Earth and Space Science

RESEARCH ARTICLE
10.1029/2020EA001152

A Framework for the Cross-Validation of Categorical Geostatistical Simulations

Przemysław Juda[1], **Philippe Renard**[1], **and Julien Straubhaar**[1]

[1]Centre for Hydrogeology and Geothermics, University of Neuchâtel, Neuchâtel, Switzerland

Key Points:
- Cross-validation framework is developed for testing simulations of categorical variables
- The methodology is generic, and competing models are based on a single score
- It requires a set of observation points and can be used with any spatial simulation method

Correspondence to:
P. Juda,
przemyslaw.juda@unine.ch

Citation:
Juda, P., Renard, P., & Straubhaar, J. (2020). A framework for the cross-validation of categorical geostatistical simulations. *Earth and Space Science*, *7*, e2020EA001152. https://doi.org/10.1029/2020EA001152

Received 24 FEB 2020
Accepted 28 JUN 2020
Accepted article online 15 JUL 2020

Abstract The mapping of subsurface parameters and the quantification of spatial uncertainty requires selecting adequate models and their parameters. Cross-validation techniques have been widely used for geostatistical model selection for continuous variables, but the situation is different for categorical variables. In these cases, cross-validation is seldom applied, and there is no clear consensus on which method to employ. Therefore, this paper proposes a systematic framework for the cross-validation of geostatistical simulations of categorical variables such as geological facies. The method is based on K-fold cross-validation combined with a proper scoring rule. It can be applied whenever an observation data set is available. At each cross-validation iteration, the training set becomes conditioning data for the tested geostatistical model, and the ensemble of simulations is compared to true values. The proposed framework is generic. Its application is illustrated with two examples using multiple-point statistics simulations. In the first test case, the aim is to identify a training image from a given data set. In the second test case, the aim is to identify the parameters in a situation including nonstationarity for a coastal alluvial aquifer in the south of France. Cross-validation scores are used as metrics of model performance and quadratic scoring rule, zero-one score, and balanced linear score are compared. The study shows that the proposed fivefold stratified cross-validation with the quadratic scoring rule allows ranking the geostatistical models and helps to identify the proper parameters.

1. Introduction

When modeling heterogeneous media, the choice of a suitable geostatistical approach is often a challenge. In the case of categorical fields (e.g., geological facies), many approaches are available, such as sequential indicator simulations (SIS), T-Progs, truncated gaussian and pluri-gaussian simulations, multiple-point statistics (MPS), object-based models, and genetic and pseudogenetic approaches (see, e.g., Chiles & Delfiner, 2012; Pyrcz & Deutsch, 2014, for a broad presentation of the methods).

Moreover, results obtained by all estimation or simulation methods depend on the model and computational parameters in a complex manner. A simple and powerful tool for statistical model selection is cross-validation: It consists of removing some data and comparing them to the predictions generated by the model. The first rigorous treatment of cross-validation was developed by Stone (1974) and Geisser (1974) simultaneously. In the former work, the term "cross-validation" appeared for the first time and was presented as the leave-one-out technique, where one data point is left out at a time and compared to the prediction; the procedure is then repeated for all points. Geisser (1975) generalized the idea of cross-validation to leaving out several points at a time. This variant of cross-validation has been later also called "multifold" (e.g., by Zhang, 1993), and currently some refer to it as "v-fold" (see, e.g., Arlot & Celisse, 2010), but the most commonly used term is "K-fold cross-validation" (Hastie et al., 2009). It consists in partitioning the data into K subsets, one subset used for testing at a time. Another simpler technique is random subsampling, also called hold-out sampling. In this method, a random subset of the data, the hold-out test set, is removed from the data set. The model is trained on the remaining part and its performance measured with the test set.

Currently, K-fold cross-validation is the technique most often used for model selection in classification problems. One of the early examples includes the work of Breiman et al. (1984). Breiman and Spector (1992) showed that leave-one-out performs worse than fivefold for model selection, which is comparable to bootstrap but less expensive computationally. Kohavi (1995) provided some more insight into different cross-validation techniques by comparing performance with the different number of folds, leave-one-out, and bootstrap methods. He found that tenfold cross-validation is a better choice even if the leave-one-out

©2020. The Authors.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

technique is computationally available, with fivefold found performing nearly as good as tenfold. He also suggested the use of stratified cross-validation: Each of the subsets should have roughly the same proportion of classes as the whole data set. Rodriguez et al. (2010) performed a study of the sensitivity of K-fold and concluded that fivefold or tenfold should be used. The goal of cross-validation is to estimate prediction error on unknown data (Hastie et al., 2009). That is why it repeats the train test split and averages the obtained errors: It strives to measure prediction error when using the whole available data set. For measuring the error, either a loss function is used, penalizing wrong predictions (the lower loss the better), or a scoring function is used, rewarding correct predictions (the higher the better). Typically, a loss function compares a single predicted value with a single true value and averages the mismatch over for multiple samples.

In a probabilistic framework, predictions take the form of predictive probability distributions, and therefore they need appropriate scoring rules which assign a score based on the predictive distribution and single true value (Gneiting & Raftery, 2007). Scoring rules assess two important features of probabilistic forecasts: sharpness and calibration. Sharpness is related to the concentration of the predictive distribution and quantifies if probabilities are spread to different values or concentrated on few values. Calibration describes statistical consistency between the distributions and observed values.

Proper and strictly proper scoring rules are an important class of the scoring rules. Let $P(X)$ be the predictive distribution and x the true (observed) value. We suppose that $Q(X)$ is the true distribution from which sample x was drawn. A reward of a forecaster is given by the score $S(P,x)$. Let us use $S(P,Q)$ for expected value of $S(P,\cdot)$ under Q. The *strictly proper* scoring rule is such that $S(Q,Q) \geq S(P,Q)$ with equality if and only if $P = Q$. A *proper* scoring rule satisfies $S(Q,Q) \geq S(P,Q)$ for all P and Q. Proper scoring rules encourage honest predictions: When the best estimate of P predicted by a forecaster is Q, the strategy to achieve the best score is to use the distribution Q. The forecaster has no interest in modifying Q, as it will not result in a better score.

In geostatistics, the first applications of cross-validation were mentioned by David (1976) and Delfiner (1976). Although Dubrule (1983) generalized cross-validation for kriging in the unique neighborhood case for large data sets, the cross-validation term has been used as a synonym of the leave-one-out technique. An interesting alternative cross-validation technique is the orthonormal residuals as introduced by Kitanidis (1991). In this technique, the data are ordered, and starting with one point, the consecutive points are predicted one by one and then added to the conditioning data set. The standardized residuals (residuals divided by kriging variances) are computed at each step, and their statistics are investigated to validate the model.

Cross-validation methods have been well established and extensively used for continuous variables and variogram-based methods (see, e.g., textbooks of Chiles & Delfiner, 2012; Cressie, 1993), but there is no similar consensus yet for the methods that should be used for categorical variables even if different applications of model testing techniques in the categorical case have been published (e.g., Allard et al., 2012; Madani et al., 2018).

In the framework of multiple-point statistics (MPS), the question of the training image (TI) selection and parameter identification has been treated using a wide range of methods mainly designed to compare some characteristics of the training image with the simulations. The question of the quality check of MPS simulations is discussed in detail in Chapter 8 of Mariethoz and Caers (2015). For example, Boisvert et al. (2007) compared the distribution of runs between the training image and the simulations. Pérez et al. (2014) focused on the frequency of patterns found in the conditioning data and the training image. Tan et al. (2014) compared the patterns of the training image with those of the simulations at different scales. Rongier et al. (2016) quantified the mismatch of connectivity and geometrical metrics between the TI and the simulations. Feng et al. (2017) used a minimal distance between the data event found in the TI and in the conditioning data. Based on any of these metrics, one can derive automated parameter selection methods (Baninajar et al., 2019; Dagasan et al., 2018). However, in real case applications the problem is not necessarily to reproduce accurately the patterns found in a training image because this image is not known precisely and is itself a parameter to be identified from the conditioning data. Al-Mudhafar (2018) mentions that he uses leave-one-out and split-sample approaches for different MPS realizations of a fluvial environment. The methods of Pérez et al. (2014) and Feng et al. (2017) are the only ones allowing to identify the training image. However, they assume that the simulation is stationary, while this is rarely the case for practical applications in which there are different trends, for example, in orientations, proportions of the facies, or even types of

patterns. We, therefore, argue that a more realistic strategy is to apply a generic cross-validation technique to identify the training image and all the other parameters when using an MPS model.

To address the issues described above, this paper aims to present a generic methodology based on K-fold cross-validation for the categorical case. It allows ranking spatial simulation methods given some observation points. The technique is based on the mean quadratic score (also called Brier score) and is especially suitable for assessing probabilistic outcomes of a simulation method. The application of the methodology is illustrated in an MPS framework, but the approach can be used with any categorical simulation method honoring conditioning data. For demonstrating the performance of the method, we show a benchmark example of training image selection and parameter selection (including the TI as one of the parameters) in a realistic nonstationary case.

2. Cross-Validation Methodology

This section presents a cross-validation methodology for geostatistical simulations. We suppose that with the simulation method, N observations are available: They are pairs of $(\mathbf{x}_n, y_n)$, $n = 1,...,N$, where $\mathbf{x}_n$ are vectors of spatial coordinates of the nth observation and $y_n \in \{1,...,M\}$ is the facies index observed at point $\mathbf{x}_n$. M is the number of possible facies. Let us denote with $\mathscr{K}$ the set of all available observations, $\mathscr{K} = \{(\mathbf{x}_n, y_n),\ n = 1...N\}$.

In this section, scoring rules for probabilistic forecasts will be introduced. They allow quantifying the performance of a stochastic method when resimulating known values. Then, K-fold cross-validation methodology will be reviewed and adapted to spatial data sets.

2.1. Scoring Rules

Scoring rules aim to quantify the quality of a probabilistic forecast by comparing it with a single true value. In our setting, repeating stochastic geostatistical simulations yields a probabilistic forecast of a geological facies for each point in the simulation domain. Some points in the simulation domain can be compared with the true, observed facies value. The observed facies value is a value from the set $\{1,...,M\}$, and the probabilistic forecast is a probability vector $\mathbf{p} = (p_1,...,p_M)$, where each vector element $p_j, j = 1,...,M$ describes probabilities of different facies.

We consider scoring rules S in the form of a collection of M functions:

$$S(\cdot, i): \mathscr{P}_M \mapsto \mathbb{R}, \qquad i = 1, ..., M, \tag{1}$$

where $\mathscr{P}_M = \{\mathbf{p} = (p_1, ..., p_M): p_1, ..., p_M \geq 0,\ p_1 + ...+p_M = 1\}$ is the probability forecasts space where p_j, $j = 1,...,M$ is the forecast probability of the outcome j and i is the index of the observed value (facies).

The scoring rules apply to a single observation, but in practice they are aggregated, and forecasts are ranked using average scores:

$$\overline{s}(\mathscr{S}) = \frac{1}{|\mathscr{S}|} \sum_{s \in \mathscr{S}} s, \tag{2}$$

where $\mathscr{S}$ is the set of scores, and $|\mathscr{S}|$ is the cardinal of $\mathscr{S}$.

It is also possible to compute the average class score or balanced mean score. Such a balanced score generalizes average class accuracy (Kelleher et al., 2015), also called balanced accuracy (Brodersen et al., 2010), used in machine learning for assessing a classifier's performance. It helps correct too optimistic estimates of classifier performance given by average accuracy and is especially useful when dealing with imbalanced data sets.

Let $\mathscr{M}$ be the set of observed facies (classes) used to compute the set of scores $\mathscr{S}$. Now let us denote $\mathscr{S}^m, m \in \mathscr{M}$, the subset of scores where the observed facies was m: $\mathscr{S}^m = \{s \in \mathscr{S},\ \text{observed value is}\ \ m\}$. The balanced mean score gives the same importance to each facies in the data set and is defined as the arithmetic mean of the average class scores, that is,

$$\hat{s}(\mathcal{S}, \mathcal{M}) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{1}{|\mathcal{S}^m|} \sum_{s \in \mathcal{S}^m} s. \tag{3}$$

2.1.1. Quadratic Score

The quadratic (or Brier) score was first introduced as a measure of quality for meteorological forecasts (Brier, 1950). It is a strictly proper scoring rule (Gneiting & Raftery, 2007) given by

$$S(\mathbf{p}, i) = -\sum_{j=1}^{M} (\delta_{ij} - p_j)^2 = 2p_i - \sum_{j=1}^{M} p_j^2 - 1,$$

with $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. The quadratic score values are between -2 and 0, the higher the better. An ideal forecaster, the one predicting the correct outcome with the probability of 1 would score 0. The worst forecaster is the one attributing probability of 1 to a wrong class and scores -2. A better forecaster which makes fewer systematic errors, for example, spreads probability equally to wrong classes, would have a better score than the one who prefers one wrong class to the others.

As Brier (1950) pointed out, the score encourages the forecaster to get the prediction exactly right. A sharp and correct prediction would score close to 0. On the other hand, the forecaster should state unbiased estimates of probability when not able to forecast perfectly. The strategy to predict the most frequent facies with certainty is penalized in comparison to unbiased strategy (referred to as climatological in weather forecasting), which just learns the probabilities from the training set. In this way, the quadratic score encourages calibrated forecasts.

2.1.2. Zero-One Score

Zero-one score is a proper but not a strictly proper scoring rule (Gneiting & Raftery, 2007) given by

$$S(\mathbf{p}, i) = \begin{cases} 1/ & |modes(\mathbf{p})| \quad \text{if } i \in modes(\mathbf{p}) \\ 0 & \text{otherwise,} \end{cases}$$

where $modes(\mathbf{p}) = \{i : p_i = \max_{j=1,\dots,M} p_j\}$ is the set of modes of $\mathbf{p}$. Mean zero-one score values are between 0 and 1 and are easily interpreted: The mean zero-one score gives a fraction of observed points (geological facies), which were given the highest probability by the simulation method.

2.1.3. Linear Score

A generalization of balanced accuracy for probabilistic forecasts is the balanced mean linear score with the linear score given by

$$S(\mathbf{p}, i) = p_i.$$

This score simply attributes the probability of the true value. The advantage of the mean linear score is its intuitive interpretation: It indicates how frequently simulations predict the correct value. It corresponds to the proportion of the correct forecast and is often used in machine learning to compare several methods. However, the linear score has some undesirable properties (Selten, 1998): It does not encourage fair predictions and is neither proper nor strictly proper. While the balanced linear score is still not proper, it forces the forecaster to try to represent rare facies (by increasing their share in the score) and remains easy to interpret.

2.1.4. Unbiased Classifier

It is often useful to compare scores of geostatistical simulations with scores of a simple reference method. As already mentioned when discussing the quadratic score, in the field of weather forecasting, the climatological forecaster serves often as a reference. It takes into account averaged historical values. In a broader sense, climatological forecasts are calibrated forecasts but lack sharpness (Gneiting & Raftery, 2007). Such a reference simulation method for spatial data would be the unbiased classifier, which looks at the proportions of facies in the complete training set and uses them to estimate the vector of probabilities (which is constant over the whole domain). The score obtained by the unbiased classifier will be referred to as the reference score.

If stratified K-fold is used (explained in the next subsection), the reference balanced mean linear score is equal to $1/M$. It is another property of the balanced mean linear score that makes it intuitive. For the mean quadratic score and mean zero-one score, the reference score will not only depend on the number of different facies M but also on the proportions of the facies in the data set.

2.2. K-Fold Cross-Validation

K-fold cross-validation consists of dividing the data set $\mathscr{K}$ into K subsets of equal sizes and performing K iterations: In each iteration, one subset is removed from $\mathscr{K}$ and becomes the validation set, while the complementary set (rest of the data) forms the training set used as conditioning data for the geostatistical method. To describe the split, we can define the partition function

$$\kappa(\cdot):\{1, ..., N\}\mapsto\{1, ..., K\}$$

that maps each point's index n to subset (iteration) index $\kappa(n)\in\{1,...,K\}$, $n = 1,...,N$. We define K disjoint sets $\mathscr{K}_1, ..., \mathscr{K}_K$:

$$\mathscr{K}_k = \{(\mathbf{x}_n, y_n):\kappa(n) = k, \quad n = 1, ..., N\}, \quad k = 1...K.$$

The union of these sets is the data set $\mathscr{K}$. The partition should be made in such a way that $|\mathscr{K}_1| = ... = |\mathscr{K}_K|$. Since geological data sets are often imbalanced (e.g., proportions of facies are strongly different, rare facies are present), it is important to use stratified cross-validation. In stratified cross-validation, the data set is split into subsets that have the same proportion of classes (facies). This translates to the following condition: $|\mathscr{K}_1^m| = ... = |\mathscr{K}_K^m|$, $m = 1,...,M$, where $\mathscr{K}_k^m$ denotes the subset of $\mathscr{K}_k$ containing only samples of category m:

$$\mathscr{K}_k^m = \{(\mathbf{x}_n, y_n):\kappa(n) = k, \quad y_n = m, \quad n = 1, ..., N\}, \quad k = 1, ..., K.$$

Moreover, if observation points are not spatially correlated, the split should be made in a random way, for instance by shuffling (randomly reordering) the data first, as depicted in Figure 1.

The training data set is the complementary data set:

$$\overline{\mathscr{K}}_k = \{(\mathbf{x}_n, y_n):\kappa(n) \neq k, \quad n = 1, ..., N\}, \quad k = 1, ..., K,$$

and it becomes the conditioning data for the geostatistical method (Figure 2). For each iteration $k = 1,...,K$, the geostatistical method produces probability vectors for each point in the validation set $\mathscr{K}_k$ and uses $\overline{\mathscr{K}}_k$ as conditioning data. Let $\hat{f}$ be the geostatistical estimator $\hat{f}(\cdot, \overline{\mathscr{K}}_k):\mathbb{R}^3\to\mathscr{P}_M$ mapping a coordinate vector to a probability vector, given $\overline{\mathscr{K}}_k$ as the conditioning set. We define the prediction $\mathbf{p}_n$ at the nth point:

$$\mathbf{p}_n = \hat{f}(\mathbf{x}_n, \overline{\mathscr{K}}_{\kappa(n)}), \; n = 1, ..., N. \tag{4}$$

We note here that in the case of stochastic simulation methods, it will be necessary to repeat the simulations at each iteration with the same conditioning data to obtain probabilities of categorical variables (geological facies) at each validation point (Figure 3).

Then a scoring function S as defined in Equation 1 is applied to all points in the validation set, and the mean score is computed either using mean (Equation 2) or balanced mean (Equation 3). The procedure is repeated for each of the K iterations so that each subset becomes a validation set once. The mean cross-validation score CV is the average of the mean scores over all iterations. If standard mean is used,

$$CV = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\mathscr{K}_k|}\sum_{(\mathbf{x},y)\in\mathscr{K}_k} S(\hat{f}(\mathbf{x}, \overline{\mathscr{K}}_k), y),$$

and in the case of balanced mean it becomes

$$CV = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{M_k}\sum_{\substack{m=1\\ |\mathscr{K}_k^m|\neq 0}}^{M_k}\frac{1}{|\mathscr{K}_k^m|}\sum_{(\mathbf{x},y)\in\mathscr{K}_k^m} S(\hat{f}(\mathbf{x}, \overline{\mathscr{K}}_k), y),$$

with M_k number of different facies in $\overline{\mathscr{K}}_k$.

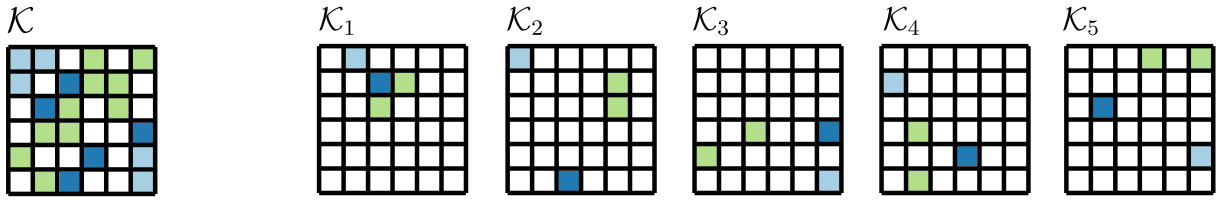

Figure 1. Example of shuffled fivefold stratified split of data ($\mathscr{K}$) in a regular grid belonging to three classes (light blue, dark blue, and green).

We can now summarize the K-fold cross-validation framework in the form of an algorithm. It takes the following as input: K—number of cross-validation iterations; $\mathscr{K}$—data set of pairs ($\mathbf{x}$,y); S—scoring rule; *balance*—TRUE or FALSE indicating whether the balanced mean should be used; n_r—number of stochastic simulation runs needed to construct the probability vector $\mathbf{p}$ (Equation 4). The algorithm is as follows:

CROSS-VALIDATE($K, \mathcal{K}, S, balance, n_r$)
1 Shuffle randomly the data in $\mathcal{K}$
2 Divide $\mathcal{K}$ into K subsets $\mathcal{K}_1, \ldots, \mathcal{K}_K$
3 **for** $k = 1$ **to** K
4 Use $\bar{\mathcal{K}}_k$ (training set) as conditioning data
5 Run geostatistical simulation n_r times
6 $\mathcal{S} = \emptyset, \mathcal{M} = \emptyset$
7 **for** $(\mathbf{x}, y) \in \mathcal{K}_k$ // for each point in validation set
8 // Construct probability vector $\mathbf{p}$ and compute the score
9 $\mathbf{p} = \hat{f}(\mathbf{x}, \bar{\mathcal{K}}_k)$
10 $\mathcal{S} = \mathcal{S} \cup \{S(\mathbf{p}, y), y\}$
11 $\mathcal{M} = \mathcal{M} \cup \{y\}$
12 **if** *balance* is TRUE
13 $s_k = \hat{s}(\mathcal{S}, \mathcal{M})$
14 **else**
15 $s_k = \bar{s}(\mathcal{S})$
16 **return** $\frac{1}{K} \sum_{k=1}^{K} s_k$

We used $\hat{s}$ to refer to mean balanced score (Equation 3) and $\bar{s}$ to mean score (Equation 2). The cross-validation requires n_r geostatistical model runs per iteration, thus Kn_r model runs in total.

3. Case Studies

The proposed cross-validation methodology can be used with any stochastic simulation method, but it will be tested on two MPS cases. Therefore, we also introduce the basic notions of MPS and the Direct Sampling

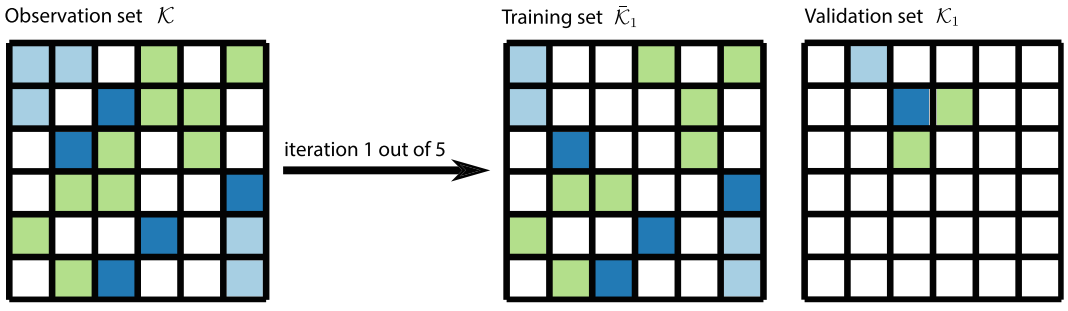

Figure 2. The split of the example data ($\mathscr{K}$) in a regular grid in the first cross-validation iteration. The first subset ($\mathscr{K}_1$) becomes the validation set, and the complementary data set (training set, $\overline{\mathscr{K}}_1$) becomes the conditioning data for the geostatistical model.

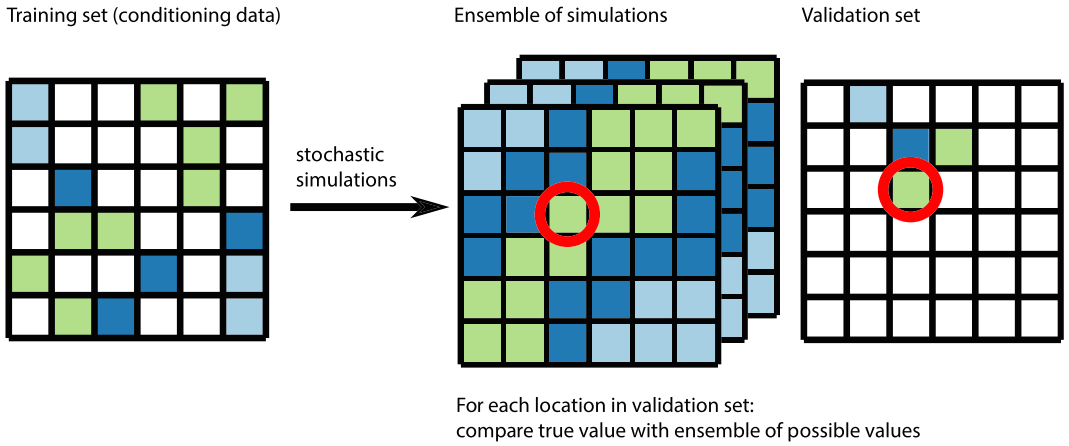

Figure 3. Example of one cross-validation iteration. The domain is simulated n_r times with conditioning data (ensemble of simulations is obtained). Each point in the validation set is compared with vector p constructed by aggregating realizations in the ensemble.

algorithm. The two examples are as follows: a training image selection problem and a parameter selection problem. Both are 2-D categorical synthetic setups.

3.1. Direct Sampling

MPS algorithms use a conceptual geological model, provided as a training image (TI). The TI is an implicit database of patterns and an example of how the simulated field should look like. Direct Sampling (Mariethoz et al., 2010) is a point-based method and constructs fields by filling the regular simulation grid point by point. It honors the conditioning data by simply putting them in the simulation grid, It can perform multivariate simulations and supports transformations, such as field rotations or affinities (Mariethoz & Kelly, 2011). In this work, DeeSse implementation of Direct Sampling is used (Straubhaar et al., 2020).

Direct Sampling is controlled by three main parameters: number of nearest neighbors (n), distance threshold (t), and maximal fraction scan (f). These parameters are crucial for the quality of simulated fields and the computing time. Their tuning may be challenging, since the results depend on the complexity of the TI as well as the interaction between the parameters and the patterns in a complicated manner (Meerschman et al., 2013). The number of nearest neighbors limits how many pixels are included in a pattern during the pattern search. Typical values range from several to 100 or more. As a rule of thumb, the more neighbors, the better the quality of simulations but also the higher the computational cost. The distance threshold specifies the maximum acceptable dissimilarity between the conditioning pattern and patterns found in the TI. It can range from 0 to 1. The value 0 means that only a perfect match is accepted, while with the value of 1, every pattern is suitable. This parameter typically ranges from 0.01 to 0.1. The maximal scan fraction indicates what fraction of the training image can be potentially scanned before the search is stopped (which can happen if no sufficiently good match was found: In such a case the best candidate found so far is accepted). The value of 1 means that the whole TI can be scanned, and a value close to 0 would mean that the scan is stopped after scanning only one node. The scan fraction helps avoid the verbatim copy of the TI and limit the computation time.

3.2. Benchmark Setup for Training Image Selection

The first example is a benchmark training image selection problem, which was first published by Pérez et al. (2014) and also used in the work of Feng et al. (2017). In this setup, there are three training images with different features: ellipsoids, sine waves, and vertical stripes (Figure 4). Pérez et al. (2014) used the same training image generator tool to construct both the reference training images and synthetic realities. In our setup a different approach for constructing synthetic realities was used. To allow more pattern variability, we first performed one unconditional DeeSse simulation with each of the TI and the following parameters: $n = 60$, $t = 0.05$, and $f = 0.25$ (Figure 5). Then, we sampled points from the synthetic realities with 10 different sampling rates ranging from 0.0025 to 0.16. In this way, we created 30 synthetic observation sets: 3 TI and 10

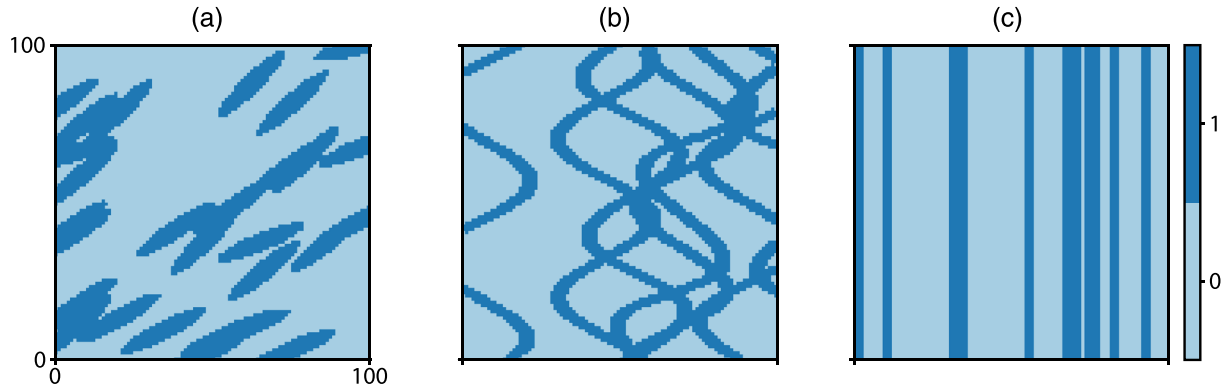

Figure 4. Three training images of type (a) ellipsoids, (b) waves, and (c) stripes used in the benchmark of training image selection (data from Pérez et al., 2014).

numbers of points ranging from 25 to 1,600. Figure 6 shows examples of such observation sets containing 1,600 points each. The sets corresponding to the different number of samples are sampled independently.

3.3. Roussillon Plain Synthetic Example

In the second example, an alluvial aquifer located in the Roussillon plain near Perpignan (France) is considered. It is a simplified 2-D version of the model build by Dall'Alba et al. (2020). The area is modeled considering four geological facies: river bed, crevasse splays, flood plains, and alluvial fans. The position of the facies in the basin is guided using a trend defined over the area (Figure 7a) as well as in the TI (Figure 8). The geological features are oriented according to the paleotopography of the region (Figure 7b). In order to test the methodology with a known reference, a synthetic reality was constructed by performing an unconditional DeeSse simulation with the following parameters: $n = 50$, $f = 0.5$, and $t = 0.05$. Then 50, 150, and 600 random samples were drawn from the area to form three synthetic observation sets (Figure 9). These numbers correspond to the number of pumping wells (50), wells with lithology information (150), and all wells (600) in the area. The example consists in selecting the best DeeSse parameters (including the TI) for each of the observations sets. Two candidate training images will be considered: the training image used to construct the observation set (reference TI, "true TI") and the analog training image (Figure 10).

4. Results

The higher the cross-validation score, the better the predictive power of the method. In our setting, the higher cross-validation score points to better simulation parameters or a more compatible training image. The results in this section were obtained using the stratified fivefold cross-validation. Three scores were compared for each run: quadratic (Brier) score, zero-one score, and balanced linear score. Since at each cross-validation iteration, simulations are repeated to construct the probability vector **p**, the technique has

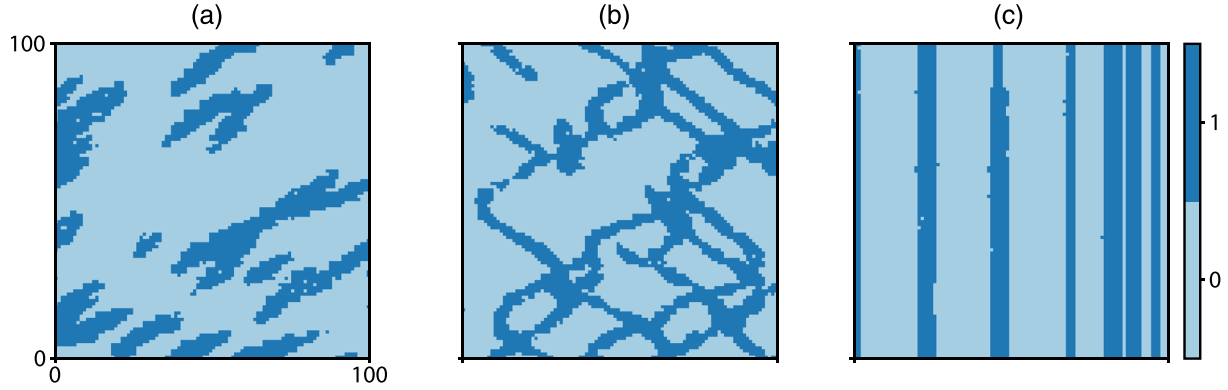

Figure 5. Three synthetic realities, each obtained by means of DeeSse simulation using training images (a, b, c).

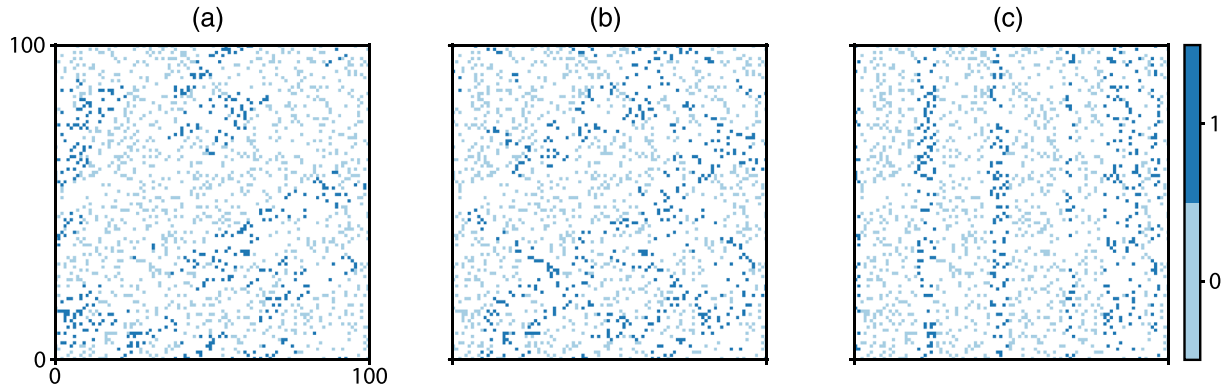

Figure 6. Example of synthetic observation sets obtained by sampling from each of the synthetic realities (a, b, c). Here each synthetic observation set contains 1,600 samples.

one hyperparameter, which was adjusted in the first test case: the number of realizations per experiment. It defines how many times a simulation is repeated to approximate the probability distribution of facies. The minimal number is 1, and it would correspond to a deterministic approach.

4.1. Training Image Selection

For each observation set, we performed the cross-validation using the original DeeSse parameters ($n = 60$, $t = 0.05$, and $f = 0.25$) and compared the cross-validation scores for each of the TI.

First, to check the influence of the number of realizations per experiment, we fixed the sampling rate to 0.005 and thus used the data sets of the three types with 50 samples each. For each of the data sets, we varied the number of realizations per iteration (ranging from 1 to 49) and recorded cross-validation scores using each of the Training Images a, b, and c and different scores: the mean quadratic score, the mean zero-one score, and the mean balanced linear score (Figure 11). The figure also reports reference scores. Even for a small number of realizations, the highest cross-validation scores are attributed to the correct training image, except for the Case b. For the quadratic scoring rule, a small number of realizations results in generally lower scores. This is expected since a small number of realizations implies that the probability densities **p** are estimated from a small sample and can result in a low score if predictions are incorrect. Adding more realizations gives more samples to estimate more accurately the spread of the probability distribution and the uncertainty resulting in a better quadratic score. In the case of the balanced linear score there is no such effect; in Case b we see an inverse trend: Score values slightly decrease with the number of realizations. This is probably related to the nature of the score, not sufficiently penalizing wrong predictions with high probabilities. The mean zero-one

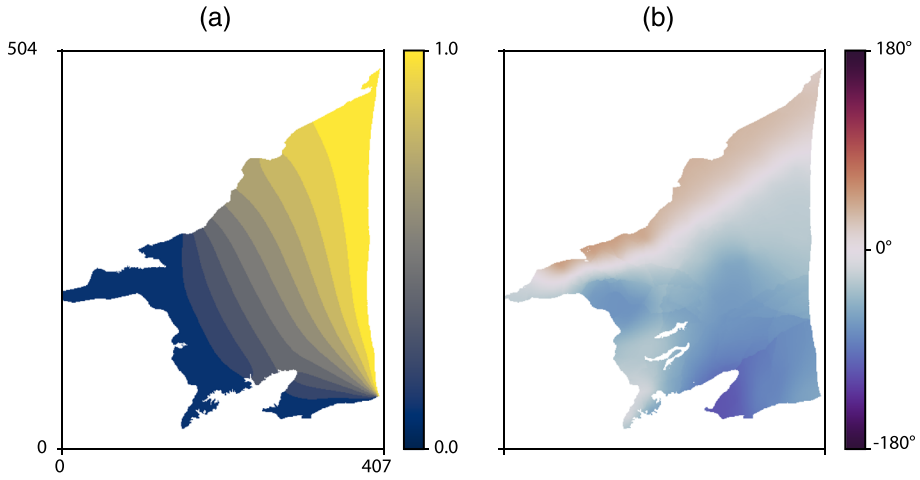

Figure 7. Area of the multivariate simulation with defined trend (a) and orientation (b).

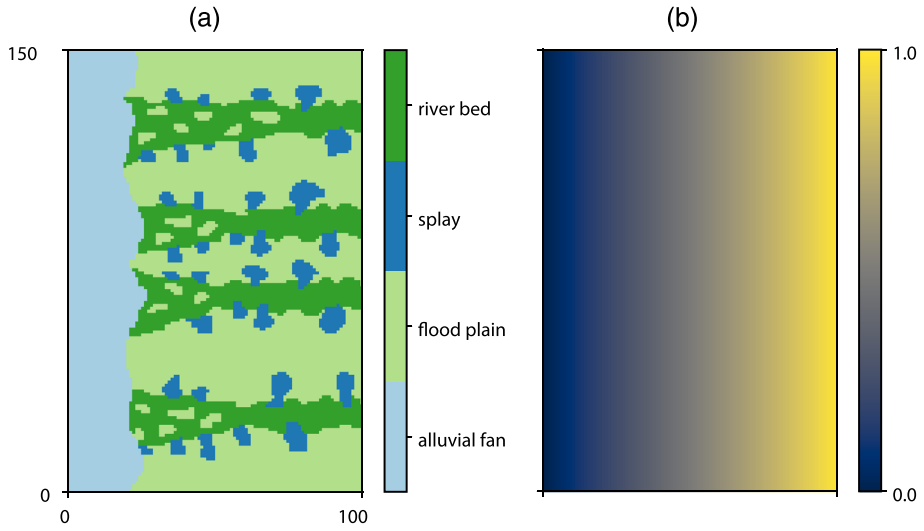

Figure 8. The reference training image (a) with the defined trend (b). The training image contains four geological facies.

score seems to be the most sensitive to the number of realizations among the three scoring methods. It provides scores that are less robust than the other methods and fluctuates erratically when the number of realizations is modified. This is not surprising because it accounts only for the maximum probability value and not for the complete probability density. Scores obtained using quadratic rule suggest that 30 realizations per iteration are sufficient to obtain robust results. Therefore, we fixed this parameter to this value for the remaining tests.

The Observation Sets a and c are characterized by a large difference between the best training image and the other training images. The most compatible TI has also a significantly higher score than the reference score. However, it does not apply to Case b. The mean quadratic score using the corresponding training image is around the reference value. Similar behavior is seen in the case of the zero-one score. The mean balanced score gives more optimistic results, higher values than the reference. While the mean quadratic score and the mean balanced linear score have correctly identified the most compatible training image in Case b, the low values (compared to the reference score) and the small difference between the different TIs suggest that the data set is too small for a reliable choice of the best training image.

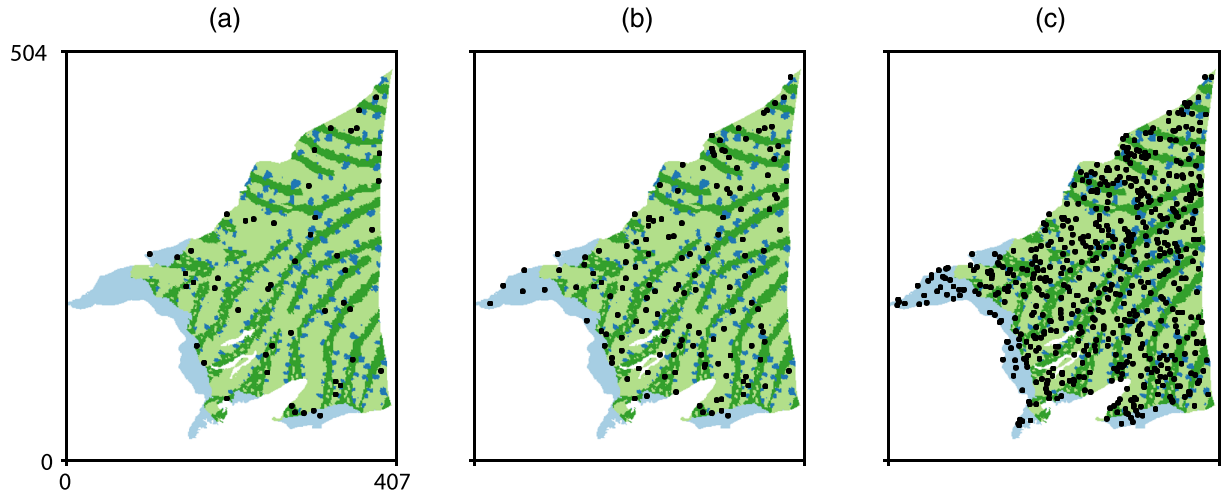

Figure 9. The synthetic reality obtained by DeeSse simulation using the reference TI and samples locations: (a) 50 wells, (b) 150 wells, and (c) 600 wells, representing synthetic observation data sets.

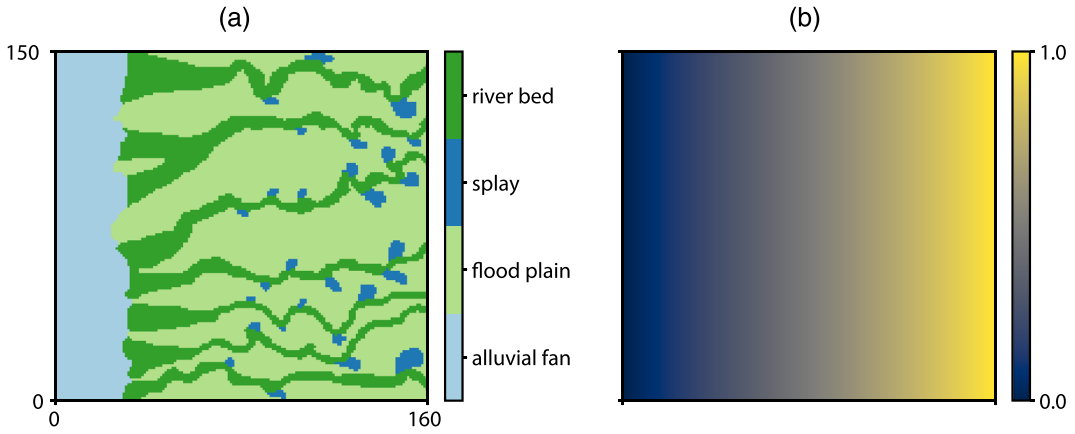

Figure 10. The analog training image (a) with its corresponding trend (b). The training image contains four geological facies.

Second, with the fixed number of realizations, cross-validation was run for each observation set. Figure 12 shows the mean quadratic scores, the mean zero-one scores, and the mean balanced linear scores for different numbers of samples in observation sets. The higher score corresponds to better TI compatibility. In all cases, the original ("true") training image was correctly identified by the mean quadratic score and the mean balanced linear score except for one data set: that of Type b with 25 samples. The mean zero-one score was not able to correctly identify the most compatible TI in the case of Type b with 25 and 50 samples. These results suggest that this sparse data set is not sufficient to identify the synthetic reality, and the mean quadratic scores close to the reference seem to confirm this statement. The mean zero-one and linear balanced scores attributed much higher values to the training image c for this observation set.

For larger observation sets, all scores tend to improve irrespectively of the TI. It might seem surprising at first but can be explained by the fact that all structures have some short-range continuity. Direct Sampling (and all interpolation or simulation methods in general) respects this short-range continuity. This results in better predictions for points in the vicinity of observed data, and there are more such points in larger observation sets and therefore the predictions are better.

4.2. Parameter Selection

To make the example closer to a real scenario, we consider the simulation parameters as unknown. The cross-validation scores were therefore computed with two candidate training images: the reference one (Figure 8) and the analog one (Figure 10) and for all combinations of the Direct Sampling parameters in Table 1. The number of realizations per iteration was fixed to 30. We note that the parameter set used to generate the synthetic reality is not present in the proposed set of combinations.

For each synthetic observation set (with 50, 150, and 600 points), cross-validation scores were obtained using three scoring methods: mean quadratic score, mean zero-one score, and mean balanced linear score. Table 2 presents the best scores for each observation set, for each TI and for each scoring method. Corresponding DeeSse parameters are also reported along with reference scores. The example DeeSse simulations with the corresponding best parameters are shown in Figure 13 (reference TI) and Figure 14 (analog TI).

In the case of 50 wells, the reference and the analog TIs received similar scores, only slightly better than the reference score (except for the optimistic balanced linear score). It suggests that this data set is too sparse to let us reliably choose the best DeeSse parameters. In the case of 150 wells, all the scores attributed the higher value to the reference training image. The mean quadratic score also is higher than in the case with 50 wells. In the case of 600 wells, all scores point to the reference training image and are significantly better than the reference score.

When looking at the example simulations with the reference training image with 50 and 150 wells (Figure 13), we observe that the parameters found by using the quadratic score and the zero-one rules result in realizations that better represent crevasse splays than those found by using the balanced linear scores.

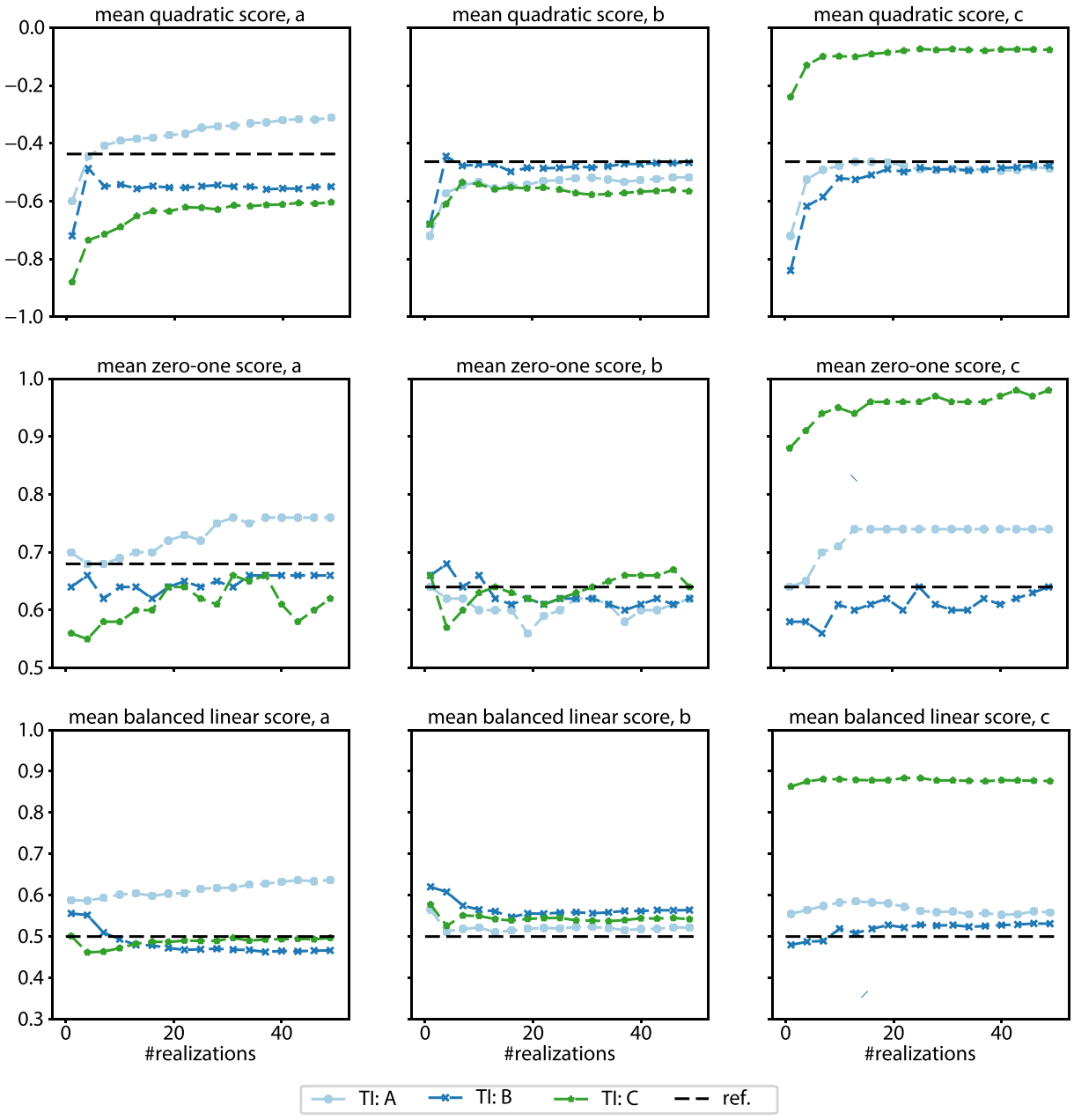

Figure 11. Sensitivity study with respect to the number of realizations per fold for each of the different types of observation sets (a, b, c). Each of the samples contained 50 observations. Mean quadratic score, mean zero-one score, and mean balanced linear score were used.

In the case of simulations with the analog training image (Figure 14), the resulting simulations are characterized by a poor reproduction of the structures of the TI. The parameters obtained from the cross-validation (a small scan fraction and a small number of neighbors) permit a high variability between the simulations but sacrifice the quality of the TI reproduction. Indeed, honoring the conditioning data requires departing from the TI which is not the one used to generate the synthetic reference.

These observations are confirmed by the analysis of Table 2. It shows that the best parameters for the simulation with the analog (wrong) training image and with a large conditioning data set correspond to values that typically provide simulations that do not reproduce precisely the patterns of the training image (e.g., the small number of neighbors). It results in more variability and uncertainty in the predictions which is favored by the quadratic score.

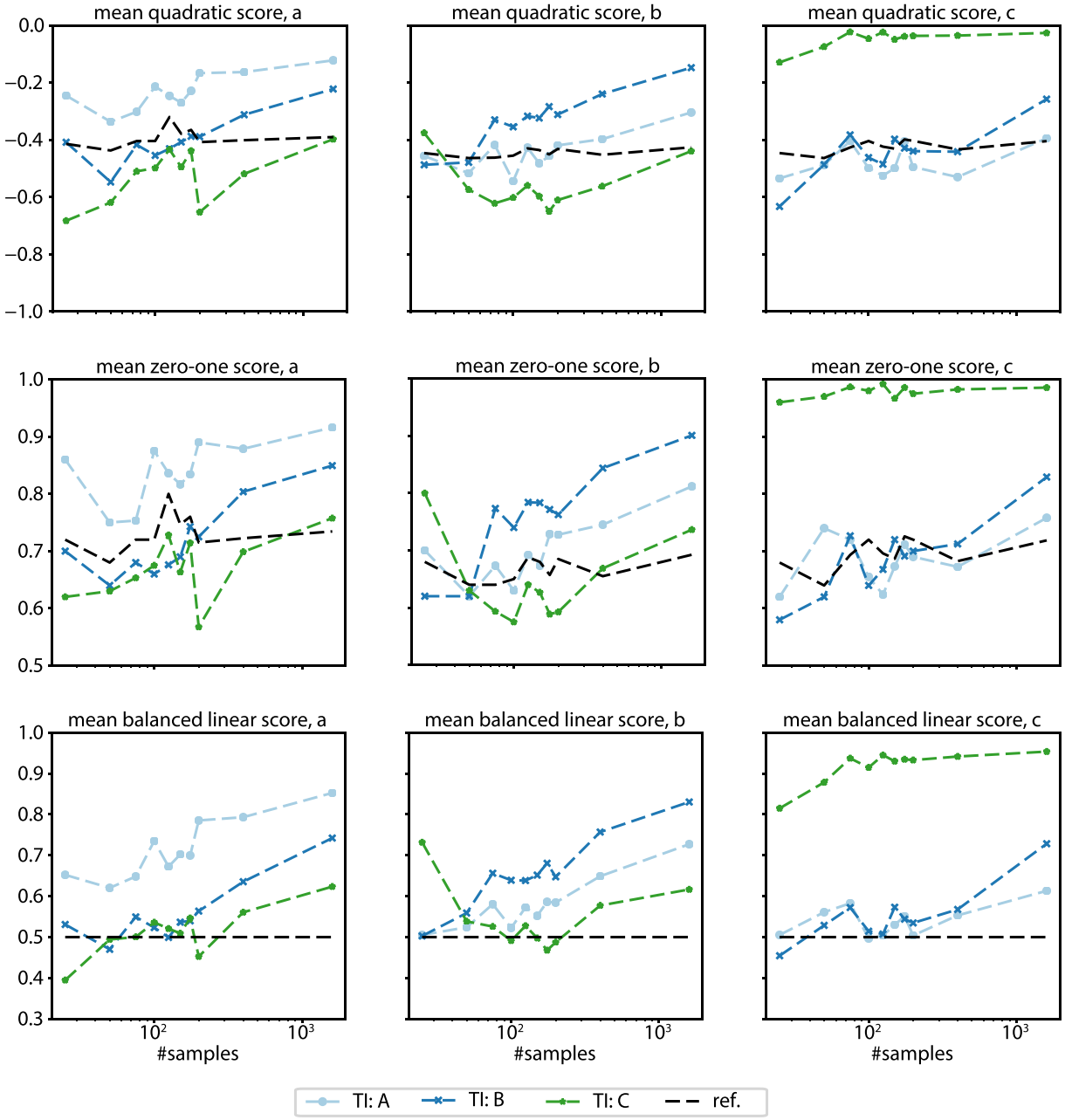

Figure 12. Cross-validation scores for synthetic observations with respect to the number of samples in the data set for each of the different types of observation sets (a, b, c). Thirty realizations per fold were generated. Mean quadratic score, mean zero-one score, and mean balanced linear score were used.

5. Discussion

The first test case shows that the methodology is able to identify correctly a training image if the amount of conditioning data is sufficient. Moreover, using the second test case, we demonstrated that the proposed methodology accounts properly for nonstationarity because it is only based on the conditioning data and the simulated values. The main novelty of our approach is that it does not compare the patterns of the training image directly with the observed data. Often the uncertainty in the training image can be important, and the best prediction can then require departing from the TI. This was already shown by Dagasan et al. (2018) but in a specific configuration, where observations were spaced regularly on a cartesian grid. The method

Table 1
DeeSse Parameters, Which Were Tested in the Parameter Selection for the Multivariate Simulation Example

Parameter	Values
f	0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.4, 0.8
n	40, 20, 10, 5
t	0.01, 0.05, 0.1, 0.2

Note. For each combination of these parameters with both training images a cross-validation score was computed.

Table 2
Best Cross-Validation Scores and Corresponding DeeSse Parameters for Each Observation Set (With Different Number of Wells), Each Candidate TI, and Each Scoring Method

Wells	Scoring Rule	TI	Score	Reference	t	f	n
50	quadratic	reference	−0.47	−0.50	0.05	0.200	40
50	quadratic	analog	−0.46	−0.50	0.05	0.200	40
50	zero-one	reference	0.72	0.68	0.05	0.400	40
50	zero-one	analog	0.72	0.68	0.05	0.020	40
50	linear	reference	0.48	0.27	0.05	0.020	40
50	linear	analog	0.48	0.27	0.05	0.005	40
150	quadratic	reference	−0.40	−0.51	0.05	0.400	40
150	quadratic	analog	−0.44	−0.51	0.01	0.400	20
150	zero-one	reference	0.73	0.67	0.05	0.800	40
150	zero-one	analog	0.71	0.67	0.01	0.050	10
150	linear	reference	0.53	0.25	0.10	0.200	40
150	linear	analog	0.48	0.25	0.10	0.020	5
600	quadratic	reference	−0.28	−0.58	0.01	0.400	20
600	quadratic	analog	−0.38	−0.58	0.01	0.050	10
600	zero-one	reference	0.82	0.60	0.01	0.400	20
600	zero-one	analog	0.74	0.60	0.01	0.050	10
600	linear	reference	0.63	0.25	0.05	0.400	20
600	linear	analog	0.55	0.25	0.01	0.050	10

Note. Corresponding best scores for optimal parameters are given.

proposed in this paper is more general. It not only tunes the parameters with respect to the data and not the training image (as in Baninajar et al., 2019) but also treats the pattern reproduction implicitly. Respecting the patterns in the data and allowing sufficient variability is essential to produce correct predictions. Recent developments (Abdollahifard et al., 2019) aim to quantify variability and pattern consistency but with respect to the TI, not to the data. The advantage of our approach is that we no longer need to compare realizations with the training image and define quality indicators as in the works of Meerschman et al. (2013) or Rongier et al. (2016). We can see that in the second test case when using the reference training image, the simulations obtained with optimal parameters are visually comparable to the synthetic reality, and patterns of the training image (like channels continuity) are well represented.

The computation time is easy to predict; the framework requires $n_r K$ model runs. For complex geological models, this cost can be significant. In such cases, cross-validation iterations should be parallelized, as they are independent. It is also possible to parallelize the computation of realizations per iteration. To further reduce computing time, a smaller number of realizations per iteration, n_r, can be chosen. It might be possible to rank models with n_r in the range from 1 to 5, but for reliable results, at least 10 realizations per iteration should be obtained. In this work, we used 30 realizations per fold, as models were relatively cheap to compute. It is widely accepted that fivefold or tenfold approach are the best choice (Kohavi, 1995; Rodriguez et al., 2010). It is possible that for sparse geological data sets other K are preferable, but we found in this paper that the fivefold approach performs reasonably well. Another way to reduce computing time is to simulate only points in the validation set (e.g., avoid generating the entire field). In such a case, additional tests should be made to assess the robustness of the method, since most of the geostatistical methods depend on the simulation path. Moreover, if a method can directly estimate the probability vector, without repeating the simulation, it will reduce computational time.

The same cross-validation strategy could be applied to continuous variables; in that case, the scoring function needs to be adapted. The continuous ranked probability score (CRPS) is a proper scoring rule and a counterpart of the quadratic (Brier) score for continuous variables (Gneiting et al., 2007). In the second test case, we used the grid search to evaluate all parameter combinations to find the optimal ones. Since the cross-validation score is a single value, it can be used as an objective function in any optimization algorithm, which could more efficiently explore the parameter space to find the best parameters of a geostatistical method.

6. Conclusion

In this paper, we propose a cross-validation framework that can be applied for ranking geostatistical stochastic simulation methods of categorical variables when an observation data set is available. The method can be used for various purposes, such as selecting the best parameter set, or the best training image, even when the simulation is not stationary. It can also be employed to compare the performances of different geostatistical algorithms.

We used a stratified fivefold approach with shuffling; our observation points were assumed to be not strongly correlated. In the case of strong correlations (e.g., a data set with many consecutive points along wells), it should be considered to group the points (e.g., per well) and then split the whole groups into cross-validation iterations.

The mean quadratic score should be used as the most reliable indicator of method performance. The mean zero-one score and balanced linear score are more intuitive, but only the mean quadratic score can correctly assess the sharpness and calibration of the model. It can also be compared to the reference score obtained by using the marginal proportions of the facies as

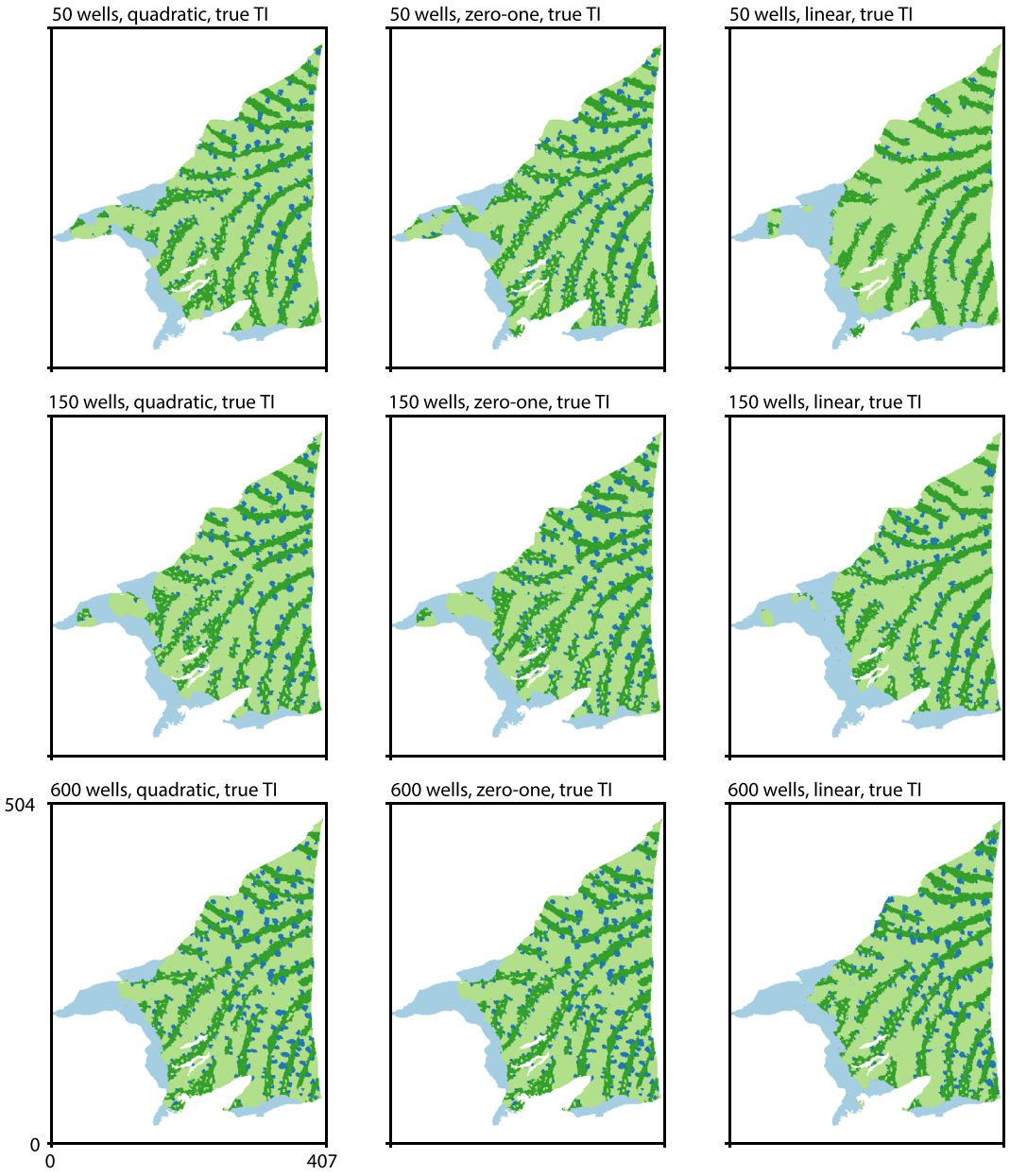

Figure 13. Example simulations with the reference TI, the best DeeSse parameters found with cross-validation for each observation set (50, 150, and 600 wells), and different scoring methods (quadratic score, zero-one score, and balanced linear score).

a predictor for all locations without accounting for the spatial correlation. A score lower than this reference indicates that the probabilities estimated by the model are biased.

The K-fold cross-validation framework is parsimonious in parameters. The number of realizations per iteration, n_r, controls the precision of results and the computation time; a value of n_r between 10 and 30 is suggested for generating a robust score.

The methodology can be applied to any stochastic simulation tool. While our examples used a regular cartesian grid, the cross-validation method can be applied to any grid: Only a set of spatial observations is required and an interface with a conditional simulation tool which returns the simulated value at given coordinates.

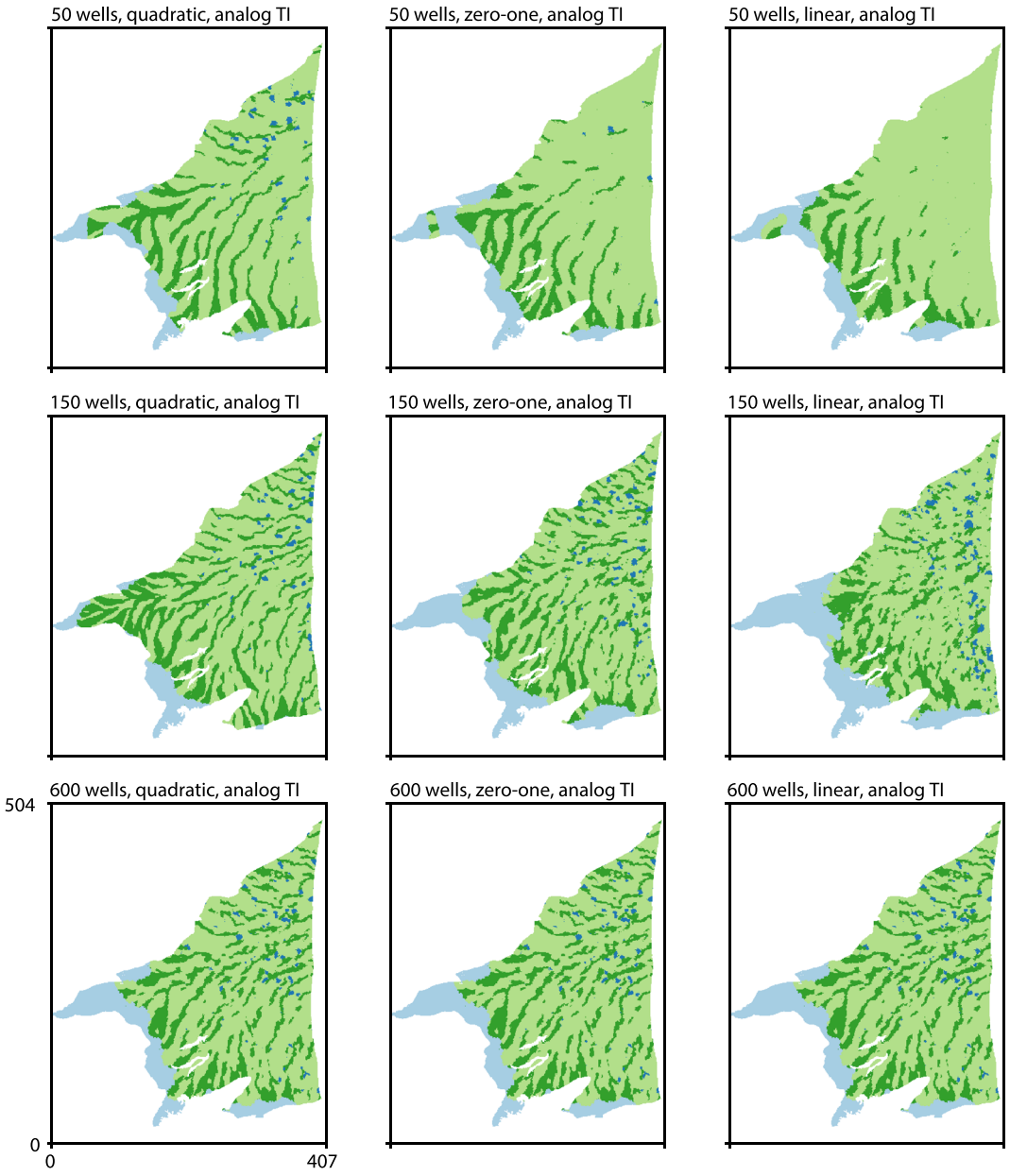

Figure 14. Example simulations with the analog TI, the best DeeSse parameters found with cross-validation for each observation set (50, 150, and 600 wells), and different scoring methods (quadratic score, zero-one score, and balanced linear score).

Data Availability Statement

All input data, the results, and the code used to generate them are available in the repository (https://doi.org/10.5281/zenodo.3901494).

Acknowledgments

We would like to thank Valentin Dall'Alba-Arnau for providing the Roussillon example data. This work was partly funded by the SNF Project 200020_182600.

References

Abdollahifard, M. J., Mariéthoz, G., & Ghavim, M. (2019). Quantitative evaluation of multiple-point simulations using image segmentation and texture descriptors. *Computational Geosciences*, *23*(6), 1349–1368.

Al-Mudhafar, W. J. (2018). Multiple-point geostatistical lithofacies simulation of fluvial sand-rich depositional environment: A case study from Zubair formation/South Rumaila oil field. *SPE Reservoir Evaluation & Engineering*, *21*(01), 39–53.

Allard, D., Comunian, A., & Renard, P. (2012). Probability aggregation methods in geoscience. *Mathematical Geosciences*, *44*(5), 545–581.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79.

Baninajar, E., Sharghi, Y., & Mariethoz, G. (2019). MPS-APO: A rapid and automatic parameter optimizer for multiple-point geostatistics. *Stochastic Environmental Research and Risk Assessment*, *33*(11-12), 1969–1989.
Boisvert, J. B., Pyrcz, M. J., & Deutsch, C. V. (2007). Multiple-point statistics for training image selection. *Natural Resources Research*, *16*(4), 313–321.
Breiman, L., Friedman, J. H., Stone, C. J., & Olshen, R. (1984). *Classification and regression trees*, *The Wadsworth Statistics/Probability Series*. Belmont, CA: Wadsworth.
Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression. The X-random case. *International Statistical Review / Revue Internationale de Statistique*, *60*(3), 291–319.
Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.
Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition* (pp. 3121–3124).
Chiles, J. P., & Delfiner, P. (2012). *Geostatistics: Modeling spatial uncertainty*, *Wiley Series in Probability and Statistics* (2nd ed.). Hoboken, NJ: Wiley.
Cressie, N. A. C. (1993). *Statistics for spatial data*, *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics* (Revised Ed.). New York: Wiley.
Dagasan, Y., Renard, P., Straubhaar, J., Erten, O., & Topal, E. (2018). Automatic parameter tuning of multiple-point statistical simulations for lateritic bauxite deposits. *Minerals*, *8*(5), 220.
Dall'Alba, V., Renard, P., Straubhaar, J., Issautier, B., Duvail, C., & Caballero, Y. (2020). 3D multiple-point statistics simulations of the Roussillon Continental Pliocene Aquifer using DeeSse. *Hydrology and Earth System Sciences Discussions*, *2020*, 1–23. https://doi.org/10.5194/hess-2020-96
David, M. (1976). The practice of kriging. In *Advanced geostatistics in the mining industry* (pp. 31–48). Springer.
Delfiner, P. (1976). Linear estimation of non stationary spatial phenomena. In *Advanced geostatistics in the mining industry* (pp. 49–68). Springer.
Dubrule, O. (1983). Cross validation of kriging in a unique neighborhood. *Journal of the International Association for Mathematical Geology*, *15*(6), 687–699.
Feng, W., Wu, S., Yin, Y., Zhang, J., & Zhang, K. (2017). A training image evaluation and selection method based on minimum data event distance for multiple-point geostatistics. *Computers & Geosciences*, *104*, 35–53.
Geisser, S. (1974). A predictive approach to the random effect model. *Biometrika*, *61*(1), 101–107.
Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*(350), 320–328.
Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 243–268.
Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*, *Springer Series in Statistics* (2nd ed.). New York: Springer.
Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. Cambridge, MA: The MIT Press.
Kitanidis, P. K. (1991). Orthonormal residuals in geostatistics: Model criticism and parameter estimation. *Mathematical Geology*, *23*(5), 741–758.
Kohavi, R. (1995).A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence—Volume 2* (pp. 1137–1143). Morgan Kaufmann Publishers Inc.
Madani, N., Maleki, M., & Emery, X. (2018). Nonparametric geostatistical simulation of subsurface facies: Tools for validating the reproduction of, and uncertainty in, facies geometry. *Natural Resources Research*, *28*, 1163–1182.
Mariethoz, G., & Caers, J. (2015). *Multiple-point geostatistics: Stochastic modeling with training images*. Chichester, West Sussex: Wiley Blackwell.
Mariethoz, G., & Kelly, B. F. (2011). Modeling complex geological structures with elementary training images and transform-invariant distances. *Water Resources Research*, *47*, W07527. https://doi.org/10.1029/2011WR010412
Mariethoz, G., Renard, P., & Straubhaar, J. (2010). The Direct Sampling method to perform multiple-point geostatistical simulations. *Water Resources Research*, *46*, W11536. https://doi.org/10.1029/2008WR007621
Meerschman, E., Pirot, G., Mariethoz, G., Straubhaar, J., Meirvenne, M. V., & Renard, P. (2013). A practical guide to performing multiple-point statistical simulations with the Direct Sampling algorithm. *Computers & Geosciences*, *52*, 307–324.
Pérez, C., Mariethoz, G., & Ortiz, J. M. (2014). Verifying the high-order consistency of training images with data for multiple-point geostatistics. *Computers & Geosciences*, *70*, 190–205.
Pyrcz, M. J., & Deutsch, C. (2014). *Geostatistical reservoir modeling* (2nd ed.). Oxford: Oxford University Press.
Rodriguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 569–575.
Rongier, G., Collon, P., Renard, P., Straubhaar, J., & Sausse, J. (2016). Comparing connected structures in ensemble of random fields. *Advances in Water Resources*, *96*, 145–169.
Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, *1*(1), 43–61.
Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111–133.
Straubhaar, J., Renard, P., & Chugunova, T. (2020). Multiple-point statistics using multi-resolution images. *Stochastic Environmental Research and Risk Assessment*, *34*, 251–273.
Tan, X., Tahmasebi, P., & Caers, J. (2014). Comparing training-image based algorithms using an analysis of distance. *Mathematical Geosciences*, *46*(2), 149–169.
Zhang, P. (1993). Model selection via multifold cross validation. *Annals of Statistics*, *21*(1), 299–313.