Following the handling editor's advice, the authors will respond to the key concerns raised in this review, particularly the fundamental issues that have been highlighted, rather than engaging with every minor point at this stage. Dr. Poulet has instructed us to consider the overarching critiques and how they impact the manuscript's core contributions.

Reviewer 2's comments
https://doi.org/10.5194/egusphere-2024-4051-RC2

RC1: 'Comment on egusphere-2024-4051', Anonymous Referee #2, 19 Mar 2025

This manuscript proposes a suit of probabilistic prediction validation measures named "FLAGSHIP" intended for use in evaluation of interpolations in the mining industry.

The presentation and writing are both very polished and I personally could not find any spelling or grammatical errors. There is a significant number of experiments on both real and synthetic data which are all adequately documented. I think some of the concepts here, such as comparing histograms and variograms, have promise, but the execution is poor. Overall the work lacks a principled approach to justifying performance metrics, the experiments are not set up in a way that can lead to insights, there are major theoretical mistakes, and the literature review on predictive performance metrics and paradigms is lacking.

[**B0**] The authors would like to thank this reviewer for the constructive comments and highlighting the positive aspects of this manuscript. In reference to poor execution, we acknowledge that we have not been sufficiently clear when it comes to stating the motivations and principal objectives of this study. Because the actual purpose of the study was not clearly articulated, it was open to interpretation and potential misunderstanding. This, along with the criticism that it "lacks a principled approach" and "the experiments are not set up in a way that can lead to insights" can be addressed by reflecting on the goals, and providing better guidance to manage readers' expectation. Indeed, work was already underway to reframe and restructure the manuscript, to address similar concerns (esp. clarity) from reviewer 1 before this critique was received. We note in particular that the issues regarding kriging/GP have been rectified, and new references have been added in the revised manuscript to strengthen the background section and discuss other approaches/paradigms in relation to this work.

[**B0a**] There are a few points we need to clarify to correct any misconception. First, this study focuses mainly on evaluating extrapolation performance (not so much interpolation) where models are required to "forward predict" into new territories. The application context and implications for the test data and validation approach are explained in the rewritten introduction. Second, this manuscript was submitted as a *methods-for-assessment* paper rather than a *models evaluation* paper. This distinction is important as it sets the tone and expectations for what is to come. Finally, many of the issues identified in the major comments have been addressed in the revised manuscript; see also claims refuted in [B4].

[**B0b**] To elaborate on the second point – According to the journal guidelines, it should describe "new standard experiments for assessing model performance or novel ways of comparing model results with observational data". Instead of comparing models to determine which is superior, the emphasis is on developing a balanced approach that adds value and insights to the analysis. In this study, we seek to develop a more complete approach that (a) meets the relevant requirements (illuminating on the three pillars of performance – global accuracy, local accuracy/spatial variability, and calibration properties of the predictive distributions for ore grade estimation) and (b) supports highly automated model assessment at scale, using standardised and interpretable measures that can be meaningfully compared across domains and target variables. An associated objective is to provide a richer understanding on different facets of model performance—this includes identifying situations where models misbehave, visualising error clusters, and testing for statistical significance between models. Regarding insights, this can come from functional enhancement, e.g., what information/patterns users are able to access, observe and interpret; not merely lessons or conclusions that can be drawn on specific models.

[**B0c**] In relation to "justifying performance metrics", the revised introduction makes clear some of the deficiencies in current practice. In Sec. 3.2, we have given further reasons for why a spatial fidelity measure is needed. It is worth pointing out that we are not suggesting the proposed measures are necessarily better than any existing alternatives; they are merely fit-for-purpose.

To do this topic justice I advise that the authors do more focused work on a smaller subset of measures and analyze them more thoroughly. I also suggest not focusing on the krigging vs GP comparison, but to instead compare variogram/kernel model choices and fitting methods withing each paradigm separately.

[**B0d**] We thank the reviewer for the advice. In relation to the first part, pursuant to [B0c] we are not arguing that the suggested measures must be used to the exclusion of everything else. In the new discussion section, we have mentioned a few potential alternatives, and indeed one can fairly represent the three pillars of performance [B0b] using a subset of the FLAGSHIP measures. Had we started with a much "smaller subset of measures" in the beginning, we would not have discovered the Wasserstein EM measure is more robust than the Jensen-Shannon histogram distance, for instance, and would subject ourselves to accusation of selection bias or cherry-picking.

[**B0e**] Regarding "not focusing on kriging vs GP comparison", we agree that it may be unhelpful reading too much into any findings regarding kriging or GP. The study should still stand irrespective of the chosen models as far as the objectives in [B0b] are concerned. It is not about what conclusions can be drawn for specific models, rather it is about the kind of observations or insights one can attain from having a richer set of assessment tools at one's disposal. We added a new section (Sec. 2.4) to clarify that it is not just differences in hyper-parameter estimation per se, but differences in kernel and neighbourhood definitions also contribute to differences between the kriging/GP models (paradigms).

This is a *methods-for-assessment* paper after all, so it would be helpful to recognise these differences for what they truly represent, as an embodiment of parameter/ configuration changes that reflect how modellers might explore different modelling options in practice. The authors also feel that shifting the focus to the modelling mechanisms (comparing variogram/kernel choices etc. beyond what is covered in Sec. 2.4) would be counterproductive as the study would stray from its original objectives [B0b].

Major Comments:

I am recommending that this work be rejected for several reasons, listed here in order of importance:

1. The experiments lack purpose. For any experiment, including computer algorithm experiments, there needs to be some prior concept of what the potential outcomes are and what different conclusions would be drawn in each case. In this manuscript we are simply presented with different measures applied to different interpolations and are told one method performs better than anther. Some principle needs to be defined for how metrics are assessed and what the experiments are meant to contribute to our understanding.

[**B1**] We acknowledge the lack of clarity has been problematic as it obfuscates the purpose of this study. The manuscript has been thoroughly revised to make its objectives much clearer (as mentioned in [B0]). To prepare readers with "some prior concept of what the potential outcomes are", we have reinforced what we seek to achieve in the introduction (Sec. 1, last paragraph) and Sec. 5 (first paragraph) before the results was presented. Further guidance is provided in the discussion (Sec. 6).

On the suggestion of aimless meandering, the measures are organised based on three principles and grouped based on three categories: the histogram, variogram and uncertainty-based measures each targets one aspect of performance, viz., the global accuracy, local accuracy (spatial variability) and calibration properties of various model predictive distributions. Regarding "[being] told one method

performs better than another], the aim is to illustrate fitness-for-purpose in the context of grade modelling for forward extrapolation in mineral deposits; not that one measure outperforms another.

For a *methods-for-assessment* paper, the goal is less about drawing conclusions for specific models, it is more about demonstrating the kind of insights one can attain from having a richer set of assessment tools at one's disposal. Regarding "what the experiments [or results] are meant to contribute to our understanding", we have now clarified what outcomes the readers can expect in line 553-556 (highlighted in paragraph 3 in Sec. 5 of the "diff").

2. There is no theoretical justification of the metrics proposed. Performance metrics are supposed to be abstract proxies for what is valued in the real world (e.g., cost of recovering minerals). There is no discussion here of how the metrics relate to the real world setting or objective.

[**B2**] The principal motivation (or empirical justification) is that the three main aspects of performance (see B1) are seldom jointly investigated and evaluated in ways that are amenable to large-scale automated processing. The measures considered in this work would support this and provide meaningful cross-site comparisons irrespective of the domain and target variable.

The variogram-based spatial fidelity measure (F) acts as a proxy for spatial variability. It can detect over-smoothing which geologists particularly object to when evaluating ore grade models. This is discussed in Sec. 3.2. More generally, the utility of these metrics, i.e., how these relate to the application setting or objectives is considered in Sec. 1 (paragraph 3).

3. The literature review on predictive performance measures is inadequate. There is an enormous amount published on this topic. The authors need to broaden their search outside of goescience and geostatistics. Consider literature in machine learning, statistics, Bayesian methods, meta-science, and philosophy of science. For things specific to spatial predictions, I know there is a lot in environmental science.

[**B3**] The literature survey has been strengthened, and relevant works are discussed in relation to this study in Sec. 1 (paragraph 1). We strived to strike the right balance cognisant of the fact that (a) this is not actually a survey article; (b) not all approaches (e.g. k-fold cross validation) are necessarily well suited to open-pit mining. For instance, partitioning the data and setting aside test points that come from the same region as the training points has its limitations. It is not going to set up a situation representative of "future-bench prediction" and will bias the results when forward extrapolation performance is of interest. The performance gap is considered in Sec. 5.3.

4. Trading off different metrics is not discussed. Some of these measures (e.g. accuracy) can be trivially maximized by simply making the prediction standard deviation as wide as possible. Others (e.g.fidelity) can be maximized by over-fitting. How are the different properties to be traded off against each other? The problem is not even acknowledged. There are plenty of existing metrics, such as cross-entropy, which have the tradeoff built in. I get that you may want to measure parts of the objective separately, but a validation framework that does not even acknowledge the trade-off issue can only mislead.

[**B4**] This is **not true**. In lines 445 and 456 (revised manuscript), we mentioned that "in general, both G [goodness] and I [interval tightness] need to be taken in account when assessing probabilistic models, because uncertainty cannot be artificially reduced at the expense of accuracy." This is what the I metric is for. In relation to spatial fidelity F , we acknowledged in line 389 that "a ratio that increases far beyond 1 is also undesirable as it signifies noise amplification." Hence, the expression $F(R)$ in eq. 22 penalises absolute deviation from 1, particularly the case where $R > 1$ which represents over-fitting. As an aside, over-fitting can be a problem in neural networks, but it is not prevalent in kriging and GP regression as they behave like weighted moving average filters.

5. The distinction between a predicted distribution and a sample from such a distribution is not reflected in the experiments. For example, one cannot simply compare the mean image obtained from OK directly with a single sample or finite sample set obtained from OK_SGS from_n (where n is small) directly, as one is a distribution and the other a realization (or set of realizations) from that distribution. Similarly, comparing a real data histogram to a histogram of a predicted mean is not meaningful. The same also applies to variogram comparison.

[**B5**] We appreciate the difference between a distribution and a realisation. This comment highlights the risk of misapprehension and not seeing the forest for the trees. The general applicability of the assessment methods (e.g. image-based visualisation of various statistics in Figs. 18-22) is NOT dependent on what models are used. It would not have mattered if the vertical axes in Figs. 18-22 are replaced with dummy labels. Although the vertical axis currently portrays variation in the number of simulation runs, it is meant to represent more broadly any modelling approaches, configuration changes (including hyperparameters tuning) that modellers might want to investigate, to assess the impact of various approximations/decisions on predictive performance. This is now mentioned in the discussion (around line 847 in the "diff"). In this sense, there is no practical limit to what users might choose to compare, whether the model interpretation is theoretically sound or not. In practice, there is good reason for not keeping hundreds or thousands of simulation runs around (especially when modelling is done at scale), one reductionist approach is to retain only the mean and variance. So, the mean and stdev are all they have at the end of the day. It would be impractical to insist users would have to regenerate hundreds of simulations again, in order to examine the cumulative distribution function of some statistic (e.g. goodness) across S simulations. It would not meet the ease-of-use criterion. It should not be misconstrued that a single (or several) realisations are on an equal footing with a proper simulation. That is not the point of these illustrations. The collection of models represents a metaphor for arbitrary variations that give rise to the different models that user would like to compare and contrast.

6. The first reviewer has already detailed the equivalence between kriging and GPs. I will agree here that the comparing them is effectively a comparison of how the variogram is fitted and how the posterior distribution is calculated/approximated. I believe much of the conclusions draw comparing these two methods are over-generalizations resulting from this lack of theoretical understanding.

[**B6**] This has been addressed elsewhere. The issues have been fixed in Sec. 2, and the observations are implementation dependent not universal (this is now indicated in the conclusion).

Specific Comments:

Lines 41: Kriging also provides a predicted mean and variance and the covariance can be calculated as well. The apparent over-smoothing is likely due to specifics of how the variograms are fitted and will be sensitive to details of how hyper parameters are fit. Note, for both kriging and GPs there are multiple methods.

Yes, absolutely.

Line 160: [*] The statement that kriging does not reproduce variability between pairs of test points seems wrong to me. [**] A mean predicted image will necessarily be smoother than the true image. The roughness of the mean prediction should not be directly interpreted as the expected roughness of the truth. The variance between two test points can be calculated simply by applying the variogram to their relative lag, thus any bias towards underestimating variance would be due to the variogram fitting process or its restrictions (e.g. stationarity and isotropy), or due to the true process being significantly non-Gaussian.

Re: [*] The kriging variance is estimated at two predicted locations (test points) independent of each other, with no spatial continuity constraint imposed between the two. This is mentioned in Prof. Michael Pyrcz's video https://www.youtube.com/watch?v=3cLqK3lR56Y&t=13m35s

Re: [**] We agree the roughness of the mean prediction should not be directly interpreted as the expected roughness of the truth. From the point of view of a *methods-for-assessment* paper, to facilitate automated evaluation at scale, the proposed methods should be model-agnostic. All model comparisons are fair game. It is not always possible to know what modelling techniques are used to create a model (proprietary models created using third party software often behave like black-boxes). Often, one only has access to just the mean prediction (and the variance estimate if lucky). So, it is a fair question to ask how different are a bunch of models in terms of spatial variability. The end-use ranges from visual interpretations to mine planning, including probabilistic inference and potentially stochastic optimisation if uncertainties are available. So, for some use cases over-smoothing matters.

Line 180: This definition is for a finite field. In general, spatial processes are defined over an infinite number of variables.

Line 191: Both kriging and GP can use isotropic or non-isotropic variograms/kernels. The way it is written here suggest that these are limitations of the methods and not of the specific implementations.

You are right. The revised manuscript now portrays these as specific implementation decisions.

Line 248: Fidelity feels like the wrong word here given that accuracy has no effect on it. Also this measure can be easily maximized by simply over-fitting thus there needs to be some discussion on how it is to be traded-off with other measures.

Line 264: This sentence makes no sense to me. What does "once the validation measure is revealed" mean? Conditioning is done on random variable outcomes, not measure types. What does "likelihood of that the model is correct mean", likelihood is the probability of data given a model as a function of the model, the shaded area is proportional to the probability of data being in that interval given the model as an assumption.

We acknowledge this was poorly written, and the interpretation of "observation likelihood given a model" is entrenched in Bayesian reasoning. In the revised manuscript, we have replaced *likelihood* with *local consensus* to reflect the consensus between model prediction and true measurement.

Line 267: The rational behind 'S' as a measure is not clear to me at all and needs elaboration.

Line 271: Likelihood is the probability density of the known data assuming the candidate model. Why is it defined as cumulative density here?

Section 3.3: I think the kappa statistics are interesting but without discussion on how they can be traded-off against each-other there is no clear way to use them. There is no discussion on what it is you are looking for from them. There is no mention of the fact that some can be easily maximized by arbitrarily over-fitting or under-fitting predictions.

As mentioned in [B4], there is an interval tightness measure to restrain (or indicate) over-fitting.

Line 418: Estimated variance does depend on the data as the variogram is estimated from it. The uniformity you see is due to the regular spacing of the data used combined with the stationarity assumption. Questions relating legitimacy should be about the stationarity assumption, which is not strictly needed for kriging or GPs.

We agree with the points regarding "regular spacing" and "stationarity". As the paper is submitted to the category of *methods-for-assessment*, we do not intend to delve more deeply into modelling processes and questions, as these take attention away from the core objectives expressed in [B0b].

Analysis of the appropriateness of the stationarity assumption should precede the interpolation.

The observed higher spatial variability of GP over SK and OK are simply due to it inferring a different variogram with higher variance at short lags.

I do not understand the logic in comparing nst with SGS or CRF.
To illustrate over-smoothing in kriging/GP and contrast with SGS/CFR; putting this beyond words.

Variance estimates from single samples are not comparable to whole posterior distributions. The apparent differences are due to misinterpreting what they produce.
Figures removed.

Line 430: The ground truth has thicker tails because it is a realization of the random variable which is being compared here to mean predictions. Again, the mistake here is to expect distribution means to have the same properties as realizations from those distributions. The correct thing to do here would be to convolve each prediction mean with its predicted standard deviation Gaussian kernel to obtain a correct posterior expected histogram.

This viewpoint is interesting and different. Essentially, it treats each prediction as a Gaussian distribution centred at the predicted mean, and takes a sum of these Gaussians over all test locations to produce the posterior expected histogram. This concept is now mentioned in the discussion in the revised manuscript.

Sections 5.1.2: Again, properties of distributions are being compared to properties of samples. SK_SGS_single is a sample from the SK distributions. Their variograms are not comparable. One could compare the SGS_single variogram with the ground truths, or the SK_SGS_from_highestnumberthatispractical with SK, but not across those groups.

We appreciate this point from a theoretical perspective. However, the techniques used to produce the models are not always known or documented in practice, so it's not possible to treat models differently or decline comparison on these grounds. That is why our methods are completely model-agnostic.

More broadly, the variation in the number of simulation runs (see vertical axis in Fig. 22) is meant to represent arbitrary changes in the model parameters or configuration. So, it would not be particularly helpful seizing on one embodiment of the idea, as users are free to choose any model they want and the context could change completely.

I am leaving out my notes for the remaining sections because they are all about the same point: properties of distributions and samples from those distributions should not be expected to be the same and are not directly comparable.