular grid**. Width of the interval of variability of **(a)** Kriged values and **(b)** Kriging error variance values associated with all models analyzed at each cell across the domain . Note that a zero variance value means that the Kriging variance is the same for all models. Large values imply a variable Kriging error variance (It is illustrating the variability of variances). **(c)** Normalized Shannon entropy of all model estimates **(d)** Kriging interpolation result for the best model parameter set (mean rank of RMSE, *DIC* and cross-validation).

in the proximity of the right boundary of the domain. Normalized information entropy values are considerably larger for the Berea dataset (attaining values consistently $>$ 0.4) than their counterparts associated with the pancake datset. This is consistent with the observation that a number of estimations differ by orders of magnitudes in these areas. Interestingly, one can also note the presence of the smallest values of the underlying field in this area. Fig. 6(d) shows the Kriging interpolation result stemming from the model parameter set, best performing in terms of mean rank in all used performance metrics. From here, these areas, colored blueish, can easily be identified.

4.4. Identifiability of variogram model parameters for uncertain variograms

We exemplify the way one can select some parameter sets as optimal upon relying on the concept of variogram uncertainty bounds through a detailed analysis of the corresponding metrics

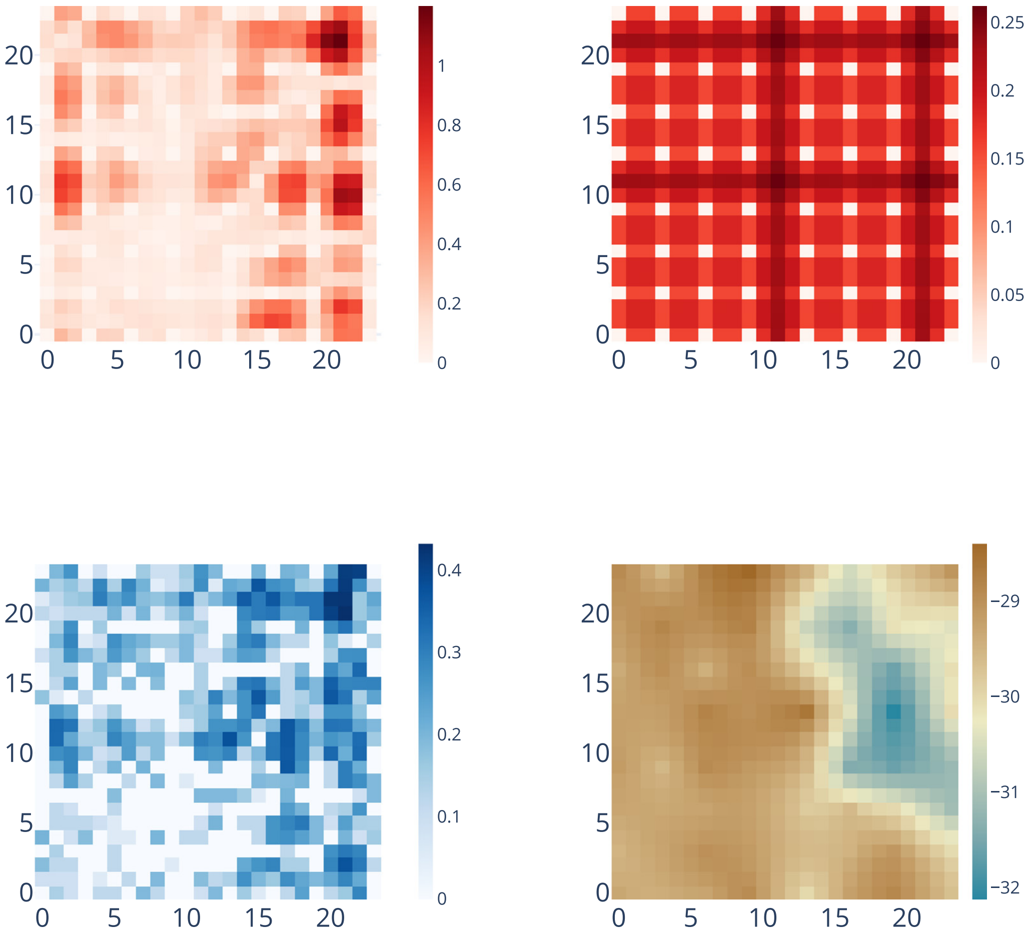


Fig. 6. Kriging estimation results for the **Berea sandstone dataset** after re-sampling on a **regular grid**. Width of the interval of variability of **(a)** Kriged values and **(b)** Kriging error variance values associated with all models analyzed at each cell across the domain . Note that a zero variance value means that the Kriging variance is the same for all models. Large values imply a variable Kriging error variance (It is illustrating the variability of variances). **(c)** Normalized Shannon entropy of all model estimates **(d)** Kriging interpolation result for the best model parameter set (mean rank of RMSE, *DIC* and cross-validation).

based on the spherical variogram model. We do so because *(a)* the model is seen to perform well for both datasets and *(b)* this is the model type selected to demonstrate automatic fitting of empirical variograms with SciKit-GStat by Mälicke (2022b).

Here, we use only the definition of the deviance in Eq. (4). As the mean deviance will be the same for all parameters, the value of *DIC* will be linearly dependent on the deviance. At the same time, the negative log-likelihood function used in a maximum likelihood approach differs only by a factor of 2 from Eq. (4). Thus, this enables us to jointly interpret the results in terms of maximum likelihood (deviance) and method-of-moment (RMSE) approaches.

Each of the aforementioned metrics is evaluated (Fig. 7) for 100 × 100 combinations of range and nugget/sill ratios. A maximum nugget to sill ratio value of 1 was used (i.e., nugget and sill have the same (absolute) value). We note that considering values of this ratio larger than unity might hamper the usefulness of geostatistical approaches, which rest on the concept of a spatial correlation structure.

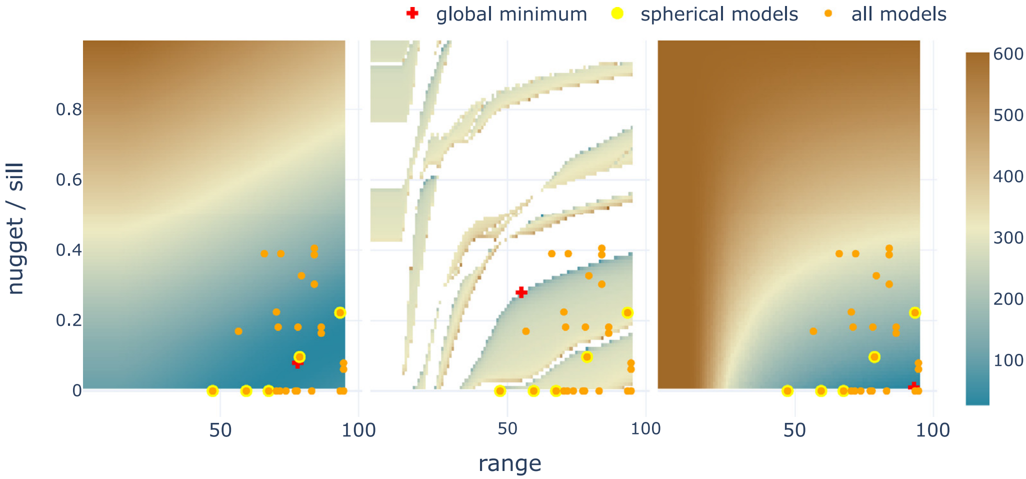


Fig. 7. Results of the parameter testing phase for 100 × 100 combinations of sill/nugget ratio and effective range for a spherical model using the **pancake** dataset. **(a)** *RMSE* (Eq. (3)) of the model fit to the uncertainty bound. Red or blue grading denotes larger or smaller metric values. **(b)** deviance value for all parameter combinations. **(c)** Leave-one-out cross-validation of the interpolated observation values. The orange symbols show the models and parametrizations used in manual fitting for all model types (the spherical model types are marked by thick yellow outline). The red cross marks the global minimum for each of the parameter tests. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We added to the coordinates of the grid locations a white noise of about 0.1‰of the grid extent. Thus, any impact on the lag classes of the empirical variogram can be neglected. This was a necessary step to circumvent the issue that the Kriging system of equations be associated with too many instances of singular matrices. This likely originated from the regular spacing of the sampling locations, as further detailed in Section 5.3.

Large areas show a satisfactory performance in terms of *RMSE* of model fit (Fig. 7(a)). Moreover, Fig. 7(a) illustrates clearly that there is in fact parameter equifinality (Beven and Binley, 1992) due to parameter interaction. We note that a global minimum is not readily identifiable across the parameter space. All of the results of the manual parametrizations here presented (orange dots) lie within the area of optimal parameter values (blue-graded region).

The deviance metric does not yield a result for several parameter sets (Fig. 7(b)); white regions. Here, the auto-correlation matrix *A* (see Eq. (6)) is singular and could not be inverted. The extent of the blue-graded areas is considerably smaller than for the *RMSE* metric. Finally, Fig. 7(c) illustrates the leave-one-out cross-validation metric for all 100 × 100 parameter combinations. Similar to the *RMSE*, all manually fitted parameter sets are contained in the blue-graded area within which good performance values of the metric are obtained. The global minimum is very close to the lower right corner (range of 93 and nugget to sill ratio of 0.01). We note that the *RMSE* and cross-validation metrics appear to be in a substantial overall agreement.

5. Discussion

Our analysis provides a clear evidence that **(a)** uncertainty of the experimental variogram should not be ignored and **(b)** the presented toolbox markedly facilitates assessment and propagation of such uncertainty onto a set of acceptable theoretical variogram models and corresponding Kriged fields. The presented test cases yield insights into the ability of different performance metrics and on the goodness of individual members of a family of acceptable variogram models (in terms of their ensuing parameters). Interestingly, ranks of individual models and parameter sets is not the same for the different metrics. We show these elements for two datasets, by propagating uncertainties into the empirical variogram, assessing acceptable theoretical variogram models and their associated parameter sets, and comparing their kriged estimates across the system (also considering cross-validation) as well as the spreading of Kriging results at unobserved locations. Overall, a clear choice of a superior variogram model type or the identification of a best parameter set cannot be identified. A key asset of the presented toolbox is that it also provides enhanced understanding about how and where these uncertainties are caused. While a variety of selection algorithms or variogram parameter optimization approaches could be considered, for the purpose of our demonstration we choose a straightforward approach and eliminate models which perform poorest with respect to each of the performance metrics.

We acknowledge that the current stage of our work and version of the associated tool is restricted to an isotropic spatial covariance. Otherwise, a variety of environmental variables/quantities exhibit an anisotropic spatial covariance, also depending on the scale at which they are considered. As an example, we mention cold front precipitation bands, topography or macropores in soils, as well as sedimentological attributes or parameters characterizing variably saturated subsurface flow. The standard approach to detect a geometric anisotropy is to use directional experimental variograms. While our method to estimate uncertainties can be readily applied to the assessment of directional experimental variograms, this task is beyond the scope of the current study.

In the following we discuss **(a)** the way the toolbox can assist interactive geostatistical analyses, **(b)** variogram estimation under uncertainty and the related model evaluation, and **(c)** the assessment of our driving hypotheses.

5.1. Interactive geostatistical analysis

Our software toolbox is built on established and well-tested software packages for numerical computing, visualization, and geostatistics. The implementation focuses on well-defined datasets. By providing clear interfaces and metadata, our API can be used to automate common tasks and build user interfaces such as those associated with the illustrated sample application. This is not only convenient but also considerably speeds up analysis workflows. As such, it empowers early stage researchers and students to dive deeper into the material and scientists and practitioners to operate on data more effectively. This will ultimately favor practical implementations of new approaches. As an example: We would not have been able to manually parameterize so many models more effectively and faster than automatic approaches if it would not have been for an interactive slider element that enables one to adjust variogram parameters on the fly.

All this convenience comes at the cost of the implementation effort. As the user is less engaged with the actual, technical implementation than in more traditional scripting approaches, the software to be employed needs to be built in a generalized way. Achieving this element, in turn, needs comprehensive tests to ensure technical correctness. Tutorials and a complete and detailed documentation are equally important. Otherwise, the user will not be able to identify misuses and errors. A website and video tutorials are available for `SKGstat-Uncertainty`. Remarkably, the essential core of all calculations is implemented within other software products, each of these being carefully chosen to entail comprehensive testing and documentation. This enables the user to focus on analysis and visualization while being confident in the technical correctness of the results.

5.2. Uncertain variogram estimation and model evaluation

Using different methods for uncertainty propagation and estimation, we evaluate uncertainty bounds for the empirical variogram associated with the two showcases illustrated. This is a key result, as by simple visual inspection it is possible to estimate variogram parameters manually, thus favoring enhanced understanding on the system behavior. We note that at least one of each available theoretical variogram model type could be parameterized to fit into the uncertainty bound, or at least very close to it. This result confirms our hypothesis **H2**, as it makes the epistemic uncertainty relate to a prior model choice way more obvious than through a classical fitting procedure targeting empirical variograms. We rely here on manual procedures, due to their ease of usage and pedagogic potential. Otherwise, we stress that the toolbox is not limited to manual parametrization. Any suitable alternative approach implemented in Python (such as, e.g., ensemble learning methods or acceptance rejection) can be readily implemented into a new chapter of the tool. An additional added value of tool resides in the observation that data management and processing chapters, as well as subsequent analysis chapters, are modular.

An important limitation to our illustrative results, though, is that only one instance of an empirical variogram was estimated. The estimation is known to be sensitive to sampling strategies, especially sampling size, binning procedures, and amount of lag classes used. While our modeling choices are based on expert knowledge, there might be a more suitable empirical variogram candidate, especially for the Berea sandstone setting. The purpose of this work, however, is to demonstrate the software package for exemplary analyses. Thus, we are confident that the illustrated insights can be adapted and transferred to focus on other critical aspects of empirical variograms, such as uncertainty bands based on systematic testing for different sample sizes.

Each selected variogram model parameter set was used in a Kriging interpolation context. As shown in Figs. 5 and 6, the corresponding interpolation results differ substantially. Considering all of the interpolation results, it was not possible to identify a unique model type (or parameter set) that clearly describes the spatial correlation structure of the field unequivocally better than all others. Otherwise, by combining insights from three different kinds of evaluation metrics, which focus on different aspects of a variogram fit, we can exclude model types. This is considered as an additional key result of our study and approach.

The Gaussian and cubic models are found to interpret adequately the empirical variogram associated with the pancake dataset (in terms of its uncertainty bounds). While *DIC* favors these two model formulations, the leave-one-out cross-validation excludes both of them regardless of their parametrizations. Inspection of the single interpolated grids based on Gaussian variogram models revealed that all of them produced considerable amounts of kriged values far outside the observation value space. We encourage the user of the application to interpret the results by considering the critical message that observation uncertainties exist and need to be comprehensively addressed. As analysis results may differ vastly, one could at least rely on insights obtained through modeling under uncertainty to exclude models (or parametrizations). This enables one to learn by rejection and enhance our knowledge from quantification of uncertainties, instead of neglecting these.

We assess sources of uncertainty that affect the different kinds of fitting procedures (least squares, maximum likelihood, manual) in very different ways and demonstrate the significance of our results in geostatistical applications . These insights, if taken into account, can assist in limiting the parameter space for a geostatistical analysis and lead to new knowledge about a field under investigation. It also bears an important implication with respect to quality and precision of the measured data. A smaller sample of highly precise observations will result in a small variogram uncertainty similar to what one could obtain through a large sample of less accurate data.

5.3. Model fitting and model parameters

A core decision taken for the application and in the exemplary study we present is the focus on manual variogram parametrization. Such a manual parametrization is a valid operational and educational choice. This is especially relevant in cases where one might select to renounce to some computing speed for a more thoughtful and detailed variogram analysis. Manual parametrization is straightforward, reflects a deep understanding of the variogram concept, and can be applied without the need for the implementation of optimization algorithms.

During the systematic exhaustive testing of variogram parameter values, the leave-one-out cross-validation calculation failed in several instances, especially for medium and small values of the effective range parameter. Due to the repetitive pattern underlying a grid, the number of distinct separating distances was significantly decreased in our examples. For the pancake sample, while the upper triangle of the distance matrix contains more than 4500 entries, these hold only 34 different distance values. The main reason for this is conceptually illustrated in Fig. 8. For the center point (in red), the Kriging equation system may be built solely from the surrounding blue points, which are all 1 or $\sqrt{2}$ units away from the center point. The two blue points at $y = 1$ are symmetrical with respect to the center. This means that the Kriging equation system is characterized by duplicated rows at

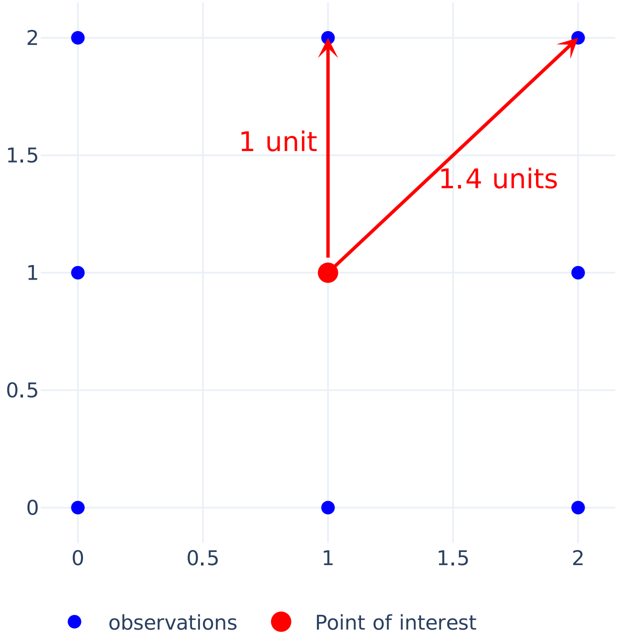


Fig. 8. Conceptual figure of an observation grid in a dimensionless Cartesian space. Blue dots correspond to observation locations. The red dot is the point of interest, which is subject to estimation in a leave-one-out cross-validation. The figure illustrates the repetitive pattern of just two different distances being used in a Kriging application. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the index of exactly these two points. This makes the Kriging matrix singular thus hampering its inversion. This was verified to happen quite often in our examples, as the Kriging algorithm built into SciKit-GStat limits the neighbor selection by the effective range of the variogram.

The same principle underpins the failure of the likelihood-based calculations observed in most cases.

In line with our first hypothesis that empirical variograms are uncertain, we present evidence that (Kriging) interpolation results cannot be simply limited to rely on a unique parametrization. Thus, geostatistical applications need to fully consider empirical variogram uncertainty bounds. Moreover, the parametrization of the variogram itself is markedly affected by different kinds of uncertainty. Our exemplary scenarios provide strong evidence of the basic assumption that propagating observation uncertainties into the variogram would lead to broad ranges of variogram parameter values that can be employed in a practical application. It is also apparent that a global minimum for a given metric cannot be identified easily. Furthermore, our results show that there is no evidence that any automatic procedure would perform better, even if only one set of parameters is considered to be valid. And finally, the best manual fit is very close to the global minimum of RMSE, even if the difference to adjacent parameter sets is considered to be significant.

While, in general terms, *RMSE* (Fig. 7(a)) and cross-validation (Fig. 7(b)) agree in identifying some sets of well-performing parameters (blue-graded), they also show some disagreements. These two figures suggest that even if a variogram model fits well to the uncertainty bound (or to the empirical variogram), cross-validation adds an additional (enriching) dimension against which the goodness of a performance should be assessed.

6. Conclusions

We introduce the toolbox *SciKit-GStat Uncertainty* and exemplify its use upon relying on sample data-sets pertaining to two different processes and scenarios. Our work leads to the following major conclusions

1. The toolbox is envisioned as a required extension of existing geostatistically-oriented computational tools and software. Our toolbox is built in a Python environment and includes the implementation of existing and new approaches to analyze, visualize, and quantitatively propagate uncertainties in variogram estimation onto kriging-based estimates and the associated variance. Its interactive nature empowers one to tackle uncertainty in a straightforward way and underpins the potential of the tool to play a key role in the context of research and educational contexts.
2. The toolbox enables one to explore the way various sources of uncertainty can imprint the results of a geostatistical analysis. Uncertainties considered through the toolbox arise from measurements (in terms of observations and location associated with these) as well as from the choice of an interpretive model and its parameters. Thus, the user can readily inspect various dimensions of uncertainty during variogram analyses. As a notable research element, we introduce and embed in *SciKit-GStat Uncertainty* the concept of replacing an empirical variogram (or semi-variance) through uncertainty bounds. This provides an original way to explore uncertainty, as imprinted onto the way one can evaluate the ability of a collection of models and ensuing parameters to perform variogram analysis upon relying on such uncertainty bounds. This is accomplished through a processing module that implements a suite of methods for the quantification of uncertainty associated with empirical variograms.
3. The software allows operating in a multi-model context and enhances our ability to interpret spatially correlated (random) fields. Exemplifying the use of our toolbox with emphasis on manual variogram parametrization enables us to emphasize the value of the toolbox in the context of a pedagogical/educational perspective. The user can then explore the benefit of resorting to the joint use of various metrics, each of them providing a specific insight on the quantification of uncertainty, to yield a comprehensive depiction of system behavior and characterization. In this context, we investigate the way coupling the concept of variogram uncertainty-bounds with the joint analysis of multiple methods and metrics can contribute to disregard some models and parametrizations over others.

CRediT authorship contribution statement

Mirko Mälicke: Conceptualization, Use-Case, Data, Methodology, Software. **Alberto Guadagnini:** Use-Case, Data, Methodology. **Erwin Zehe:** Use-Case, Methodology.

Appendix A. Structural risk minimization

Another available option to assess model parameter performance is derived form structural risk minimization. The work of Vapnik and Chervonenkis (1974) is focused on classification problems and support vector machines, its basic idea can be transferred to the scenario we consider. This is consistent with the observation that model selection can be viewed as a classification problem driven by uncertainty. The methodology introduced by Vapnik and Chervonenkis (1974) is designed to balance training errors and an expected over-fitting. Similar to an information criterion, the so-called capacity of the parameter space is employed, which in turn should measure model complexity through the minimization of:

$$J(\vec{\Theta}) = \varepsilon_{train}(\vec{\Theta}) + \lambda H(\vec{\Theta}) \tag{11}$$

Here, $\varepsilon_{train}(\vec{\Theta})$ is a measures of training error and $H(\vec{\Theta})$ is a regularization term. The latter penalizes models with a higher level of complexity, in terms of parametrization. The value of the weight *lambda*, needs to be set by the user. We adapt this concept by interpreting the parametrization of a variogram model as the training of our model and combine it with the *pD* as described in Section Section 3.4.3. The scientific demo application evaluates $\varepsilon_{train}(\vec{\Theta})$ either with the *RMSE* (see Section 3.4.1) or the MAE as suggested by Vapnik and Chervonenkis (1974):

$$MAE = \sum_{i=1}^{N} min(l - \gamma_i', \quad \gamma_i' - u) \tag{12}$$

Where γ_i' the modeled semi-variance at the *i*th lag class. As such, $MAE := 0$ for $l < \gamma' < u$.

With reference to the regularization term $H(\vec{\Theta})$ in Eq. (11) one can set it either to Eq. (7) or Eq. (9).

The toolbox implements all combinations to evaluate Eq. (11), but the exemplary demo application does not make use of these metrics.

Appendix B. Data and code

The pancake dataset is available with the SciKit-GStat package (Mälicke, 2022b). The source code, including the pancake data sample, is available on Github[4]. The Berea sandstone data sample can be obtained from the original publication (Tidwell and Wilson, 1997).

The source code for the *SciKit-GStat Uncertainty* extension is available on Github.[5] This repository includes a copy of the used data samples. The primary distribution of the software package is a docker image.[6]

The demo application is not open source. It can be reached at https://geostat.hydrocode.de.

References

Akaike, H., 1973. Information theory and an extension of the likelihood ratio principle. In: Petrov, B., Csaki, F. (Eds.), Proceedings of the Second International Symposium of Information Theory, Vol. 257. p. 281.

Arthur, A.M., Robinson, I., 2015. A critique of field spectroscopy and the challenges and opportunities it presents for remote sensing for agriculture, ecosystems, and hydrology. In: Neale, C.M.U., Maltese, A. (Eds.), Remote Sensing for Agriculture, Ecosystems, and Hydrology XVII, Vol.9637. International Society for Optics and Photonics. SPIE, pp. 29–39. http://dx.doi.org/10.1117/12.2201046.

Bayer, M., 2012. Sqlalchemy. In: Brown, A., Wilson, G. (Eds.), The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few more Fearless Hacks. aosabook.org, URL http://aosabook.org/en/sqlalchemy.html.

Beven, K., Binley, A., 1992. The future of distributed models: Model calibration and uncertainty prediction. Hydrol. Process. 6 (3), 279–298. http://dx.doi.org/10.1002/hyp.3360060305, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/hyp.3360060305 URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.3360060305.

Burgess, T., Webster, R., 1980. Optimal interpolation and isarithmic mapping of soil properties: I the semi-variogram and punctual kriging. J. Soil Sci. 31 (2), 315–331.

Chapman, L., Bell, C., Bell, S., 2017. Can the crowdsourcing data paradigm take atmospheric science to a new level? A case study of the urban heat island of London quantified using Netatmo weather stations. Int. J. Climatol. 37 (9), 3597–3605. http://dx.doi.org/10.1002/joc.4940, arXiv:https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.4940 URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.4940.

Cressie, N., Hawkins, D.M., 1980. Robust estimation of the variogram: I. J. Int. Assoc. Math. Geol. 12 (2), 115–125. http://dx.doi.org/10.1007/BF01035243.

Delbari, M., Afrasiab, P., Loiskandl, W., 2009. Using sequential Gaussian simulation to assess the field-scale spatial uncertainty of soil water content. Catena 79 (2), 163–169.

Dell'Oca, A., Guadagnini, A., Riva, M., 2020. Interpretation of multi-scale permeability data through an information theory perspective. Hydrol. Earth Syst. Sci. 24 (6), 3097–3109. http://dx.doi.org/10.5194/hess-24-3097-2020, URL https://hess.copernicus.org/articles/24/3097/2020/.

Der Kiureghian, A., Ditlevsen, O., 2009. Aleatory or epistemic? Does it matter? Struct. Saf. 31 (2), 105–112.

Dowd, P., 1984. The variogram and kriging: robust and resistant estimators. In: Geostatistics for Natural Resources Characterization. Springer, pp. 91–106.

Emery, X., Peláez, M., 2011. Assessing the accuracy of sequential Gaussian simulation and cosimulation. Comput. Geosci. 15 (4), 673–689.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, Aki Rubin, D.B., 2014. Bayesian Data Analysis, third ed. Chapman and Hall/CRC.

[4] https://github.com/mmaelicke/scikit-gstat; last accessed: 25.10.2022.

[5] https://github.com/hydrocode-de/skgstat_uncertainty; last accessed 25.10.2022.

[6] https://github.com/hydrocode-de/skgstat_uncertainty/pkgs/container/skgstat_uncertainty; last accessed: 25.10.2022.

Genton, M.G., 1998. Highly robust variogram estimation. Math. Geol. 30 (2), 213–221.

Handcock, M.S., Stein, M.L., 1993. A Bayesian analysis of kriging. Technometrics 35 (4), 403–410.

Höge, M., Wöhling, T., Nowak, W., 2018. A primer for model selection: The decisive role of model complexity. Water Resour. Res. 54 (3), 1688–1715.

Hora, S.C., 1996. Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. Reliab. Eng. Syst. Saf. 54 (2–3), 217–223.

Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Mach. Learn. 110 (3), 457–506.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76 (2), 297–307.

Jackisch, C., Germer, K., Graeff, T., Andrä, I., Schulz, K., Schiedung, M., Haller-Jans, J., Schneider, J., Jaquemotte, J., Helmer, P., Lotz, L., Bauer, A., Hahn, I., Šanda, M., Kumpan, M., Dorner, J., de Rooij, G., Wessel-Bothe, S., Kottmann, L., Schittenhelm, S., Durner, W., 2020. Soil moisture and matric potential – an open field comparison of sensor systems. Earth Syst. Sci. Data 12 (1), 683–697. http://dx.doi.org/10.5194/essd-12-683-2020, URL https://essd.copernicus.org/articles/12/683/2020/.

Journel, A.G., 1994. Modeling uncertainty: some conceptual thoughts. In: Geostatistics for the Next Century. Springer, pp. 30–43.

Journel, A.G., Huijbregts, C.J., 1976. Mining geostatistics.

Lark, R.M., 2000. Estimating variograms of soil properties by the method-of-moments and maximum likelihood. Eur. J. Soil Sci. 51 (4), 717–728. http://dx.doi.org/10.1046/j.1365-2389.2000.00345.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2389.2000.00345.x URL https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2389.2000.00345.x.

Lloyd, C., Atkinson, P.M., 2001. Assessing uncertainty in estimates with ordinary and indicator kriging. Comput. Geosci. 27 (8), 929–937.

Loritz, R., Gupta, H., Jackisch, C., Westhoff, M., Kleidon, A., Ehret, U., Zehe, E., 2018. On the dynamic nature of hydrological similarity. Hydrol. Earth Syst. Sci. 22 (7), 3663–3684. http://dx.doi.org/10.5194/hess-22-3663-2018, URL https://hess.copernicus.org/articles/22/3663/2018/.

Mälicke, M., 2022a. hydrocode-de/skgstat_uncertainty: Version 1.3. Zenodo, http://dx.doi.org/10.5281/zenodo.6545079.

Mälicke, M., 2022b. SciKit-GStat 1.0: a SciPy-flavored geostatistical variogram estimation toolbox written in Python. Geosci. Model Dev. 15 (6), 2505–2532. http://dx.doi.org/10.5194/gmd-15-2505-2022, URL https://gmd.copernicus.org/articles/15/2505/2022/.

Mälicke, M., Hassler, S.K., Blume, T., Weiler, M., Zehe, E., 2020. Soil moisture: variable in space but redundant in time. Hydrol. Earth Syst. Sci. 24 (5), 2633–2653. http://dx.doi.org/10.5194/hess-24-2633-2020, URL https://hess.copernicus.org/articles/24/2633/2020/.

Mälicke, M., Möller, E., Schneider, H.D., Müller, S., 2021. mmaelicke/scikit-gstat: A Scipy Flavoured Geostatistical Variogram Analysis Toolbox. Zenodo, http://dx.doi.org/10.5281/zenodo.4835779.

Matheron, G., 1963. Principles of geostatistics. Econ. Geol. 58 (8), 1246–1266.

Montero, J.-M., Fernández-Avilés, G., Mateu, J., 2015. Spatial and spatio-temporal geostatistical modeling and kriging. John Wiley & Sons.

Mowrer, H.T., 1997. Propagating uncertainty through spatial estimation processes for old-growth subalpine forests using sequential Gaussian simulation in GIS. Ecol. Model. 98 (1), 73–86.

Müller, S., Schüler, L., Zech, A., Heß e, F., 2021. Gstools v1.3: A toolbox for geostatistical modelling in Python. Geosci. Model Dev. Discuss. 2021, 1–33. http://dx.doi.org/10.5194/gmd-2021-301, URL https://gmd.copernicus.org/preprints/gmd-2021-301/.

Neuper, M., Ehret, U., 2019. Quantitative precipitation estimation with weather radar using a data- and information-based approach. Hydrol. Earth Syst. Sci. 23 (9), 3711–3733. http://dx.doi.org/10.5194/hess-23-3711-2019, URL https://hess.copernicus.org/articles/23/3711/2019/.

Nowak, M., Verly, G., 2005. The practice of sequential Gaussian simulation. In: Geostatistics Banff 2004. Springer pp. 387–398.

Pardo-Igúzquiza, E., Dowd, P., 2001. Variance–covariance matrix of the experimental variogram: assessing variogram uncertainty. Math. Geol. 33 (4), 397–419.

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

Riva, M., Neuman, S.P., Guadagnini, A., Siena, M., 2013. Anisotropic scaling of Berea sandstone log air permeability statistics. Vadose Zone J. 12 (3), http://dx.doi.org/10.2136/vzj2012.0153, vzj2012.0153. arXiv:https://acsess.onlinelibrary.wiley.com/doi/pdf/10.2136/vzj2012.0153 URL https://acsess.onlinelibrary.wiley.com/doi/abs/10.2136/vzj2012.0153.

Riva, M., Panzeri, M., Guadagnini, A., Neuman, S.P., 2011. Role of model selection criteria in geostatistical inverse estimation of statistical data-and model-parameters. Water Resour. Res. 47 (7).

Russian, A., Dentz, M., Gouze, P., 2017. Self-averaging and weak ergodicity breaking of diffusion in heterogeneous media. Phys. Rev. E 96, 022156. http://dx.doi.org/10.1103/PhysRevE.96.022156, URL https://link.aps.org/doi/10.1103/PhysRevE.96.022156.

Schiavo, M., Riva, M., Guadagnini, L., Zehe, E., Guadagnini, A., 2022. Probabilistic identification of preferential groundwater networks. J. Hydrol. 610, 127906. http://dx.doi.org/10.1016/j.jhydrol.2022.127906, URL https://www.sciencedirect.com/science/article/pii/S0022169422004814.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (4), 583–639.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van der Linde, A., 2014. The deviance information criterion: 12 years on. J. R. Stat. Soc. Ser. B Stat. Methodol. 76 (3), 485–493.

Tidwell, V.C., Wilson, J.L., 1997. Laboratory method for investigating permeability upscaling. Water Resour. Res. 33 (7), 1607–1616. http://dx.doi.org/10.1029/97WR00804, arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/97WR00804.

Tidwell, V.C., Wilson, J.L., 1999. Permeability upscaling measured on a block of Berea sandstone: Results and interpretation. Math. Geol. 31 (7), 749–769.

Tidwell, V.C., Wilson, J.L., 2002. Visual attributes of a rock and their relationship to permeability: A comparison of digital image and minipermeameter data. Water Resour. Res. 38 (11), 43–1–43–13. http://dx.doi.org/10.1029/2001WR000932, arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2001WR000932.

Todini, E., 2001. Influence of parameter estimation uncertainty in Kriging: Part 1-theoretical development. Hydrol. Earth Syst. Sci. 5 (2), 215–223.

Vapnik, V., Chervonenkis, A.Y., 1974. The method of ordered risk minimization, i. Avtomat. I Telemekh. 8, 21–30.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods 17, 261–272. http://dx.doi.org/10.1038/s41592-019-0686-2.

Webster, R., Oliver, M., 1993. How large a sample is needed to estimate the regional variogram adequately? In: Geostatistics Tróia'92. Springer, pp. 155–166.

Zehe, E., Becker, R., Bárdossy, A., Plate, E., 2005. Uncertainty of simulated catchment runoff response in the presence of threshold processes: Role of initial soil moisture and precipitation. J. Hydrol. 315 (1), 183–202. http://dx.doi.org/10.1016/j.jhydrol.2005.03.038, URL https://www.sciencedirect.com/science/article/pii/S0022169405001873.

Zehe, E., Blöschl, G., 2004. Predictability of hydrologic response at the plot and catchment scales: Role of initial conditions. Water Resour. Res. 40 (10).

Zehe, E., Loritz, R., Edery, Y., Berkowitz, B., 2021. Preferential pathways for fluid and solutes in heterogeneous groundwater systems: self-organization, entropy, work. Hydrol. Earth Syst. Sci. 25 (10), 5337–5353. http://dx.doi.org/10.5194/hess-25-5337-2021, URL https://hess.copernicus.org/articles/25/5337/2021/.

Zimmermann, B., Zehe, E., Hartmann, N.K., Elsenbeer, H., 2008. Analyzing spatial data: An assessment of assumptions, new methods, and uncertainty using soil hydraulic data. Water Resour. Res. 44 (10), 1–18. http://dx.doi.org/10.1029/2007WR006604.