Monitoring Generative Models in Subsurface Modeling via Generalized Analysis of Distance

Gabrielle Brandemburg[1], Matheus do Ó[1], Julio Lopes[1], Renata Nascimento[1], and Erick Talarico [2]

[1]TecGraf Institute, PUC-Rio
Rio de Janeiro – RJ – Brazil

[2]Petrobras
Rio de Janeiro – RJ – Brazil

{gabriellebrand,matheussantos,jnlopes,rlins}@tecgraf.puc-rio.br

erick.talarico@petrobras.com.br

Abstract. *Deep learning have gained widespred adoption for generating complex geological facies, but quantitatively monitoring training performance remains challenging. The standard benchmark, Fréchet Inception Distance (FID), relies on a black-box feature extractor built for natural images and assumes Gaussian data. The Analysis of Distance (ANODI) score instead grounds evaluation in geostatistics, decomposing quality into pattern reproduction and spatial uncertainty without distributional assumptions, but lacks explicit support for multi-channel, arbitrary-sized datasets. We propose Generalized ANODI (GENODI), extending ANODI accordingly, and validate it on a Latent Diffusion Model trained on 2D geological facies. Results show GENODI correlates with FID while offering granular per-channel insights and a domain-relevant early stopping criterion.*

1. Introduction

Geological facies modeling is a fundamental task in subsurface characterization and reservoir modeling, as it provides spatial representations of geological structures and their associated uncertainty. For decades, geostatistical methods have been the focus for data generation, starting with two-point statistics techniques, such as Sequential Indicator Simulation (SIS) [Alabert 1987] and Truncated Gaussian Simulation (TGS) [Matheron et al. 1987], to explicitly capture and quantify spatial uncertainty. Object-Based Modeling (OBM) [Haldorsen and Damsleth 1990] was developed to overcome the limitations of two-point methods in reproducing geometrically complex structures, yet it struggles to condition the model to hard data. Multiple Point Statistics (MPS) emerged as a conceptual evolution of the previous techniques aiming to unite the ability to replicate complex geological architectures while honoring conditioning constraints in the hybrid MPS methods. MPS utilizes a training image as a source for expected spatial patterns and joint variability. Several algorithms have been proposed in the literature, notably SIMPAT and FILTERSIM are pattern-based methods that replace conditional probabilities by distance functions to identify the most similar pattern in a database [Tahmasebi 2018].

More recently, deep learning has emerged as a complementary approach to traditional geostatistical methods for facies modeling. Generative Adversarial

Networks (GANs) have been widely used to generate realistic realizations of subsurface facies. Notably, Song et al. [Song et al. 2022] introduced the GANSim framework, demonstrating the ability of GANs to reproduce complex spatial patterns while supporting robust conditional generation. However, GAN-based approaches may suffer from training instability and mode collapse. To overcome these limitations, recent research has increasingly embraced Diffusion Models, particularly Latent Diffusion Models (LDMs). Studies such as Lee et al. [Lee et al. 2025] demonstrate that diffusion-based architectures offer enhanced training stability, greater structural diversity, and superior fidelity when honoring conditioning data, establishing a new state of the art in generative reservoir modeling.

Despite these advances, quantitatively evaluating the quality and geological fidelity of generative models remains challenging as concluded in [Betzalel et al. 2022]. The Analysis of Distance (ANODI) [Tan et al. 2014] score was developed to rank geostatistical simulation algorithms, particularly those based on Multiple Point Statistics. Using distance-based measures, ANODI defines the "best" algorithm as one that simultaneously maximizes the pattern reproduction (minimizing the distance between the realization and the training image) while maximizing spatial uncertainty (maximizing the distance between individual realizations).

In this work, we investigate the use of ANODI as an auxiliary score for monitoring generative model training by treating each epoch checkpoint as a distinct generative model. Because ANODI calculates distance scores between two model distributions, it can be applied progressively throughout training by comparing the first epoch with subsequent checkpoints. This approach provides a quantitative trajectory of model performance across training epochs.

Although the original ANODI framework conceptually addresses categorical and continuous data, its explicit implementation is detailed only for binary images and a single training image. To resolve this gap, we propose Generalized ANODI (GENODI), an extension capable of handling datasets of arbitrary size and channel count. GENODI preserves the explainability of the original score by computing individual scores for each channel while maintaining a low computational overhead during the scoring phase. To facilitate decision-making, individual channel scores are also aggregated into a single scalar that serves as a criterion for early stopping or simply best-epoch selection.

Building upon these principles, GENODI adapts the core score formulation to accommodate multi-image and multi-channel contexts. In summary, the main contributions of the GENODI method are as follows:

- Extension of ANODI to multi-channel spatial data;
- Support for training datasets of arbitrary size;
- Channel-level evaluation for improved interpretability;
- Provides a domain-aligned score for training monitoring and early stopping.

2. Related Work

Evaluating sample quality in generative models relies on several methodologies, each grounded in distinct theoretical premises. The most widely adopted benchmark for this purpose is the *Fréchet Inception Distance* (FID) [Heusel et al. 2017], which considers both the real dataset and the generated samples as multivariate Gaussian distributions and

calculates the 2-Wasserstein distance between them. Intuitively, this quantity measures the minimal effort required to transform one probability distribution into the other. Assuming Gaussian distributions yields a closed-form expression for the 2-Wasserstein distance, making its calculation computationally straightforward.

FID does not operate in the pixel space. Instead, it processes images through the *Inception v3* network [Szegedy et al. 2016], utilizing feature representations extracted from its deep layers. This reliance on Inception v3 is justified by its pre-training on ImageNet, which endows the network with robust feature extraction capabilities for natural images. The mean vector and covariance matrix of these deep features are estimated for both real and generated data and then used to compute the 2-Wasserstein distance.

Although the FID is the standard method, it operates as a black box and relies on feature representations optimized for natural images limited to three channels. To address this domain mismatch, the *Fréchet Autoencoder Distance* (FAED) [Buzuti and Thomaz 2023] replaces the Inception v3 network with a Vector-Quantized Variational Autoencoder (VQ-VAE). Unlike FID, which uses feature encodings from a supervised model, FAED employs an unsupervised architecture trained directly on the target domain data, yielding latent representations tailored to the specific application. The latent encodings from the VQ-VAE are then used to compute the 2-Wasserstein distance in the same manner as FID. Like FID, FAED also assumes that the data underlying features are normally distributed.

The *Analysis of Distance* differs fundamentally from the previous mentioned methods. ANODI encodes images based on the frequencies of extracted patterns, represented as normalized frequency histograms, and does not rely on deep learning feature extractors. The distances between these empirical distributions are then computed without making any normality assumptions. Originally introduced to evaluate spatial simulation algorithms, ANODI grounds its assessment on two core geostatistical criteria:

1. **Spatial Uncertainty**: The degree of spatial diversity among generated realizations (where greater distance indicates higher diversity).
2. **Pattern Reproduction**: The fidelity with which generated realizations replicate the patterns of the training data (where smaller distance indicates higher reproduction accuracy).

A primary advantage of ANODI is its domain-aligned explainability: it computes multiple distinct quantities corresponding to explicit evaluation premises instead of a single number. Although ANODI originated in geostatistics for MPS, its underlying mathematical formulation is domain-agnostic and applicable to image datasets. The classical ANODI pipeline consists of four main steps:

1. Multi-resolution decomposition;
2. Pattern extraction and clustering;
3. Computation of pattern frequency histograms;
4. Calculation of uncertainty, reproduction, and total ratio scores.

3. Methodology

This section details the algorithmic and mathematical formulation of the proposed method. Building upon the core premises of ANODI, GENODI extends the original

framework to support datasets of arbitrary size and multi-channel image representations. We first present the single-channel binary formulation and subsequently generalize it to multi-channel and multi-resolution scenarios.

3.1. Cluster-Based Histogram of Patterns

The algorithm begins by extracting spatial patterns (as described in [Tan et al. 2014]) from the realizations using a standard sliding window defined by a fixed square template size W and stride S. Because the number of patterns grows exponentially with the template size, a compact representation of pattern frequencies is required. To achieve this, a distance transform is applied to the extracted binary patterns, and the resulting distance maps are grouped using an unsupervised clustering algorithm (e.g., k-means). Clustering is applied to the distance maps instead of raw binary patterns because the distance transform encodes richer spatial features, which provides superior discriminative power, as stated by Honarkhah and Caers in [Honarkhah and Caers 2010]. The clustering model is fitted exclusively on patterns extracted from the reference training dataset. The procedures for pattern extraction and clustering are shown in Algorithm 1.

Require: Training dataset $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$, sliding window size W, stride S, and cluster count K
Ensure: Clustering model $\mathcal{C}$
1: Initialize an empty set of patterns $\mathcal{P}_\mathcal{T} \leftarrow \emptyset$
2: **for** $t_i \in \mathcal{T}$ **do**
3: $\quad \mathcal{P}_i \leftarrow \text{ExtractPatterns}(t_i, W, S)$
4: $\quad \mathcal{P}_i \leftarrow \text{DistanceTransform}(\mathcal{P}_i)$
5: $\quad \mathcal{P}_\mathcal{T} \leftarrow \mathcal{P}_\mathcal{T} \cup \mathcal{P}_i$
6: **end for**
7: $\mathcal{C} \leftarrow \text{FitClusteringModel}(\mathcal{P}_\mathcal{T}, K)$ {e.g., using k-means}
8: **return** $\mathcal{C}$

Algorithm 1. Pattern extraction and clustering.

Through this pipeline, any given sample is mapped into a normalized frequency vector of dimension K (where K denotes the number of clusters), capturing the proportion of patterns assigned to each cluster. Following [Tan et al. 2014], this compact representation is referred to as the *Cluster-based Histogram of Patterns* (CHP). The CHP computation is described in the Algorithm 2.

3.2. Score Definitions

The ANODI framework defines its evaluation scores as distances between image representations within the CHP space. To quantify these distances, both ANODI and GENODI utilize the Jensen-Shannon (JS) divergence, a symmetric and bounded divergence measure for probability distributions. For two discrete probability distributions p and q defined over the finite set $\mathcal{X} = \{1, \ldots, K\}$, the JS divergence is formulated as

$$m = \frac{p + q}{2}, \tag{1}$$

$$\mathrm{JS}(p, q) = \frac{1}{2} \sum_{i \in \mathcal{X}} \left(p_i \log \frac{p_i}{m_i} + q_i \log \frac{q_i}{m_i} \right), \tag{2}$$

Require: A sample s, and clustering model $\mathcal{C}$
Ensure: A cluster-based histogram of patterns for s

$s' \leftarrow$ zero vector of dimension K
$\mathcal{P} \leftarrow \text{ExtractPatterns}(s, W, S)$
$\mathcal{P} \leftarrow \text{DistanceTransform}(\mathcal{P})$
for p_i in $\mathcal{P}$ **do**
    $c_i \leftarrow \mathcal{C}(p_i)$
    $s'[c_i] \leftarrow s'[c_i] + 1$
end for
$s' \leftarrow s' / \left(\sum_{i=1}^{K} s'[i]\right)$
return s'

Algorithm 2. CHP function for a single sample.

where p_i and q_i denote the i-th cluster probabilities of p and q, and m_i represents their mixture distribution.

Rooted in information entropy, the JS divergence quantifies the dissimilarity between probability distributions, serving a role analogous to the 2-Wasserstein distance employed in FID and FAED [Heusel et al. 2017, Buzuti and Thomaz 2023]. Because the entries of a CHP vector represent normalized relative frequencies that sum up to 1, they can be seen as approximations to discrete probability distributions over the K clusters of patterns.

To formally define the evaluation scores, consider two generative models, $\mathcal{A}$ and $\mathcal{B}$, both trained on a reference dataset $\mathcal{T} = \{t_i\}_{i=1}^{T}$, where each t_i represents an image. Each model generates a set of N realizations, denoted by $R_{\mathcal{A}} = \{a_i\}_{i=1}^{N}$ and $R_{\mathcal{B}} = \{b_i\}_{i=1}^{N}$, respectively.

In ANODI and GENODI, the *uncertainty* score quantifies the internal diversity of a generative model by computing the average pairwise distance among its generated realizations. This score measures the dispersion of the generated samples within the pattern feature space. It follows that

$$\text{unc}\,(R_{\mathcal{A}}) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} d(\text{CHP}(a_i), \text{CHP}(a_j)), \tag{3}$$

where $d(\cdot, \cdot)$ represents a distance function, defined in our framework as the Jensen-Shannon divergence $\text{JS}(\cdot, \cdot)$ and $\text{CHP}(\cdot)$ is the histogram computation defined in Algorithm 2.

The *reproduction* score evaluates the degree to which generated realizations replicate the spatial patterns present in the training dataset. Similar to spatial uncertainty, pattern reproduction is framed as a distance measure; however, instead of computing intra-set dispersion, it quantifies the dissimilarity between the generated realizations and the reference training data. Unlike the original ANODI formulation, which assumes a single reference image, GENODI accounts for the entire training dataset. As a possible extension, we may compute the minimum distance between each realization and individual training samples, although this is highly sensitive to outliers. Despite the

fact that an outlier is a valid member in the training dataset, this sample itself does not adequately describe the dataset tendency. Specifically, a realization near an outlier sample would yield an artificially low distance, falsely indicating high pattern reproduction despite deviating from the overall dataset distribution.

To mitigate this issue, GENODI aggregates the pattern frequencies from all realizations in $R_{\mathcal{A}}$ and all training images in $\mathcal{T}$ into a *Global CHP* distribution vector for each. The Global CHP is presented in Algorithm 3.

Require: A set of samples S, and clustering model $\mathcal{C}$
Ensure: Global histogram $\mathcal{G}$
initialize global CHP vector $\mathcal{G} \leftarrow$ zero vector of dimension K
for s_i in S **do**
  $s_i' \leftarrow$ zero vector of dimension K
  $\mathcal{P}_i \leftarrow \text{ExtractPatterns}(s_i, W, S)$
  $\mathcal{P}_i \leftarrow \text{DistanceTransform}(\mathcal{P}_i)$
  for p_j in $\mathcal{P}_i$ **do**
    $c_j \leftarrow \mathcal{C}(p_j)$
    $s_i'[c_j] \leftarrow s_i'[c_j] + 1$
  end for
  $\mathcal{G} \leftarrow \mathcal{G} + s_i'$
end for
$\mathcal{G} \leftarrow \mathcal{G} / \left(\sum_{j=1}^{K} \mathcal{G}[j]\right)$
return $\mathcal{G}$

Algorithm 3. Global CHP function.

The reproduction score is obtained by computing the final distance between them in order to capture the general tendency of the training dataset distribution.

$$\text{rep}\,(R_{\mathcal{A}}) = d(\text{GlobalCHP}(R_{\mathcal{A}}), \text{GlobalCHP}(\mathcal{T})) \tag{4}$$

Intuitively, lower scores reflect higher pattern reproduction, indicating that the generated realizations tend to match the spatial pattern distribution of the reference dataset.

3.3. Computing the Ratios

To synthesize model performance into a single quantitative value, we combine uncertainty and reproduction into a composite measure. Since a desirable model behavior corresponds to high uncertainty (spatial diversity) and low score for pattern reproduction (high fidelity), the *total ratio* for a model $R_{\mathcal{A}}$ is defined as

$$\text{tot}\,(R_{\mathcal{A}}) = \frac{\text{unc}\,(R_{\mathcal{A}})}{\text{rep}\,(R_{\mathcal{A}})}. \tag{5}$$

Consequently, based on ANODI premises, maximizing the total ratio reflects the improvement of generative quality by simultaneously promoting pattern reproduction and spatial uncertainty.

To compare two distinct generative models, $R_{\mathcal{A}}$ and $R_{\mathcal{B}}$, we define the GENODI score as the ratio between their total ratios.

$$\mathrm{GEN}\left(R_{\mathcal{A}}, R_{\mathcal{B}}\right) = \frac{\mathrm{tot}\left(R_{\mathcal{A}}\right)}{\mathrm{tot}\left(R_{\mathcal{B}}\right)} \tag{6}$$

A score $\mathrm{GEN}\left(R_{\mathcal{A}}, R_{\mathcal{B}}\right) > 1$ indicates that model $R_{\mathcal{A}}$ has a better rank than $R_{\mathcal{B}}$, whereas a score less than 1 indicates superior rank for $R_{\mathcal{B}}$. To avoid numerical instability or disproportionately large values when $\mathrm{tot}\left(R_{\mathcal{B}}\right)$ is small, a logarithmic transformation can optionally be applied to the comparative score.

The complete procedure for the single-channel GENODI score is detailed in Algorithm 4. The component ratios and the overall GENODI score are calculated following Equations 5 and 6.

Require: Training dataset $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$, realizations sets $R_{\mathcal{A}} = \{a_1, a_2, \ldots, a_N\}$, $R_{\mathcal{B}} = \{b_1, b_2, \ldots, b_N\}$, and clustering model $\mathcal{C}$
Ensure: GENODI score for models $\mathcal{A}$ and $\mathcal{B}$
1: $u_{\mathcal{A}} \leftarrow \mathrm{unc}\left(R_{\mathcal{A}}\right)$ {using CHP}
2: $u_{\mathcal{B}} \leftarrow \mathrm{unc}\left(R_{\mathcal{B}}\right)$ {using CHP}
3: $r_{\mathcal{A}} \leftarrow \mathrm{rep}\left(R_{\mathcal{A}}\right)$ {using Global CHP}
4: $r_{\mathcal{B}} \leftarrow \mathrm{rep}\left(R_{\mathcal{B}}\right)$ {using Global CHP}
5: $tot_{\mathcal{A}} \leftarrow u_{\mathcal{A}}/r_{\mathcal{A}}$
6: $tot_{\mathcal{B}} \leftarrow u_{\mathcal{B}}/r_{\mathcal{B}}$
7: **return** $tot_{\mathcal{A}}/tot_{\mathcal{B}}$

Algorithm 4. GENODI computation for single-channel.

4. Generalization to Multiple Channels and Resolutions

The score formulation presented so far applies exclusively to single-channel image data at a fixed resolution. In this section, we extend the framework to accommodate multi-channel and multi-resolution scenarios. For this reason, we expand our notation by introducing indices g and c to denote the resolution grid level and channel index, respectively. Thus, $R_{\mathcal{A}}^{g,c} = \{a_i^{g,c}\}_{i=1}^{N}$ represents the set of realizations evaluated at resolution level g for channel c. The channel- and resolution-specific GENODI score is evaluated directly on these subsets.

$$\mathrm{GEN}\left(R_{\mathcal{A}}^{g,c}, R_{\mathcal{B}}^{g,c}\right) = \frac{\mathrm{tot}\left(R_{\mathcal{A}}^{g,c}\right)}{\mathrm{tot}\left(R_{\mathcal{B}}^{g,c}\right)}$$

4.1. Multi-Resolution Analysis

Extracting patterns using a sliding window captures local spatial correlations among neighboring pixels but fails to account for long-range spatial dependencies. To mitigate this limitation, the image resolution is progressively reduced via downsampling (e.g., interpolation), followed by pattern extraction at each scale. Downsampling expands the spatial scope of the sliding template, enabling the capture of non-local spatial structures.

Evaluating images across G resolution levels yields G distinct GENODI scores for a given channel c. To aggregate these multi-scale evaluations into a unified score, we compute a weighted average across resolutions. Letting $g = 1$ to denote the original resolution and $g = G$ the coarsest level, the multi-resolution GENODI score for channel c is defined as

$$\mathrm{GEN}\left(R_{\mathcal{A}}^{[G],c}, R_{\mathcal{B}}^{[G],c}\right) = \sum_{g=1}^{G} w_g \mathrm{GEN}\left(R_{\mathcal{A}}^{g,c}, R_{\mathcal{B}}^{g,c}\right), \tag{7}$$

where the bracket notation $[G]$ indicates aggregation across all G resolution scales. The weighting coefficients w_g can be flexibly assigned under the normalization constraint $\sum_{g=1}^{G} w_g = 1$. In the single-resolution case ($G = 1$), setting $w_1 = 1$ recovers the baseline formulation in Equation 6. Following the multi-resolution strategy introduced in [Tan et al. 2014], we adopt exponentially decaying weights $w_g = 2^{-g}$ in our experimental evaluation.

4.2. Multi-Channel Aggregation

Many application domains rely on multi-channel representations, where individual channels encode distinct spatial or semantic properties (e.g., categorical facies in geological modeling or satellite spectral bands). Unlike single channel evaluations, a C-channel dataset yields C separate GENODI values. To aggregate these channel-wise evaluations into a unified score, we employ a weighted harmonic mean, using weights inversely proportional to the relative channel proportions in the training dataset. We define the overall multi-channel, multi-resolution GENODI score as

$$\mathrm{GEN}\left(R_{\mathcal{A}}, R_{\mathcal{B}}\right) = \left(\sum_{c=1}^{C} \frac{w_c}{\mathrm{GEN}\left(R_{\mathcal{A}}^{[G],c}, R_{\mathcal{B}}^{[G],c}\right)}\right)^{-1}, \tag{8}$$

where $w_c = (1/p_c)/\left(\sum_{j=1}^{C}(1/p_j)\right)$ is the normalized weight for channel c, and $p_c \in (0, 1]$ represents the fraction of total pixels belonging to channel c across the training dataset $\mathcal{T}$. This formulation ensures that the weights satisfy, satisfying $\sum_{c=1}^{C} wp_c = 1$. For concise notation, we abbreviate $\mathrm{GEN}\left(R_{\mathcal{A}}^{[G],[C]}, R_{\mathcal{B}}^{[G],[C]}\right)$ as $\mathrm{GEN}\left(R_{\mathcal{A}}, R_{\mathcal{B}}\right)$.

The choice of a harmonic mean is motivated by its inherent sensitivity to low values, a property widely leveraged in quantities such as the F_1-score. In imbalanced multi-channel datasets, the inverse weighting ensures that strong performance on majority classes cannot mask poor spatial reconstruction on rare or underrepresented channels. Finally, for a single-channel input ($C = 1$ and $w_1 = 1$), this formulation reduces directly to the baseline expression in Equation 6.

4.3. Merging scores

It may also be of interest to evaluate only the uncertainty or reproduction of the models instead of their total ratios. Both scores can be defined as ratios between two different

models.

$$\mathrm{unc}\left(R_{\mathcal{A}}, R_{\mathcal{B}}\right) = \frac{\mathrm{unc}\left(R_{\mathcal{A}}\right)}{\mathrm{unc}\left(R_{\mathcal{B}}\right)} \tag{9}$$

$$\mathrm{rep}\left(R_{\mathcal{A}}, R_{\mathcal{B}}\right) = \frac{\mathrm{rep}\left(R_{\mathcal{A}}\right)}{\mathrm{rep}\left(R_{\mathcal{B}}\right)} \tag{10}$$

A ratio $\mathrm{unc}\left(R_{\mathcal{A}}, R_{\mathcal{B}}\right) > 1$ implies that model $R_{\mathcal{A}}$ ranks higher than $R_{\mathcal{B}}$ for the spatial uncertainty criterion. In contrast, $R_{\mathcal{A}}$ achieves superior pattern reproduction compared to $R_{\mathcal{B}}$ when $\mathrm{rep}\left(R_{\mathcal{A}}, R_{\mathcal{B}}\right) < 1$.

Since these definitions receive two arguments, they do not conflict with the previous ones. In fact, $\mathrm{unc}\left(R_{\mathcal{A}}, R_{\mathcal{B}}\right)\left(\mathrm{rep}\left(R_{\mathcal{A}}, R_{\mathcal{B}}\right)\right)^{-1} = \mathrm{GEN}\left(R_{\mathcal{A}}, R_{\mathcal{B}}\right)$, and the original ANODI paper adopts this definition. We opt to define it differently to separate the contributions of each model to their final ratio, being able to express their individual values for uncertainty, reproduction, and total ratio. In the following experiment section, the results for uncertainty and reproduction were computed based on (9) and (10), respectively.

5. Experiment

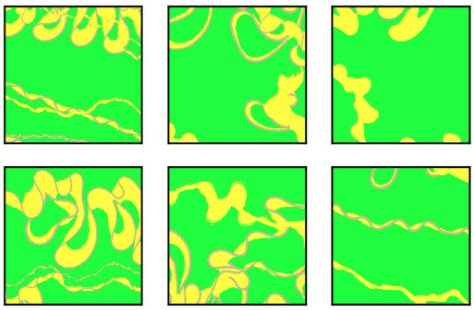

Figure 1. Random samples from GAN River I dataset of 3 facies used for training the generative LDM and GENODI model. Facies names are Channel (grey), Point Bar (yellow) and Background (green).

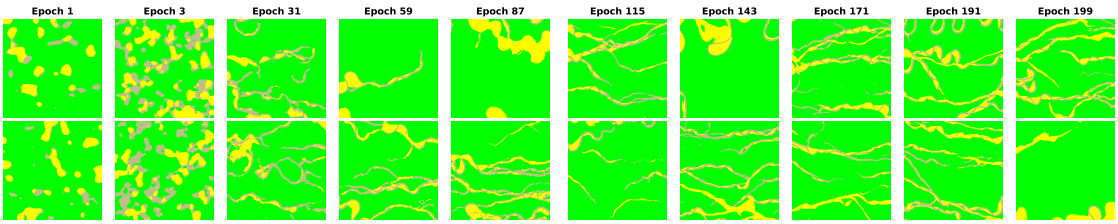

Figure 2. Random generated samples along training epochs.

5.1. Dataset specification and implementation details

To demonstrate the use of GENODI as a scoring method for training monitoring, we trained a 2D Latent Diffusion Model based on Lee et al [Lee et al. 2025] architecture with the GAN River-I [Sun et al. 2023] dataset with 3-facies. GAN River-I is a widely

used dataset for machine learning and geostatistical methods that wish to recreate complex geometries of realistic facies distributions in subsurface reservoirs. It is composed by 25 3D cubes with a total of 16,000 2D layers of 256x256 pixels produced by a process-based simulator of a meandering fluvial system. Samples from the facies set are illustrated in Figure 1.

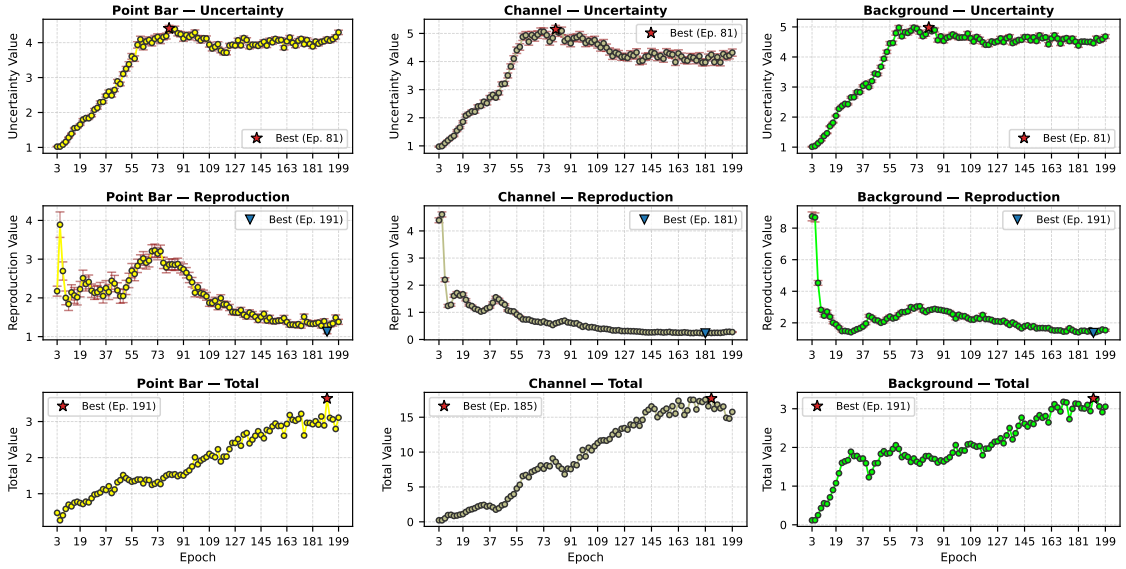

Figure 3. Uncertainty, Reproduction and Total ratio curve evolution from GENODI Score computed every epoch. Uncertainty and Reproduction plots the result of the mean and standard deviation (in red) resulted from 5 different GENODI executions.

The Diffusion Model was trained for 200 epochs with 12,800 layers from GAN River-I. 1,000 samples were generated for every 2 epochs to compute GENODI and FID scores. Although standard FID evaluation typically requires 10,000 to 50,000 samples to fully eliminate small-sample bias, generating 1,000 samples per evaluation step represents the maximum feasible limit for a diffusion model in a frequent training monitoring context, offering a practical compromise between statistical reliability and acceptable sampling times.

For the GENODI model parameters we used a square template size of 12x12 pixels with stride 6 and a pyramid resolution of 3 levels as described in the ANODI paper. The number of clusters was fixed to 50 and batch size of 1024 for Mini Batch k-means from Scikit Learn [Pedregosa et al. 2011].

GENODI model training, represented in Algorithm 1, is processed on the CPU with parallel processes. For score computation, represented in Algorithm 4, the CHP construction is executed in CPU with no parallel optimization and calculation of uncertainty and reproduction is done in GPU. To compute FID, we used the TorchMetrics [Detlefsen et al. 2022] implementation. Because the Inception-v3 feature extractor requires continuous 3-channel RGB inputs, we explicitly mapped our 3 categorical facies to pure color channels: Point Bar to red $(255, 0, 0)$, Channel to green $(0, 255, 0)$, and Background to blue $(0, 0, 255)$. For datasets with more than three categories, mapping the encodings to a continuous colormap would be required.

5.2. Multi-Execution Stability Analysis

To evaluate metric stability against the stochastic initialization of the clustering step, GENODI was evaluated across $E = 5$ independent executions for each epoch. Thanks to the probabilistic seeding mechanism of the k-means++ algorithm [Arthur and Vassilvitskii 2007] used within MiniBatchKMeans, the spatial clustering demonstrated low sensitivity to random initialization. As illustrated by the narrow error bars in Figure 3, the metric exhibits minimal variance across runs, yielding an average standard deviation of $\sigma \leq 0.108$ for spatial uncertainty and $\sigma \leq 0.122$ for pattern reproduction across all facies. This suggests that the chosen hyperparameter configuration provides a stable evaluation space.

5.3. Results

Figure 2 shows two random samples generated across training epochs, whereas Figure 3 displays the curves of uncertainty, reproduction and total ratios. Because uncertainty is quantified by the distance between realizations, greater distances correspond to increased realization diversity. Pattern reproduction is defined by the distance between the realization set and the training set, meaning that lower distances reflect higher pattern fidelity relative to the training dataset.

During the initial training epochs, the model generates unstructured artifacts that lack both spatial diversity and fidelity to the training distribution, resulting in the lowest overall scores. The pattern reproduction for the Point Bar facies deteriorates throughout the early stages of training, reaching a penalty peak near epoch 73 before exhibiting rapid improvement. Conversely, Background reproduction initially improves but begins to degrade after epoch 31, ultimately hitting a similar penalty peak around epoch 73. A plausible morphological explanation for this behavior is that the Point Bar and Background facies in the training dataset are characterized by smooth, rounded boundaries, whereas the samples generated during this intermediate training phase display sharper, noisier edges. This morphological mismatch may also explain why the reproduction scores for both classes never drop below 1.0. Despite a relative improvement in later epochs (indicated by a continuous decrease in distance values), the generated geometries never achieve a closer structural alignment with the training patterns than the highly unstructured baseline model of Epoch 1. An alternative or additional hypothesis is that the chosen template size or other GENODI hyperparameters are sub-optimal for this specific dataset, warranting further investigation into parameter sensitivity. Nevertheless, during these early epochs, the steady increase in the spatial uncertainty curve compensates for the reproduction penalties, driving a continuous rise in the total ratio.

Across all classes, the spatial uncertainty scores reach a maximum between epochs 60 and 80 before stabilizing. This stabilization coincides with the inflection point where pattern reproduction begins to consistently improve, thereby contributing positively to the total GENODI score. Although the diffusion model was configured for a maximum of 200 training epochs, visual inspection of the learning curves indicates that all classes enter a local plateau from epoch 171 onward. This trajectory suggests that the model might not have fully converged and could potentially achieve superior geological fidelity with an extended training regimen.

A comparison between the weighted harmonic mean of all classes from GENODI

and FID is illustrated in Figure 4. For FID score, the lower the value the better is the model quality. Although FID was originally designed for natural images, it provided a satisfactory evaluation when applied to geological data. FID shares a similar curve format with GENODI's reproduction. Both starts with a rapid decrease in the beginning, even though FID has a more stable decrease than GENODI's reproduction. They also share a small local peak around epoch 43 and epoch 87.

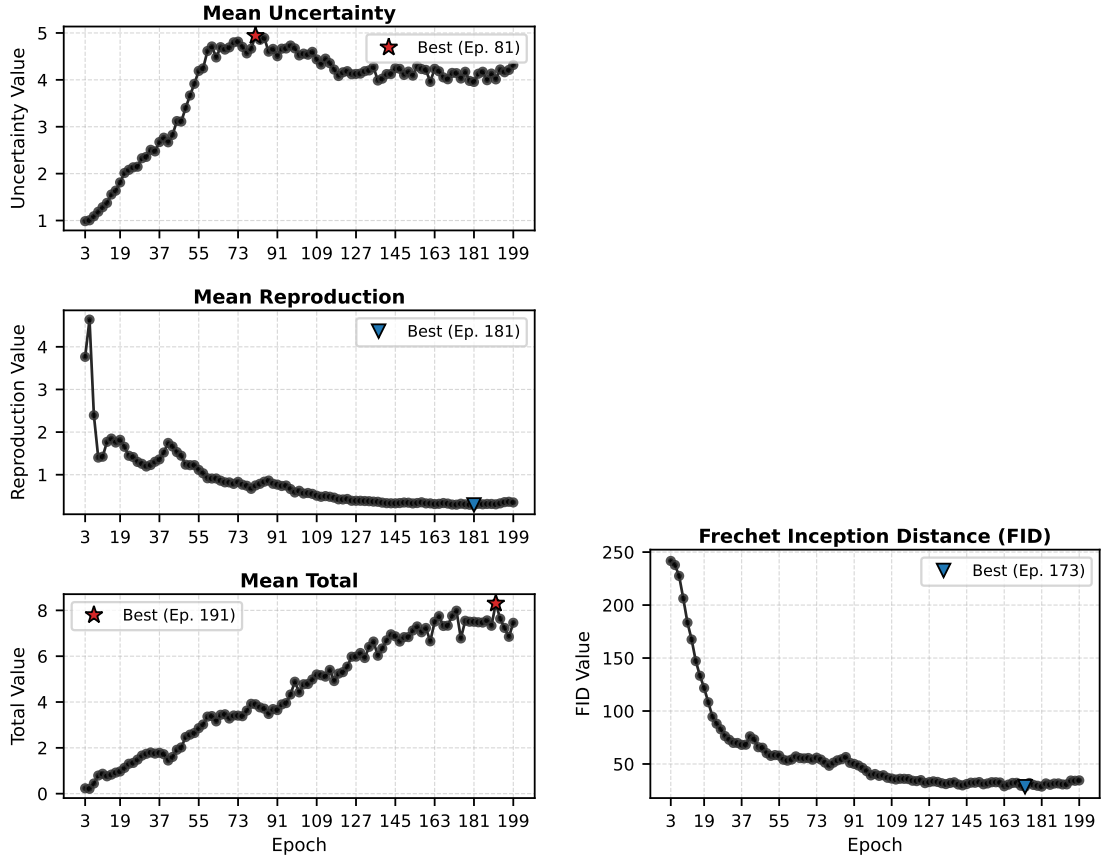

Figure 4. Weighted harmonic mean of GENODI score on the left and FID score on the right across training epochs.

5.4. A brief correlation analysis

Looking to both GENODI and FID graphs, we can see a correlation between the curves. As expected by their constructions, GENODI tends to increase and FID to decrease. But does this always happen?

To show this correlation we show a dispersion graph in Figure 5. Each point corresponds to a check-in point and its x and y coordinates indicates its GENODI score and FID score, respectively. The colormap indicates epoch progression. As we can see, it forms a curve that respects the color gradient. This shows that both are moving in opposite directions, an anti-correlation.

Although the dispersion graph helps to visually see the anti-correlation, it does not give an evident guarantee about it. For this purpose, we compute the Spearman coefficient between the scores along the epochs. The Spearman coefficient between two functions

tries to measure if (and how much) they are monotonically correlated. The score can be any real number between -1 and 1, in which -1 indicates a high anti-correlation and 1 indicates a high correlation.

Because the Spearman correlation coefficient depends on the number of available epochs, we evaluated its uncertainty using 95% confidence intervals. We resampled the original 100 epochs 10,000 times with replacement ($n = 100$) and calculated the Spearman coefficient for each bootstrap sample. Table 1 shows the resulting confidence intervals for both 100 and 1,000 realizations per epoch.

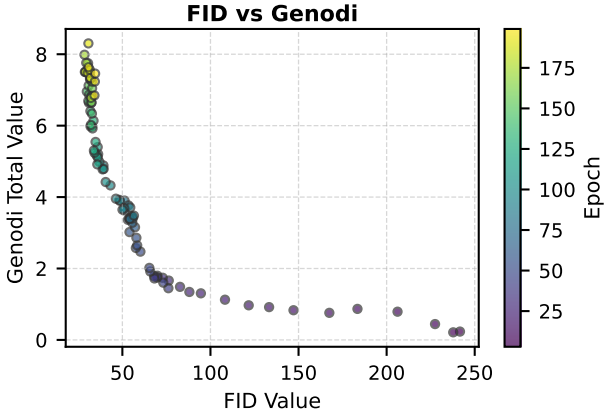

Figure 5. Dispersion graph for GENODI and FID scores. Each point is an epoch, indicated by the color.

Table 1. Confidence interval for Spearman coefficient

Realizations per epoch	Min value	Max value
100	-0.9811	-0.9383
1,000	-0.9863	-0.9420

Both numbers of realizations produce similar intervals, providing empirical support that both metrics are highly anti-correlated. These results can justify the dispersion graph behavior and the anti-correlation between the metrics. Since FID is known to be biased at lower realization counts, obtaining a consistent Spearman coefficient helps to enforce the belief of anti-correlation.

5.5. Pure-noise and mode-collapse controls

In order to evaluate GENODI score under extreme conditions, we computed the metric for two baseline sets against the Epoch 1 reference model from the training experiment. The first set contains 1,000 identical clones of a single test dataset sample, representing a severe mode-collapse scenario. The second set consists of 1,000 samples generated from a uniform random distribution, simulating a high-entropy generator. Results are shown in Table 2.

As expected, the mode-collapse uncertainty is exactly zero since all samples are the same. The reproduction score is reasonable but not optimal, as the Global CHP from a single cloned sample is essentially compared against the Global CHP of the entire training dataset. Conversely, the pure-noise set yields a heavily penalized (high) reproduction score, as unstructured noise severely diverges from geological training patterns. Uncertainty also achieves a low score because independent noise realizations produce statistically homogeneous pattern histograms, minimizing the pairwise Jensen-Shannon divergence.

Table 2. GENODI Score for Mode-Colapse and Pure-Noise sets

Realization Set	Class	Uncertainty	Reproduction	Total
Mode Collapse	Point Bar	0.000	1.301	0.000
	Channel	0.000	0.875	0.000
	Background	0.000	1.639	0.000
Pure Noise	Point Bar	0.699	46.388	0.015
	Channel	0.643	23.396	0.027
	Background	0.607	59.210	0.010

6. Conclusion

In this work, we introduced GENODI (Generalized Analysis of Distance), an explainable quality score for monitoring and evaluating generative models on non-natural, multi-channel datasets. By generalizing the original ANODI framework, GENODI accommodates arbitrary dataset sizes and multiple image channels without relying on pre-trained black-box networks or assuming Gaussian data distributions.

Our experimental evaluation on a Latent Diffusion Model trained with the GAN River-I geological dataset demonstrated that GENODI effectively captures training dynamics. While showing strong correlation with the standard FID score, GENODI provides granular explainability by decoupling performance into separate uncertainty and reproduction metrics for each channel.

While our empirical hyperparameters ($K = 50, W = 12$) proved highly stable for the evaluated dataset, conducting a comprehensive sensitivity analysis over the template size and cluster count remains an important avenue for future work to establish generalized default parameters across diverse geological domains.

Future work includes expanding the GENODI framework to 3D volumetric data, applying multi-dimension distance transform to categorical data instead of binary distance transform for each class map, resulting in only one clustering model for all image channels. We also propose to explore different metrics for measuring distance between subsets such as Hausdorff distance to obtain the pattern reproduction score.

Acknowledgment

The authors would like to thank the Tecgraf Institute (PUC-Rio) and Petróleo Brasileiro S.A. (PETROBRAS) for their technical, financial, and computational support in this research.

References

Alabert, F. G. (1987). Stochastic imaging of spatial distributions using hard and soft information. Master's thesis, Stanford University, Stanford, CA. Stanford Center for Reservoir Forecasting (SCRF).

Arthur, D. and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.

Betzalel, E., Penso, C., Navon, A., and Fetaya, E. (2022). A study on the evaluation of generative models.

Buzuti, L. F. and Thomaz, C. E. (2023). Fréchet autoencoder distance: A new approach for evaluation of generative adversarial networks. *Computer Vision and Image Understanding*, 235:103768.

Detlefsen, N. S., Borovec, J., Schock, J., Jha, A. H., Koker, T., Di Liello, L., Stancl, D., Quan, C., Grechkin, M., and Falcon, W. (2022). Torchmetrics - measuring reproducibility in pytorch. *Journal of Open Source Software*, 7(70):4101.

Haldorsen, H. H. and Damsleth, E. (1990). Stochastic modeling. *Journal of Petroleum Technology; (USA)*, 42:4.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Honarkhah, M. and Caers, J. (2010). Stochastic simulation of patterns using distance-based pattern modeling. *Mathematical Geosciences*, 42(5):487–517.

Lee, D., Ovanger, O., Eidsvik, J., Aune, E., Skauvold, J., and Hauge, R. (2025). Latent diffusion model for conditional reservoir facies generation. *Computers amp; Geosciences*, 194:105750.

Matheron, G., Beucher, H., de Fouquet, C., Galli, A., Guerillot, D., and Ravenne, C. (1987). Conditional simulation of the geometry of fluvio-deltaic reservoirs. In *SPE Annual Technical Conference and Exhibition*, pages SPE–16753. Society of Petroleum Engineers.

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., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Song, S., Mukerji, T., Hou, J., Zhang, D., and Lyu, X. (2022). Gansim-3d for conditional geomodeling: Theory and field application. *Water Resources Research*, 58(7):e2021WR031865. e2021WR031865 2021WR031865.

Sun, C., Demyanov, V., and Arnold, D. (2023). Gan river-i: A process-based low ntg meandering reservoir model dataset for machine learning studies. *Data in Brief*, 46:108785.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tahmasebi, P. (2018). *Multiple Point Statistics: A Review*, pages 613–643. Springer International Publishing, Cham.

Tan, X., Tahmasebi, P., and Caers, J. (2014). Comparing training-image based algorithms using an analysis of distance. *Mathematical Geosciences*, 46.