

Article

Evaluating the Uncertainty and Predictive Performance of Probabilistic Models Devised for Grade Estimation in a Porphyry Copper Deposit

Raymond Leung ^{*}, Alexander Lowe  and Arman Melkumyan 

Rio Tinto Sydney Innovation Hub, Faculty of Engineering, The University of Sydney, Sydney, NSW 2008, Australia; alexander.lowe@sydney.edu.au (A.L.); arman.melkumyan@sydney.edu.au (A.M.)

* Correspondence: raymond.leung@sydney.edu.au

Abstract: Probabilistic models are used to describe random processes and quantify prediction uncertainties in a principled way. Examples include geotechnical and geological investigations that seek to model subsurface hydrostratigraphic properties or mineral deposits. In mining geology, model validation efforts have generally lagged behind the development and deployment of computational models. One problem is the lack of industry guidelines for evaluating the uncertainty and predictive performance of probabilistic ore grade models. This paper aims to bridge this gap by developing a holistic approach that is autonomous, scalable and transferable across domains. The proposed model assessment targets three objectives. First, we aim to ensure that the predictions are reasonably calibrated with probabilities. Second, statistics are viewed as images to help facilitate large-scale simultaneous comparisons for multiple models across space and time, spanning multiple regions and inference periods. Third, variogram ratios are used to objectively measure the spatial fidelity of models. In this study, we examine models created by ordinary kriging and the Gaussian process in conjunction with sequential or random field simulations. The assessments are underpinned by statistics that evaluate the model's predictive distributions relative to the ground truth. These statistics are standardised, interpretable and amenable to significance testing. The proposed methods are demonstrated using extensive data from a real copper mine in a grade estimation task and are accompanied by an open-source implementation. The experiments are designed to emphasise data diversity and convey insights, such as the increased difficulty of future-bench prediction (extrapolation) relative to in situ regression (interpolation). This work enables competing models to be evaluated consistently and the robustness and validity of probabilistic predictions to be tested, and it makes cross-study comparison possible irrespective of site conditions.

Keywords: probabilistic models; sequential simulations; ore grade estimation; future bench extrapolation; prediction uncertainty; probability calibration; spatial fidelity; methods for assessment; performance statistics



Academic Editor: Wei Gao

Received: 2 May 2025

Revised: 6 June 2025

Accepted: 10 June 2025

Published: 17 June 2025

Citation: Leung, R.; Lowe, A.; Melkumyan, A. Evaluating the Uncertainty and Predictive Performance of Probabilistic Models Devised for Grade Estimation in a Porphyry Copper Deposit. *Modelling* **2025**, *6*, 50. <https://doi.org/10.3390/modelling6020050>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

This paper is principally concerned with model validation [1], where the primary goal is assessing whether a given model provides a fair representation of an orebody system and reflects the subterranean geochemical properties in a mineral deposit. Although modelling techniques (geostatistical prediction and conditional simulation) feature strongly in this work, this study is oriented towards assessment methods more so than model evaluation. This subtle distinction is important as it sets the tone and expectations for what is to come.

Instead of comparing models to determine which is superior, there is greater emphasis on developing a balanced approach that adds value and insights to the analysis. The general problem involves comparing model predictions $(\hat{\mu}, \hat{\sigma})$ with observational data (μ_0) , where $\hat{\mu}$ and $\hat{\sigma}$ denote the estimated mean and standard deviation for some target variable, and μ_0 denotes the ground truth. The objective is to present standard experiments—using different measures and procedures—to arrive at a broad-based understanding of model performance. For probabilistic models, performance may be assessed in terms of the global accuracy (e.g., histogram distances), local accuracy (spatial variability) and probability calibration (uncertainty aspect) of the predictive distributions. More details on these will be presented in due course. The main point is that these three aspects are seldom jointly investigated and evaluated in ways that are amenable to large-scale automated processing, particularly in relation to mineral resources estimation. For instance, Tutmez [2] focuses solely on global accuracy [3], but the chosen measures (RMSE and variance accounted for) are not standardised and are not interpretable outside of the study area. This renders cross-site and cross-species comparisons impossible if the models are to be tested at a different location or used to predict the grade of another mineral. In Singh et al. [4], the coefficient of determination, $R^2 \in [0, 1]$, is used to assess global accuracy; however, the predicted spatial distribution is assessed visually, so it is not scalable. Mery et al. [5] use pluri-Gaussian and multi-Gaussian joint simulations to model rock-type uncertainty and grades. For local analysis, they relied on visual assessment for spatial correlation. As an alternative, variograms can be used, but a detailed examination still requires considerable effort [6]. For global accuracy, scatter plots are used in that study to check that marginal and bivariate distributions are reproduced, whereas scatter plots are used by Emery and Ortiz [7] to diagnose conditional bias for univariate distributions. Mery et al. [5] examined model calibration by computing probability intervals on the true values for different theoretical probabilities and then comparing these with actual proportions of data that belong to these intervals using leave-one-out cross-validation. Similar concepts [8] based on theoretical and sample quantiles or the probability of occurrence vs. proportion observed in the test dataset are used in [9,10]. Together, these ideas broadly inform the quantitative measures we propose to use to target the three pillars of model performance. To be clear, this study should not be characterised as presenting more of the same. Fundamentally, it is about improving ease of use through standardised measures or statistics, enabling consistent model comparisons at a scale beyond a single site or geographic domain, as well as functional enhancements (adding extra layers to the analysis) to develop a richer and more complete understanding of model performance. Our experiments feature a new modality for focusing attention on identifying situations where models have underperformed. This is particularly suitable for comparing a large number of models or model configurations. A tangible outcome is a collection of measures (abbreviated as FLAGSHIP) that assess performance based on the fidelity, local consensus, accuracy, goodness, synchronicity, histogram, interval tightness and precision of the predictive distributions. These concepts will be described in Section 3.

1.1. Performance Criteria and Application Context

To consolidate understanding, we take a step back to clarify the problem scope and application context. In relation to probabilistic models, it is helpful to know what they can represent and what the writers have in mind. Probabilistic models are useful for describing a wide range of stochastic processes and natural phenomena in the geosciences. For instance, Monte Carlo techniques are used in the field of landslide hazard assessment to account for estimation uncertainties and spatial variability of geological, geotechnical, geomorphological and seismological parameters by treating the target quantities as statisti-

cal distributions [11]. In the field of subsurface and hydrostratigraphic modelling, borehole and geophysical data are often combined to improve lithological and structural understanding in a study area [12]. One difficulty is maintaining consistency between different pieces of information while taking into account various spatial and geological factors when assigning uncertainties to such an interpretation. In Madsen et al. [13], the uncertainty is estimated by generating random realisations of the subsurface from a 3D geological model. This requires geostatistical simulation of each hydrostratigraphic layer using contact points (boundary interpretations) specified by geologists. Another area of research focuses on dynamic stochastic models and Bayesian inference in very-high-dimensional space. For instance, Bacci et al. [14] considered different approaches for sampling distribution, while others applied Bayesian fusion to multiple data sources to minimise uncertainty [15]. This list represents only a slice of recent work conducted in probabilistic modelling. Nonetheless, it shows there is active interest in using probabilistic models to describe stochastic processes in the wider geoscientific community. Accompanying model development is the need to assess performance to determine how reasonable the models are. For discourse on the topic of errors and uncertainties in the geosciences, readers are referred to [16,17]. The application targeted in this study is mineral resource estimation and grade modelling in particular. For this application, the assessment scope is defined by field-specific performance criteria, which are underpinned by the three pillars: global accuracy, local accuracy (spatial fidelity in terms of spatial correlation and variability) and calibration properties. Having described what performance criteria are relevant and where the probabilistic models will be used, the remaining questions are why grade models are worthy of consideration and how the assessment will be carried out.

The motivations and validation approach are crucial for understanding the attributes that set this study apart. In open-pit mining, the training data typically comes from geochemical assays which measure the concentration of elements, such as copper, using X-ray fluorescence. These assay samples are taken from production blastholes at a given bench/elevation after they are drilled (but before blasting and excavation). Hence, the spatial domain is three-dimensional, whereas it is common for surficial geochemical surveys [18] in environmental monitoring to be two-dimensional. A model is trained by learning from this sparse data to predict the ore grade at unknown locations. These models can facilitate precision mining and tracking of material movement. Improved knowledge of the deposit can improve grade control and reduce incidents like ore dilution, where low-grade waste is excavated and transferred inadvertently to high-grade stockpiles [19]. This use case of probabilistic modelling is depicted in Figure 1. Generally, there is a strong incentive for understanding the geochemical properties of an orebody beyond the currently drilled area [20]. The ability to predict with a quantifiable degree of confidence the grade distribution in future benches is also necessary for robust mine planning. These probabilistic predictions can affect block sequencing, scheduling decisions and optimisation of mining operations downstream [21,22]. Thus, the ability to accurately predict the grade distribution in the bench below, which has not yet been excavated, or in adjacent areas within the current bench is highly valued. This work mainly focuses on extrapolation performance where models will forward-predict into new territories. For this reason, k-fold cross-validation is considered inappropriate (unless in situ regression performance is of interest), as the unseen dataset aside for testing would come from the same region as the training data and increase bias. Instead, the hold-out data should come from the bench below or adjacent areas that have not yet been drilled. Hence, future-bench model extrapolation performance is typically evaluated retrospectively, when sanctioned data (the pseudo-ground truth) becomes available. This delay usually does not pose a problem, as the previously excavated materials are often stockpiled—meaning there is a time buffer for

uncertainty updates provided there is adequate tracking—and numerous mining processes are happening concurrently, sometimes with significant lag in between. The model assessment result is still useful, as it can at the very least influence future decisions, for instance, justifying increased sampling in more uncertain and geologically complex areas in future drilling campaigns.

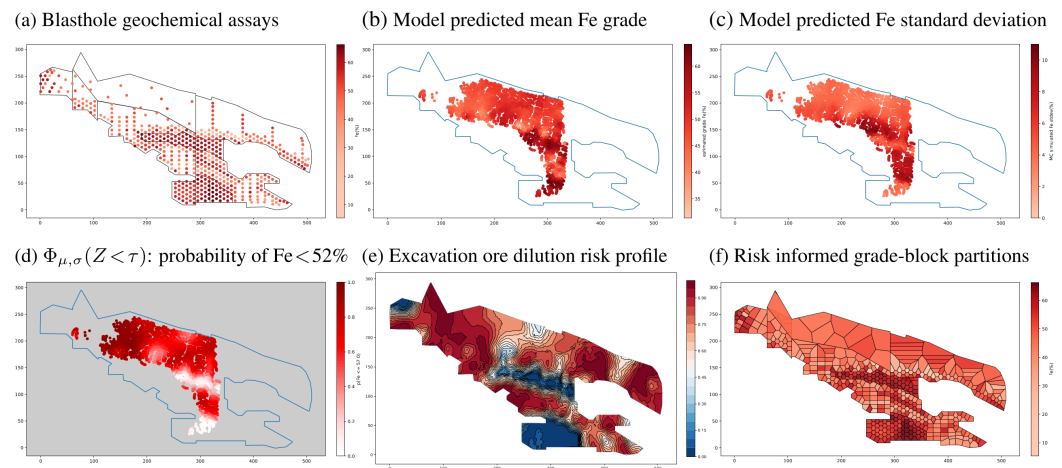


Figure 1. A motivating example. For open-pit mining at iron ore deposits, (a) sparse assay measurements are taken from blastholes to facilitate ore grade probabilistic modelling. (b,c) These show the estimated mean Fe concentration and standard deviation in a local region soon to be excavated. The value of having a probabilistic model is that it provides a reliable and objective description of ore/waste distribution in spite of sampling errors and epistemic uncertainty. This allows operators to assess risks, such as ore dilution in (d), if a volume of low-grade material is dug up at a [red] location and transported to a high-grade destination. (e) A high-fidelity probabilistic model makes informed decision-making possible. Its applications include high-precision large-scale tracking of material movement, as well as (f) grade-block partitioning and reconfiguration during mine planning and the ability to reroute material to different destinations on demand [19].

1.2. Objectives and Organisation

To complete this overview, we note that this paper has a dual objective. First, it seeks to develop a consistent approach that supports highly automated model assessment at scale, using standardised and interpretable measures that can be meaningfully compared across domains and target variables. Second, it adds layers to the analysis to provide a more complete understanding of different facets of model performance—this includes identifying situations where models misbehave, visualising error clusters, and testing for statistical significance between models. Associated with each prediction is an implied probability that the true value lies within some interval according to the predictive distribution. To evaluate the performance of univariate probabilistic models, we use a suite of histogram-, variogram- and uncertainty-based measures to systematically assess the global accuracy, local accuracy and calibration properties of the predictive distributions. For grade estimation in an ore deposit, model extrapolation performance is of special interest. Aside from high sampling cost and pit-level causality, the application setting also has some bearing on the chosen validation method. As forward prediction capability is assessed, using hold-out data sourced from future benches (new territories where prediction is required) is preferred over k-fold cross-validation, as the latter would set aside unrepresentative test data taken from the same region as the training data. In terms of organisation, Section 2 describes the model candidates considered in this work. This encompasses simple and ordinary kriging and different forms of Gaussian process regression, computed with and without Gaussian simulations. Section 3 introduces the proposed measures. These include classical histogram distances, a spatial fidelity measure derived from variograms and a reframing

of established uncertainty measures based on the notion of ‘synchronicity’. Section 4 describes the experiments, dataset and geological setting. Section 5 presents results and extensive analysis covering a dozen inference periods and approximately ten domains. The experiments are designed to mimic the staged progression of mining operations in an open-pit mine and subject the models to diverse data distributions. Section 6 summarises the main findings and contributions from this study.

2. Geostatistical Modelling

A fundamental viewpoint of probabilistic models is that the target attribute at each point $x \in \mathbb{R}^d$ is described by a probability distribution rather than a single value. Ascribing to the theory of random functions, the observed value is considered a random realisation of a stochastic process, where the pattern of variation with respect to x can only be described in a statistical sense—typically by the mean and correlation structure in the signal. In this work, the attribute represents the concentration/grade of copper in an ore deposit and the points x correspond to locations in 3D space. The present proposal does not impose restrictions on what modelling techniques can be used. The only requirement is that the models should output a predictive distribution that describes the probable values a continuous random variable may take at each inference location. This can be satisfied, for instance, by a Bayesian model that provides posterior estimates for the mean and standard deviation. In engineering geology, geostatistical approaches based on kriging [23] are generally seen as the methods of choice for spatial interpolation. They are used to estimate the spatial distribution of a regionalised variable in applications such as soil sampling where the available measurements are sparse. Kriging essentially provides a weighted moving average prediction, and it is formulated as the best linear unbiased estimator [24]. This theory is better known as Gaussian processes (GPs) in machine learning circles, where it is characterised as a kernel-based, nonparametric, probabilistic regression (Bayesian inference) technique. Although these perspectives appear somewhat different, it is important to stress that kriging and GP share the same conceptual foundations [25]. In particular, both use the Gaussian process to model the relationship between the input and output variables. The key difference lies in how the hyperparameters are inferred. In kriging, the range of spatial dependence is inferred by fitting data to a semi-variogram. In a GP, the same information is deduced from the covariance function or kernel length-scale parameters which are optimised automatically by maximising the marginal likelihood [26]. Christianson et al. [27] compared the kriging variogram and GP likelihood approaches and found GP predictions to be at least as accurate as ordinary kriging and to offer better uncertainty quantification, particularly with the parameters identifiability (robustness) and reproducibility (scalability) in mind. A point of interest for this study is to see if such findings can be corroborated. However, we emphasise that the general applicability of the assessment framework (to be described in Section 3) is not dependent on the chosen models. To familiarise readers with the modelling theory and computations, the technical details [26,28–39] are collated and described from the GP and kriging perspectives in Part 1 of the Supplementary Material. This supplement also describes sequential Gaussian simulation (SGS) and Cholesky correlated random field simulation (CRF) which will be mentioned in this work.

Model Configurations, Conditional Dependence and Approximations

Table 1 shows eight model candidates that will be considered during performance analysis. SK and OK indicate that only simple kriging and ordinary kriging are used. For GP(L), the mean $\mu(\mathbf{X})$ in Equation (S3) (equations and sections prefixed with S refer to contents in Part 1 of the Supplementary Material) is computed from a local neighbourhood

$N(\mathbf{x}_*)$ centred at $\mathbf{x}_*$, whereas for GP(G), a global mean, $\mu(\mathbf{X}) = \text{constant}$, is assumed. Candidates with the suffix -SGS combine kriging or GP with sequential Gaussian simulation; likewise, GP-CRF combines GP with correlated random field simulation as described in Sections S.1.3.1 and S.1.3.2. The approach is consistent with [40]. For each random realisation s , we generate a sequence of simulated values $\{\tilde{y}_{*,\pi_s(i)}^{(s)}, i = 1, \dots, m\}$ by sampling from a univariate conditional cumulative density function (ccdf) $F(\mathbf{x}_{*,\pi_s(i)}; y | (n + i - 1))$, using effectively the posterior mean and variance estimates $(\hat{\mu}_{*,\pi_s(i)}, \hat{\sigma}_{*,\pi_s(i)})$ from kriging/GP for $Y(\mathbf{x}_{*,\pi_s(i)})$. The conditional $| (n + i - 1)$ implies using the n data points in D plus $i - 1$ previously simulated values, $\tilde{y}_{*,\pi_s(1)}^{(s)}(\mathbf{x}_{*,\pi_s(1)})$ through to $\tilde{y}_{*,\pi_s(i)}^{(s)}(\mathbf{x}_{*,\pi_s(i)})$, following a random path with rearranged indices given by π_s . In practice, this conditional dependence is limited in scope by a spatial neighbourhood, $N_{\mathbf{x}_{*,\pi_s(i)}}$, which is imposed to mitigate computation cost. Hence, it would be more accurate to express the ccdf as $F(\mathbf{x}_{*,\pi_s(i)}; y | N_{\mathbf{x}_{*,\pi_s(i)}}(n + i - 1))$ to highlight the neighbourhood restriction in simulation s at step i , which applies to all -SGS model candidates.

For modelling without simulations, viz., SK, OK and GP(L), the neighbourhood restriction still applies. However, there is no permutation and no dependence on other estimated points; thus, the ccdf may be modified as $F(\mathbf{x}_{*,i}; y | N_{\mathbf{x}_{*,i}}(n))$ —this is a subtle but important distinction. Although kriging and GP are often viewed as equivalent, their computational differences with respect to hyperparameters estimation and other implementation differences make them practically different in this study. First, with respect to the kernels, all kriging models (SK, SK-SGS, OK and OK-SGS) use an isotropic kernel described by a range and a Matérn shape parameter (r and ν), whereas all GP models employ a Matérn 3/2 kernel ($\nu = 3/2$) with heterogeneous length scales (l_x, l_y, l_z), amplitude and noise parameters (a, σ_v). Second, with respect to the neighbourhoods ($N_{\mathbf{x}_{*,i}}$), kriging uses spherical search (with an equal share of points selected from each octant where possible), while GP applies a search ellipsoid in rotated space and returns at least a minimum (configurable) number of points in sparse regions. The final remark is that while normal score transformation (nst) is always applied in SGS and CRF simulations, for the candidates SK, OK, GP(L) and GP(G), modelling is performed on both raw data and normal-score-transformed data. The latter will be annotated with ‘nst’ for clarity, and the moments are estimated for a function of a random variable. Suppose the nst values are denoted $y = g(z)$. Having obtained $\text{Var}[g(Z)] \equiv \hat{\sigma}_Y^2$, the mean is given approximately by $\hat{\mu}_Z = g^{-1}(\mathbb{E}[Y])$, and the variance in Z , $\hat{\sigma}_Z^2$, is obtained as shown in (1) using Taylor series approximation [41] in preference to Monte Carlo simulation.

$$\text{Var}[g(Z)] = g'(\mu)[g'(\mu) - \mu g''(\mu)]\text{Var}[Z] + \text{higher order terms} \implies \hat{\sigma}_Z^2 \approx \frac{\text{Var}[g(Z)]}{g'(\mu)^2} \quad (1)$$

This is purely a pragmatic decision. For simulated models, the realisations are always back-transformed [40] before the means and variance are computed.

While these differences and approximations may seem like a nuisance and might even be discouraged for model performance ranking in a zero-sum game, that is not the point of this study. For the purpose of demonstrating the assessment methods, there is no need for the models to be optimally constructed or symmetrically configured. Indeed, these variations allow us to emulate how modellers might explore different options in practice, change the modelling parameters or implementation intentionally to assess the impact of those decisions on different facets of model performance.

Table 1. Model candidates for probabilistic copper grade estimation.

Abbreviation	Description	Cross-Reference
1 SK	Simple Kriging	Section S.1.2 (S9) (S11)
2 SK-SGS	Simple Kriging + Sequential Gaussian Simulation	Section S.1.2, Section S.1.3.1
3 OK	Ordinary Kriging	Section S.1.2 (S10) (S12)
4 OK-SGS	Ordinary Kriging + Sequential Gaussian Simulation	Section S.1.2, Section S.1.3.1
5 GP(L)	Gaussian Process Regression (local neighbourhood mean)	Section S.1.1 (S3) (S4)
6 GP(L)-SGS	Gaussian Process + Sequential Gaussian Simulation	Section S.1.1, Section S.1.3.1
7 GP(G)	Gaussian Process Regression (with stationary/global mean)	Section S.1.1 (S3) (S4)
8 GP(G)-CRF	Gaussian Process Spatially Correlated Random Field	Section S.1.1, Section S.1.3.2

3. Performance Measures

This section describes the performance measures used for model evaluation. The first category comprises histogram distance measures that reflect global accuracy in the mean grade estimates. The second category includes variogram-based measures which capture spatial correlation; these are meant to reflect spatial fidelity (local variability) in the model predicted mean. The third category is made up of uncertainty-based measures, which assess the goodness of probabilistic models using both the mean and standard deviation estimates, $\hat{\mu}(\mathbf{x}_*)$ and $\hat{\sigma}(\mathbf{x}_*)$, and the ground truth (actual grade from hold-out test data) $\mu_0(\mathbf{x}_*)$.

3.1. Histogram-Based Measures

The general goal is to measure discrepancies between the mean prediction and ground-truth histograms. For simplicity, let $\mathbf{p}$ and $\mathbf{q}$ be the probability mass functions (pmf) corresponding to the model predicted mean and ground-truth vectors, $\hat{\boldsymbol{\mu}} = [\hat{\mu}(\mathbf{x}_{*i})]_{i=1:m}$ and $\boldsymbol{\mu}_0 = [\mu_0(\mathbf{x}_{*i})]_{i=1:m}$, respectively. To avoid cherry-picking, four options are considered to reflect a range of perspectives based on hypothesis testing, information theory, set theory and the Monge–Kantorovich optimal transportation/distribution morphing problem [42].

3.1.1. Probabilistic Symmetric Chi-Square Measure

The probabilistic symmetric χ^2 histogram distance [43] is a symmetric variant of the regular χ^2 distance. It represents twice the *triangular discrimination* defined by Topsøe [44]:

$$h_{\text{psChi}} = 2 \sum_x \frac{|p(x) - q(x)|^2}{p(x) + q(x)} \quad (2)$$

3.1.2. Jensen–Shannon Divergence

The Jensen–Shannon divergence, h_{JS} in (3), represents a symmetric and smoothed form of the KL divergence [45].

$$\begin{aligned}
 h_{\text{JS}} &= \frac{1}{2} \left[\sum_x p(x) \log \left(\frac{2p(x)}{p(x) + q(x)} \right) + \sum_x q(x) \log \left(\frac{2q(x)}{p(x) + q(x)} \right) \right] \\
 &= \frac{1}{2} [KL(\mathbf{p} \parallel \mathbf{m}) + KL(\mathbf{q} \parallel \mathbf{m})] \\
 &\text{where } KL(\mathbf{p} \parallel \mathbf{q}) = \sum_x p(x) \log(p(x)/q(x)) \text{ and } \mathbf{m} = (\mathbf{p} + \mathbf{q})/2
 \end{aligned} \quad (3)$$

The second line expresses h_{JS} in terms of the Kullback–Leibler divergence, which is also known as relative entropy. Using a base-2 logarithm, this measure satisfies the bounds $0 \leq h_{\text{JS}} \leq 1$, attaining zero when $\mathbf{p} = \mathbf{q}$.

3.1.3. Ruzicka Distance

The Ruzicka distance, h_{Ruz} , is defined by Ruzicka similarity S_{Ruz} . Given two probability mass functions, p and q ,

$$h_{\text{Ruz}} = 1 - S_{\text{Ruz}} = 1 - \frac{\sum_x \min\{p(x), q(x)\}}{\sum_x \max\{p(x), q(x)\}} \quad (4)$$

S_{Ruz} may be interpreted as the intersection between p and q over the union of p and q and abbreviated as IoU [46]. It generalises the Jaccard similarity index from $\{0, 1\}^m$ to $\mathbb{R}^m$. The Ruzicka distance is bounded by $0 \leq h_{\text{Ruz}} \leq 1$.

3.1.4. Wasserstein Distance

The Wasserstein distance, W_1 , also called the Earth-mover's distance, EM distance or Kantorovich optimal transport distance, is a similarity metric that may be interpreted as the minimum energy cost of moving and transforming a pile of dirt in the shape of one probability distribution into the other [47]. The cost is quantified by the distance and amount of probability mass being moved. It might be preferred over JS divergence, as the Kantorovich–Mallows–Monge–Wasserstein metric represents the Lipschitz distance between probability measures and has to be K-Lipschitz continuous. When the measures are uniform over a set of discrete elements, the problem is also known as minimum weight bipartite matching. Formally, the k -Wasserstein distance between probability distributions P and Q is defined as an infimum over joint probabilities

$$h_{\text{EM}}(P, Q) \equiv W_k(P, Q) = \inf_{\gamma \in \Pi(P, Q)} \mathbb{E}_{(x, y) \sim \gamma} [d(x, y)] \quad (5)$$

where $\Pi(P, Q)$ is the set of all joint distributions whose marginals are P and Q . In general, it requires solving a linear assignment problem. However, in one dimension, it may be computed simply using order statistics. In particular,

$$W_k(P, Q) = \left(\frac{1}{m} \sum_i |\tilde{p}_{(i)} - \tilde{q}_{(i)}|^k \right)^{1/k} \quad (6)$$

where $\tilde{p}_{(i)}$ and $\tilde{q}_{(i)}$ refer to the i th element in the sorted sequence of $\hat{\mu} = [\hat{\mu}(x_{*j})]_{j=1:M}$ and $\mu_0 = [\mu_0(x_{*j})]_{j=1:M}$. Therefore, it does not require quantisation or conversion of $\hat{\mu}$ and μ_0 into histograms or pmf.

3.2. Variogram-Based Measures

A general issue with global accuracy measures is that they ignore spatial correlation. RMSE treats prediction errors as uncorrelated, and histogram distances are no different. Liemohn et al. [48] argued that these measures are incomplete and cannot be solely relied upon as they do not indicate whether the distortion is random or systematic. Therefore, a balanced model assessment framework needs to also examine spatial association. This partially justifies the introduction of a variogram-based measure. A related reason is that geologists involved in model validation are particularly sensitive to the loss of texture (local patterns of variability) in model predictions. Hence, spatial fidelity is viewed as an important assessment criterion.

A variogram ratio statistic is proposed as a basis for measuring the loss of spatial fidelity in a model. This partly stems from the widespread use of semi-variograms in geostatistics. The variogram curve $\gamma(d)$, see Oliver et al. [33], measures the inverse correlation between points as a function of their separating distance, d . A key feature of this curve is the *sill*, which refers to the maximum height of $\gamma(d)$ at large distances as samples become

uncorrelated. Hence, when two variograms are compared and the sill associated with a model is lower than the sill of a reference, it is indicative of smoothing or a reduction in spatial variability. Formally, this may be represented by the ratio between two variogram curves as shown in (7):

$$r_{\text{model}}(d) = \frac{\gamma_{\text{model}}(d)}{\gamma_{\text{reference}}(d)} \quad (7)$$

Percentile statistics such as the median or lower/upper quantiles can be computed from $r_{\text{model}}(d)$ to indicate the average loss in spatial fidelity (equivalently, attenuation in signal power). This allows a visual diagnostic tool to be converted into a quantitative measure. In practice, the sill may also be raised when overfitting occurs, e.g., if the result from one simulation is considered. Thus, a ratio that increases far beyond 1 is also undesirable as it signifies noise amplification. For this reason, the following convex function (symmetrical about $R = 1$) is proposed as a proxy measure for spatial fidelity:

$$\text{Spatial Fidelity } F(R) = \sqrt{1 - |\min\{R, 2\} - 1|}, \text{ where } R = \text{median } r_{\text{model}}(d) \quad (8)$$

This needs to be interpreted judiciously with respect to a suitable reference, such as verification data (actual ground-truth values at the predicted locations) or training data from which the kriging variogram or GP kernel hyperparameters are learned.

3.3. Uncertainty-Based Measures

All models (listed in Table 1) will provide an estimate $(\hat{\mu}_j, \hat{\sigma}_j)$ at inference location $\mathbf{x}_{*j}$ and be compared with the ground truth $\mu_{0,j}$. Under the Gaussian assumption, the second-order statistics $\{(\hat{\mu}_j(\mathbf{x}_{*j}), \hat{\sigma}_j(\mathbf{x}_{*j}))\}_{j=1:m}$ are sufficient for characterising the conditional distribution $p(\mathbf{f}_* | \mathbf{X}_*, \mathbf{X}, \mathbf{f})$. Since it is clear that we are dealing with predicted points, henceforth the $\mathbf{x}_*$ notation will be dropped. To assess how reasonable these estimates are given verification data $\{\mu_{0,j}\}_{j=1:m}$, it is useful to convert $(\mu_{0,j}, \hat{\mu}_j, \hat{\sigma}_j)$ into Z scores via $z_j = (\mu_{0,j} - \hat{\mu}_j)/\hat{\sigma}_j$. In Figure 2, the black dots represent true values. Panels (a,b) illustrate the case where the model mean *underestimates* the true value ($\hat{\mu} < \mu_0$), while panels (c,d) illustrate the case where the true value is *overestimated* ($\hat{\mu} > \mu_0$). The shaded area in (a) represents the coverage probability, p . Its complement, $1 - p$, describes the local consensus between the model and the true measurement. This corresponds to the area under the tail sections in (b). To distinguish overestimation from underestimation, we define a signed scoring function s called *synchronicity*.

$$\begin{aligned} \text{Synchronicity } S &\equiv s(\hat{\mu}, \hat{\sigma} | \mu_0) \\ &= \begin{cases} 2 \times [1 - \Phi(z)] & \text{if } \mu_0 \geq \hat{\mu} \quad [\text{underestimating}] \\ -2\Phi(z) & \text{otherwise} \quad [\text{overestimating}] \end{cases} \end{aligned} \quad (9)$$

$$= 2 \times [\mathbb{I}(\hat{\mu} \leq \mu_0) \cdot (1 - \Phi(z)) - (1 - \mathbb{I}(\hat{\mu} \leq \mu_0)) \cdot \Phi(z)] \quad (10)$$

where $z = (\mu_0 - \hat{\mu})/\hat{\sigma}$ and $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt$ denotes the CDF of the standard normal distribution. This may be written more compactly using an indicator function in (10). The local consensus is simply the magnitude of s , as shown in (11).

$$\text{Local Consensus } L \equiv l(\hat{\mu}, \hat{\sigma} | \mu_0) = |s(\hat{\mu}, \hat{\sigma} | \mu_0)| \quad (11)$$

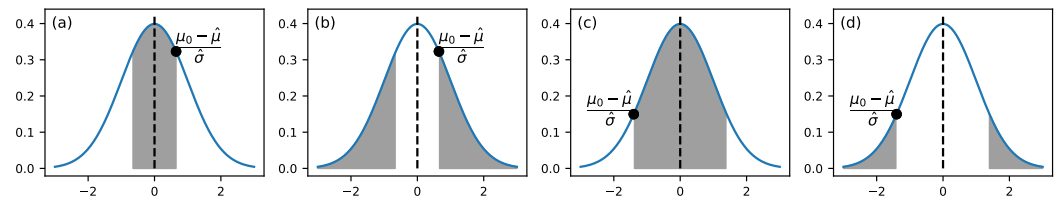


Figure 2. Illustration of coverage probability in (a,c) and consensus between model and observation in (b,d).

3.3.1. Goodness of Model Predicted Uncertainty

One criterion for assessing local uncertainty prediction accuracy is based on the Deutsch [8] goodness statistic. By construction, there is a probability p (identical to the coverage probability described in Section 3.3) that the true value of the random variable falls within a symmetric p -probability interval (PI) bounded by the $p_L = (1 - p)/2$ and $p_U = (1 + p)/2$ quantiles, $Q_L \equiv Q_{(1-p)/2}$ and $Q_U \equiv Q_{(1+p)/2}$, of the estimated conditional distribution function [9]. As a special case, when $p = 0.5$, Q_L and Q_U correspond to the lower and upper quartiles, $Q_{0.25}$ and $Q_{0.75}$, respectively. Given the true measurements $\{\mu_{0,j}\}_{j=1:m}$ at inference locations $\{x_{*j}\}_{j=1:m}$, one is interested in $\bar{\kappa}(p)$, the fraction of true values that are bounded by the PI interval with probability p . Concretely, the expression for $\bar{\kappa}(p)$ in (12) computes the empirical mean over all test locations as a function of p .

$$\bar{\kappa}(p) = \frac{1}{m} \sum_{j=1}^m \kappa_j(p), \quad (12)$$

$$\kappa_j(p) = \begin{cases} 1 & \text{if } \hat{Q}_{(1-p)/2}(j) < Y_j < \hat{Q}_{(1+p)/2}(j) \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

In practice, when the random process (fluctuations about the posterior mean function) is modelled as Gaussian, symmetrical intervals are obtained following Z score transformation, so effectively, $\hat{Q}_{(1-p)/2}(j) = -\hat{Q}_{(1+p)/2}(j)$. The mean proportion K is given by the integral $K = \int_0^1 \bar{\kappa}(p) dp$.

3.3.2. Accuracy of the Estimated Distribution

The average of $[\bar{\kappa}(p) \geq p]$ over p in (14) is known as *distribution accuracy*, $A_{\bar{\kappa}}$.

$$\text{Accuracy } A_{\bar{\kappa}} = \int_0^1 \mathbb{I}_{\bar{\kappa}}(p) dp, \quad (14)$$

$$\mathbb{I}_{\bar{\kappa}}(p) \equiv \mathbb{I}(\bar{\kappa}(p), \xi) = \begin{cases} 1 & \text{if } \bar{\kappa}(p) \geq (1 - \xi)p \\ 0 & \text{otherwise} \end{cases} \quad \text{for a slack variable } \xi \in [0, 0.1]$$

3.3.3. Precision of the Estimated Distribution

Precision measures the narrowness of the model estimated distribution. It is only defined for accurate probability distributions. A p -probability interval that recalls more than $p\%$ of true values is accurate but not precise. Optimal precision means the p -PI contains the true values exactly $p\%$ of the time. On this basis, the precision of the estimated distribution is defined by Deutsch [8] as

$$\text{Precision } P = 1 - 2 \int_0^1 \mathbb{I}_0(p) [\bar{\kappa}(p) - p] dp \quad (15)$$

The precision is only meaningful when there is accuracy, in other words, when the estimated proportions $\bar{\kappa}(p)$ are consistently above the expected proportions p . This can be checked from the accuracy plot of $\bar{\kappa}(p)$ vs. p in the bottom half of Figure 3.

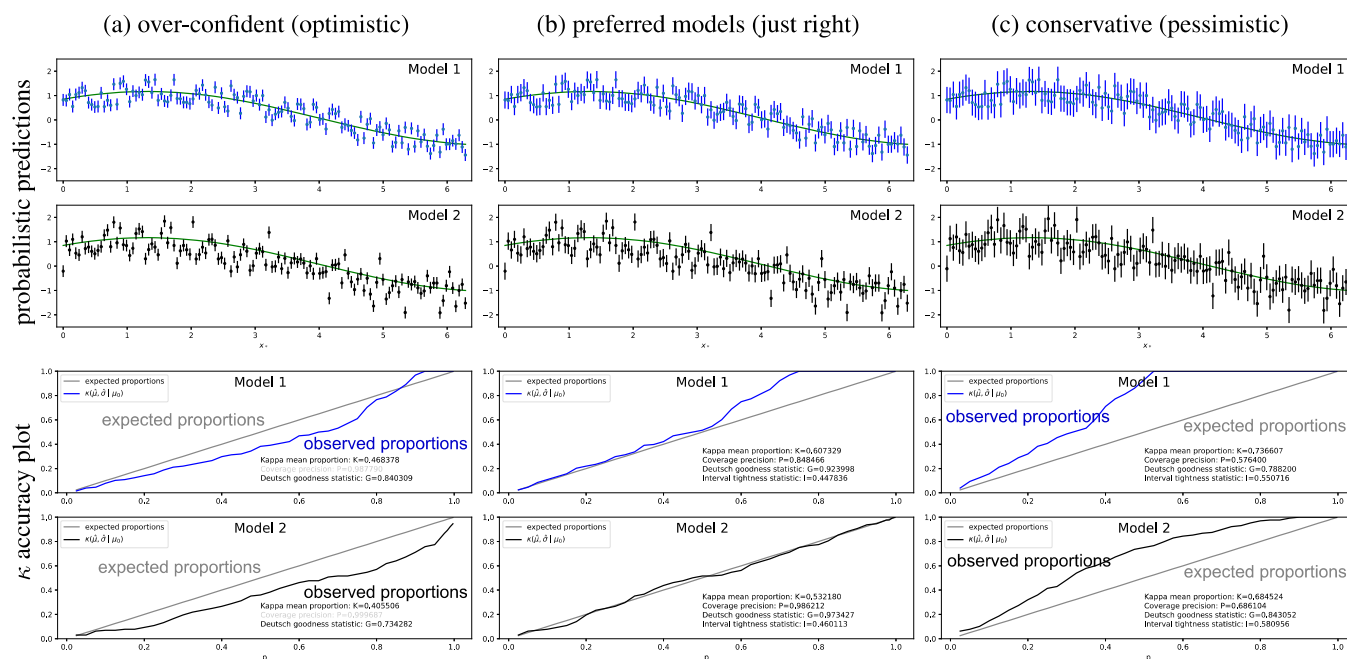


Figure 3. Evaluating the uncertainty and predictive performance of two synthesised models. Top: probabilistic predictions. Bottom: κ accuracy plots. From left to right, (a–c) show what can be expected from the optimistic, preferred and conservative models.

3.3.4. Prediction Uncertainty Goodness Statistic

The closeness between the estimated and theoretical proportions is quantified by G in (16):

$$\text{Goodness } G = 1 - \int_0^1 [3\mathbb{I}_0(p) - 2][\bar{\kappa}(p) - p]dp \quad (16)$$

The G statistic indicates the closeness of points to the bisector of the κ -accuracy plot. Unlike the accuracy and precision, this also considers instances where $\bar{\kappa}(p) < p$. $G = 1$ when $\bar{\kappa}(p) = p \forall p \in [0, 1]$. $G = 0$ when none of the true values are contained in any PIs. The choice of weights indicates that $\bar{\kappa}(p) < p$ is more consequential. The penalty for $\bar{\kappa}(p) < p$ (when observed proportions are below expectation) is twice that for $\bar{\kappa}(p) > p$.

3.3.5. Width of Prediction Uncertainty

For models with similar goodness statistics, one would prefer a model where the p -probability interval is as narrow as possible. A model (or conditional cumulative distribution function) that consistently provides narrow and accurate PIs should be preferred over another that provides wide and accurate PIs. Different notions of spread such as entropy, variance or inter-quartile range can be used. Goovaerts [49] proposed using the interval width in (17) to measure the average tightness of the p -PIs subject to containment of the true value

$$\bar{W}(p) = \frac{1}{m\bar{\kappa}(p)} \sum_{j=1}^m \kappa_j(p) [\hat{Q}_{(1+p)/2}(j) - \hat{Q}_{(1-p)/2}(j)] \quad (17)$$

3.3.6. Prediction Uncertainty Tightness Statistic

The average width of $\bar{W}(p)$ over p can be defined in an analogous manner to A . However, it is more difficult to interpret since it is highly dependent on the data. To make the tightness scale more meaningful, the average uncertainty interval is normalised by the process standard deviation σ_Y observed in the ground truth as shown in (18).

$$\text{Interval tightness } I = \frac{1}{\sigma_Y} \int_0^1 \bar{W}(p) dp \quad (18)$$

In general, both G and I need to be taken into account when assessing probabilistic models, because uncertainty cannot be artificially reduced at the expense of accuracy [8].

3.3.7. Connections

The calculation of $\kappa_j(p)$ can be reframed in terms of $s(\hat{\mu}_j, \hat{\sigma}_j | \mu_{0,j})$ or $l(\hat{\mu}_j, \hat{\sigma}_j | \mu_{0,j})$. Instead of searching for $Q_{(1-p)/2}$ and $Q_{(1+p)/2}$ for various p , there exists a critical value p^* at which $z_{0,j} = (\mu_{0,j} - \hat{\mu}_j) / \hat{\sigma}_j$ lies just on the edge of $[q_L(j, p^*), q_U(j, p^*)]$. This is precisely the purpose of $l(\hat{\mu}_j, \hat{\sigma}_j | \mu_{0,j})$ which converts each input $(\hat{\mu}_j, \hat{\sigma}_j; \mu_{0,j})$ into a Z score $z_{0,j}$ and maps either $q_L(j, p^*)$ or $q_U(j, p^*)$ to $1 - p^*$. Since the interval grows with p , $\kappa_j(p) = 1$ for all $p \geq p^*$, where $p^* = 1 - l(\hat{\mu}_j, \hat{\sigma}_j | \mu_{0,j})$. In the next subsection, we will reinforce the general concepts with an example and demonstrate the efficacy of the uncertainty-based statistics using synthetic data where the ground truth is known.

3.4. Illustration

From left to right, the panels in Figure 3 illustrate three scenarios. The columns labelled (a), (b) and (c) correspond to optimistic, preferred and conservative settings, respectively. If we restrict our attention to (a), the probabilistic predictions and uncertainty interpretations (κ accuracy plots) occupy the top and bottom halves, respectively. For the predictions, the green smooth curve represents the ground truth. Each prediction at location x_* consists of the mean and uncertainty, which are represented by a dot and vertical bar that signifies $\pm \hat{\sigma}$. This length is somewhat arbitrary; its purpose is to emphasise that we have a predictive distribution. The current choice, $[\hat{\mu} - \hat{\sigma}, \hat{\mu} + \hat{\sigma}]$, corresponds to $[Q_{(1-p)/2}, Q_{(1+p)/2}]$, where $p \approx 0.68$. In blue and black, we have two noisy models. They differ in terms of how much their mean predictions gravitate toward the actual mean, $\mu_0(x_*)$. Model 1 is simulated using a uniform distribution, so its estimated means are more spread out. Model 2 is simulated from a normal distribution, so its mean predictions, $\hat{\mu}(x_*)$, tend to be concentrated around $\hat{\mu}_0(x_*)$ but its tail values extend further out. (In other words, for model 1, $\hat{\mu}(x_*)$ is drawn from a uniform distribution of the form $\mu_0(x_*) + U([-b/2, b/2])$. For model 2, $\hat{\mu}(x_*)$ is drawn from a Gaussian, specified by $\mu_0(x_*) + \text{bias} + \text{const} \cdot \mathcal{N}(0, 1)$. The estimated values are contrived. The attention is purely on $(\hat{\mu}, \hat{\sigma}, \mu_0)$ and the uncertainty-based statistics. It would not have mattered whether it represents extrapolation or not, as no actual training data were used by these fictitious models.) Both models in (a) are considered over-confident, as only a small fraction of the p probability intervals contain the actual mean. This can be seen in the κ accuracy plots which show the observed truth containment ratios, described by $\bar{\kappa}(p)$ in (12), consistently below the expected proportions (p) for most values of p . The ideal situation is depicted in (b) where $\bar{\kappa}(p)$ is close to p and the models (especially model 2) live up to expectations. The models in (c) are considered pessimistic because $\bar{\kappa}(p)$ far exceeds p , as can be seen from the κ accuracy plots.

Inspection of Table 2 confirms that these observations are reflected in the statistics. The local consensus (L) and proportion (K) are lowest in the over-confident scenario (a) where the p -probability intervals captured fewer true values than what was expected. As this situation is remedied in the preferred scenario (b), the precision (P) and goodness (G)

statistics markedly improved. For instance, G increased from 0.734 to 0.973 for model 2. The interval tightness measure (I) is highest/worst in the pessimistic scenario (c), while the precision and goodness statistics also suffered. For instance, P plummeted to 0.576 and 0.686 for models 1 and 2, respectively. These findings are consistent with our expectations. They reveal the strengths and weaknesses of models and show promise for large-scale model evaluation using real-world data.

Table 2. Uncertainty-based statistics for the example in Figure 3.

Model	Setting	$\hat{\mu}$ Distribution [†]	$\hat{\mu}$ Bias	$\hat{\sigma}$ Amplitude [†]	Consensus (L)	Proportion (K)	Precision (P)	Goodness (G)	Tightness (I)
a.1	optimistic	uniform	0	0.15	0.430	0.468	–	0.840	0.336
b.1	preferred	uniform	0	0.225	0.576	0.607 [↑]	0.848 [✓]	0.923 [↑]	0.447
c.1	pessimistic	uniform	0	0.35	0.711	0.736 [↑]	0.576 [↓]	0.788 [↓]	0.550
a.2	optimistic	normal	−0.2	0.15	0.369	0.405	–	0.734	0.331
b.2	preferred	normal	−0.2	0.225	0.494	0.532 [↑]	0.986 [✓]	0.973 [↑]	0.460
c.2	pessimistic	normal	−0.1	0.35	0.656	0.684 [↑]	0.686 [↓]	0.843 [↓]	0.580

[†] $\hat{\mu}$ distribution controls the central tendency and spread of the predicted mean. $\hat{\sigma}$ amplitude scales the prediction interval to emulate different model behaviours.

4. Experiments

This section describes the geological setting, data attributes, design and implementation of the experiments. Analysis will be presented separately in Section 5.

4.1. Geological Setting

The data used in our experiments were obtained from the Bingham Canyon (Kennecott) open-pit copper mine located in Utah. It is classified as a porphyry skarn-hosted copper deposit which describes a copper orebody formed from hydrothermal fluids that originate from a magma chamber. Predating or associated with those fluids are multiple intrusions and vertical dikes of diorite to quartz monzonite composition with porphyritic textures. This basically refers to the appearance of large crystals set in a fine-grained or glassy groundmass on the surface of igneous rocks, which gives rise to its name. Metasomatism further explains how the rocks undergo compositional and mineralogical transformations associated with chemical reactions triggered by the reaction of fluids which invade the protolith [50]. Detailed descriptions of its geomorphology and mineralogical properties can be found in [51,52]. A major orebody characteristic is that successive envelopes of hydrothermal alteration typically enclose a core of disseminated ore minerals in a complex system of hairline fractures and veins known as *stockwork* (see [52] Figures 2, 3, 9 and 10). This mineralisation produces pit maps with >1%, >0.7%, >0.35% and >0.15% copper gradation extending from the inner to the outer zones.

4.2. Data Attributes

The input used for modelling consists of the location x and grade y of blasthole assay measurements taken roughly 20 ± 5 m apart. The sampling, assaying techniques and geological interpretations are elaborated in [53]. This data is grouped spatiotemporally, resulting in 11 geological domains and 12 inference periods. Each geological domain is represented by a four-digit code, LGPR, which represents the limb zone, grade zone, porphyry zone and rock type, respectively. These are determined by geologists based on stratigraphy, lithology and other relevant information that control mineralisation and ore/waste boundaries; see [51,52]. The spatial structure of these domains can be seen in Figure 4. An important property is that geochemical diversity is localised and reflected through the grade distribution in these domains. This can be seen in Figure 5. From a modelling perspective, variations in the grade distribution (in terms of skewness, dispersion

and shape) across different domains are useful as they mitigate selection bias and allow the robustness of models to be properly tested.

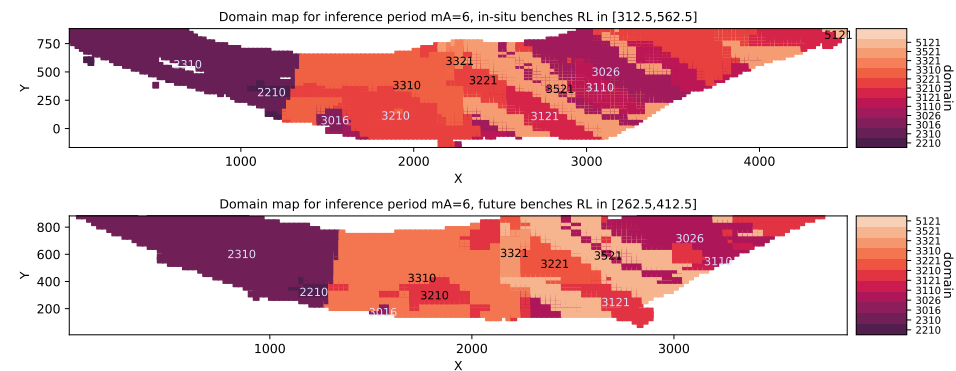


Figure 4. The orebody is partitioned into different geological domains to facilitate copper grade modelling. The panels show correlated spatial structure and subtle changes at different RL elevations. Top image shows blocks associated with active mining operations, i.e., benches that may benefit from in situ regression. Bottom image shows blocks associated with future-bench prediction which require extrapolation. It should be noted that the actual spatial coordinates are shifted so that the minimum coordinates of the study area are close to the origin in the Cartesian coordinates system. This applies also to the RL elevation (ft) to anonymise the data due to commercial sensitivity.

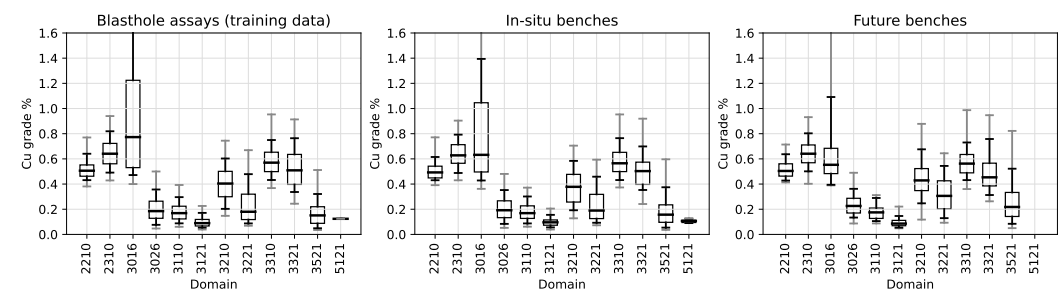


Figure 5. Copper grade distribution across different domains for inference period $m_A = 6$. In each box-plot, the outer (faint) and inner (dark) whiskers represent the 2.15/97.85 and 8.87/91.13 percentiles, whereas the horizontal bar and box edges represent the median and lower/upper quartiles, respectively. From an economic perspective, porphyry orebodies can be mined profitably from Cu concentrations as low as 0.15–0.3%. The left, middle and right plots pertain to blasthole training data, blocks that require in situ regression and future-bench prediction, respectively. As expected, in situ distributions more closely resemble the training data than future-bench distributions.

4.3. Experimental Design

The experiments were designed to emulate the staged observations and progression of mining operations in a real mine. For *future-bench prediction*, each inference period (m_A) signals the intent that the probabilistic models will use data gathered prior to the month of m_A to predict into new locations relevant to mine planning for the next three months (for instance, the months of April, May and June if $m_A = 4$). These new locations represent regions below or adjacent to the current bench; thus, blasthole measurements will not be available since these benches have not yet been drilled or developed. A less technically challenging problem is in situ *regression* which requires interpolating the grade within current benches or operating areas where excavation activities might be planned; the distinction is that this does not require extrapolation into new territories. The number of blasthole samples available for training (n) and inference locations that require grade prediction (m) are both highly variable; some statistics are shown in Table 3.

Table 3. Sample size statistics for in situ interpolation and future-bench extrapolation.

	Number of Blasthole Samples for Training [†]		Number of Locations Requiring Prediction ^{†,‡}	
	In Situ	Future Bench	In Situ	Future Bench
Lower quartile	133	481	96	26
Median	694	1112	680	138
Upper quartile	2338	2293	1968	401
Total count [*]	148,302	160,459	125,004	29,689

[†] All figures refer to quantiles per domain, per inference period unless otherwise stated. [‡] Includes only instances where validation measurements are available. ^{*} Sum over all domains and inference periods.

4.4. Implementation

The `eup3m.git` repository provides a Python implementation of all the algorithms described in this paper. The `run_experiments.py` script executes one single experiment at a time given an inference period/month (m_A) and geological domain (g_D) as input, using the standard configuration parameters specified in `rtcma_utils.py`. A bash script is used to run a complete set of experiments asynchronously on a machine with 30 CPUs, iterating over m_A and g_D to produce the full results. Each individual experiment has a model construction phase and performance analysis phase (refer to the eight model candidates in Table 1 and statistical measures described in Section 3). The measures implemented in `rtcma_evaluation_metrics.py` utilise the `stats`, `spatial` and `special` libraries in `scipy`. The GP approaches implemented in `gstatsim3d_gaussian_process.py` utilise the `scipy.linalg` and `scikit-learn` packages. The kriging approaches implemented in `gstatsim3d_kriging.py` utilise the `scikit-gstat` package and extends existing functionalities in `GStatSim` to support 3D data and an irregularly spaced inference grid. For sequential simulation, the selection of random paths is domain- and inference-period-dependent; however, the sequence remains the same for the SK-SGS, OK-SGS and GP-SGS models in each simulation run, s . To achieve consistent and reproducible results, a SHA256 hash is computed for each (m_A , g_D) pair to initialise the state of a random generator, then N_S values are drawn to obtain $N_S = 128$ random seeds which will subsequently determine the order π_s in which the $\{x_{*,j}\}_{j=1:m}$ points are visited as described in Section S.1.3.1 in Supplementary S1.

5. Results

It is worth reiterating that the purpose of this study is to develop methods for assessment that are fit for purpose for grade modelling and forward extrapolation in mineral deposits. This comprises measures that reflect relevant aspects of performance (viz., global accuracy, spatial correlation and calibration properties of the predictive distributions) and assessment procedures that can operate autonomously at scale to support cross-site comparisons irrespective of the domain and target variable. Although the kriging/GP and simulation techniques add substance to model comparisons, the modelling processes and configurations represent entities that lie outside the scope of the proposal. Fundamentally, this contribution is about the type of insights that can be obtained from having a richer set of tools rather than the conclusions that can be drawn for specific models.

The ensuing analysis is organised in two parts to target two related objectives. The first is to familiarise the reader with the proposed statistics and assess how they respond to real data. To reinforce concepts and develop real insight, we will devote our attention to in-depth analysis of two domains in one inference period, examining histogram distances, variogram ratios visually and uncertainty-based measures quantitatively. The second objective is to systematically evaluate the uncertainty and predictive performance of the chosen probabilistic models and interpret the results across all domains and inference

periods. This is conducted with the view of adding extra layers to the analysis to provide a more complete understanding of different facets of model performance.

As a preview, synchronicity will be rendered as a distortion map to visualise potential error clusters (Section 5.1.6); readers will be introduced to an image-based statistics visualisation modality that highlights the conditions a model may struggle in (Section 5.2); significance testing will be performed using standardised and interpretable measures (Section 5.2.7); and the performance gap between in situ regression (interpolation) and future-bench prediction (extrapolation) will be considered in Section 5.3. Table 4 provides a roadmap highlighting the specific objectives designated for each part of the analysis.

Table 4. Analysis roadmap.

Section	Objectives
5.1	Preliminary analysis to illustrate the spatial characteristics of copper grade model predictions in two domains.
5.1.1	- Examine data diversity and correlation between histogram distance measures (global accuracy).
5.1.2–5.1.4	- Examine variogram ratios and differences between models (local accuracy/spatial variability).
5.1.5	- Examine the accuracy and interval tightness of predictive distributions (κ -plots, calibration/uncertainty).
5.1.6	- Use synchronicity measure to render local consensus/distortion maps to identify error clusters.
5.2	Comprehensive analysis across all domains and inference periods. Goals: (1) Demonstrate an approach that supports automated model assessment and meaningful comparison across different domains and configurations; (2) add layers to the analysis to provide a more complete understanding of different facets of model performance.
5.2.1–5.2.6	- Introduce new assessment modality: image-based visualisation of standardised statistics. This focuses attention on instances where models underperform, whether by inspection or anomaly detection.
5.2.7–5.2.8	- Compute confidence interval and test for statistical significance between models.
5.3	- Measure the difficulty of future-bench prediction (extrapolation) relative to in situ regression (interpolation).
5.4	- Consider the convergence and effects of sequential Gaussian simulations.
6	Discuss and reflect on the main features of the proposal and what it enables.

5.1. Analysis 1: Specific Domains Within a Single Inference Period

The analysis throughout Section 5.1 pertains to future-bench prediction. Two geological domains (2310 and 3521) with vastly different geochemical characteristics were selected for analysis from one inference period ($m_A = 4$). The copper concentration values reported in the blasthole training data and ground truth (for estimated locations in future benches) are depicted in the left and right columns in Figure 6, respectively. The reason for including these is to show explicitly the known data $\{x_i, y_i\}_{i=1:m}$ used for fitting variograms, learning kriging weights or GP kernel hyperparameters (on the left) and actual grades for predicted points $\{x_{*j}\}_{j=1:m}$ (on the right) where verification measurements are available. Focusing on domain 2310 first, Figure 7 shows the mean grade predicted by all eight models. A couple of observations stem from these results. First, techniques that rely on (or assume) a stationary mean, such as SK and GP(G), tend to produce predictions that are too smooth compared with the ground truth in Figure 6. Second, simulations—whether SGS or CRF—can improve the spatial fidelity of the predictions. These behaviours are amplified in domain 3521, which exhibits high-grade mineralisation in the northwestern tip. The same observations on oversmoothing and the benefits of simulation can be seen more clearly in Figure 8. In particular, SK and GP(G) both significantly underestimate the Cu peaks.

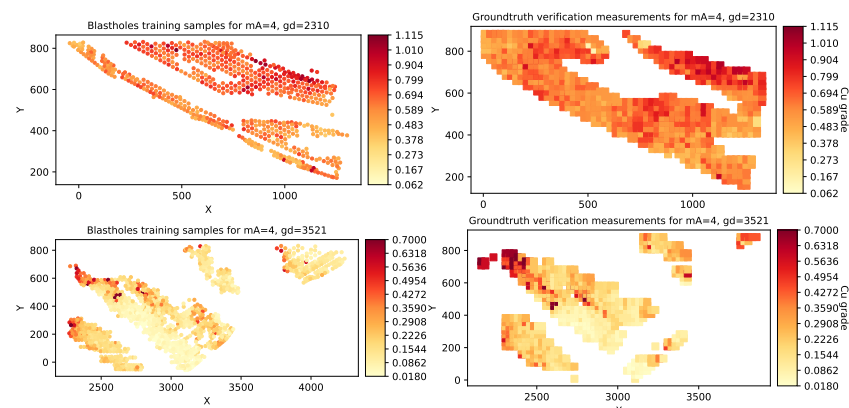


Figure 6. Visualisation of copper grade for blasthole training data (left) and ground truth at predicted locations (right) for two domains, 2310 (top) and 3521 (bottom), in inference period $m_A = 4$.

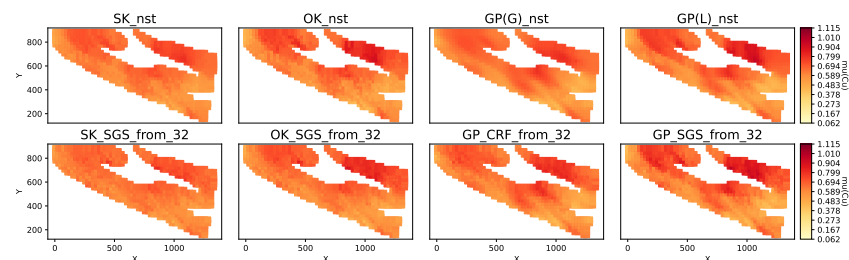


Figure 7. Mean copper grade predicted by models for domain $g_D = 2310$ and inference period $m_A = 4$.

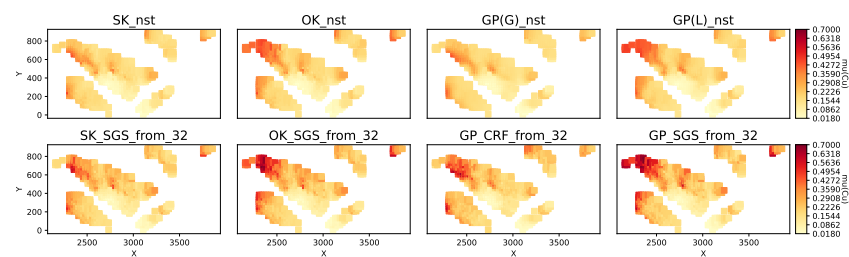


Figure 8. Mean copper grade predicted by models for domain $g_D = 3521$ and inference period $m_A = 4$.

Since the model candidates provide probabilistic predictions, it would be instructive to examine the variance estimates in Figures 9 and 10. In the mining geology context, the kriging/GP variance largely reflects the epistemic uncertainty—due to a lack of training data or spatial sampling—rather than the aleatoric uncertainty which is concerned with inherent grade variability within the deposit [54]. In Figures 9 and 10, it can be seen that the kriging/GP standard deviations (in the top row) show very little spatial variation. With SGS/CRF simulations, the bottom row reveals heterogeneity in the uncertainty when the unsampled values are taken together in space rather than one by one. These findings confirm the benefits of conditional simulations, particularly in relation to remedying the variance deficit (Section S.1.3).

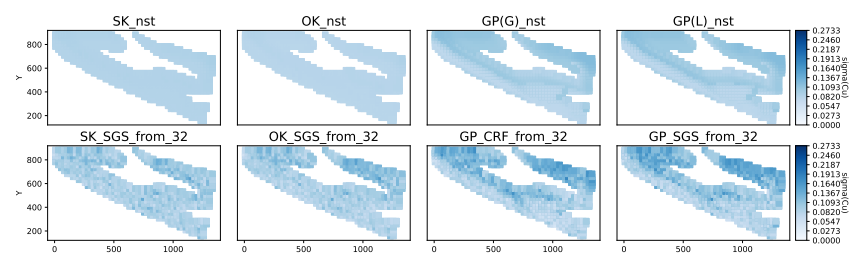


Figure 9. Copper grade standard deviation predicted by models for domain $g_D = 2310$ and inference period $m_A = 4$.

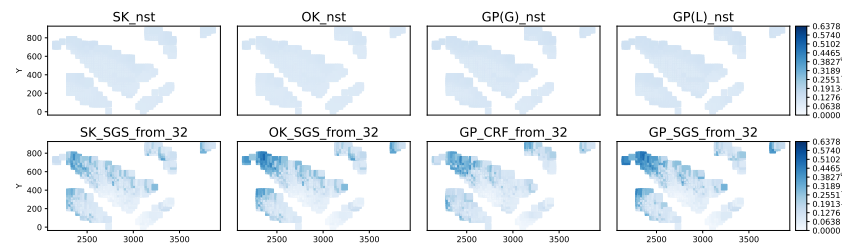


Figure 10. Copper grade standard deviation predicted by models for domain $g_D = 3521$ and inference period $m_A = 4$.

5.1.1. Histograms

To assess global accuracy, histograms are rendered in Figures 11 and 12. In these bar graphs, the ground truth and mean model predictions are represented by black hollow and blue filled columns, respectively. Visually, GP-SGS and OK-SGS can be seen to provide a closer approximation to the ground-truth probability mass function (pmf), whereas the range is more compressed in the case of SK and GP(G), resulting in inadequate coverage of the tail(s).

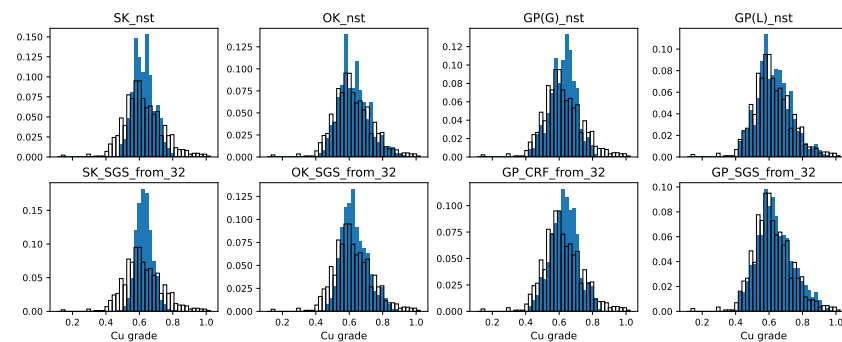


Figure 11. Copper grade histograms for $g_D = 2310$ and $m_A = 4$. Black hollow: ground truth. Blue: model predictions.

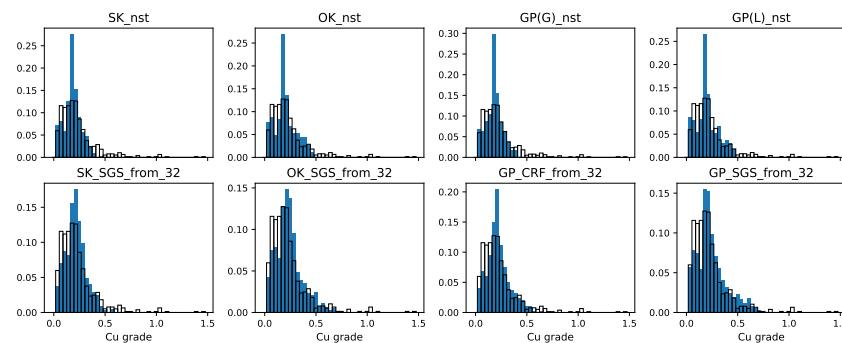


Figure 12. Copper grade histograms for $g_D = 3521$ and $m_A = 4$. Black hollow: ground truth. Blue: model predictions.

The probabilistic symmetric χ^2 , Jensen–Shannon, Ruzicka and Wasserstein histogram distances (described in Sections 3.1.1–3.1.4) are computed and presented in Table 5. Although the trend varies somewhat depending on the domain, a common observation is a general improvement in the histogram rank (equivalently, reduction in histogram distances) when SGS or CRF is coupled with GP. Overall, the computed histogram distances are consistent with our graph-based interpretations.

In Figure 13, the cross-plots show that h_{psChi} and h_{JS} are linearly correlated ($\rho > 0.99$), while h_{Ruz} is strongly correlated with both h_{psChi} and h_{JS} ($\rho \in [0.84, 0.97]$). Since the Jensen–Shannon divergence can be interpreted as an information difference between two distributions and is bounded, we suggest it should be included in general assessments

alongside the Wasserstein histogram measure, h_{EM} , as the latter is not dependent on quantisation and is less sensitive to sample size.

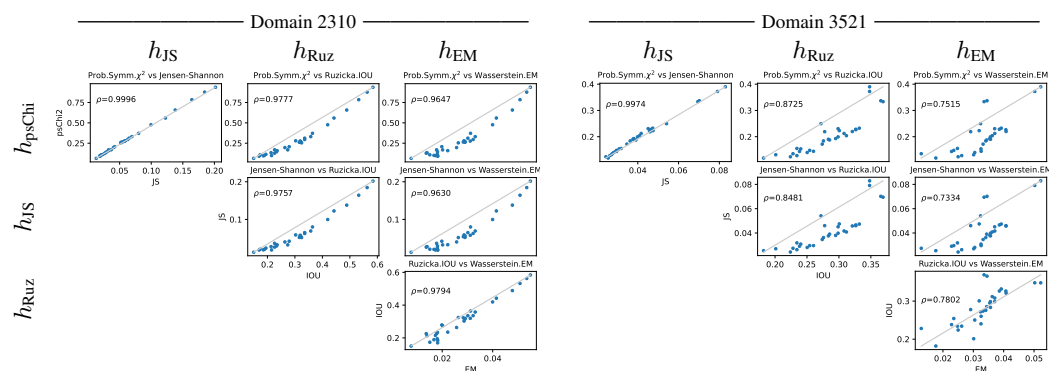


Figure 13. Histogram distances cross-plots for $g_D = 2310$ and $g_D = 3521$ in $m_A = 4$.

Table 5. Histogram distances for mean model predictions relative to ground truth.

Model	Domain 2310					Domain 3521				
	h_{psChi}	h_{JS}	h_{Ruz}	h_{EM}	Rank	h_{psChi}	h_{JS}	h_{Ruz}	h_{EM}	Rank
SK_nst	0.5594	0.1225	0.4422	0.0414	7	0.3723	0.0791	0.3480	0.0502	7
OK_nst	0.2071	0.0422	0.2792	0.0199	3	0.3368	0.0702	0.3656	0.0345	6
GP(L)_nst	0.1273	0.0266	0.2261	0.0137	2	0.3335	0.0695	0.3690	0.0335	5
GP(G)_nst	0.3303	0.0693	0.3578	0.0331	6	0.3900	0.0829	0.3480	0.0522	8
SK-SGS (from 32)	0.8774	0.1845	0.5629	0.0535	8	0.2198	0.0458	0.3210	0.0408	4
OK-SGS (from 32)	0.3119	0.0650	0.3356	0.0302	5	0.1929	0.0380	0.2949	0.0354	3
GP-SGS (from 32)	0.0948	0.0192	0.1851	0.0182	1	0.1902	0.0376	0.2835	0.0357	2
GP-CRF (from 32)	0.2763	0.0580	0.3194	0.0311	4	0.1876	0.0391	0.2854	0.0344	1

5.1.2. Variograms

Variograms are presented separately for models in the SK, OK, GP(L) and GP(G) family along with their SGS/CRF counterparts in Figure 14. To be clear, the northwest, northeast, southwest and southeast quadrants in each half represent the SK/SK-SGS, OK/OK-SGS, GP(G)/GP-CRF and GP(L)/GP-SGS families, respectively. Within a given domain, the variogram plots can be compared directly between families since the scales are the same. Two reference curves—black-solid for the *ground truth* $\{x_{*j}, y_{*j}\}_{j=1:m}$ and black-dashed for *blasthole* training data $\{x_i, y_i\}_{i=1:n}$ —are included to indicate the range of spatial variability the models should strive to achieve. Additionally, a grey curve representing a black-box long-range prediction model and a lilac curve representing GP(L) are included in each plot to provide a benchmark.

Focusing on domain 2310 first, the northwest quadrant shows that simple kriging is not competitive with GP(L), in fact, SK-SGS generally performs far worse than the long-range model which itself is inferior to all other models. In the southwest quadrant, the pink curve representing GP(G) outperforms the long-range model. Henceforth, we use the notation γ_N to denote the variogram for a model/reference N . With sequential Gaussian simulation, the average variogram for a single realisation (orange $\blacktriangle$) matches $\gamma_{blastholes}$ and from two realisations, $\gamma_{GP-CRF \text{ from } 2}$ (orange $\blacktriangleleft$) matches $\gamma_{groundtruth}$ for the most part. As the number of simulations (s) increases, $\gamma_{GP-CRF \text{ from } s}$ becomes smoother and approaches $\gamma_{GP(G)}$. The patterns for ordinary kriging are very similar, except $\gamma_{OK-SGS \text{ from } s} < \gamma_{GP-CRF \text{ from } s}$ for smaller lags.

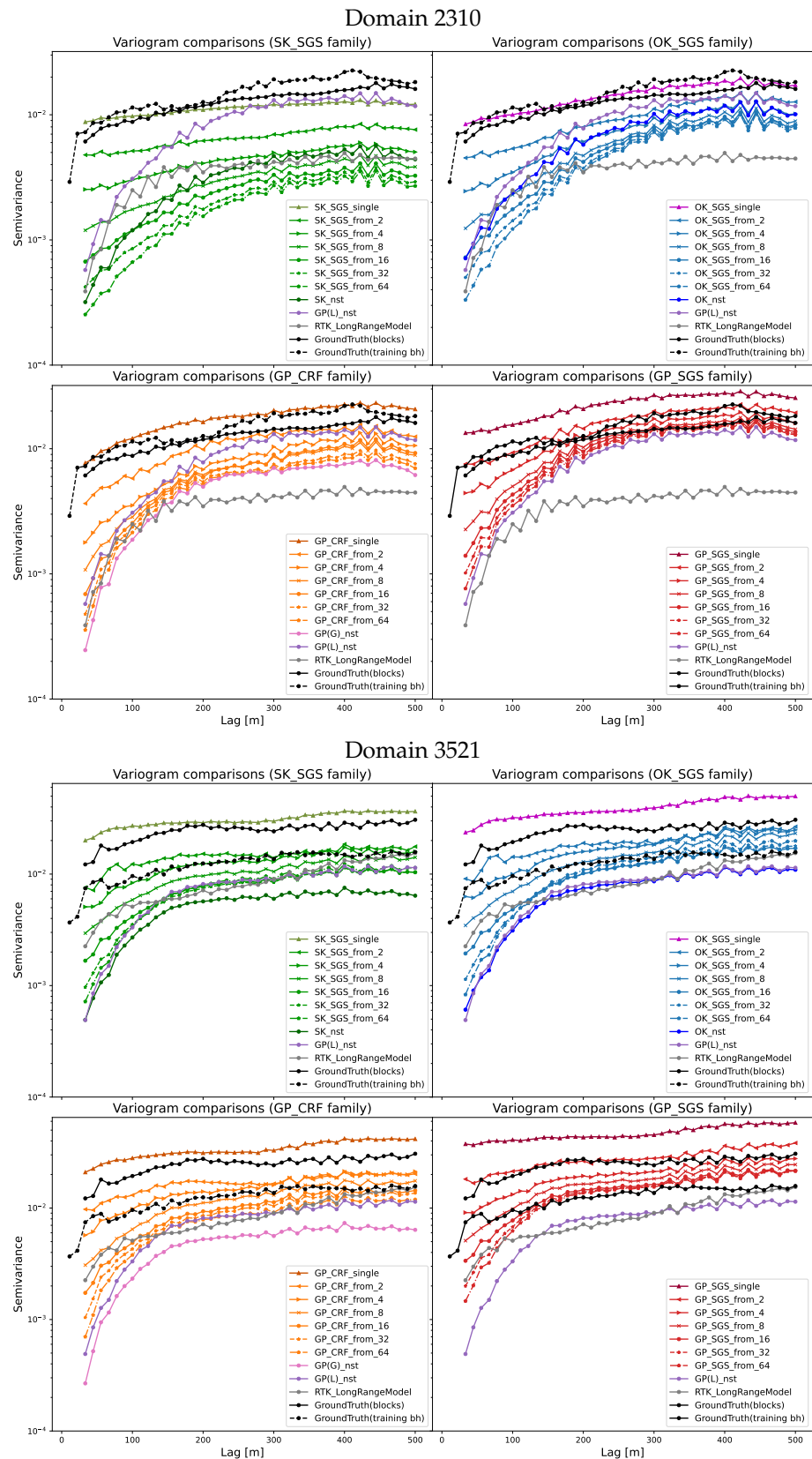


Figure 14. Copper grade variograms for $g_D = 2310$ and $g_D = 3521$ in $m_A = 4$. The quadrants group together models in the following families: (northwest) SK/SK-SGS, (northeast) OK/OK-SGS, (southwest) GP(G)/GP-CRF, (southeast) GP(L)/GP-SGS.

5.1.3. Insights from the Variograms

In the northeast quadrant of Figure 14 (domain 2310), one observes that the blue curves representing $\gamma_{\text{OK-SGS}}$ from s are generally above $\gamma_{\text{long-range}}$ but, more importantly, underneath the lilac curve representing $\gamma_{\text{GP(L)}}$. This indicates GP(L) has the highest spatial fidelity among the base candidates. In the southeast quadrant, it can be seen that sequential simulations can further propel $\gamma_{\text{GP-SGS}}$ above $\gamma_{\text{GP(L)}}$. This finding is highly significant. It shows that while GP(L) can handle mean regression for a non-stationary process through the use of local neighbourhoods, the local variance estimates based on these neighbourhoods do not adequately capture the covariance (full inter-sample dependencies) of the underlying process. The loss, in terms of longer-range spatial correlation, can be replenished through sequential Gaussian simulation (see Section S.1.3.1 and chain rule in (S.13)), which effectively propagates conditional information through random paths.

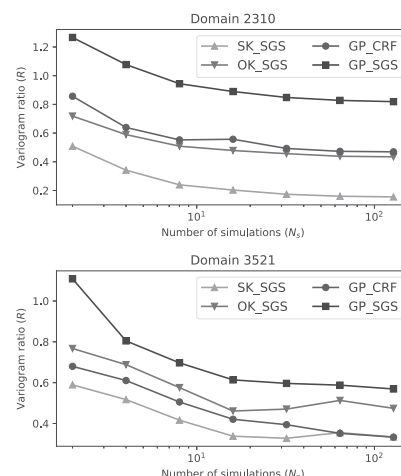
The lessons are similar for domain 3521 in Figure 14, except GP(L) on its own is close to but not necessarily better than the long-range model. We believe this is due to $\gamma_{\text{blastholes}} < \gamma_{\text{groundtruth}}$, viz., the training data used for learning is smoother than the ground truth. This perhaps makes the goal of matching the spatial variability in the ground truth unattainable and the task of future-bench prediction more difficult. The key observation is that SGS is needed to elevate the performance of OK and GP into the target band encompassed by $\gamma_{\text{blastholes}}$ and $\gamma_{\text{groundtruth}}$. A related observation is that GP-SGS pushes the curves higher than OK-SGS.

5.1.4. Practicality

It takes considerable effort to visually compare variograms even for eight models in a single domain. This becomes cumbersome and error-prone when there are over a hundred (g_D, m_A) combinations to assess, as is the case in our later experiments. The variogram ratio (R) and spatial fidelity (F) statistics formulated in Section 3.2 provide a practical measure of model quality, taking into account local spatial correlation in the regression results. These statistics are reported in Table 6. The key finding is that the spatial fidelity of the base models can be boosted with limited rounds of sequential simulation. As a case in point, consider domain 3521 in Table 6. With four rounds of SGS, the F value has increased from 0.575 to 0.829 and from 0.587 to 0.896 for OK and GP(L), respectively. Although spatial fidelity drops off with further rounds of simulations (as illustrated graphically in Table 6), the benefit is sustained for GP-SGS and GP-CRF even after 32 iterations— F values remain at 0.772 and 0.627, respectively, which are substantially higher than 0.575 and 0.587.

Table 6. Variogram ratios (R) and spatial fidelity (F) statistics for $g_D = 2310$ and $g_D = 3521$ in $m_A = 4$.

	Domain 2310		Domain 3521	
	R	F	R	F
SK_nst	0.2775	0.5267	0.2328	0.4824
OK_nst	0.5945	0.7710	0.3311	0.5754
GP(L)_nst	0.7568	0.8699	0.3451	0.5874
GP(G)_nst	0.4525	0.6726	0.2217	0.4708
SK-SGS (from 4)	0.3415	0.5843	0.5171	0.7190
OK-SGS (from 4)	0.5894	0.7677	0.6882	0.8295
GP-SGS (from 4)	1.0774	0.9605	0.8046	0.8969
GP-CRF (from 4)	0.6391	0.7994	0.6105	0.7813
SK-SGS (from 32)	0.1736	0.3998	0.3275	0.5722
OK-SGS (from 32)	0.4565	0.6756	0.4705	0.6859
GP-SGS (from 32)	0.8481	0.9209	0.5964	0.7722
GP-CRF (from 32)	0.4929	0.7020	0.3941	0.6277



5.1.5. Prediction Accuracy and Uncertainty Intervals

This section examines the accuracy and interval of the predictive distributions both qualitatively and quantitatively. In domains 2310 and 3521, we observed that all model curves have similar shapes in the accuracy and interval plots, $\bar{\kappa}(p)$ and $\bar{W}(p)/\sigma_Y$. Therefore, it suffices to illustrate the general behaviour through one model family, viz., ordinary kriging and OK-SGS. Results depicted in Figure 15 are typical of all models within the respective domains. The main finding that can be distilled from these plots is that the models are very accurate in domain 2310, as is evident from its high precision (closeness between the observed and expected proportions, $\bar{\kappa}(p)$ and p). Based on the interpretations of Section 3.4, the fact that $\bar{\kappa}(p)$ exceeds p more substantially in domain 3521 (note: p denotes the expected ground-truth capture probability) suggests that these models are more conservative; this is reflected by lower precision (P) and goodness (G) in domain 3521.

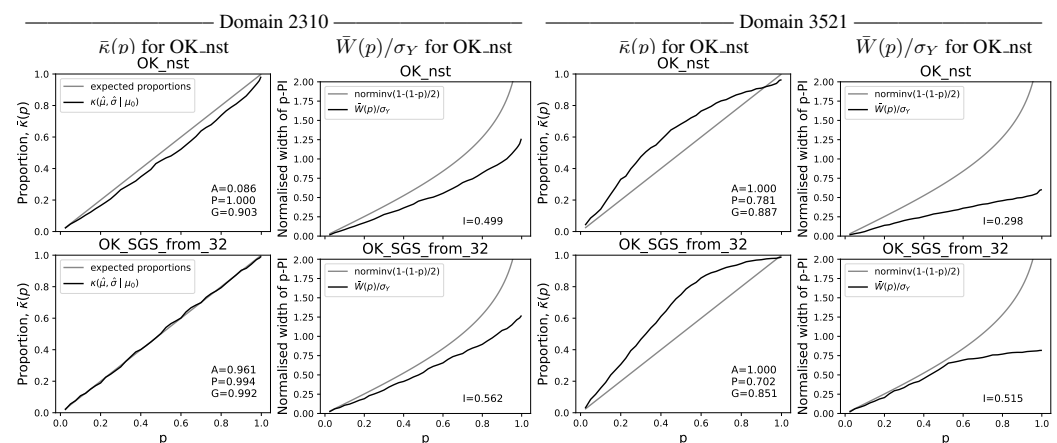


Figure 15. Copper grade predictive distribution accuracy and uncertainty interval plots. Selected results for $g_D = 2310$ and $g_D = 3521$ in $m_A = 4$.

A detailed reading of the statistics shown in Table 7 shows that the results are mixed and no dominant model can be established between these two domains. Regarding the individual indicators, the following comments can be made. The accuracy, $A_{\bar{\kappa}}$, is quite sensitive even when a slack variable $\xi = 0.05$ is employed—see (14). It can change rapidly from 0 to 1, see domain 2310 OK_nst vs. GP(G)_nst, due to the hard constraint it imposes on the observed vs. expected proportion comparison. As the number of simulation runs increases, the precision and goodness statistics (P and G) both get slightly worse due to stochasticity. This pattern runs contrary to the local consensus trend (L). Philosophically, the notion of p -probability intervals, which is built on the notion of ground-truth capture and what a model promises and actually delivers, might not be a great criterion to judge models on. On current evidence, these p -PI statistics seem to be quite limited in their capacity at differentiating models. However, we reserve final judgement on their utility as the scope of our analysis is quite limited at this point. This issue will be revisited in part two of our analysis (Section 5.2).

Looking at the base models in Table 7, what is clear is that GP models produce higher consensus scores (L) than kriging models in both domains. For example, the consensus scores for GP(L) and GP(G) [$L = 0.496, 0.518$] are higher than those for SK and OK [$L = 0.481, 0.451$] in domain 2310, and the same can be said for domain 3521. The consensus scores also show SGS/CRF improve prediction performance, and this effect increases with more simulation runs. Significantly, this improvement is geared toward bringing the predictions closer to the ground truth (μ_0) rather than mere containment of the ground truth within a prediction interval [9]. This distance-based interpretation follows

from (10)–(11) where local consensus L is defined in terms of synchronicity $s(\hat{\mu}, \hat{\sigma} | \mu_0)$, which is driven by Z scores, $z = (\mu_0 - \hat{\mu})/\hat{\sigma}$.

Table 7. Uncertainty-based statistics for $g_D = 2310$ and $g_D = 3521$ in $m_A = 4$. Specifically, L , A , P , G and I denote the local consensus, accuracy, precision, goodness and interval tightness of the probabilistic predictions.

	Domain 2310					Domain 3521				
	L	$A_{0.05}$	P	G	I	L	$A_{0.05}$	P	G	I
SK_nst	0.4817	0.6172	1.0000	0.9631	0.5182	0.6231	1.0000	0.7480	0.8686	0.2918
OK_nst	0.4517	0.0859	0.9999	0.9030	0.4992	0.6075	1.0000	0.7808	0.8866	0.2984
GP(L)_nst	0.4963	0.9062	0.9953	0.9857	0.5752	0.6353	1.0000	0.7272	0.8618	0.3071
GP(G)_nst	0.5184	1.0000	0.9628	0.9812	0.5548	0.6294	1.0000	0.7361	0.8634	0.2965
SK-SGS (from 32)	0.4967	0.9297	0.9987	0.9913	0.5652	0.6506	1.0000	0.6978	0.8484	0.4246
OK-SGS (from 32)	0.5010	0.9609	0.9936	0.9925	0.5617	0.6485	1.0000	0.7022	0.8509	0.5152
GP-SGS (from 32)	0.5071	0.7383	0.9753	0.9771	0.6531	0.6411	1.0000	0.7172	0.8585	0.5311
GP-CRF (from 32)	0.5265	0.9961	0.9462	0.9723	0.6208	0.6255	0.9727	0.7477	0.8730	0.4779
ine SK-SGS (from 128)	0.5116	0.9805	0.9754	0.9863	0.5659	0.6728	1.0000	0.6535	0.8266	0.4273
OK-SGS (from 128)	0.5142	0.9688	0.9707	0.9844	0.5661	0.6698	1.0000	0.6598	0.8298	0.5187
GP-SGS (from 128)	0.5232	0.9609	0.9530	0.9759	0.6475	0.6650	0.9961	0.6693	0.8346	0.5370
GP-CRF (from 128)	0.5434	1.0000	0.9127	0.9563	0.6085	0.6592	0.9961	0.6810	0.8404	0.4491

5.1.6. Synchronicity as a Visualisation Tool

It is worth highlighting the potential of the synchronicity measure, $S \stackrel{\text{def}}{=} s(\hat{\mu}, \hat{\sigma} | \mu_0)$, for model evaluation from a spatial perspective. By construction, $S > 0$ (resp. $S < 0$) when the predicted mean underestimates (resp. overestimates) the true grade. These instances are rendered in red (resp. blue) in Figures 16 and 17. Larger deviations from the ground truth are indicated by a darker shade. Following this convention, these figures can serve effectively as local distortion maps. Specifically, the blue cluster in the northwest corner in Figure 16 show areas where the OK model underperformed by way of overestimating the Cu grade. The relative strength of the GP(G) model is highlighted by lighter patches at the corresponding location. Moving over to the label marked ‘A’ in the east, GP(L) can be seen to provide a better estimation than GP(G) whereby the intensity of the red patches (underestimation) is reduced. There is perhaps nowhere more obvious than in Figure 17, where the problem of underestimation at the northwestern tip is conspicuous in all base models and the magnitude of the prediction error is significantly reduced through SGS/CRF simulation.

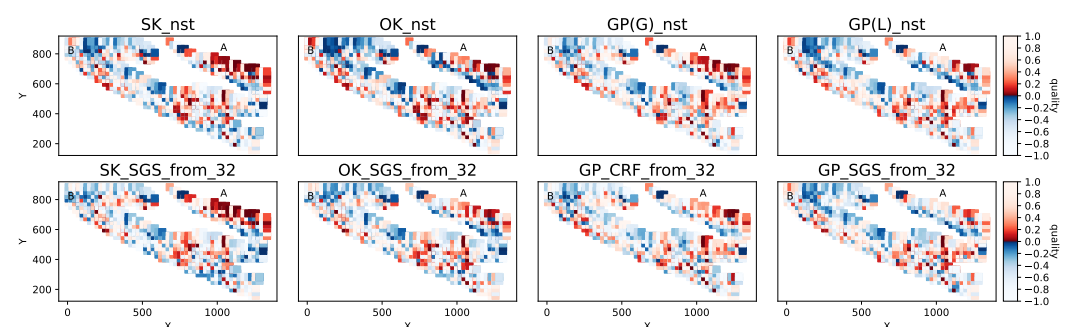


Figure 16. Synchronicity of grade predictions with regard to the ground truth for $g_D = 2310$ and $m_A = 4$.

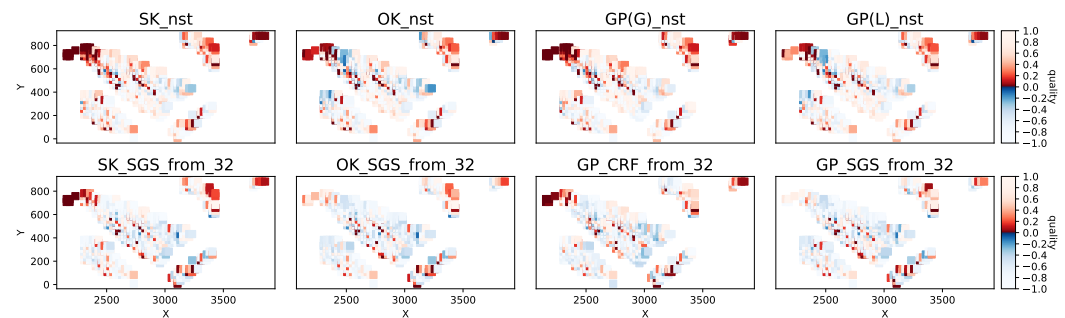


Figure 17. Synchronicity of grade predictions with regard to the ground truth for $g_D = 3521$ and $m_A = 4$.

5.2. Analysis 2: Performance of Probabilistic Models Across All Domains and Inference Periods

Consolidating the points in the preliminary analysis, this section now looks at the broader picture across all inference periods and domains. This is prompted by a desire to minimise selection bias and determine the stability of models under varying conditions. A key motivation is to obtain statistically significant results so that findings arising from random chance can be effectively ruled out. Readers can expect to see qualitative and quantitative analysis on future-bench prediction performance, including a statistical comparison with in situ regression in Section 5.3. The chief strategy advocated in this paper is to view various statistics from an image perspective, whereby models and conditions (inference period and domain) are represented by the vertical and horizontal axes, respectively. This takes inspiration from microplates [55], a standard screening tool used in clinical diagnostic testing such as enzyme-linked immunosorbent assay (ELISA), whereby antigen–antibody interactions are detected within a 2D array. This has been used in biochemistry to study enzyme diversity in soils [56]. A similar setup that exploits this attention mechanism is equally well suited to large-scale simultaneous comparisons in geostatistics.

5.2.1. Histogram Distances

As an example, the Jensen–Shannon and Wasserstein histogram distances, h_{JS} and h_{EM} , are visualised as images in the left and right halves of Figure 18, respectively. In these arrays, the rows represent models which are grouped along family lines into four categories: SK, OK, GP(G) and GP(L). The columns represent geological domains (see outer x labels). Furthermore, successive inference periods (m_A) are interleaved within each domain (see inner x labels). Looking at h_{JS} , these results may be interpreted in two ways. At a macro-level, the GP(G) and GP(L) families, represented by the third and fourth blocks down the y axis, appear much darker than the rest. These indicate lower distortion in the predicted grade histograms relative to the ground truth. The relevant group statistics are summarised in Table 8. At a granular level, differences between row 0 (SK) and row 1 (SK_nst) illustrate the importance of normal score transformation in simple kriging. Focusing on higher-level trends, more pervasive distortion can be seen in the SK and OK families, as is evident from the prevalence of brighter pixels in the first and second blocks.

Table 8. Histogram distance summary statistics for future-bench prediction over domains and inference periods.

Model Family	Abbrev	h_{JS} Mean	h_{EM} Mean
Simple kriging	SK/SK-SGS	0.4607	0.1678
Ordinary kriging	OK/OK-SGS	0.3524	0.1126
Gaussian process (global mean)	GP(G)/GP-CRF	0.2937	0.0753
Gaussian process (local mean)	GP(L)/GP-SGS	0.2802	0.0691

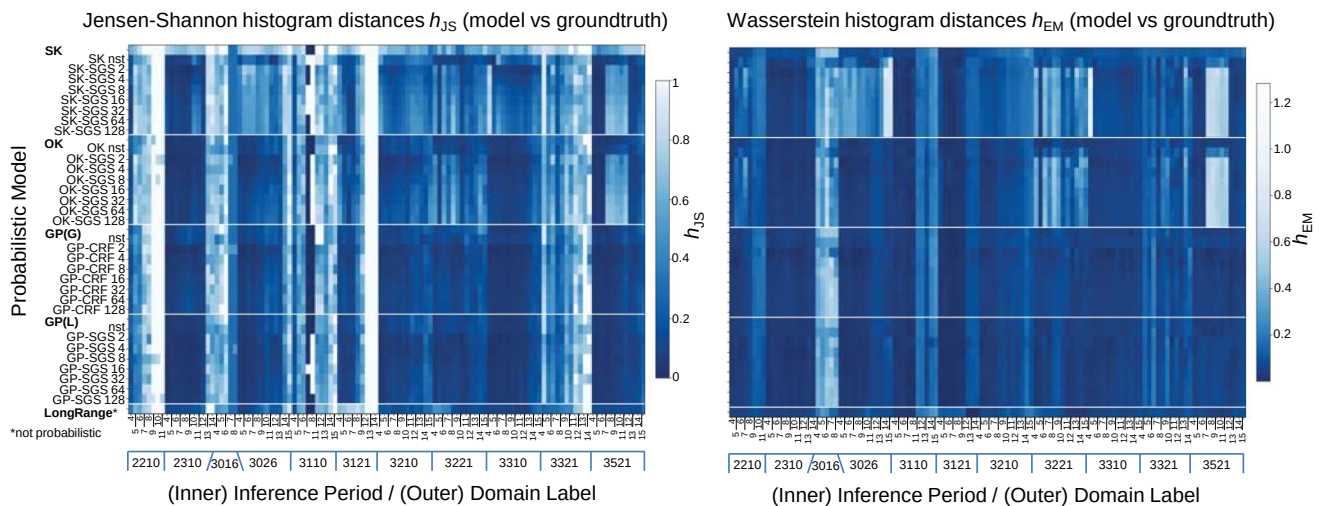


Figure 18. View of (left) Jensen–Shannon and (right) EM histogram distances for future-bench prediction across domains and inference periods.

5.2.2. Influential Factors

An investigation of the bright pixel columns in Figure 18 (left)—instances where SK and OK apparently underperformed—reveals two contributing factors. The first is that histogram distance measures (not just h_{JS} but also h_{psChi} and h_{Ruuz}) are sensitive to discretisation and number of inference points (n_I) used in a given ground-truth comparison; this is not a modelling artefact. The second is a divergence between the training data and ground-truth distributions. By way of an example, the first phenomenon is evident from the white patches that appear in the 2210 columns in Figure 18 (left), and this coincides with $n_I \leq 16$ from $m_A = 7$ to $m_A = 11$ in Table 9. This indicates a drop in the efficacy of h_{JS} as the sample size decreases. For domain 2310, the number of inference points is similarly small for $m_A = 13$ and 14; we see a similar drop-off as h_{JS} becomes unreliable. On the contrary, h_{EM} (see corresponding columns in Figure 18 (right)) is quite insensitive to sample size.

For domain 3016, the number of inference points is once again very small (mostly $n_I \leq 10$ in Table 9). However, h_{JS} and h_{EM} are both large; this indicates the degradation in performance is genuine. Looking at the distribution of the training data in Figure S6 (see Part 2 of the Supplementary Material), we hypothesise that this is due to the spread (almost uniform distribution) observed in this domain. Prediction is more difficult when the entropy of the measured data is high. This may indicate volatility in the grade distribution (an intrinsic property in certain parts of the deposit) or incorrect domaining (epistemic uncertainty attributable to data sparseness and boundary uncertainty).

Table 9. Sample size statistics for certain domains (g_D) and inference periods (m_A) involved in future-bench prediction.

g_D	m_A	n_T	n_I	g_D	m_A	n_T	n_I	g_D	m_A	n_T	n_I	g_D	m_A	n_T	n_I	g_D	m_A	n_T	n_I
2210	4	76	66		10	123	5		6	86	10		15	2120	10		9	208	28
	5	66	30		11	123	4		7	86	2	3321	4	14	49		10	224	36
	6	114	37	2310	13	2653	19		8	86	2		5	61	101		11	227	12
	7	109	16		14	2654	19	3026	12	2050	111		6	59	104		12	242	16
	8	130	15	3016	4	67	28		13	2088	75		7	143	108		13	238	4
	9	121	9		5	81	8		14	2113	35		8	195	68		14	242	4

n_T and n_I denote number of training and inference points.

Domain 3026 is afflicted by the same issues (discretisation and few inference points) as domain 2210. The bad behaviours observed in $m_A = 14$ and 15 correlate directly with sample size in Table 9. The most striking results for h_{JS} occur in domains 3110, 3121 and 3321. For domain 3110, a slight performance drop-off is observed in $m_A = 11$ and beyond. Examination of the training data distribution and ground-truth grade distribution in Figure S6 reveals fundamental differences between the two. The higher grade values present in the ground truth were beyond anything seen in the training data; thus, they are quite unexpected and hard to predict. For domain 3121, the significant elevation in h_{JS} from $m_A = 12$ to $m_A = 14$ is due to the propensity of samples lying outside the grade range observed in the training data. From Figure S6, it can be verified the training data and ground-truth distributions hardly intersect; hence, their JSD similarity is only 23.9%.

For domains 3210, 3221, 3310 and 3521, there is general consensus between h_{JS} and h_{EM} . The GP models all performed well with respect to both measures. For domain 3321, the moderately elevated h_{JS} from $m_A = 4$ to $m_A = 6$ is due to the small number of training samples available (see Table 9) and the slight mismatch between the training data and ground-truth distributions. The much elevated h_{JS} from $m_A = 9$ and beyond is due to the small number of inference points and sensitivity to histogram discretisation. For domain 3521, a persistent cluster of poor performance is observed for kriging models from $m_A = 7$ to $m_A = 11$. These instances were found to occur when there is a highly positive-skewed, long-tail ground-truth distribution coinciding with a more narrow training data distribution with a more left-leaning mode; this is illustrated in Figure S6. This issue also affects domain 3221 to some extent. These technically more challenging situations for kriging-based models can be seen clearly and unambiguously as light colour patches in the Wasserstein image in Figure 18 (right). This makes it imperative to include h_{EM} in global accuracy assessment if the confounding effects due to sample size or discretisation are to be suppressed.

5.2.3. Spatial Fidelity

The same techniques are used to examine variogram ratios and spatial fidelity across all periods and domains. What is different about Figure 19 is the appearance of dotted cells. These represent instances where a variogram cannot be reliably computed (when the number of inference points $n_I < 30$) and are used to avoid confusion with bright pixels (more extreme ratios) which are undesirable. In Figure 19 (left), variogram ratios in the ranges $[0, 1)$ and $[1, 2]$ are rendered in red and purple, respectively, with the colour intensity transitioning from light to dark as the ratios get closer to 1, which is the ideal.

The results in Figure 19 (left) reinforce the findings described in Section 5.1.2 in two important respects. First, SGS/CRF simulation increases the variogram ratios across all domains and inference periods irrespective of the model family: SK, OK, GP(G) or GP(L). The effect is strongest when $s = 2$ and decreases with further simulations (s). However, the benefits are sustained the longest in GP(L); notably, GP-SGS($s = 32$) has higher spatial fidelity than the GP(L)_nst base model. Second, the GP(L) family achieves the highest spatial fidelity among all model candidates, especially when combined with SGS. This can be seen from Figure 19 (right) where the pixels in the lowest block (corresponding to the GP(L) family) are on average the darkest (F being the closest to 1). This conclusion is supported by the summary statistics in Table 10, where the standard error (SE) further demonstrates that the SK and OK modelling results are more variable.

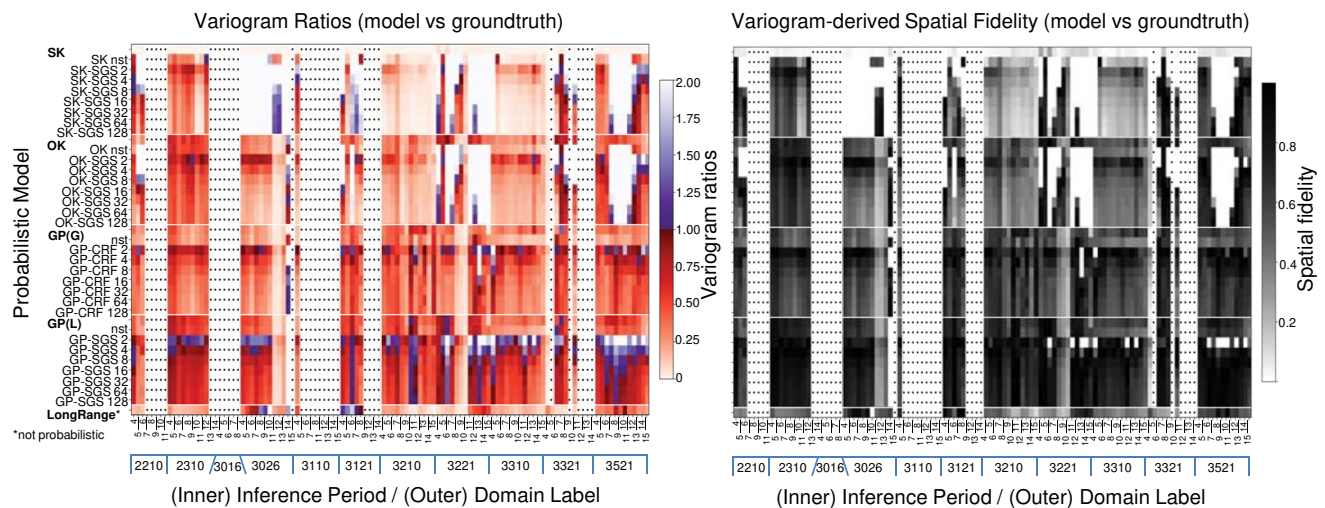


Figure 19. View of (left) variogram ratios R and (right) spatial fidelity F for future-bench prediction across domains and inference periods.

Table 10. Variogram ratio (R) and spatial fidelity (F) summary statistics for future-bench prediction over domains and inference periods.

Model Family	Abbrev	R Mean (SE)	F Mean (SE)
Simple kriging	SK/SK-SGS	0.2983 (0.0112)	0.4272 (0.0125)
Ordinary kriging	OK/OK-SGS	0.4787 (0.0102)	0.6234 (0.0110)
Gaussian process (global mean)	GP(G)/GP-CRF	0.6114 (0.0070)	0.7675 (0.0055)
Gaussian process (local mean)	GP(L)/GP-SGS	0.7132 (0.0081)	0.8231 (0.0069)

5.2.4. Accuracy and Precision

Turning our attention now to uncertainty-based measures, this section examines the accuracy and precision of the predictive distributions across all periods and domains. To keep this brief, we restrict our comments to peculiar cases and general trends. Accuracy is depicted in Figure 20 (left). It turns out that Deutsch's notion of accuracy conveys something important about SGS/CRF. Recall that accuracy relates to ground-truth capture by p -probability intervals (see Section 3.3.2). It considers what a model promises and actually delivers with respect to the proportion of samples it expects to cover. The prominent white horizontal strips in Figure 20 (left) show that the SGS/CRF models fail to live up to this expectation when $s = 2$ and $s = 4$. In this study, we find that at least 8 to 16 simulations are required to obtain a roughly calibrated probabilistic model for future-bench grade prediction in a porphyry copper deposit. Next, moving on to lesser issues, the vertical strips in domain 2210 (from $m_A = 8$ to $m_A = 11$) and domain 3121 (from $m_A = 12$ to $m_A = 14$) coincide with few inference points, according to Table 9. In the case of 3121, the relevant periods each had only two samples. For the precision image in Figure 20 (right), the lightly coloured columns in domain 3110, $m_A \in \{7, 8\}$, are similarly explained by virtue of having only one test sample. The summary statistics in Table 11 show the GP(L) and GP(G) families achieve the highest overall accuracy across all domains and inference periods, while precision is similar across all families (between 0.851 and 0.868).

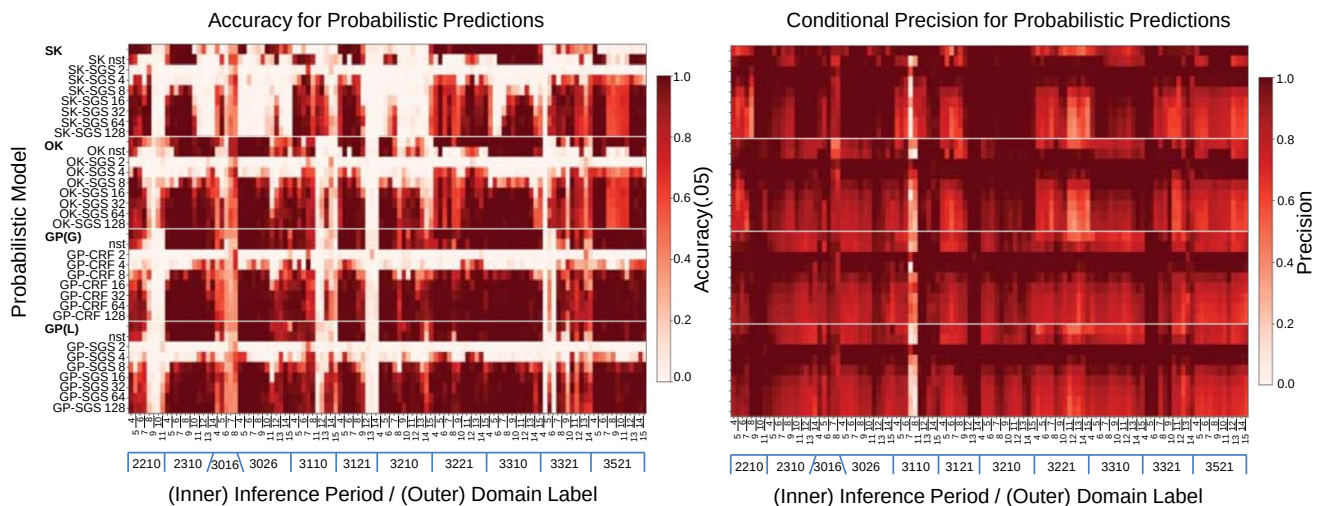


Figure 20. View of (left) accuracy A and (right) precision P for future-bench prediction across domains and inference periods.

Table 11. Accuracy (A) and precision (P) summary statistics for future-bench prediction over domains and inference periods.

Model Family	Abbrev	A Mean (SE)	P Mean (SE)
Simple kriging	SK/SK-SGS	0.5245 (0.0159)	0.8680 (0.0058)
Ordinary kriging	OK/OK-SGS	0.7173 (0.0133)	0.8672 (0.0050)
Gaussian process (global mean)	GP(G)/GP-CRF	0.8084 (0.0119)	0.8637 (0.0041)
Gaussian process (local mean)	GP(L)/GP-SGS	0.8127 (0.0117)	0.8510 (0.0048)

Group averages exclude SGS/CRF $s=2$ and $s=4$, viz., epochs long before convergence.

5.2.5. Consensus and Goodness

The local consensus (L) and goodness (G) of the predictive distributions across all periods and domains are shown in Figure 21. An immediate observation is that L and G are generally correlated. Looking at Figure 21 (right), aside from the statistics being more variable for the kriging base models (see first two rows in the SK and OK blocks), the image is quite unremarkable. Looking at the group statistics in Table 12, the consensus statistic suggests GP(L) is best and GPs are to be preferred (with $L_{GP(L)} = 0.543$ and $L_{GP(G)} = 0.536$) over ordinary kriging (with $L_{OK} = 0.513$ and $L_{SK} = 0.452$). The message from the goodness statistic is similar but subtly different—it places GP(L) and GP(G) as equal (with $G_{GP(L)} = 0.799$ and $G_{GP(G)} = 0.797$) and ordinary kriging as a close alternative (with $G_{OK} = 0.786$).

Table 12. Consensus (L) and goodness (G) summary statistics for future-bench prediction over domains and inference periods.

Model Family	Abbrev	L Median	$[q_L, q_U]$	G Mean (SE)
Simple kriging	SK/SK-SGS	0.4527	[0.2346, 0.6219]	0.7149 (0.0076)
Ordinary kriging	OK/OK-SGS	0.5137	[0.2795, 0.6764]	0.7868 (0.0065)
Gaussian process (global mean)	GP(G)/GP-CRF	0.5366	[0.2855, 0.6990]	0.7974 (0.0062)
Gaussian process (local mean)	GP(L)/GP-SGS	0.5432	[0.2996, 0.7053]	0.7997 (0.0059)

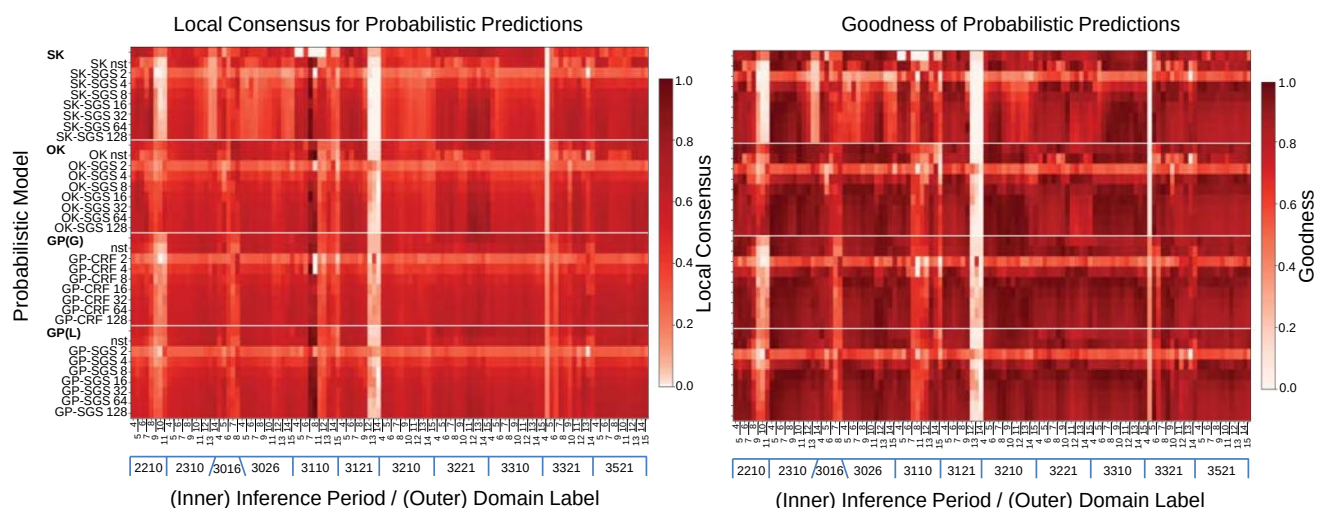


Figure 21. View of (left) consensus L and (right) goodness G for future-bench prediction across domains and inference periods.

5.2.6. Interval Tightness

The interval tightness (I) of the predictive distributions across all periods and domains is shown in Figure 22. From the extensive white-out regions, where the width of the prediction interval is large, one can reasonably infer that simple kriging produces the least confident (most uncertain) predictions. This can be confirmed from the group statistics in Table 13, which also show GP(G) produces the narrowest predictions.

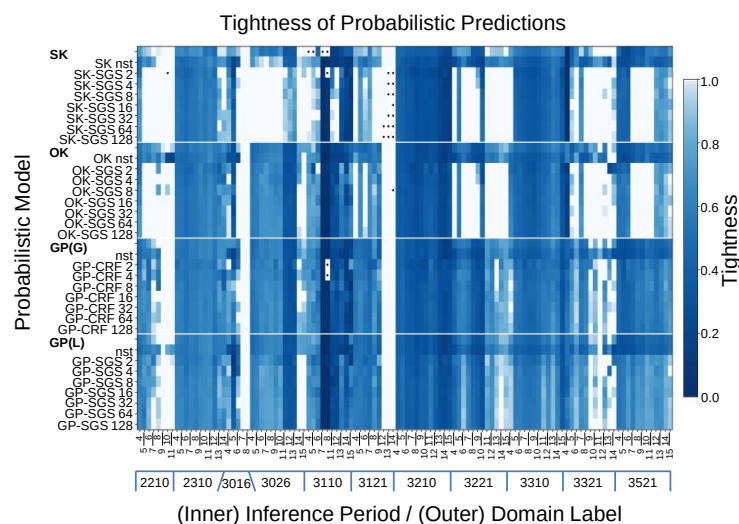


Figure 22. View of interval tightness I for future-bench prediction across domains and inference periods.

Table 13. Interval tightness (I) summary statistics for future-bench prediction over domains and inference periods.

Model Family	Abbrev	I Mean (SE)
Simple kriging	SK/SK-SGS	0.6898 (0.0096)
Ordinary kriging	OK/OK-SGS	0.6408 (0.0087)
Gaussian process (global mean)	GP(G)/GP-CRF	0.5787 (0.0072)
Gaussian process (local mean)	GP(L)/GP-SGS	0.6180 (0.0073)

5.2.7. Statistical Significance

The dependent t -test is applied to the histogram, fidelity, accuracy, precision, interval tightness, goodness and consensus scores (H , F , A , P , T , G and L) to establish the significance of the results. In general, the null hypothesis asserts that the mean score for model family ψ (where $\psi \in \{\text{SK}, \text{OK}, \text{GP}(\text{G})\}$) is greater than or equal to the mean for the GP(L) family. Thus, the null and alternative hypotheses may be written as $H_0(X, \psi) : \mu_X^\psi \geq \mu_X^{\text{GP}(\text{L})}$ and $H_a(X, \psi) : \mu_X^\psi < \mu_X^{\text{GP}(\text{L})}$. When applied to scores that ought to be maximised, viz., $X \in \{F, A, P, G, L\}$, a true H_a indicates the GP(L) family has superior performance. For scores that ought to be minimised, the inequality signs are reversed such that $H_0(Y, \psi) : \mu_Y^\psi < \mu_Y^{\text{GP}(\text{L})}$ and $H_a(Y, \psi) : \mu_Y^\psi \geq \mu_Y^{\text{GP}(\text{L})}$ for $Y \in \{H, I\}$. The p -values are reported in Table 14 along with the 95% confidence intervals for the difference (viz., $X^\psi - X^{\text{GP}(\text{L})}$ or $Y^\psi - Y^{\text{GP}(\text{L})}$) under the alternative hypothesis that the two are unequal.

Table 14. Significance testing of statistical scores for future-bench prediction over all domains and inference periods.

Family ψ	Histogram $H = h_{\text{EM}}$		Spatial Fidelity F		Accuracy A		Precision P	
	p	CI	p	CI	p	CI	p	CI
SK/SGS	<0.001	[0.1670, 0.1940]	<0.001	[−0.4221, −0.3696]	<0.001	[−0.3212, −0.2551]	>0.99	[0.0221, 0.0396]
OK/SGS	<0.001	[0.0627, 0.0816]	<0.001	[−0.2241, −0.1753]	<0.001	[−0.1142, −0.0764]	>0.99	[0.0128, 0.0247]
GP(G)/CRF	<0.001	[0.0058, 0.0211]	<0.001	[−0.0683, −0.0429]	0.1685	[−0.0132, 0.0045]	>0.95	[0.0084, 0.0161]
Reference	μ	SE	μ	SE	μ	SE	μ	SE
GP(L)/SGS	0.0691	(0.0026)	0.8231	(0.0069)	0.8127	(0.0117)	0.8510	(0.0048)
Family ψ	Interval I		Goodness G		Consensus L			
	p	CI	p	CI	$p^\dagger$	CI [†]		
SK/SGS	<0.001	[0.0624, 0.0905]	<0.001	[−0.0950, −0.0746]	<0.001	[−0.0805, −0.0503]		
OK/SGS	<0.001	[0.0127, 0.0336]	<0.001	[−0.0188, −0.0070]	0.0019	[−0.0343, −0.0066]		
GP(G)/CRF	>0.99	[−0.0470, −0.0337]	0.1961	[−0.0075, 0.0029]	0.0671	[−0.0239, 0.0032]		
Reference	μ	SE	μ	SE	median	$[q_L, q_U]$		
GP(L)/SGS	0.6180	(0.0073)	0.7997	(0.0059)	0.5432	[0.2996, 0.7053]		

[†] The more conservative Welch's t -test is used assuming unequal population variance.

5.2.8. Interpretations

A direct translation of the results in Table 14 is as follows. At a statistical significance (p -value) of 0.05, the alternative hypothesis, $H_a(h_{\text{EM}}, \psi)$ is accepted for all models $\psi \in \{\text{SK}, \text{OK}, \text{GP}(\text{G})\}$. This means, in terms of global distortion in the predictive mean, that the performance of GP(L) is superior to SK, OK and GP(G). With regard to spatial fidelity, the alternative hypothesis, $H_a(F, \psi)$ is also accepted for all models. Not only is the spatial fidelity of GP(L) higher than SK, OK and GP(G) but the confidence intervals indicate that GP(L) is superior by a large margin. To estimate their respective differences, dividing the CI midpoints [−0.3958, −0.1995, −0.0556] by mean (F) = 0.8231 for GP(L), one arrives at an average loss in spatial fidelity of 48%, 24% and 6.7% if the SK, OK and GP(G) models are used with SGS/CRF simulation in place of GP(L)/SGS.

With regard to accuracy, $H_a(A, \psi)$ is accepted for SK and OK but rejected for GP(G). This means that the accuracy of the predictive distributions generated by GP(L) is superior to SK and OK but is not significantly different to GP(G) given a p -value of 0.1685, with zero contained in the CI [−0.0132, 0.0045]. The alternative hypothesis $H_a(P, \psi)$, on the other hand, is rejected for all models. This implies that the precision of the GP(L) predictive distributions is not superior to SK, OK and GP(G). However, GP(L) is inferior only by a small margin with a combined CI of [0.008, 0.039]. Because the precision score is conditioned on having an accurate distribution where only instances of $\bar{\kappa}(p) > p$ are counted [here,

p represents proportions as defined in Section 3.3.3], the goodness statistic G is generally considered a more prudent measure. Since $H_a(G, \psi)$ is accepted for SK and OK but rejected for GP(G) at a p -value of 0.196, GP(L) adheres more closely to p -probability interval ground-truth containment expectations than either SK or OK, and the differences between GP(L) and GP(G) are insignificant. This finding is corroborated by the consensus score, as $H_a(L, \psi)$ is also accepted for SK and OK but rejected for GP(G) at a p -value of 0.067. Finally, the GP(L) prediction intervals are narrower for all models except GP(G) since $H_a(I, \psi)$ is accepted for both SK and OK.

Collectively, these significance tests indicate that GP(L)/GP-SGS—Gaussian process regression using local neighbourhood mean with sequential Gaussian simulation—is superior to both simple kriging (SK/SK-SGS) and ordinary kriging (OK/OK-SGS) approaches. The confidence intervals for $X^\psi - X^{\text{GP(L)}}$ quantify the margin of superiority, and the evidence from Table 14 is extremely strong against SK/SK-SGS on all scores and very strong against OK/OK-SGS with respect to histogram (H), fidelity (F) and accuracy (A) and moderate with respect to goodness (G) and consensus (L). The t -tests also confirm the performance of GP(G)/GP-CRF—Gaussian process regression using stationary global mean with Cholesky random field simulation—is close to GP(L)/GP-SGS with respect to H , A , G and L . In fact, GP(G)/GP-CRF prediction intervals tend to be narrower. The main reason for preferring GP(L)/GP-SGS is that it achieves higher spatial fidelity based on the F score, which is informed by variogram considerations, as discussed in Sections 3.2 and 5.1.2.

5.3. Comparison with In Situ Regression

Experimental results for in situ regression (i.e., performing interpolation instead of extrapolation) were separately compiled. The same procedures were followed, and thus the same analysis and graphics seen in Sections 5.1.1–5.2.7 were produced and included in Supplementary S2 (figures and tables therein carry the S prefix). Image-based views of the statistics across domains and inference periods are shown in Figures S8–S12. At a high level, similar patterns emerge albeit with greater clarity. The main features can be seen in Table 15, which compares the summary statistics for in situ regression with future-bench prediction. This table shows the average scores for in situ regression and expresses differences as percentage changes relative to the average scores for future-bench prediction. [For brevity, standard errors are omitted. These details can be found in Tables S2–S6.] An insight from the F scores is that the spatial fidelity gaps are smaller between OK/SGS, GP(G)/CRF and GP(L)/SGS for in situ regression; however, GP(L)/SGS really excels and the gaps widen under future-bench prediction. For the remaining discussion, it is instructive to focus on the last row for GP(L)/SGS in Table 15. The reductions in the histogram and interval scores ($\Delta H \approx -45\%$ and $\Delta I \approx -28\%$) show improvement in mean grade distribution resemblance and contraction in the prediction interval; the latter in particular points to a more confident model. These, together with associated improvements in the fidelity, accuracy, goodness and consensus scores ($\Delta F \approx +5.2\%$, $\Delta A \approx +19.6\%$, $\Delta G \approx +6.1\%$, $\Delta L \approx +9.6\%$), indicate how much easier in situ regression is compared with future-bench prediction. The level of difficulty associated with a prediction task is too often omitted from model analysis; this is something to be mindful of.

Significance testing was also carried out on the in situ regression results. A comparison of Table 14 (the future-bench results) with Table S7 (the in situ results) confirms the relative merits of GP(L)/SGS over OK/SGS and SK/SGS remain unchanged. Minor differences exist with respect to the alternative hypotheses $H_a(A, \text{GP(G)/CRF})$ and $H_a(L, \text{GP(G)/CRF})$ which were accepted with p -values of 0.0012 and 0.0137. This suggests GP(L)/SGS has a slightly better accuracy and consensus scores than GP(G)/CRF at a significance level of 0.05 when the confidence intervals in Table S7 are taken into account.

Table 15. Performance comparison with in situ regression. Statistical scores for in situ regression are shown. Parentheses show percentage change $\Delta_{\text{future} \rightarrow \text{in-situ}}$ as a general improvement relative to future-bench prediction. Note: The increase in difficulty going from in situ regression to future-bench prediction is given by $\Delta_{\text{in-situ} \rightarrow \text{future}} = -\Delta_{\text{future} \rightarrow \text{in-situ}} / (1 + \Delta_{\text{future} \rightarrow \text{in-situ}})$. Figures are aggregated over all domains and inference periods.

Family ψ	Histogram	$H = h_{\text{EM}}$	Spatial Fidelity F	Accuracy A	Precision P	Interval I	Goodness G	Consensus L
SK/SGS	0.1186	(−29.3%)	0.5149 (+20.5%)	0.6142 (+17.1%)	0.8840 [†] (+1.84%)	0.6964 –	0.7876 (+10.1%)	0.4731 (+4.50%)
OK/SGS	0.0790	(−29.8%)	0.7175 (+15.0%)	0.9051 (+26.1%)	0.8310 (−4.17%)	0.5924 (−7.55%)	0.8407 (+6.85%)	0.5708 (+11.1%)
GP(G)/CRF	0.0409	(−45.6%)	0.8482 (+10.5%)	0.9666 (+19.5%)	0.8303 (−3.86%)	0.4402 (−23.9%)	0.8546 (+7.17%)	0.5853 (+9.07%)
GP(L)/SGS	0.0382	(−44.7%)	0.8656 (+5.16%)	0.9723 (+19.6%)	0.8109 (−4.71%)	0.4445 (−28.0%)	0.8487 (+6.12%)	0.5954 (+9.60%)

[†] Conditional on having an accurate model. Computed using only 61% of the samples in the case of SK.

5.4. Effects of Simulation

The issue of how sequential or CRF simulation affects the predictive performance of probabilistic models has received little attention in the geoscientific literature. This section seeks to provide some answers by asking whether SGS/CRF simulation actually improves the base models and in what way. Table 16 shows the *sample-weighted* performance statistics for future-bench prediction across all domains and inference periods. First, the accuracy of predictive distributions is examined to infer the convergence behaviour of SGS and CRF. The abrupt improvement from $A < 0.17$ ($s = 4$) to $A > 0.85$ ($s = 8$) shows that approximately eight simulation runs is required to produce valid predictions. This can be verified by inspecting the absolute synchronicity columns which show the lower and upper quartiles, $|S|_{0.25}$ and $|S|_{0.75}$, start exceeding 0.25 and 0.75, respectively, when $s \geq 8$. The goodness measure (G) achieves its maximum when $s = 8$, whereas the local consensus (L) keeps increasing and attains a value superior to the base model after $s = 32$ iterations. This suggests a reasonable number of simulations to choose is between $s = 8$ and $s = 64$ for this dataset. For both GP-CRF and GP-SGS, the accuracy, precision and goodness statistics are comparable to the GP(G) and GP(L) base models after $s = 32$ simulation runs. Hence, GP(L) and GP(G) both generate competent probabilistic predictions. The main reason for preferencing GP-SGS or GP-CRF is that their mean predictions achieve higher spatial fidelity than their corresponding base model, as is evident from the F column. This issue is further investigated in Part 3 of the Supplementary Material.

Table 16. Sample-weighted performance statistics for future-bench prediction.

Model	Histogram h_{JS}	Spatial Fidelity F	Abs. Synchronicity $ S _{0.25}$	Abs. Synchronicity $ S _{0.75}$	Consensus L	Accuracy $A_{\xi=0.05}$	Precision P	Goodness G	Interval I
GP(G)_nst	0.1260	0.4856	0.3930	0.8229	0.5892	0.9585	0.8111	0.8954	0.4675
GP(G)_CRF_from_2	0.0557	0.8617	0.0009	0.5166	0.2795	0.0047	0.9999	0.5578	0.5170
GP(G)_CRF_from_4	0.0658	0.7424	0.1343	0.7035	0.4317	0.1440	0.9983	0.8605	0.5508
GP(G)_CRF_from_8	0.0796	0.6740	0.2748	0.7643	0.5138	0.8532	0.9481	0.9498	0.5693
GP(G)_CRF_from_16	0.0904	0.6303	0.3495	0.7948	0.5608	0.9606	0.8708	0.9280	0.5658
GP(G)_CRF_from_32	0.0990	0.6063	0.3931	0.8111	0.5866	0.9768	0.8211	0.9053	0.5687
GP(G)_CRF_from_64	0.1052	0.5851	0.4170	0.8208	0.6020	0.9800	0.7909	0.8908	0.5631
GP(G)_CRF_from_128	0.1094	0.5745	0.4297	0.8257	0.6103	0.9803	0.7748	0.8831	0.5615
GP(L)_nst	0.0896	0.6368	0.3935	0.8273	0.5933	0.9618	0.8032	0.8918	0.4748
GP(L)_SGS_from_2	0.0576	0.8247	0.0012	0.5177	0.2811	0.0036	0.9999	0.5611	0.5814
GP(L)_SGS_from_4	0.0579	0.8499	0.1504	0.7001	0.4369	0.1680	0.9980	0.8705	0.6202
GP(L)_SGS_from_8	0.0606	0.8354	0.2909	0.7622	0.5208	0.8662	0.9382	0.9490	0.6323
GP(L)_SGS_from_16	0.0687	0.8105	0.3651	0.7915	0.5665	0.9533	0.8596	0.9226	0.6372
GP(L)_SGS_from_32	0.0732	0.7924	0.4065	0.8066	0.5914	0.9653	0.8116	0.9006	0.6404
GP(L)_SGS_from_64	0.0793	0.7791	0.4304	0.8137	0.6067	0.9741	0.7820	0.8866	0.6426
GP(L)_SGS_from_128	0.0815	0.7718	0.4454	0.8197	0.6153	0.9781	0.7651	0.8786	0.6417

This represents a cropped version of Table S8 in Supplementary S2.

6. Discussion

As mentioned in the Introduction, this paper prioritises its focus on assessment methods rather than drawing specific conclusions on model performance. Its core objective is to develop a consistent approach that supports highly automated model assessment at scale and add layers to the analysis to provide a more complete understanding of different facets of model performance. A key to achieving this objective is using standardised and interpretable measures that can be meaningfully compared across domains and target variables. An image-based view of the relevant statistics also enables a multitude of models to be compared simultaneously irrespective of what modelling processes are used. This allows modellers to focus attention on identifying conditions (columns) that present the most difficulties and investigate situations where models might misbehave, which is useful in large-scale model evaluation. This modality enables modellers to observe, analyse and interpret latent patterns. Although the vertical axis in Figures 18–22 currently portrays a variation in the number of simulation runs, this is meant to represent a metaphor more broadly for any modelling approaches, configuration settings, approximations or parameter changes that modellers might want to investigate, to assess the impact of various decisions on model performance. Other dimensions considered include testing for the statistical significance between models and visualising error clusters. An example of the latter is shown in Figure 23. In mining, it is not uncommon for correlated errors to appear in the prediction residuals. Contributing factors include domain delineation errors, unexpected spatial discontinuities, rapid changes in the geology, limited data for the models to learn from (which may result in poor estimates of the hyperparameters) and also drifts in the target distribution when extrapolation is required over large distances.

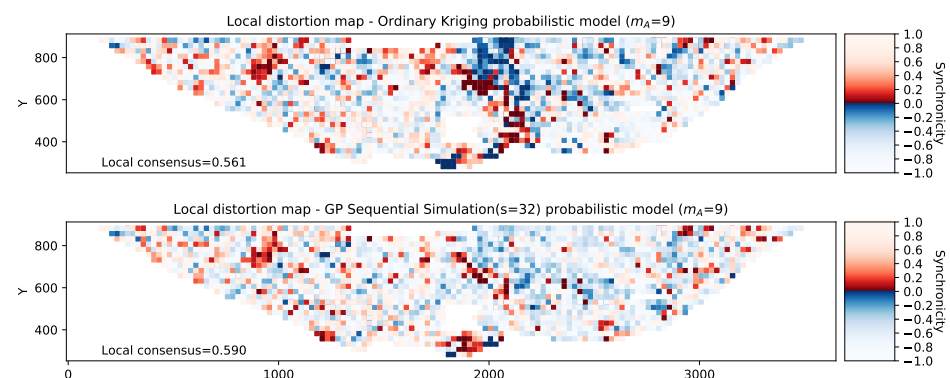


Figure 23. Synchronicity map comparing model predictions with the ground truth across multiple domains. Dark patches indicate a high degree of inconsistency between prediction and ground truth. Red and blue colours indicate underestimation and overestimation, respectively.

It is worth stating that the FLAGSHIP measures are not necessarily better than other alternatives, they are merely fit for purpose. For instance, a selection that includes the Wasserstein EM histogram distance (H), the variogram-based spatial fidelity measure (F), goodness of the predictive distribution (G)—and perhaps interval tightness (I)—should adequately reveal the global accuracy, local variability and calibration properties of a probabilistic model. What is more important is that all three aspects are fairly represented. For H , a diffused histogram may be computed for model predictions by accumulating Gaussians with posterior mean and variance ($\hat{\mu}(x_*), \hat{\sigma}^2(x_*)$) over all predicted points instead of accumulating votes $\delta(y - \hat{\mu}(x_*))$ as performed currently, if increased computation can be tolerated. As a loss measure, the negative log probability of the target under the model [26], $-\log p(y_* | D, x_*) = \frac{1}{2} \log(2\pi\hat{\sigma}_*^2) + \frac{(y_* - \hat{\mu}(x_*))^2}{2\hat{\sigma}_*^2}$, is also a viable alternative to the synchronicity measure (S); however, it is unbounded and not directly connected with p -probability

intervals. Other approaches such as looking at the CDF of the linear correlation coefficient or mean error percentage across s simulations [57,58] can also be considered. However, these might be more limited to comparing different simulation techniques.

To summarise, a couple of observations emanate from this work. First, systematic evaluation is crucial for understanding the uncertainty and predictive performance of probabilistic models and reducing reliance on manual interpretation. Too often, practitioners focus only on global accuracy, neglecting aspects such as uncertainty and local correlation (or vice versa). This can lead to an incomplete and sometimes flawed understanding of the strengths and deficiencies of models. To address this imbalance, this work advocates a comprehensive approach based on FLAGSHIP—an acronym for fidelity, local consensus, accuracy, goodness, synchronicity, histogram, interval and precision—which assesses the pmf, variogram and uncertainty properties relating to the models. A key benefit with FLAGSHIP is that its statistical scores are standardised and interpretable. For instance, the Jensen–Shannon and Rudzica histogram distances are both bounded between 0 and 1 and have information- and set-theoretic interpretations. This makes it possible to compute averages for these quantities, and others such as the fidelity (F) and consensus (L) scores, meaningfully across geological domains and inference periods. More importantly, it facilitates comparison across commodities and mine sites. The interpretation of the FLAGSHIP statistics is universal and independent of geochemistry. In contrast, conventional measures such as MSE can vary considerably depending on the location and target attribute; it does not have a clear meaning standing on its own. As a case in point, the concentration of copper and molybdenum are measured in units of wt% and ppm which are incompatible, so their MSE cannot be pooled together. (When the statistical scores for two sets of observations are not on equal footing, it is possible for one to be dominant and masquerade changes (improvement or deterioration) in the other.) However, using the L score, sensible comparisons can be made. This is also the reason for incorporating σ_Y normalisation in our definition of interval tightness.

The second point is that FLAGSHIP enables significance testing via large-scale model evaluation and allows modelling performance to be contextualised. For a well-balanced study, a sufficiently large dataset with varied characteristics (such as distribution diversity) should be used where possible to minimise selection bias and present challenges to the models. A well-recognised problem in the mining industry is that model evaluation takes tremendous time and effort. One novel aspect of this work is the conversion of the variogram from a visual diagnostic tool into a quantitative measure of spatial fidelity. When hypothesis testing is applied to the FLAGSHIP measures, users can establish if there are statistical differences between the models and quantify these differences using confidence intervals as seen in Sections 5.2.7 and 5.2.8. Model performance is often reported without much thought on how demanding the problem or data is. This is especially true for future-bench prediction, where there are no protocols or standardised measures for articulating how challenging the geology or modelling task is. This opacity is a source of frustration, as it is difficult to assess whether a promising approach would be efficacious in a different situation without some benchmark. In Section 5.3, we showed that it is possible to quantify the decline in model performance or infer the increase in difficulty, moving from in situ regression (interpolation) to future-bench prediction (extrapolation). Collectively, these could form the basis of one or more objective measures to help communicate geological modelling difficulties and, by extension, draw attention to challenging areas with a view of deploying additional drilling, sensing or adaptive sampling to reduce uncertainty and optimise mining operations in an intelligent, automated and cost-effective way [59]. In particular, the synchronicity score can generate local distortion maps for probabilistic

predictions, as demonstrated in Figure 23. The FLAGSHIP measures can be used within a study or compared between studies since the scores are normalised.

Recommendations

Based on this study, some general considerations for evaluating the uncertainty and predictive performance of probabilistic models are collated in Table 17. These are intended to fill a gap in the absence of industry standards, to support highly automated model assessment performed at scale and across multiple sites and to provide a richer understanding of model/data deficiencies in a potentially complex geological (grade modelling) environment.

Table 17. EUP³M recommendations for evaluating the uncertainty and predictive performance of probabilistic models.

Material	Obtain a sufficiently large dataset with target attribute diversity to minimise selection bias and challenge the models.	Section 4.2
Design	Experiments should reflect observational and modelling constraints in practice. For instance, the data available in each inference period defines the scope of our regression/prediction tasks in a manner that emulates staged data acquisition and progression of mining activities in a real mine.	Section 4.3
Measures	Employ representative measures (such as the FLAGSHIP statistics) to investigate the global accuracy, local correlation and uncertainty-based properties of the models relative to the ground truth. [FLAGSHIP considers the spatial fidelity, local consensus, accuracy, goodness, synchronicity, histogram distance, interval tightness and precision of the predictive distributions.]	Sections 3.1–3.3.6, (2)–(18)
Analysis	Perform one or more of the following according to needs: (a) Compute summary statistics to assess group performance: e.g., aggregate values by model family, average over domains or time periods (see Tables 8–13). (b) Observe general trends and variation in individual models: Perform large-scale simultaneous comparison across models and conditions using image-based views of the relevant statistics to identify instances where models may have underperformed (see Figures 18 and 19). (c) Establish statistical significance and confidence interval: e.g., perform hypothesis testing using <i>t</i>-tests and interpret the results using <i>p</i>-values and CIs. (d) Contextualise model performance: e.g., compare in situ regression with future-bench prediction to articulate the difficulty of extrapolation relative to interpolation. Pairwise comparison can also reveal the benefits of modelling with additional data.	Sections 5.2.1–5.2.6 Section 5.2 Sections 5.2.7 and 5.2.8 Section 5.3

7. Conclusions

Although this paper began with a description of geostatistical models, its core contribution and objectives remain firmly on developing measures and novel ways for assessing and comparing the predictive performance of probabilistic models with observational data. Section 2 reviewed the theories that underpin Gaussian process and kriging regression and outlined the procedures for sequential Gaussian and Cholesky random field simulations (SGS and CRF). Section 3 examined three categories of geostatistics: (a) histogram measures that reflect the global accuracy in the mean estimates, such as the probabilistic symmetric χ^2 , Jensen–Shannon, Ruzicka and Wasserstein distances; (b) variogram measures that target spatial correlation and local variability in the model predicted mean; and (c) uncertainty-based measures that assess the performance of probabilistic models using both the mean and standard deviation estimates, $\hat{\mu}(x_*)$ and $\hat{\sigma}(x_*)$, and ground truth or actual grade, $\mu_0(x_*)$. An example was presented using synthetic data to develop the basic intuition, before the measures were applied to real models and data obtained from a porphyry copper deposit. Section 4 described the geological setting, data attributes, general considerations and implementation of the experiments. It explained the importance of having diversity in the data and the distinction between future-bench prediction and in situ regression. Section 5 provided in-depth analysis, focusing initially on the efficacy of the histogram,

variogram and uncertainty measures in two domains within one inference period. Its scope was subsequently expanded to encompass the entire dataset—this includes up to 11 domains and 12 inference periods—to eliminate selection bias and ensure the results would be fair, representative and statistically significant.

The proposed measures and analytic approach provided insights and clarity. One observation in relation to histogram distance, H , is that the JS divergence, Ruzicka and p.s. χ^2 distances are sensitive to discretisation. They may give the false impression that a model is underperforming when few inference points are involved. This confounding effect can be suppressed by using the Wasserstein distance, since it does not involve quantisation and can be computed directly from order statistics. Another benefit of viewing the H statistics as an image is that it focuses attention on difficult cases. Targeted investigation subsequently revealed that instances of poor predictive performance (see light blue pixels in Figure 18) can generally be explained by a significant mismatch between the training data and ground-truth grade distribution or by insufficient training data for certain domains/periods in this study. In terms of insights, inspection of the variogram curves and automatic determination of variogram ratios uncovered a general trend, viz., GP-SGS produces results with higher spatial fidelity, F , than the GP(L) base model. This finding indicates that while GP(L) can model a random process with non-stationary mean using samples in the local neighbourhood of the inference points, it does not adequately capture mid-range or long-range spatial correlation; therefore, it benefits from sequential simulation which propagates mid- to long-range conditional dependence according to the chain rule in (S.13).

The lessons pertaining to the uncertainty measures are that Deutsch's accuracy, A , is useful for indicating SGS/CRF convergence whilst P , G and I convey the conditional precision, goodness and tightness of the model predictive distributions. The synchronicity measure, S , was described in connection with the concept of p -probability intervals and is used to judge the performance of probabilistic models. The goodness criterion is whether the ground-truth containment intervals live up to expectation, that is, how close the observed proportions, $\bar{\kappa}(p)$, are to the expected proportions, p . The proposed consensus measure, L , while related to G , is more discerning as it is a decreasing function of the Z score, $(\hat{\mu} - \mu_0)/\hat{\sigma}$. An important reason for computing the synchronicity, S , from which the local consensus is derived, is that it can be rendered as a distortion map to identify areas where overestimation or underestimation occurred. Collectively, the FLAGSHIP statistics provide a standardised approach that is amenable to large-scale simultaneous comparisons between many models. Unlike other measures such as the RMSE, these statistics can be aggregated/averaged meaningfully across spatial and temporal domains or even compared between different target variables (such as copper and molybdenum). For significance testing, t -tests show the GP-SGS and GP-CRF models considered in this study are superior to simple and ordinary kriging and their SGS counterparts based on the FLAGSHIP statistics. We hasten to point out that this finding is implementation-dependent and not universal. Finally, the performance gap of future-bench prediction (extrapolation) relative to in situ regression (interpolation) was quantified to contextualise the increased difficulty of the inference task. In summary, this work described a systematic approach for evaluating the uncertainty and predictive performance of univariate probabilistic models using the FLAGSHIP statistics. This culminated in a set of recommendations (see Table 17) for assessing, comparing and validating probabilistic models to serve various needs, including a path towards standardisation.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/modelling6020050/s1>. Supplementary S1—Geostatistical modelling techniques; Supplementary S2—Additional experiment results; Supplementary S3—Spatial fidelity improvement attributed to sequential simulations.

Author Contributions: Conceptualisation: R.L., A.L., A.M. Methodology: R.L., A.L., A.M. Investigation: R.L. Formal analysis: R.L. Software: R.L., A.L. Validation: A.L. Data curation: R.L. Writing—original draft preparation: R.L. Writing—review and editing: A.L., A.M. Funding acquisition: A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Australian Centre for Robotics and Rio Tinto.

Data Availability Statement: An open-source implementation of the algorithms described in this article is available from [GitHub](#) [60] and archived in [Zenodo](#). The `eup3m.git` repository provides anonymised test data, Python code for model construction and statistical analysis, a bash script to replicate the experiments and a Jupyter notebook to reproduce key figures. These are further described in the README.md file.

Acknowledgments: Rio Tinto Kennecott Copper is thanked for providing the data used in this study.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Thacker, B.H.; Doebling, S.W.; Hemez, F.M.; Anderson, M.C.; Pepin, J.E.; Rodriguez, E.A. *Concepts of Model Verification and Validation*; Technical Report; Los Alamos National Laboratory: Los Alamos, NM, USA, 2004. [\[CrossRef\]](#)
2. Tutmez, B. Use of hybrid intelligent computing in mineral resources evaluation. *Appl. Soft Comput.* **2009**, *9*, 1023–1028. [\[CrossRef\]](#)
3. Maniteja, M.; Samanta, G.; Gebretsadik, A.; Tsae, N.B.; Rai, S.S.; Fissaha, Y.; Okada, N.; Kawamura, Y. Advancing Iron Ore Grade Estimation: A Comparative Study of Machine Learning and Ordinary Kriging. *Minerals* **2025**, *15*, 131. [\[CrossRef\]](#)
4. Singh, R.K.; Ray, D.; Sarkar, B.C. Mineral deposit grade assessment using a hybrid model of kriging and generalized regression neural network. *Neural Comput. Appl.* **2022**, *34*, 10611–10627. [\[CrossRef\]](#)
5. Mery, N.; Emery, X.; Cáceres, A.; Ribeiro, D.; Cunha, E. Geostatistical modeling of the geological uncertainty in an iron ore deposit. *Ore Geol. Rev.* **2017**, *88*, 336–351. [\[CrossRef\]](#)
6. Ortiz, J.M.; Emery, X. Geostatistical estimation of mineral resources with soft geological boundaries: A comparative study. *J. S. Afr. Inst. Min. Metall.* **2006**, *106*, 577–584.
7. Emery, X.; Ortiz, J. Estimation of mineral resources using grade domains: Critical analysis and a suggested methodology. *J. S. Afr. Inst. Min. Metall.* **2005**, *105*, 247–255.
8. Deutsch, C.V. Direct assessment of local accuracy and precision. In *Geostatistics Wollongong*; Baafi, E.Y., Schofield, N.A., Eds.; Quantitative Geology and Geostatistics Series; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1997; Volume 96, pp. 115–125.
9. Fouedjio, F.; Klump, J. Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. *Environ. Earth Sci.* **2019**, *78*, 38. [\[CrossRef\]](#)
10. Fouedjio, F.; Scheidt, C.; Yang, L.; Achtziger-Zupančič, P.; Caers, J. A geostatistical implicit modeling framework for uncertainty quantification of 3D geo-domain boundaries: Application to lithological domains from a porphyry copper deposit. *Comput. Geosci.* **2021**, *157*, 104931. [\[CrossRef\]](#)
11. Refice, A.; Capolongo, D. Probabilistic modeling of uncertainties in earthquake-induced landslide hazard assessment. *Comput. Geosci.* **2002**, *28*, 735–749. [\[CrossRef\]](#)
12. Tacher, L.; Pomian-Srzednicki, I.; Parriaux, A. Geological uncertainties associated with 3-D subsurface models. *Comput. Geosci.* **2006**, *32*, 212–221. [\[CrossRef\]](#)
13. Madsen, R.B.; Høyer, A.S.; Andersen, L.T.; Møller, I.; Hansen, T.M. Geology-driven modeling: A new probabilistic approach for incorporating uncertain geological interpretations in 3D geological modeling. *Eng. Geol.* **2022**, *309*, 106833. [\[CrossRef\]](#)
14. Bacci, M.; Sukys, J.; Reichert, P.; Ulzega, S.; Albert, C. A comparison of numerical approaches for statistical inference with stochastic models. *Stoch. Environ. Res. Risk Assess.* **2023**, *37*, 3041–3061. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Seillé, H.; Thiel, S.; Brand, K.; Mulè, S.; Visser, G.; Fabris, A.; Munday, T. Bayesian fusion of MT and AEM probabilistic models with geological data: Examples from the eastern Gawler Craton, South Australia. *Explor. Geophys.* **2024**, *55*, 486–505. [\[CrossRef\]](#)
16. Pérez-Díaz, L.; Alcalde, J.; Bond, C.E. Introduction: Handling uncertainty in the geosciences: Identification, mitigation and communication. *Solid Earth* **2020**, *11*, 889–897. [\[CrossRef\]](#)

17. Lindi, O.T.; Aladejare, A.E.; Ozoji, T.M.; Ranta, J.P. Uncertainty quantification in mineral resource estimation. *Nat. Resour. Res.* **2024**, *33*, 2503–2526. [CrossRef]
18. Chlingaryan, A.; Leung, R.; Melkumyan, A. Augmenting stationary covariance functions with a smoothness hyperparameter and improving Gaussian Process regression using a structural similarity index. *Math. Geosci.* **2024**, *56*, 605–637. [CrossRef]
19. Leung, R.; Seiler, K.; Hill, A. Data Analytics for Open-Pit Mining: Examining Vehicle Interactions, Material Movement and Compositional Uncertainty with Bucket Inference and Monte Carlo Simulation. 2023. Available online: <https://t.ly/B1xr8> (accessed on 1 May 2025).
20. Leung, R.; Lowe, A.; Chlingaryan, A.; Melkumyan, A.; Zigman, J. Bayesian surface warping approach for rectifying geological boundaries using displacement likelihood and evidence from geochemical assays. *ACM Trans. Spat. Algorithms Syst.* **2022**, *8*, 1–23. [CrossRef]
21. Seiler, K.M.; Palmer, A.W.; Hill, A.J. Flow-Achieving Online Planning and Dispatching for Continuous Transportation with Autonomous Vehicles. *IEEE Trans. Autom. Sci. Eng.* **2022**, *19*, 457–472. [CrossRef]
22. Samavati, M.; Essam, D.; Nehring, M.; Sarker, R. Production planning and scheduling in mining scenarios under IPCC mining systems. *Comput. Oper. Res.* **2020**, *115*, 104714. [CrossRef]
23. Cressie, N. The origins of kriging. *Math. Geol.* **1990**, *22*, 239–252. [CrossRef]
24. Matheron, G. Principles of geostatistics. *Econ. Geol.* **1963**, *58*, 1246–1266. [CrossRef]
25. Williams, C.K.; Rasmussen, C.E. Gaussian processes for regression. *Adv. Neural Inf. Process. Syst.* **1995**, *8*, 514–520.
26. Williams, C.K.; Rasmussen, C.E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006; Volume 2.
27. Christianson, R.B.; Pollyea, R.M.; Gramacy, R.B. Traditional kriging versus modern Gaussian processes for large-scale mining data. *Stat. Anal. Data Min. ASA Data Sci. J.* **2023**, *16*, 488–506. [CrossRef]
28. Melkumyan, A.; Ramos, F.T. A sparse covariance function for exact Gaussian process inference in large datasets. In Proceedings of the 21st International Joint Conference on Artificial Intelligence, Hainan, China, 25–26 April 2009; pp. 1936–1942.
29. Melkumyan, A.; Ramos, F.T. Multi-kernel Gaussian processes. In Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011; pp. 1408–1413.
30. Shekaramiz, M.; Moon, T.K.; Gunther, J.H. *A Note on Kriging and Gaussian Processes*; Technical report; Information Dynamics Laboratory, Electrical and Computer Engineering Department, Utah State University: Logan, UT, USA, 2019.
31. Olea, R.A.; Pawlowsky, V. Compensating for estimation smoothing in kriging. *Math. Geol.* **1996**, *28*, 407–417. [CrossRef]
32. Journel, A.G.; Kyriakidis, P.C.; Mao, S. Correcting the smoothing effect of estimators: A spectral postprocessor. *Math. Geol.* **2000**, *32*, 787–813. [CrossRef]
33. Oliver, M.A.; Webster, R. *Basic Steps in Geostatistics: The Variogram and Kriging*; Springer International Publishing: Cham, Switzerland, 2015.
34. Ortiz, J.M. *Introduction to Sequential Gaussian Simulation*; Technical report; Queen’s University: Kingston, ON, Canada, 2020.
35. Hansen, T.M. Entropy and information content of geostatistical models. *Math. Geosci.* **2021**, *53*, 163–184. [CrossRef]
36. Bai, T.; Tahmasebi, P. Sequential Gaussian simulation for geosystems modeling: A machine learning approach. *Geosci. Front.* **2022**, *13*, 101258. [CrossRef]
37. Asghari, O.; Soltni, F.; Hassan, B. The comparison between sequential gaussian simulation (SGS) of Choghart ore deposit and geostatistical estimation through ordinary kriging. *Aust. J. Basic Appl. Sci.* **2009**, *3*, 330–341.
38. McLennan, J. *The Effect of the Simulation Path in Sequential Gaussian Simulation*; Technical Report 4:115; Centre for Computational Geostatistics Report; University of Alberta: Edmonton, AB, Canada, 2002.
39. Yang, Y.; Wang, P.; Brandenburg, S.J. An algorithm for generating spatially correlated random fields using Cholesky decomposition and ordinary kriging. *Comput. Geotech.* **2022**, *147*, 104783. [CrossRef]
40. Deutsch, C.V.; Journel, A.G. *GSLIB Geostatistical Software Library and User’s Guide*; Oxford University Press: New York, NY, USA, 1997. Available online: <http://claytonvdeutsch.com/wp-content/uploads/2019/03/GSLIB-Book-Second-Edition.pdf> (accessed on 9 June 2025).
41. Hendeby, G.; Gustafsson, F. *On Nonlinear Transformations of Gaussian Distributions*; Technical Report SE-581; Linköpings Universitet: Linköping, Sweden, 2007.
42. Chizat, L.; Peyré, G.; Schmitzer, B.; Vialard, F.X. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *J. Funct. Anal.* **2018**, *274*, 3090–3123. [CrossRef]
43. Deza, E.; Deza, M.M. *Encyclopedia of Distances*; Springer: Berlin/Heidelberg, Germany, 2009. [CrossRef]
44. Topsøe, F. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inf. Theory* **2000**, *46*, 1602–1609. [CrossRef]
45. Nielsen, F. On the Jensen–Shannon symmetrization of distances relying on abstract means. *Entropy* **2019**, *21*, 485. [CrossRef] [PubMed]
46. Cha, S.H. Taxonomy of nominal type histogram distance measures. In Proceedings of the American Conference on Applied Mathematics, Harvard, MA, USA, 24–26 March 2008; pp. 325–330.

47. Rubner, Y.; Tomasi, C.; Guibas, L.J. A metric for distributions with applications to image databases. In Proceedings of the Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), Bombay, India, 4–7 January 1998; IEEE: New York, NY, USA, 1998; pp. 59–66. [\[CrossRef\]](#)
48. Liemohn, M.W.; Shane, A.D.; Azari, A.R.; Petersen, A.K.; Swiger, B.M.; Mukhopadhyay, A. RMSE is not enough: Guidelines to robust data-model comparisons for magnetospheric physics. *J. Atmos. Sol.-Terr. Phys.* **2021**, *218*, 105624. [\[CrossRef\]](#)
49. Goovaerts, P. Geostatistical modelling of uncertainty in soil science. *Geoderma* **2001**, *103*, 3–26. [\[CrossRef\]](#)
50. Leshner, C.E.; Spera, F.J. Thermodynamic and transport properties of silicate melts and magma. In *The Encyclopedia of Volcanoes*; Elsevier: Amsterdam, The Netherlands, 2015; pp. 113–141.
51. Porter, J.P.; Schroeder, K.; Austin, G. Geology of the Bingham Canyon porphyry Cu-Mo-Au deposit, Utah. In *Geology and Genesis of Major Copper Deposits and Districts of the World: A Tribute to Richard H. Sillitoe*; Hedenquist, J.W., Harris, M., Camus, F., Eds.; Society of Economic Geologists: Littleton, CO, USA, 2012; Chapter 6, pp. 136–137. [\[CrossRef\]](#)
52. Redmond, P.B.; Einaudi, M.T. The Bingham Canyon porphyry Cu-Mo-Au deposit. I. Sequence of intrusions, vein formation, and sulfide deposition. *Econ. Geol.* **2010**, *105*, 43–68. [\[CrossRef\]](#)
53. Hayes, R.; McInerney, S. Rio Tinto Kennecott Mineral Resources and Ore Reserves. Technical Report; ASX Notice. 2022. Available online: <https://minedocs.com/28/Kennecott-NRS-MR-06202023.pdf> (accessed on 23 April 2025).
54. Heuvelink, G.B.; Pebesma, E.J. Is the ordinary kriging variance a proper measure of interpolation error. In Proceedings of the Fifth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Melbourne, Australia, 10–12 July 2002; RMIT University: Melbourne, Australia, 2002; pp. 179–186.
55. Piletska, E.V.; Piletsky, S.S.; Whitcombe, M.J.; Chianella, I.; Piletsky, S.A. Development of a new microtiter plate format for clinically relevant assays. *Anal. Chem.* **2012**, *84*, 2038–2043. [\[CrossRef\]](#)
56. Marx, M.C.; Wood, M.; Jarvis, S.C. A microplate fluorimetric assay for the study of enzyme diversity in soils. *Soil Biol. Biochem.* **2001**, *33*, 1633–1640. [\[CrossRef\]](#)
57. Talebi, H.; Sabeti, E.H.; Azadi, M.; Emery, X. Risk quantification with combined use of lithological and grade simulations: Application to a porphyry copper deposit. *Ore Geol. Rev.* **2016**, *75*, 42–51. [\[CrossRef\]](#)
58. Cáceres, A.; Emery, X. Conditional co-simulation of copper grades and lithofacies in the Rio Blanco-Los Bronces copper deposit. In Proceedings of the IV International Conference on Mining Innovation MININ, Santiago de Chile, Chile, 23–25 June 2010; pp. 311–320.
59. Leung, R.; Hill, A.J.; Melkumyan, A. Automation and Artificial Intelligence Technology in Surface Mining: A Brief Introduction to Open-Pit Operations in the Pilbara. *IEEE Robot. Autom. Mag.* **2023**. [\[CrossRef\]](#)
60. Leung, R.; Lowe, A. EUP³M: Evaluating Uncertainty and Predictive Performance of Probabilistic Models—Python Code for Model Construction and Statistical Analysis. GitHub. 2025. Available online: <https://github.com/raymondleung8/eup3m> (accessed on 23 April 2025).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.