



Calibration Is Not Enough: TAP-Bathy v1.0, a Three-Axis Protocol for Evaluating Spatial Uncertainty in Bathymetry

Xinye Cui ¹, Houpu Li ¹, Xin Xiong ¹, Shaofeng Bian ¹, and Guojun Zhai ²

¹Naval University of Engineering, Wuhan 430033, China

²Key Laboratory of Geological Survey and Evaluation (Ministry of Education), China University of Geosciences, Wuhan 430074, China

Correspondence: Houpu Li (1210051025@nue.edu.cn)

Abstract. Calibration is the standard yardstick for spatial uncertainty quantification (UQ), but it is insufficient. On the Amazon Fan, ordinary-kriging variance achieves excellent calibration (ECE = 4.1%) yet carries negligible spatial information ($r_{\text{pooled}} = +0.16$). On the Blake Escarpment, kriging variance yields below-random block prioritization under the locked evaluation protocol. Motivated by this disconnect, we present TAP-Bathy v1.0 (Three-Axis Protocol for Bathymetric Uncertainty Evaluation) for evaluating spatial UQ: calibration, an exact within/cross-patch decomposition of the error–uncertainty correlation, and operational prioritization at a fixed resurvey budget. Applying the protocol to deep-sea bathymetric reconstruction across three terrains reveals three failure modes that any single axis misses—calibrated-but-uninformative uncertainty, statistically significant but operationally negligible within-patch signals, and cross-region ranking reversal via Simpson’s paradox. Learned σ carries reliable cross-patch information ($r_{\text{cross}} = 0.08\text{--}0.50$, locked evaluation protocol; confidence intervals at patch level) and concentrates $1.5\text{--}3\times$ the random error at 10% block-level resurvey budgets and remains above-random at 25%. A deep ensemble produces a near-identical primary endpoint, with the broad failure-mode pattern robust to the epistemic-uncertainty estimator, though secondary operational capture is unstable on the small Region 009 evaluation set. A pre-registered prospective test on an unseen fan/plain region yielded $r_{\text{cross}} = 0.358$ inside the predicted interval, consistent with retained cross-patch informativeness on new terrain. An eight-region prospective test cleanly falsified the strong (sign-reversal) form of a terrain-coupling prediction. The contribution is TAP-Bathy v1.0, a reproducible, scale-aware evaluation protocol: we show how to assess spatial UQ, not that one estimator universally wins.

Keywords. bathymetry; uncertainty quantification; spatial validation; deep learning; ordinary kriging; pre-registration; model evaluation

1 Introduction

Global bathymetric grids such as GEBCO and GMRT (Ryan et al., 2009) are foundational Earth-observation products: they underpin tsunami propagation models, ocean-circulation simulations, habitat mapping, and marine spatial planning (Weatherall et al., 2015; Tozer et al., 2019). Yet across the majority of the ocean floor, these grids are not measurements but interpolations, produced by blending sparse ship-track soundings with satellite-altimetry-derived gravity, yielding a spatial resolution that far



exceeds the density of the underlying data (Smith and Sandwell, 1997). The GEBCO 2026 grid is released at 15-arcsecond
 25 (~450 m at the equator) native resolution; for our patch-based processing we resample to a uniform 100 m pixel spacing,
 yielding 64×64 -pixel patches of ~6.4 km across (Section 2). In regions where the nearest ship sounding is tens of kilometers
 away, the interpolated depth at any given pixel is a statistical guess whose error is unknown and unquantified, with no per-pixel
 estimate of its magnitude.

The Seabed 2030 initiative aims to map the entire ocean floor at high resolution by 2030 (Mayer et al., 2018). Achieving
 30 this requires not only collecting new data but deciding *where* to collect it: given finite survey resources, which unsurveyed or
 poorly surveyed regions would benefit most from new multibeam measurements? This is fundamentally a spatial uncertainty
 quantification (UQ) problem. If a bathymetric grid could report, at every pixel, a well-calibrated, spatially informative, and
 operationally actionable estimate of its own error, resurvey planning could be guided by evidence rather than by geometric
 heuristics or ad-hoc terrain proxies.

35 Two families of methods can, in principle, provide such pixel-level uncertainty: geostatistical interpolation, where ordinary
 kriging produces a prediction variance from the fitted variogram and remains the standard engine in bathymetric workflows,
 and learned reconstruction, where a neural network with a heteroscedastic output head predicts both a residual correction and
 an aleatoric uncertainty. Both yield spatially resolved σ fields, and both serve here as concrete cases for the evaluation question
 we raise, not as competitors whose ranking is the object of study.

40 Evaluating the quality of a spatial uncertainty field turns out to be more subtle than reporting a single calibration metric.
 As we demonstrate below, a σ field can be well-calibrated in aggregate (low expected calibration error, ECE) yet carry no
 spatial information (a σ field that varies without tracking the error field), or it can be statistically correlated with error at
 the pixel level yet add no detectable value at the operational scale of resurvey planning (because the signal is confined to
 within-survey-block variation that budget-constrained planning cannot exploit). Any single axis of evaluation (calibration,
 45 correlation, or prioritization), taken alone, can mislead. We therefore argue that spatial uncertainty must be evaluated along
 multiple complementary axes, and we develop **TAP-Bathy v1.0** (Three-Axis Protocol for Bathymetric Uncertainty Evaluation)
 to do so. We identify several distinct ways in which single-axis evaluation can produce a misleading verdict, and show that a
 real held-out prediction can fall prey to them simultaneously.

Our work draws on three lines of prior research.

50 *Bathymetric uncertainty.* Standard practice relies on kriging variance from fitted variograms (Cressie, 1993; Goovaerts,
 1997), with operational standards such as IHO S-44 (International Hydrographic Organization, 2022) and algorithms like
 CUBE (Calder and Mayer, 2003) providing grid-level uncertainty. These methods offer principled uncertainty fields but are
 limited by stationarity assumptions. Direct comparisons of geostatistical and learned uncertainty under a common multi-axis
 protocol have been absent.

55 *Learned predictive uncertainty.* Deep networks produce spatially resolved uncertainty via heteroscedastic heads (Kendall
 and Gal, 2017; Nix and Weigend, 1994), MC Dropout (Gal and Ghahramani, 2016), or deep ensembles (Lakshminarayanan
 et al., 2017). Recent surveys distinguish aleatoric from epistemic uncertainty (Hüllermeier and Waegeman, 2021; Gawlikowski
 et al., 2023) and review the broader family of methods including Laplace approximations (Daxberger et al., 2021) and Bayesian



neural networks (Wilson and Izmailov, 2020). A key concern is that these estimates can degrade under distribution shift (Ovadia et al., 2019). In bathymetry, learned approaches have shown promise for depth reconstruction (Xie et al., 2023; Liu et al., 2024), but uncertainty evaluation has been limited to aggregate metrics without spatial decomposition or operational testing.

Spatial validation. Block cross-validation (Roberts et al., 2017) and the Modifiable Areal Unit Problem (Openshaw, 1984) formalize the dependence of evaluation on spatial aggregation scale. The within/between-group decomposition of correlation (Simpson, 1951) provides the statistical foundation for evaluating at which spatial scale uncertainty carries information. Calibration metrics alone (Naeini et al., 2015; Guo et al., 2017; Kuleshov et al., 2018) and proper scoring rules (Gneiting and Raftery, 2007) assess aggregate reliability but cannot distinguish informative from uninformative uncertainty fields.

This paper presents and assesses TAP-Bathy v1.0, a named and versioned method for evaluating spatial uncertainty in bathymetry. Our contributions are two-fold. The **methodological contribution** is TAP-Bathy v1.0, a three-axis protocol for evaluating spatial UQ—calibration, an exact within/cross-patch decomposition of the error–uncertainty correlation, and operational prioritization at a fixed resurvey budget—with the argument that no single axis suffices. The **empirical contribution** applies the protocol to three terrains spanning distinct geomorphic regimes, comparing learned (MC Dropout) and geostatistical (ordinary kriging) uncertainty. Supporting checks include pre-registered comparisons of deep-ensemble versus MC-Dropout uncertainty, a prospective test of UQ generalization on an unseen region (blind001), a blind prediction (008), and an eight-region prospective test of kriging-variance alignment (Section 5.2). Full pre-registration documents, lock manifests, and canonical-reveal scripts are archived in the TAP-Bathy v1.0 Zenodo release.

2 Data and study regions

2.1 Data sources

The primary data source is the difference between two releases of the GEBCO global bathymetric grid: GEBCO 2021 (the “prior”) and GEBCO 2026 (the “target”). Both are released at 15-arcsecond native resolution (~ 450 m at the equator), but the 2026 release incorporates five additional years of multibeam, single-beam, and crowd-sourced bathymetry that were not available in 2021. For our patch-based processing, both grids are resampled to a uniform 100 m pixel spacing, giving 64×64 -pixel patches of ~ 6.4 km $\times$ 6.4 km. The temporal transition from 2021 to 2026 provides an operationally realistic “holdout”: it reflects the kind of improvement that a new survey campaign would produce, and it is strictly out-of-sample with respect to the 2021 prior. We refer to it as a temporal-spatial holdout: “temporal” because GEBCO 2026 post-dates the 2021 training prior, and “spatial” because the TID-based mask restricts evaluation to pixels where the data type upgraded from interpolation (TID 40) to direct multibeam (TID 11). The 2026 compilation remains an interpolated grid rather than direct seabed measurements; it serves as a higher-quality proxy, not a true ground-truth.

For each grid cell we extract the type identification (TID) metadata of the underlying source data from the GEBCO gridded data releases (GEBCO Compilation Group, 2021; GEBCO Bathymetric Compilation Group 2026, 2026). Our evaluation uses a temporal holdout: pixels whose TID in GEBCO 2021 is 40 (interpolated or predicted, typically from satellite altimetry and/or gravity data) and whose TID in GEBCO 2026 has been upgraded to 11 (direct multibeam). These are locations where a GEBCO



grid cell that was previously filled by interpolation has been superseded by a new multibeam survey, providing a real-world proxy for the information gain that a resurvey campaign would produce. The holdout set is strictly out-of-sample with respect to the GEBCO 2021 prior, since the 2021 grid at these locations contained only interpolated values. Pixels that are land, lake, or otherwise non-seabed ($TID \geq 70$) are excluded.

2.2 Study regions

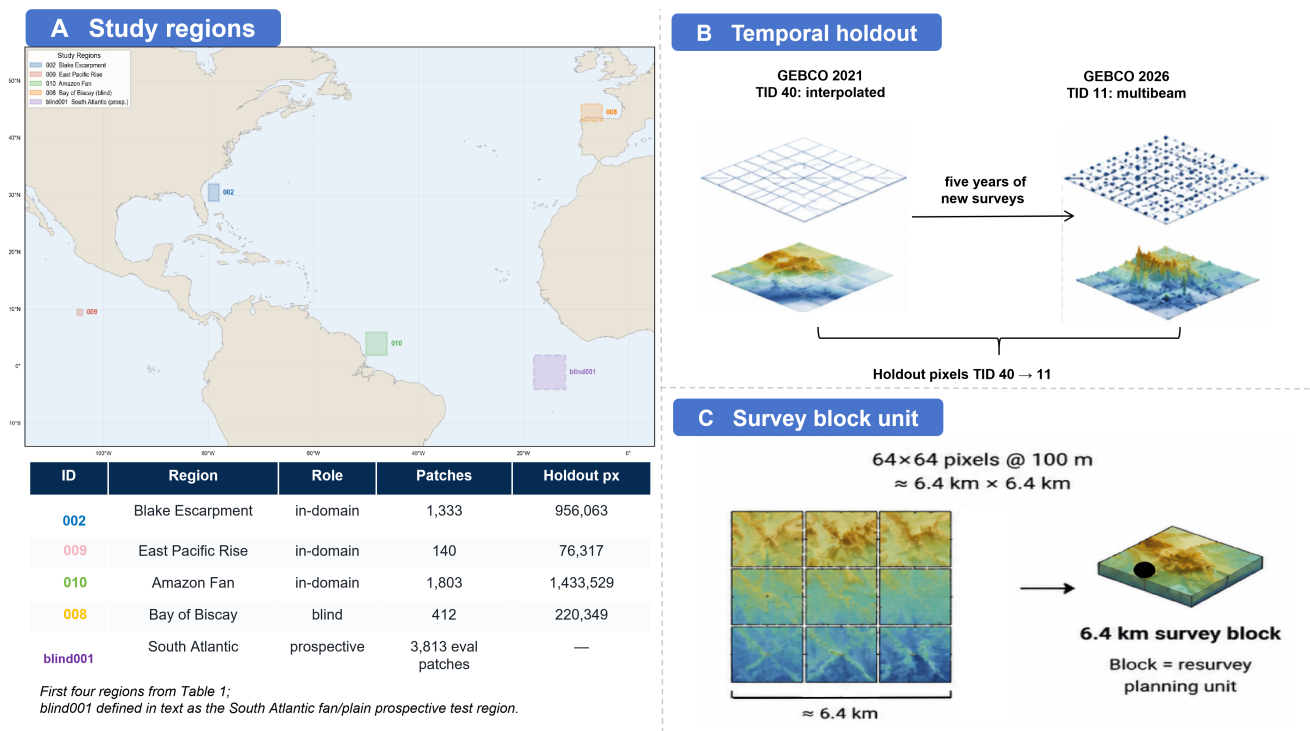


Figure 1. Data and evaluation design. (A) Study regions used for in-domain evaluation (002 Blake Escarpment, 009 East Pacific Rise, 010 Amazon Fan), blind testing (008 Bay of Biscay), and prospective validation (blind001 South Atlantic). (B) Temporal-spatial holdout definition: GEBCO 2021 TID 40 interpolated pixels upgraded to GEBCO 2026 TID 11 multibeam pixels define the out-of-sample evaluation support. (C) The operational resurvey unit is a 64×64-pixel block at 100 m spacing, approximately 6.4 km × 6.4 km; all prioritization results rank blocks rather than individual pixels.

We use four study regions selected to span distinct geomorphic regimes, chosen by terrain type rather than by geographic proximity (Figure 1, Table 1). Regions 002, 009, and 010 serve as the in-domain set on which the three-axis protocol was developed; region 008 is a blind prediction region, pre-registered before any model training or evaluation (Section 4.5.3). An additional eight regions (spanning the Atlantic, Pacific, and Indian Oceans) were used solely for a prospective test of kriging-variance alignment (Section 5.2); they do not carry DL uncertainty results and are not part of the three-axis protocol evaluation.



Table 1. Study regions. All patches are 64×64 prior-DEM pixels ($\sim 6.4 \text{ km} \times 6.4 \text{ km}$). Holdout pixels are those with a validated TID transition from 2021 to 2026. Patch counts are the number of test-set patches containing holdout pixels. Holdout pixel counts refer to the Kr-common evaluation subset (50% calibration split, seed 42, retaining only patches common to both DL and kriging); see Section 2 for the full split protocol.

Region	Location	Terrain	Lat. range	Lon. range	Patches	Holdout px
002	W. North Atlantic	Cont. escarpment	29.0–32.0°N	80.0–78.0°W	1,333	956,063
009	E. Pacific	Mid-ocean ridge	9.0–10.0°N	105.0–104.0°W	140	76,317
010	W. Eq. Atlantic	Deep-sea fan	2.0–6.0°N	50.0–46.0°W	1,803	1,433,529
008	NE Atlantic	Shelf/canyon	43.0–46.0°N	9.0–5.0°W	412	220,349

Region 002 (Blake Escarpment, western North Atlantic) is a steep carbonate-platform margin off the southeastern United States. The escarpment is incised by submarine canyons, producing strong local depth gradients (depth range ~ 20 – $1,000 \text{ m}$ across the sampled area) and a roughness distribution with a pronounced high tail: mean local roughness 1.55 m (5×5 -pixel standard deviation of prior elevation, $\sim 0.5 \text{ km}$ window at 100 m pixel spacing), P99/P50 roughness ratio 12.1, and observation density corresponding to a mean holdout-to-observation distance of $31.3 \text{ px} \approx 3.1 \text{ km}$. The canyon-incised topography concentrates roughness in a minority of pixels (P99 roughness 10.4 m vs. median 0.86 m), while the intervening smooth plateau dominates the mean.

Region 009 (East Pacific Rise, eastern equatorial Pacific) spans a fast-spreading mid-ocean ridge axis at 9 – 10°N . The terrain is uniformly rough (mean local roughness 8.1 m , the highest among the three in-domain regions) with moderate spatial variability (roughness CV 0.93). Depth ranges from approximately $1,700$ to $3,400 \text{ m}$. It has the highest observation density among the three in-domain regions (91.7% TID 11 in GEBCO 2021) and the shortest mean holdout-to-observation distance ($19.5 \text{ px} \approx 2.0 \text{ km}$).

Region 010 (Amazon Fan, western equatorial Atlantic) is a low-relief deep-sea sedimentary environment off the coast of northern Brazil. The terrain is moderately smooth (mean local roughness 2.1 m , CV 1.11) with sparse ship-track coverage (mean observation spacing $\sim 36.9 \text{ px} \approx 3.7 \text{ km}$). Depth ranges from near sea level at the coast to approximately $4,000 \text{ m}$ at the abyssal plain.

An additional eight regions (spanning the Atlantic, Pacific, and Indian Oceans; the data supplement lists their locations and terrain types) were used solely for a prospective test of kriging-variance alignment (Section 5.2); they do not carry DL uncertainty results and are not part of the three-axis protocol evaluation. Their selection criteria, pre-registered predictions, and outcomes are given in Section 5.2.

Region 008 (Bay of Biscay / North Iberian margin, northeastern Atlantic) combines a continental shelf break with submarine canyon systems and an abyssal-plain transition. We pre-registered it as a blind prediction region before any model training or data inspection (Section 4.5.3), selecting it to provide a discriminative test of candidate terrain predictors of UQ quality.



125 Region blind001 (South Atlantic, 4°S–2°N, 12–18°W) is a fan/plain terrain with intermediate error–roughness coupling (see data supplement) and the largest evaluation set in the study ($N = 3813$ patches). blind001 was held out from all ProbabilisticMCUNet training and is used exclusively for a pre-registered prospective test of DL UQ generalization (Section 4.5.2). Its kriging-variance alignment also contributes to the eight-region prospective test of Section 5.2.

2.3 Preprocessing and patch generation

130 Each region is divided into 64×64 -pixel patches on the 100 m grid. Throughout this paper, “patch,” “block,” and “survey block” are used interchangeably to refer to a 64×64 -pixel spatial unit ($\sim 6.4 \text{ km} \times 6.4 \text{ km}$); this is the operational unit at which resurvey decisions are made (Seabed 2030 planning operates at the scale of survey blocks, not individual soundings). Within each patch, the prior elevation, a binary observation mask, and the residual $R = Z_{\text{target}} - Z_{\text{prior}}$ are extracted. The observation mask simulates ship-track transects: random linear tracks are generated within the area of existing GEBCO 2021
 135 multibeam coverage (TID 11 pixels), constrained to 15% areal coverage and excluding all holdout pixels (see leakage audit, Section 3.2). At each track pixel, the GEBCO 2026 depth is provided as a training observation; the residual at these pixels is $R = M \cdot (Z_{\text{target}} - Z_{\text{prior}})$. This setup mimics the operational scenario where a sparse set of recent multibeam soundings is available to correct an older interpolated grid. Only holdout pixels (those with a validated TID transition from 40 in 2021 to 11 in 2026) are used for evaluation.

140 The holdout pixels within each region are subdivided at the patch level into a 50% calibration split and a 50% evaluation split (random seed 42). The calibration split is used exclusively to fit the post-hoc rescaling scalar s for each model (Section 3.4); all metrics (ECE, r , capture rates) are computed on the evaluation split. When comparing DL and kriging, only patches that are present in both methods’ evaluation split are retained (the Kr-common subset; Section 3.8.4), ensuring identical pixel sets for all metrics. The spatial decomposition identity (Equation 4) requires this shared patch set to hold exactly.

145 3 Methods

3.1 Problem formulation

We consider the setting where a gridded bathymetric DEM $Z_{\text{prior}} \in \mathbb{R}^{H \times W}$ (produced by interpolating sparse ship-track soundings, typically combined with satellite altimetry) is to be improved using a set of more recent, higher-quality depth measurements. The prior Z_{prior} is a GEBCO 2021 tile; the target Z_{target} is the corresponding GEBCO 2026 tile, whose updated multi-
 150 beam compilation serves as a higher-accuracy proxy for true depth over the region. The prediction task is to estimate Z_{target} from the prior, conditioned on a binary observation mask $M \in \{0, 1\}^{H \times W}$ indicating which pixels in the prior were derived from ship-track soundings (as opposed to satellite altimetry or interpolation). The residual $R = Z_{\text{target}} - Z_{\text{prior}}$ represents the correction that the model must learn to predict.



We require not only a point prediction of the residual but a spatially resolved uncertainty estimate $\sigma \in \mathbb{R}_+^{H \times W}$ that quantifies, pixel by pixel, the expected magnitude of the prediction error. The evaluation framework for assessing the quality of these uncertainty estimates is detailed in Section 3.8.

3.2 Temporal leakage audit

The temporal holdout requires that no 2026 multibeam information enters training through any channel. We conducted a systematic loss-side and input-side audit. An early pipeline version inadvertently computed the loss over all pixels (including holdout), producing a spuriously optimistic 23.7% apparent improvement on Region 009; the fix restricts both the loss and the input residual to non-holdout pixels using the binary mask $v_i = 1 - h_i$. Six input-pathway checks verify that observation mask, residual, prior, normalization, holdout overlap, and loss exclude holdout pixels; these are formalized as 12 parameterized unit tests (`tests/test_no_temporal_leakage.py`) applicable to any new region. This audit establishes that the evaluation set is genuinely unseen during training, a necessary precondition for the UQ quality claims that follow.

3.3 Model: ProbabilisticMCUNet v1.0

The learned estimator is ProbabilisticMCUNet v1.0¹, a probabilistic U-Net variant that jointly predicts a residual correction $\hat{\mu}$ and an aleatoric uncertainty $\hat{\sigma}_{\text{alea}}$ from three input channels: the prior DEM, a binary observation mask, and the target residual (zero on holdout pixels). The architecture is a standard 4-level U-Net encoder-decoder (Ronneberger et al., 2015) (~31M parameters) with features = 64, GroupNorm, and Dropout2d ($p = 0.1$); the complete layer specification is provided in the data supplement. The final 1×1 convolution outputs $\hat{\mu}$ (residual correction) and $\log \hat{\sigma}_{\text{alea}}^2$ (clamped to $[-7.0, 7.0]$). The predicted depth is $\hat{\mu} + Z_{\text{prior}}$; both outputs are denormalized to meters using per-sample prior statistics.

Epistemic uncertainty is estimated via MC Dropout with $T = 30$ forward passes (saturated at $T \geq 5$, Supplementary S7): $\hat{\sigma}_{\text{epi}} = \text{std}_t(\hat{\mu}^{(t)})$, $\hat{\sigma}_{\text{total}} = \sqrt{\hat{\sigma}_{\text{alea}}^2 + \hat{\sigma}_{\text{epi}}^2}$. $\hat{\sigma}_{\text{alea}}$ dominates (89–96% of variance on 002/009). The model minimizes the Gaussian NLL of the residual targets, masked to exclude holdout pixels (Section 3.2): $\mathcal{L} = \frac{1}{\sum_i v_i} \sum_i v_i \left[\frac{(r_i - \hat{\mu}_i)^2}{2\hat{\sigma}_{\text{alea},i}^2} + \log \hat{\sigma}_{\text{alea},i} \right]$, with $v_i = 1 - h_i$. No auxiliary loss terms are used. Epistemic-uncertainty estimation through MC Dropout is the single method evaluated; the deep-ensemble comparison (Section 4.5.1) tests one alternative estimator. The model has been validated on three terrain types plus one blind region; behavior on unrepresented terrains is unknown.

3.4 Training and evaluation protocol

Each region is divided into 64×64 patches on a regular grid with a stride of 48 pixels (~6.4 km patch size, ~4.8 km stride at the 100 m pixel spacing), producing a 16-pixel overlap between adjacent patches that provides spatial context for the model. Simulated ship-track observation masks are generated within GEBCO 2021 TID 11 (multibeam) areas, excluding holdout pixels (TID 2021=40 $\rightarrow$ 2026=11 transition). The per-patch holdout mask ensures each holdout pixel is evaluated exactly once despite

¹Our architecture is a U-Net with Monte-Carlo dropout and a heteroscedastic output head; the name is unrelated to the MCUNet of Lin et al. (Lin et al., 2020) for efficient TinyML inference. Version v1.0 corresponds to the code archived in the TAP-Bathy v1.0 Zenodo release (see Code availability).

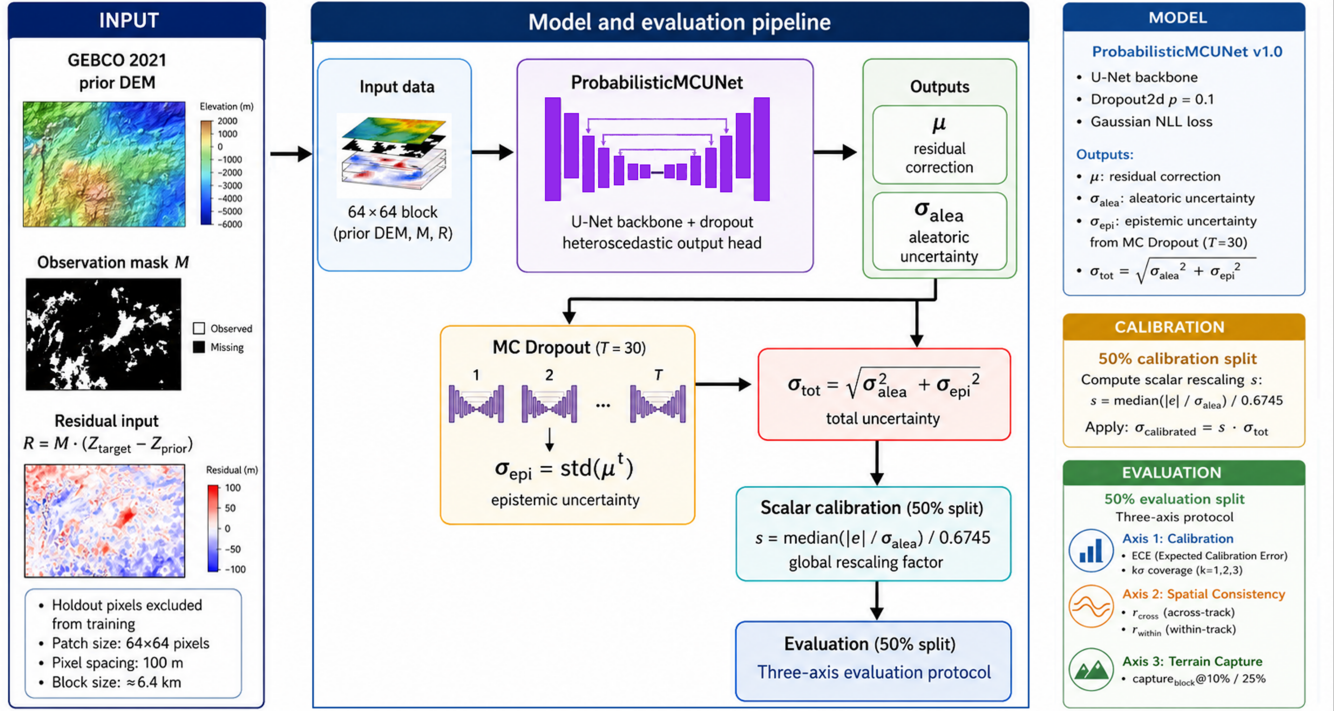


Figure 2. Model and evaluation pipeline. Inputs (prior DEM, observation mask, residual) feed into ProbabilisticMCUNet, which outputs a residual correction $\hat{\mu}$ and aleatoric uncertainty σ_{alea} . MC Dropout with $T = 30$ forward passes produces epistemic uncertainty σ_{epi} . Total uncertainty $\sigma_{\text{tot}} = \sqrt{\sigma_{\text{alea}}^2 + \sigma_{\text{epi}}^2}$ is rescaled on a 50% calibration split and then evaluated on the held-out 50% through the three-axis protocol (Section 3.8). Lower row: robustness and generalization checks that follow the same pipeline.

the overlapping patch grid. Holdout pixels constitute the primary evaluation set; both models are tested on depth differences that post-date their training data, providing a real-world proxy for predictive accuracy under ongoing survey improvement.

185 All models train for 100 epochs (AdamW (Loshchilov and Hutter, 2019), cosine-annealed LR $10^{-3} \rightarrow 10^{-6}$, batch size 32, ~ 2 GPU-hours per seed on one RTX 4090). Three independent seeds (42, 123, 456) per region quantify seed variability. Training uses only non-holdout pixels; holdout locations are unseen during training. For any uncertainty field σ used in an evaluation path, the sole post-hoc correction is a multiplicative scalar $s(\sigma) = \text{median}_{\text{cal}}(|e|/\sigma)/0.6745$, fitted per region on the 50% calibration split and applied unchanged to the held-out 50% evaluation split. The locked headline path uses σ_{tot} for DL;
 190 σ_{alea} -only quantities are reported as diagnostics where explicitly labeled. No distributional recalibration is applied.

3.5 Kriging baseline

The geostatistical baseline is Ordinary kriging with a spherical variogram model, implemented identically across all regions. To maintain a level comparison, the kriging setup deliberately avoids region-specific tuning beyond the variogram parameters:



Table 2. Fitted spherical variogram parameters for the three in-domain regions.

Region	Nugget (m ²)	Sill (m ²)	Range (px)	Range (km)
002	2.6	37.2	24.0	2.4
009	40.8	1702.6	28.4	2.8
010	5.7	68.0	12.2	1.2

the same neighborhood size, the same global variogram assumption, and the same post-hoc scalar fitting procedure are used for every region.

For each region, an empirical semivariogram is computed from the residuals of the GEBCO 2021 prior relative to the GEBCO 2026 target at the training (observed) pixel locations. A spherical model

$$\gamma(h) = \begin{cases} c_0 + c \left[\frac{3}{2} \frac{h}{a} - \frac{1}{2} \left(\frac{h}{a} \right)^3 \right], & h \leq a \\ c_0 + c, & h > a \end{cases} \quad (1)$$

is fitted by weighted least squares (weights $\propto 1/h^2$) using `gstools.AutoFitVariogram` with default convergence tolerance; the range initial guess is set from the empirical correlation length. This yields estimates of the nugget c_0 , partial sill c , and range a . The fitted parameters vary substantially across terrains (Table 2): the range spans 12.2–28.4 pixels (1.2–2.8 km), and the sill spans 37.2–1702.6 m², reflecting differences in both spatial correlation structure and residual magnitude across the three regions.

For each holdout pixel, the kriging prediction $\hat{z}$ and prediction variance $\hat{\sigma}_{\text{krig}}^2$ are computed from the $K = 100$ nearest training observations using the global variogram. The neighborhood size $K = 100$ is large enough to approximate global kriging for practical purposes while remaining computationally tractable for per-pixel evaluation over $\sim 10^6$ holdout pixels.

The raw kriging variance is rescaled by a multiplicative scalar s_{krig} , fitted on the calibration split using the same procedure as for DL ($s = \text{median}(|e|/\sigma)/0.6745$). This ensures that any differences in calibration or informativeness between DL and kriging reflect genuine differences in the spatial structure of their uncertainty estimates, not differences in the post-hoc adjustment protocol.

3.6 Roughness baseline

Terrain roughness (the local variability of the prior DEM) provides a purely geometric, no-training baseline for spatial error prediction. For each pixel we compute the standard deviation of the GEBCO 2021 prior within a 5×5 window ($\sim 0.5 \text{ km} \times 0.5 \text{ km}$ at the 100 m pixel spacing). This roughness field σ_{rough} is deterministic per region and requires no model training, no variogram fitting, and no post-hoc recalibration. It serves two roles: an operationally relevant heuristic baseline (a survey planner who ranks blocks by terrain complexity is implicitly using a roughness proxy) and a diagnostic for what the learned σ_{alea} encodes, since a model that merely re-encodes roughness would add no information beyond this baseline. The 5×5 window size is arbitrary; sensitivity to window size is discussed qualitatively in Supplementary S2.



3.7 Deep ensemble: a second epistemic-uncertainty method

220 To test whether the three-axis findings are specific to MC Dropout, we train a deep ensemble of $M = 5$ ProbabilisticMCUNet members (Lakshminarayanan et al., 2017) per region with independent seeds (0–4), identical architecture (Section 3.3) and training recipe (Section 3.4). σ_{alea} , σ_{epi} , and σ_{tot} follow the same mean-variance aggregation as MC Dropout (data supplement). The ensemble passes through the identical three-axis pipeline (same calibration split, same metric definitions). The comparison was pre-registered (Supplementary S4).

225 3.8 TAP-Bathy v1.0: Three-Axis UQ Evaluation Protocol

The TAP-Bathy v1.0 protocol described below emerged from diagnostic analysis of the three in-domain regions (002, 009, 010). It is a rational reconstruction of the evaluation dimensions that proved jointly necessary to prevent misleading conclusions, not an a priori design. Every metric definition, budget level, and evaluation split was specified after observing the in-domain data, with the exception of the blind001 prospective test (Section 4.5.2), the 008 blind prediction thresholds (Section 4.5.3), the
 230 deep-ensemble comparison (Section 4.5.1), and the kriging-alignment prospective test (Section 5.2), which were pre-registered before any training or evaluation on those regions. The protocol is presented as a methodological contribution (a template for bathymetric UQ evaluation that future studies can apply prospectively), while the empirical results that follow are post-hoc findings protected by multi-seed replication (3 seeds $\times$ 3 regions) and block-bootstrap confidence intervals.

3.8.1 Axis 1: Calibration: does σ mean what it claims?

235 Calibration is assessed through three complementary measures:

- (1) **Expected Calibration Error (ECE)**. Standardized absolute errors $z_i = |e_i|/(s \cdot \sigma_i)$ should follow a half-normal distribution if σ_i is correctly scaled and errors are Gaussian. We evaluate the empirical CDF of $\{z_i\}$ at $K = 30$ evenly spaced levels $k_j \in [0.1, 3.0]$, comparing the fraction of pixels with $z_i \leq k_j$ to the Gaussian-expected half-normal CDF $\text{erf}(k_j/\sqrt{2})$:

$$240 \quad \text{ECE} = \frac{1}{K} \sum_{j=1}^K \left| \frac{1}{N} \sum_{i=1}^N \mathbf{1}[z_i \leq k_j] - \text{erf}\left(\frac{k_j}{\sqrt{2}}\right) \right|. \quad (2)$$

This is the mean absolute deviation between the empirical and theoretical coverage curves, evaluated as a reliability diagram in expectation. We report ECE as a percentage.

- (2) **Empirical $k\sigma$ coverage**. The fraction of holdout pixels satisfying $|e_i| \leq k \cdot s \cdot \sigma_i$ for $k \in \{1, 2, 3\}$. For Gaussian errors, these are 68.3%, 95.4%, and 99.7% respectively.

- 245 (3) **Reliability diagrams** (Appendix Figure D1). The full coverage curve across all k -levels, plotted against the Gaussian CDF, with multi-seed bands for DL and bootstrap confidence bands for kriging.



Three-Axis Protocol for Spatial UQ

Input: predicted $\sigma(x)$ + holdout error $|e(x)|$, grouped by 6.4 km blocks

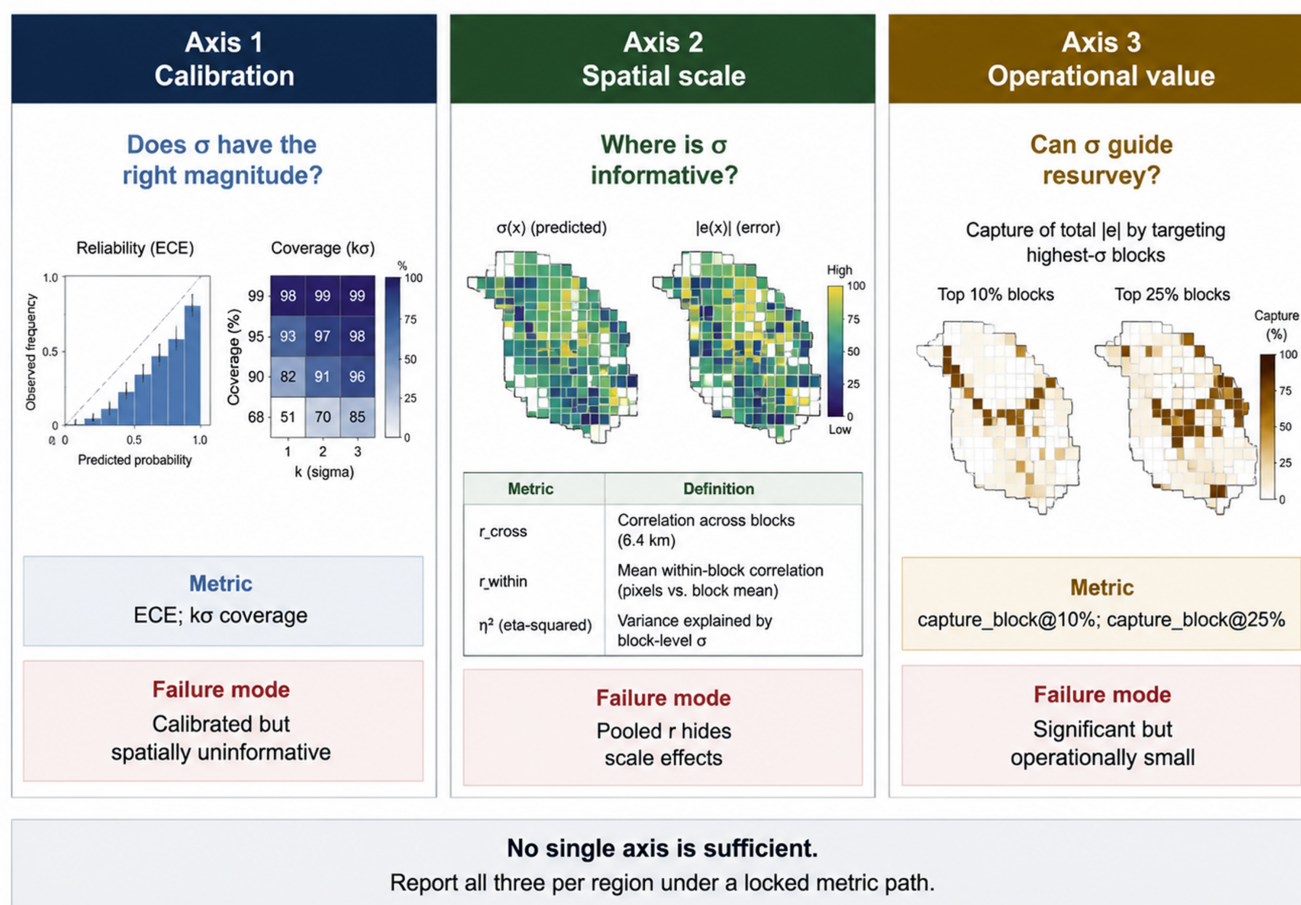


Figure 3. TAP-Bathy v1.0. A spatial σ field is evaluated along three independent axes, each diagnosing a distinct failure mode that the other axes miss. The protocol emerged from diagnostic analysis and is presented as a rational reconstruction; its empirical validation is limited to $n = 3$ regions.



All calibration metrics are computed *after* the per-model, per-region rescaling scalar s is fitted on the 50% calibration split and applied to the held-out 50% evaluation split.

For the ensemble-vs-MC comparison (Section 4.5.1), where both methods produce $\sigma_{\text{tot}} = \sqrt{\sigma_{\text{alea}}^2 + \sigma_{\text{epi}}^2}$ and the comparison
 250 requires identical metric paths, we use a stricter variant. After the same 50/50 calibration split and scalar fitting, the evaluation-set standardized residuals $z_i = |e_i|/(s \cdot \sigma_{\text{tot},i})$ are sorted by magnitude and partitioned into $B = 10$ approximately equal-count rank bins. Within each bin b , the fraction of pixels with $z_i \leq 1.0$ is compared to the Gaussian expectation of 0.6827:

$$\text{RCD} = \frac{1}{B} \sum_{b=1}^B \left| \frac{1}{|\mathcal{D}_b|} \sum_{i \in \mathcal{D}_b} \mathbf{1}[z_i \leq 1.0] - 0.6827 \right| \times 100\%. \quad (3)$$

Because the bins condition on the realized residual rank, even a perfectly calibrated half-normal sample would not yield
 255 68.3% coverage in every bin: low- z bins are expected to be near 100% and high- z bins near 0%, giving an absolute value of $\sim 41\%$ rather than 0. **RCD** (rank-conditioned coverage deviation; RCD in equations) is therefore a deliberately severe, rank-conditioned diagnostic whose absolute magnitude is not comparable to the conventional ECE of Equation 2 (Table 4); it is used only for the *relative* Δ between two methods evaluated under an identical code path.

3.8.2 Axis 2: Spatial Informativeness: where is σ informative?

260 We decompose the pooled-pixel Pearson correlation $r_{\text{pooled}} = \text{Corr}(|e|, \sigma)$ into cross-patch (r_{cross}) and pooled within-patch (r_{within}) components using the exact identity:

$$r_{\text{pooled}} = r_{\text{cross}} \cdot \sqrt{\eta_{\sigma}^2 \eta_e^2} + r_{\text{within}} \cdot \sqrt{(1 - \eta_{\sigma}^2)(1 - \eta_e^2)}, \quad (4)$$

where η_{σ}^2 and η_e^2 are the between-patch variance fractions of σ and $|e|$ respectively. The derivation follows from the ANOVA
 265 decomposition of total covariance into between-patch and within-patch components (the cross-terms vanish because within-patch deviations sum to zero within each patch), and is given in full in A. The three components use unified definitions: r_{within} is the pooled within-patch correlation from concatenated per-patch demeaned residuals; r_{cross} is the n_k -weighted² Pearson correlation of patch-mean $\bar{\sigma}_k$ against patch-mean $\bar{e}_k$; and η^2 is the n_k -weighted between-patch variance fraction. Under these definitions the identity holds to machine precision.

The decomposition serves two purposes. First, it *localizes the spatial scale* of the σ signal: $r_{\text{cross}} \gg r_{\text{within}}$ indicates patch-
 270 scale informativeness (the model knows which 6.4 km blocks are hard), while $r_{\text{within}} \gg r_{\text{cross}}$ would indicate pixel-scale discrimination within blocks. Second, it *explains sign reversals*: a negative r_{cross} with large η^2 can yield negative r_{pooled} despite positive r_{within} (as observed for kriging on Region-002, Section 4.3).

² n_k -weighted: each patch k contributes to the covariance and variance sums in proportion to its pixel count n_k . This ensures that large patches do not dominate the correlation estimate and that the decomposition identity holds exactly. The same weighting applies to all between-patch quantities throughout the paper.



3.8.3 Axis 3: Operational Prioritization: can σ guide resurvey decisions?

The operational test simulates a resurvey campaign with a fixed budget. We report two complementary capture metrics that differ in their ranking granularity.

Pixel-level capture ($\text{capture}_{\text{pixel}}$) ranks all holdout pixels individually by σ (descending), selects the top $B\%$ (budget levels $B \in \{10, 25\}\%$ of evaluation pixels), and computes the fraction of total absolute holdout error captured:

$$\text{capture}_{\text{pixel}}(B\%) = \frac{\sum_{i \in \text{top-}B\%} |e_i|}{\sum_i |e_i|}. \quad (5)$$

This metric is reported in Table 6 for direct comparability with the per-pixel r metrics of Axis 2.

Block-level capture ($\text{capture}_{\text{block}}$) ranks survey blocks (6.4 km patches) by their patch-mean σ (descending), selects blocks in order until $B\%$ of evaluation pixels are accumulated, and computes the fraction of total absolute holdout error captured. This is the operationally relevant quantity: resurvey campaigns operate at the scale of survey blocks, not individual soundings, and DL σ 's error-relevant information is concentrated at the cross-patch scale ($r_{\text{cross}} \gg r_{\text{within}}$, Section 4.3). Block-level capture is the primary operational metric and is reported in Table 7, Fig. 4, and Fig. 6; it is also used in the ensemble comparison (Table 8) and the blind001 prospective test (Table 9).

The random baseline is exactly $B\%$ (expectation under random selection) for both metrics. Operational budgets of 10–25% reflect the fraction of unsurveyed seabed that can realistically be re-measured under current resource constraints; the question of which fraction to prioritize is the operationally decisive one for Seabed 2030 planning.

All capture rates are reported per-region first, global aggregate second. Cross-region pooling can exhibit Simpson's paradox: the global ranking may reverse every per-region ranking when between-region variance in the error- σ relationship inflates the global score of one method (demonstrated for roughness vs. DL in Section 4.4). Per-region verification is required before interpreting any global aggregate.

To test whether DL patch-scale and kriging pixel-scale uncertainty signals combine synergistically, we evaluate a two-stage prioritization strategy: patches are ranked by their mean DL σ under the corresponding evaluation path, and within the single partially filled patch that results, pixels are ordered by kriging σ . The difference between two-stage and pure DL patching is:

$$\Delta = \frac{f_{p^*} \left(E_{p^*}^{(\text{kr-ranked})} - E_{p^*}^{(\text{dl-ranked})} \right)}{\sum_i |e_i|}, \quad (6)$$

where f_{p^*} is the fraction of pixels taken from the last patch. Even with perfect within-patch ranking, Δ is structurally bounded by the error mass of a single partial patch divided by the total error, typically < 0.5 percentage points at the budgets tested (the bound $\Delta \leq f_{p^*} E_{p^*} / E_{\text{total}}$ is derived in the accompanying code release). This structural bound, not the quality of the within-patch signal, limits any potential two-stage gain.

3.8.4 Protocol application

All metrics are computed on the identical set of holdout pixels common to both DL and kriging evaluations—the intersection of the two methods' evaluation coverage. We refer to this shared-pixel set as the **Kr-common subset** (the intersection constrained



Table 3. RMSE (m) on temporal holdout pixels. DL values are 3-seed mean $\pm$ standard deviation; kriging is deterministic.

Region	Terrain	Prior	DL	kriging
002	Continental slope	22.11	17.0 ± 0.6	17.5
009	Mid-ocean ridge	63.68	63.2 ± 0.8	59.9
010	Deep-sea fan	28.28	27.1 ± 0.4	24.2

by kriging coverage). Concretely, it is the 50% evaluation split of patches for which kriging produces valid predictions at every holdout pixel. This guarantees that (i) the decomposition identity of Equation 4 holds by construction, and (ii) any difference between DL and kriging metrics reflects a genuine difference in σ quality rather than pixel-set composition. The full-test-patch r_{pooled} values are reported in table notes for completeness; r_{within} and r_{cross} are insensitive to patch-set inclusion because they condition on patch identity.

DL metrics are reported as mean $\pm$ standard deviation across three independently trained seeds per region (seeds 42, 123, 456), re-using the same checkpoints for all protocol axes. kriging and roughness metrics are deterministic point estimates; where applicable, 95% confidence intervals are computed by block bootstrap at the patch level (5,000 resamples). All reported r correlations and their Fisher- z confidence intervals use the number of patches as the effective sample size n , never the number of pixels (which would inflate the degrees of freedom by 2–3 orders of magnitude). All per-pixel metrics exclude the 50% calibration split used for scalar fitting.

The ProbabilisticMCUNet v1.0 implementation, training pipeline, canonical reveal driver, metric functions, and all evaluation scripts are archived in the TAP-Bathy v1.0 Zenodo release at 10.5281/zenodo.23669317 under the MIT license. Model checkpoints, reveal outputs, and the four pre-registration documents with SHA-256 lock manifests and canonical-reveal scripts are archived at the same Zenodo record (see repository README for the complete lock–train–reveal timeline with Git commit hashes). Pinned software dependencies are provided as an `environment.yml` file.

4 Results

4.1 Point estimate accuracy

Before evaluating uncertainty quality, we establish that both estimators produce meaningful depth predictions on the temporal holdout. Table 3 reports the root-mean-square error (RMSE) of the predicted depth relative to the GEBCO 2026 updated-grid proxy, compared against the GEBCO 2021 interpolation (the prior).

DL improves over the prior on the slope (-5.1 m, $\sim 8.5\sigma$ of the seed variation) and marginally on the fan (-1.2 m, $\sim 3\sigma$), but ties the prior on the ridge (-0.5 m, $< 1\sigma$). The ridge result is expected: mid-ocean ridge bathymetry at 100 m resolution is dominated by the large-scale depth trend that the prior already captures, and the 15% ship-track observation mask provides limited additional information for residual prediction. kriging achieves lower RMSE than DL on the ridge (59.9 vs. 63.2 m) and the fan (24.2 vs. 27.1 m). This is not a weakness for the UQ comparison: the fact that kriging provides a better or comparable



Table 4. Calibration metrics. s_{alea} is the post-hoc rescaling factor ($\text{median}(|e|/\sigma)/0.6745$). Lower ECE and coverage closer to 68.3% indicate better calibration.

Region	Est.	ECE (%)	$\text{cov}_{1\sigma}$ (%)	s_{alea}
002	DL	5.5 ± 0.3	65.0 ± 0.6	0.92 ± 0.02
	Kr	13.2	59.2	N/A
009	DL	5.3 ± 0.2	61.9 ± 0.2	2.73 ± 0.42
	Kr	3.9	67.5	N/A
010	DL	4.9 ± 0.4	64.9 ± 0.3	1.06 ± 0.13
	Kr	4.1	65.8	N/A

point estimate in two of three regions makes the uncertainty comparison more stringent. The learned uncertainty is not being favored by a uniformly superior point predictor.

Point-estimate accuracy is reported to establish that both uncertainty fields accompany competitive depth predictions; the remainder of this section evaluates the quality of those uncertainty fields.

4.2 Calibration: does σ mean what it claims?

We evaluate calibration along the three measures defined in Section 3.8: ECE, empirical 1σ coverage, and full reliability diagrams. All metrics are computed after per-model, per-region post-hoc rescaling by the scalar s , fitted exclusively on the 50% calibration patch split and evaluated on the held-out 50% (Section 3.8.4). Multi-seed DL results report mean $\pm$ standard deviation across three independently trained checkpoints per region; kriging values are point estimates with 95% block-bootstrap confidence intervals (patch-level, 5,000 resamples).

The learned estimator achieves consistent ECE across all three regions (4.9–5.5%), with tight across-seed variation (± 0.2 – 0.4 pp). The post-hoc scalar s_{alea} is near unity for the slope (0.92) and fan (1.06), indicating that the raw heteroscedastic output transfers its scale to the temporal holdout on these terrains. On the ridge (009), $s_{\text{alea}} = 2.73$ with substantial seed-to-seed variation (CV = 15%); the learned σ_{alea} systematically underestimates the error magnitude on ridge terrain, requiring a larger correction. Despite this seed-dependent magnitude variability, the ECE remains stable across seeds (± 0.2 pp), confirming that per-model recalibration effectively decouples scale from calibration quality (Section 3.4).

Empirical 1σ coverage is systematically below the Gaussian expectation of 68.3% for DL in all three regions (61.9–65.0%), despite low ECE. The 2σ and 3σ coverage are also below their Gaussian targets (85.6% on 002 to 91.3% on 009 for 2σ vs. Gaussian 95.4%; 91.6% on 002 to 98.2% on 009 for 3σ vs. Gaussian 99.7%), with the deficit largest on the slope (002) and smallest on the ridge (009). This is the signature of heavy-tailed prediction errors: the under-coverage persists across all σ multiples, consistent with a distribution whose tails decay more slowly than a Gaussian. Standardized residuals $z_i = e_i/(s \cdot \sigma_i)$ are strongly heavy-tailed in all regions (excess kurtosis: +741, +7, and +17 for 002, 009, and 010 respectively, vs. 0 for a Gaussian). A small number of overconfident pixels ($\sim 0.1\%$) drives the extreme kurtosis in region 002; trimming these does not



materially change the $|e|$ - σ correlation or the calibration conclusions (Supplementary S1). We present full reliability diagrams (Appendix Figure D1) as the primary calibration evidence; ECE and 1σ coverage are reported as compact summaries alongside. Figure 4 provides a unified three-axis overview of the main results. We avoid describing DL σ as simply "well-calibrated": the low ECE supports reasonable average calibration, but the systematic under-coverage at all σ levels and extreme kurtosis in 002 indicate that tail risk remains non-negligible.

kriging achieves excellent ECE on the ridge and fan (3.9% on 009, 4.1% on 010), outperforming DL on these two regions. On the fan (010), the σ field varies (CV = 18.3%) but its variation aligns only weakly with error; the calibrated ECE is good but the σ field carries negligible spatial information (Section 4.3). On the slope (002), kriging calibration is poor (ECE = 13.2%, $\text{cov}_{1\sigma} = 59.2\%$). The mechanism behind these patterns is examined in Section 5.2.

4.3 Spatial informativeness: where is σ informative?

We apply the spatial decomposition of Axis 2 (Equation 4) to both estimators. Table 5 reports r_{pooled} , r_{within} , and r_{cross} for all three regions. All quantities are computed on the identical pixel set, the intersection of DL predictions and kriging holdout pixels (Kr-common evaluation subset, Section 3.8.4), for two reasons. First, comparability: both methods are evaluated on exactly the same holdout pixels, so differences in r reflect genuine differences in σ quality rather than pixel-set composition. Second, decomposition self-consistency: the identity of Equation 4 requires that r_{pooled} , r_{within} , and r_{cross} share a common patch set; otherwise the algebraic relationship does not hold. The full-test-patch r_{pooled} values are reported in the table note for completeness; r_{within} and r_{cross} are insensitive to patch-set inclusion because they condition on patch identity. The unified definitions (r_{within} pooled-concatenated, r_{cross} n -weighted) follow Section 3.8 and replace earlier mixed-definition computations (Section 3.8.2).

DL informativeness is concentrated at the cross-patch scale. In all three regions, the pooled-pixel correlation is dominated by cross-patch variation: $r_{\text{cross}} = 0.08$ – 0.50 (DL, locked evaluation protocol) versus $r_{\text{within}} = 0.16$ – 0.18 (Table 5). A within-patch variogram analysis (Supplementary S3) confirms σ is not patch-constant (it has 0.5–2.6 km sub-patch structure) but this fine-scale variation contributes little to error alignment, which is dominated by between-patch differences. DL σ predicts which patch is hard. The cross-patch signal is consistent across terrain types and robust across training seeds. A 20-split sensitivity analysis (Supplementary S6) confirms that the choice of calibration split does not materially affect r_{cross} (CV $\leq 7.3\%$ on the two large regions).³

Region 009 merits caution. The small evaluation set of 70 patches (50% of 140 total holdout patches under the calibration split) amplifies seed-to-seed variability across all three correlations (3-seed pre-split σ_{alea} path, the only path with multi-seed data): $r_{\text{pooled}} = +0.284 \pm 0.084$ (CV = 30%), $r_{\text{within}} = +0.162 \pm 0.085$ (CV = 53%), and $r_{\text{cross}} = +0.484 \pm 0.096$ (CV = 20%). For comparison, 002 has CVs of 1.3%, 7.6%, and 2.1% for the same three quantities, and 010 has 3.0%, 42%, and 4.1%. On 009, the within-patch correlation spans +0.07 to +0.28 across only three seeds and is essentially unmeasured; the pooled correlation's relative uncertainty is an order of magnitude larger than on 002 or 010; the cross-patch per-seed range of 0.23 is 5–6 $\times$ wider than on the other two regions. With all three components showing this level of seed instability, 009 provides no

³All r_{cross} values use the locked evaluation path (σ_{tot} , 50/50 split, MC seed 42). Pre-split σ_{alea} values and full cross-walk are in the data supplement.



Three-Axis Protocol: Results Across Three Terrains

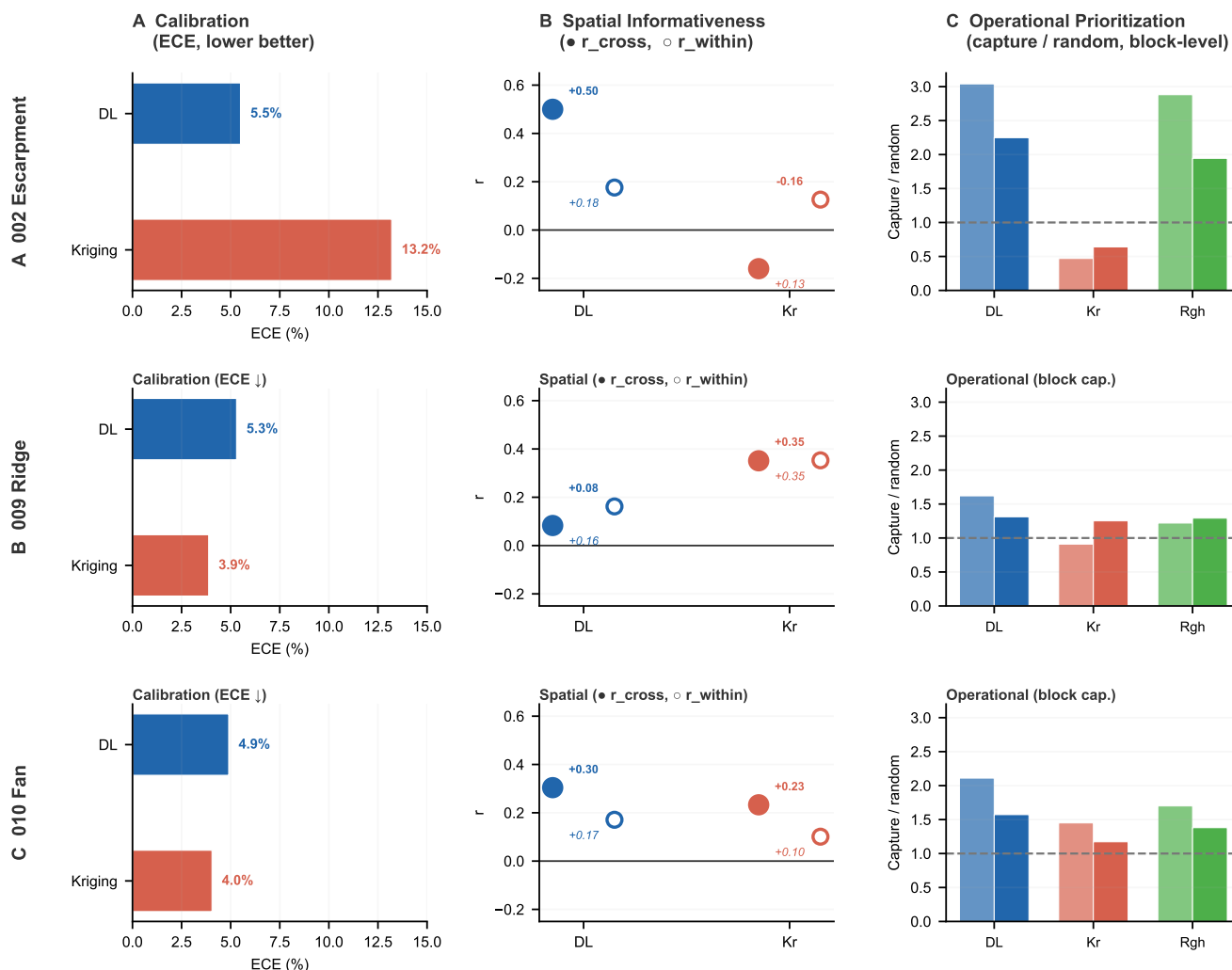


Figure 4. Three-axis protocol applied to three terrains. Panels A–C (rows): 002 Blake Escarpment, 009 East Pacific Rise, 010 Amazon Fan. Columns: (left) Calibration: ECE for DL (blue) and kriging (orange), lower is better; all three panels share a 0–15% x-axis. (center) Spatial informativeness: r_{cross} (filled circles) and r_{within} (open circles) for each method; horizontal line at zero. (right) Operational prioritization: block-level error capture divided by the random expectation at 10% (light) and 25% (full) resurvey budgets; dashed line at 1× (random baseline). All values from the locked evaluation protocol (Section 3.4).



Table 5. Pooled, within-patch, and cross-patch $|e|-\sigma$ Pearson correlations. All quantities computed on the identical patch set (Kr-common holdout pixels) to guarantee self-consistency of the decomposition (Equation 4); the three components satisfy the identity within $\Delta < 10^{-7}$ for each DL seed and $\Delta < 10^{-15}$ for kriging. r_{within} = pooled within-patch; r_{cross} = n -weighted patch-mean. DL values are 3-seed mean $\pm$ seed standard deviation; kriging values are point estimates with 95% block-bootstrap CIs (patch-level, 5,000 resamples) in brackets.

Region	Est.	r_{pooled}	r_{within}	r_{cross}
002	DL	$+0.617 \pm 0.008$	$+0.176 \pm 0.013$	$+0.500$
	Kr	$-0.080 [-0.124, -0.039]$	$+0.126 [+0.104, +0.152]$	$-0.160 [-0.220, -0.100]$
009	DL	$+0.284 \pm 0.084$	$+0.162 \pm 0.085$	$+0.083$
	Kr	$+0.353 [+0.287, +0.424]$	$+0.353 [+0.284, +0.421]$	$+0.351 [+0.164, +0.550]$
010	DL	$+0.437 \pm 0.013$	$+0.171 \pm 0.072$	$+0.304$
	Kr	$+0.163 [+0.144, +0.186]$	$+0.101 [+0.082, +0.123]$	$+0.233 [+0.185, +0.293]$

Note: All values use the locked Kr-common evaluation subset, the 50/50 calibration split, and σ_{tot} (locked evaluation path, MC seed 42). DL r_{pooled} and r_{within} are 3-seed mean $\pm$ SD; DL r_{cross} is reported for seed 42 ($\text{CV} \leq 7.3\%$ across seeds on 002/010; 3-seed range for 009 = 0.23). Kriging values are point estimates with 95% patch-bootstrap CIs (5,000 resamples). Exploratory pre-split σ_{alea} values and split/variogram sensitivity analyses are provided in the data supplement (Supplementary Sx).

individually conclusive evidence about DL UQ spatial structure. We treat 009 as directionally consistent with the cross-region pattern (DL $r_{\text{cross}} \gg r_{\text{within}}$) but individually inconclusive; the 002 and 010 results, each with substantially tighter precision on r_{pooled} and r_{cross} , provide the primary evidence.

Even in 010, the best-measured region for r_{pooled} ($\text{CV} = 3.0\%$) and r_{cross} ($\text{CV} = 4.1\%$), the within-patch correlation has $\text{CV} = 42\%$. This is consistent with the interpretation that DL within-patch UQ signal is intrinsically weak and poorly measured across all terrains, not merely an artifact of small sample size on 009. The weak within-patch alignment constrains the operational resolution: σ discriminates difficult 6.4 km blocks but not pixels within them, so it is suited to block-level resurvey ranking rather than within-block guidance.

kriging on 002: pooled inversion. On the slope, kriging's r_{pooled} is negative (-0.080 , CI entirely below zero). The decomposition reveals why: the n -weighted cross-patch correlation is significantly negative ($r_{\text{cross}} = -0.160$, CI $[-0.220, -0.100]$) while the within-patch correlation is positive ($r_{\text{within}} = +0.126$, CI $[+0.104, +0.152]$). Deep-water patches far from ship tracks receive high kriging variance (the geometric signal), but these patches have low actual error because the smooth slope terrain is easy to predict regardless of observation density. Because between-patch variance dominates ($\eta_{\sigma}^2 = 0.84$, $\eta_e^2 = 0.59$), the negative cross-patch term determines the sign of r_{pooled} . Reporting only r_{pooled} would suggest that kriging σ is uniformly uninformative; the decomposition shows that it carries a measurable within-patch signal whose operational value is assessed in Section 4.4.

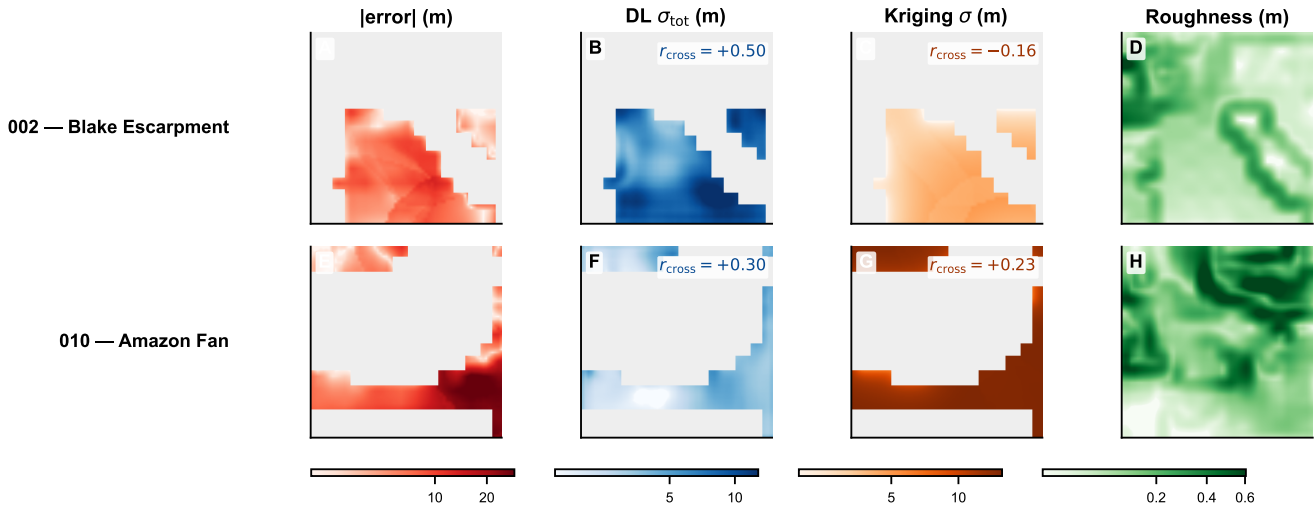


Figure 5. Spatial fields on the common Kr-common holdout evaluation mask. Rows: Blake Escarpment (region 002, top) and Amazon Fan (region 010, bottom). Columns: absolute prediction error, learned total uncertainty (σ_{tot} from MC Dropout with $T = 30$), kriging uncertainty, and local terrain roughness (5×5 std of prior DEM). Light-gray pixels are invalid (outside the unified holdout $\cap$ eval-split $\cap$ Kr-common mask) and excluded from the error, DL σ_{tot} , and kriging σ columns; roughness is computed over the full 64×64 patch as a terrain property. Per-column color limits are shared between regions and computed from the 2nd–98th percentiles of valid pixels. Each region displays one representative non-outlier block: $\geq 20\%$ valid holdout coverage, nearest to the median block-level $|\text{error}|$ among eligible blocks. Panel letters A–H; r_{cross} annotations (B, C, F, G) use the locked evaluation protocol (Section 3.8).

On the ridge (009), kriging performs comparably to DL on all informativeness metrics, and is the only region where kriging r_{cross} is significantly positive. This is also the only region where the variogram range (28.4 px) exceeds the mean observation spacing (19.5 px), giving kriging variance sufficient dynamic range ($\text{CV} \approx 34\%$) to track error variation. However, 009 has the smallest evaluation set (70 patches after the calibration split), and DL’s σ_{alea} ranking (pre-split diagnostic; headline uses σ_{tot}) shows the highest seed-to-seed variability here (Spearman $\rho = 0.76$ vs. 0.92–0.94 for 002/010), making it the least conclusive of the three regions.

The r_{cross} values reported in Table 5 use one of several valid evaluation paths; the alternative paths (locked evaluation, split-sensitivity mean, MC-ablation $T=30$) and their numerical reconciliation are documented in the data supplement.

The ECE–informativeness disconnect is clearest on the fan (010), where kriging achieves good calibration ($\text{ECE} = 4.1\%$, Table 4) but its r_{pooled} (+0.163) is substantially below DL’s (+0.437). The σ field varies ($\text{CV} = 18.3\%$) but this variation carries limited spatial information, showing that calibration and informativeness are distinct axes. This is the *calibration–informativeness disconnect* that motivates the multi-axis protocol: ECE alone cannot distinguish a well-calibrated but uninformative uncertainty field from one that is both calibrated and informative. The operational consequences of this disconnect are examined in the next section.



Table 6. Diagnostic pixel-level capture. Percentage of total holdout $|e|$ captured by top-ranked pixels ($\text{capture}_{\text{pixel}}$). The operationally relevant block-level metric is reported in Fig. 4 and Fig. 6. Random expectation equals the budget fraction. DL values are 3-seed range; kriging and roughness are deterministic.

Region	Est.	10% budget	25% budget
002	DL	30.2–31.4%	56.2–58.7%
	kriging	5.3%	16.7%
	Roughness	24.8%	41.7%
	Random	(10%)	(25%)
009	DL	14.8–18.0%	32.8–36.2%
	kriging	14.0%	37.2%
	Roughness	9.2%	25.7%
	Random	(10%)	(25%)
010	DL	19.0–22.1%	38.9–39.3%
	kriging	15.5%	30.2%
	Roughness	20.8%	35.1%
	Random	(10%)	(25%)

4.4 Operational prioritization: can σ guide resurvey decisions?

The third axis evaluates whether σ can guide a practical resurvey decision: given a fixed measurement budget, does ranking locations by σ capture more total error than random selection or a geomorphometric proxy? Both budgets correspond to operationally plausible resurvey fractions: under the resource constraints that motivate Seabed 2030, only a small fraction of interpolated seabed can realistically be re-measured, which makes the question of *which* fraction to prioritize the operationally decisive one. Table 6 reports the fraction of total absolute holdout error captured by the top-10% and top-25% pixels ranked by each method's σ ($\text{capture}_{\text{pixel}}$), alongside the deterministic roughness baseline (5×5 local standard deviation of the GEBCO 2021 prior) and the random expectation. DL values are 3-seed ranges; kriging and roughness are deterministic. All captures are computed per region on the Kr-common evaluation subset, the same patch set used for the spatial decomposition in Section 4.3 (identical holdout pixels, 50% random calibration split). Global aggregates are reported alongside per-region values but are subject to Simpson's-paradox reversal (see Roughness paragraph below) and should not be interpreted without per-region verification.

Table 6 reports pixel-level capture rates for comparability with the per-pixel r metrics of Axis 2. Because DL σ_{alea} aligns with error primarily at the cross-patch scale ($r_{\text{cross}} \gg r_{\text{within}}$, Section 4.3), the operational interpretation is block-level: the σ field identifies which 6.4 km survey blocks to prioritize, not which individual soundings within a block, even though σ itself resolves sub-patch spatial structure (Supplementary S3). A pixel-level capture rate should be understood as a diagnostic whose operational conclusions are validated against the patch-scale two-stage analysis (below).



Table 7. Block-level operational prioritization: percentage of total holdout $|e|$ captured by ranking 6.4 km survey blocks by their patch-mean σ ($\text{capture}_{\text{block}}$). Random expectation equals the budget fraction. Values computed from the Kr-common evaluation subset (seed 42).

Region	Est.	10% budget	25% budget
002	DL	30.4% (3.0×)	56.2% (2.2×)
	kriging	4.7% (0.5×)	16.0% (0.6×)
	Roughness	28.8% (2.9×)	48.6% (1.9×)
	Random	(10%)	(25%)
009	DL	16.2% (1.6×)	32.8% (1.3×)
	kriging	9.1% (0.9×)	31.3% (1.3×)
	Roughness	12.2% (1.2×)	32.3% (1.3×)
	Random	(10%)	(25%)
010	DL	21.1% (2.1×)	39.3% (1.6×)
	kriging	14.5% (1.4×)	29.3% (1.2×)
	Roughness	22.0% (2.2×)	35.5% (1.4×)
	Random	(10%)	(25%)

DL provides actionable patch-level prioritization. On the slope (002), DL captures 2–3× the random expectation: ~30% of total error at 10% budget versus 10% random, and 56.2% at 25% budget (block-level; Table 7). The 3-seed range is tight, confirming that the operational value of the learned difficulty ranking is seed-stable. On the fan (010), DL captures 19–22% at 10% budget and ~39% at 25%, again substantially above random. On the ridge (009), DL’s advantage is more modest (15–18% at 10%, 33–36% at 25%) and should be interpreted with caution given the small evaluation set (70 patches; Section 4.3). The prioritization advantage mirrors the cross-patch r pattern of Section 4.3: strongest where DL r_{cross} is highest (002: +0.50, locked path), weakest where it is lowest (009: +0.083).

kriging prioritization can be worse than random. On the slope (002), kriging variance captures only 5.3% of total error at 10% budget, below the 10% random baseline. This is the operational consequence of the negative cross-patch correlation documented in Section 4.3: ranking pixels globally by kriging variance selects deep-water, low-error patches ahead of the rough, high-error terrain that dominates the prediction error. At the pixel level, kriging on the ridge (009) appears competitive at the 25% budget (37%, pixel-level diagnostic; Table 6) and captures 14% at 10%. The operationally relevant block-level metrics (Table 7) are less favorable: kriging captures 31.3% at 25% (close to DL’s 32.8%) but only 9.1% at 10% (below random). Given 009’s small evaluation set (70 patches; Section 4.3), this single-region comparison should be interpreted as a terrain-specific observation rather than a general claim. On the fan (010), kriging captures 16% at 10% and 30% at 25%, better than random but below DL.

Roughness is a strong heuristic baseline. The 5×5 local standard deviation captures 25% of error at 10% budget on 002 and 21% on 010, outperforming kriging in both regions. DL beats roughness on a per-region basis in all three regions at the



25% budget, and in two of three at 10% (010 is tied: DL 19–22% vs. roughness 20.8%). Global cross-region aggregates invert this ranking at the 10% budget (roughness 25.2% vs. DL 23.8%) because roughness-based pixel ranking, when pooled across regions, disproportionately selects pixels from the region where roughness is most informative. Per-region reporting avoids this Simpson’s-paradox mechanism (Section 3.8).

455 Cross-scale two-stage provides no detectable gain. A natural question is whether a two-stage strategy (ranking patches by DL σ and then ranking pixels within each selected patch by kriging σ) combines the strengths of both estimators. We tested this on all 9 region–seed combinations at both budget levels. The difference between two-stage and pure DL patching is below 0.5 percentage points in all 18 cases (maximum 0.38 pp, region 009 at 10% budget). The reason is structural: at any practical budget, only one patch is partially filled; all preceding patches are taken in their entirety, and within-patch ordering
 460 contributes nothing to fully taken patches. kriging’s within-patch signal ($r_{\text{within}} = +0.10$ – $+0.35$, Section 4.3) is statistically real but is diluted to < 0.5 pp globally because it applies to only one partial patch per budget level.⁴ This is a case of statistically significant informativeness without operational value, and we return to its methodological implications in Section 5.

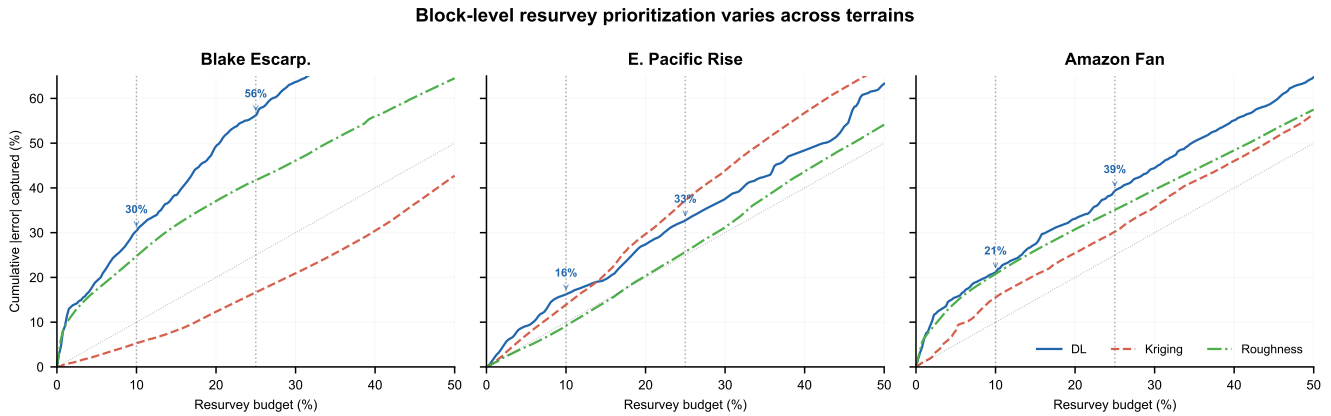


Figure 6. Block-level resurvey prioritization across three terrains. Cumulative $|e|$ captured vs. resurvey budget (0–50% of blocks). Gray dashed: random baseline. Vertical lines at 10% and 25% budget with DL capture annotated. DL (blue) ranks blocks by patch-mean σ ; kriging (orange) and roughness (green) shown for comparison.

Across the three terrains, learned σ is the only estimator that delivers consistent above-random prioritization at the patch scale: it concentrates roughly half of the total holdout error (56.2% on the slope, 39% on the fan) into the top 25% of survey
 465 blocks, and never falls below random in any region or budget. Classical kriging variance, by contrast, is unreliable for this purpose: competitive only on the ridge, and below-random on the slope—the very terrain where resurvey prioritization matters most.

⁴Let p^* be the last (partially filled) patch at budget B . Only p^* benefits from within-patch ranking; all preceding patches are taken in their entirety. The gain $\Delta = f_{p^*}(E_{p^*}^{(\text{kr-ranked})} - E_{p^*}^{(\text{dl-ranked})}) / \sum_i |e_i|$ is structurally bounded by $f_{p^*} E_{p^*} / E_{\text{total}}$, which is < 0.5 pp at the budgets tested.



Table 8. Deep ensemble vs. MC-Dropout (seed 42). Δ = ensemble – MC. Both use the locked evaluation path. Primary endpoint: r_{cross} on Region 002.

Region	r_{cross} (ens / MC / Δ)	RCD (%) (ens / MC / Δ)	capture _{block} @10% (ens / MC / Δ)
002 (4 members)	0.506 / 0.500 / +0.005	42.3 / 41.6 / +0.8	66.8 / 65.8 / +1.0
009 (5 members)	0.145 / 0.083 / +0.062	40.5 / 41.7 / –1.3	3.6 / 12.8 / –9.2
010 (5 members)	0.368 / 0.304 / +0.065	41.5 / 41.2 / +0.3	22.2 / 19.8 / +2.3

The practical implication is scoped but concrete: DL uncertainty can guide *which* 6.4 km survey blocks to re-measure, but, given the patch-scale nature of its signal (Section 4.3), not which pixels within a block, and the present evidence supports this only for terrains represented in training.

4.5 Robustness and prospective checks

The three-axis findings above used MC Dropout as the single epistemic-uncertainty method on three in-domain terrains. We report three pre-registered checks that test whether these findings are specific to the method, the terrains, or the evaluation choices.

4.5.1 Deep ensemble: method-robustness

A pre-registered deep ensemble ($M = 5$, identical backbone and recipe; Section 3.7) tests whether the failure-mode patterns are artifacts of dropout sampling. Under the identical locked three-axis pipeline, the primary endpoint (r_{cross} on Region 002) differs by +0.005 (Table 8), near zero relative to estimation uncertainty. RCD deltas are below 2 pp across all regions, and capture@10% deltas show no consistent sign. The largest swing (–9.2 pp on Region 009) is attributed to the ridge’s small evaluation set (70 patches). One ensemble member (seed 0, Region 002) diverged reproducibly on the aleatoric variance head; per the pre-registered divergence criterion, Region 002 uses four members (Supplementary S4). The verdict: the broad failure-mode pattern is robust to the epistemic-uncertainty estimator within this backbone and recipe; secondary operational capture on small evaluation sets is unstable.

4.5.2 Blind001 prospective test: terrain-generalization

blind001 (South Atlantic fan/plain, $N = 3813$ evaluation patches) was held out from all ProbabilisticMCUNet training. We pre-registered a primary prediction interval $r_{\text{cross}} \in [+0.35, +0.65]$, anchored on blind001’s error-roughness coupling (intermediate between 009 and 010; data supplement). The primary endpoint HIT: $r_{\text{cross}} = 0.358 \pm 0.018$ (3-seed mean), Fisher- z 95% CI [0.33, 0.38]. The 3-seed mean falls inside the pre-registered interval; the lower CI bound (0.33) marginally crosses the lower boundary. Secondary endpoints returned MISS: capture_{block}@10% of 22.7% marginally missed the 25% threshold; the ECE endpoint was invalidated by a lock artifact (pre-registration threshold calibrated to conventional ECE, locked reveal uses



Table 9. Blind001 prospective test against pre-registered thresholds (locked evaluation path).

Metric	Value (3-seed mean $\pm$ SD)	Threshold	Verdict
r_{cross} (primary)	0.358 ± 0.018	$[+0.35, +0.65]$	HIT
Fisher- z CI half-width	0.028	< 0.10	CONCLUSIVE
RCD	$41.8 \pm 0.1\%$	$< 8\%$ (conventional)	MISS*
$\text{capture}_{\text{block}}@10\%$	$22.7 \pm 1.5\%$	$> 25\%$	MISS

*Lock artifact: pre-registration threshold calibrated to conventional ECE (Table 4); locked reveal uses RCD. All three in-domain anchors yield RCD $\sim 41\%$.

A post-hoc conventional-ECE check gives 5.21%, within the $< 8\%$ threshold.

RCD; Table 9 note). This is, to our knowledge, the first adequately-powered prospective evidence that DL bathymetric UQ retains positive cross-patch informativeness on an unseen terrain. The evidence is limited to one region and one architecture.

4.5.3 Region 008: a cautionary example

Region 008 (Bay of Biscay, $N = 206$) was pre-registered as a blind test before any model training. Both DL and kriging
 495 produce inconclusive r_{pooled} (DL -0.04 , kriging $+0.04$; both CIs span zero), falsifying the pre-registered prediction of a strong
 DL advantage. The 008 analysis passed through three data-cleaning rounds, each producing a contradictory conclusion (details
 in data supplement): the three rounds differed in sample size (103 vs. 206 patches), metric (r_{cross} vs. r_{pooled}), and bias handling,
 demonstrating that with small N and unstandardized protocols, evaluation choices alone can manufacture contradictory UQ
 verdicts. A post-hoc power analysis confirms: detecting $r = 0.20$ with 80% power requires $N \approx 193$; the observed effects
 500 ($|r| \approx 0.04$) are near zero. The cautionary lesson is that the protocol's value lies in preventing such sensitivity, not in the 008
 result itself. The three rounds collapsed to a single finding— $N = 206$ insufficient—once a common protocol was applied.

This is not presented as a result about 008 bathymetry per se, but as a methodological exhibit: UQ evaluation is extremely
 sensitive to metric definition, sample selection, and bias handling. Without a pre-specified, multi-axis protocol, analysis choices
 can manufacture spurious “significant” findings. The three-axis protocol (ECE + r with spatial decomposition + operational
 505 priority) is designed to prevent this by locking all definitions and evaluation splits in advance. The 008 experience (a genuinely
 prospective test of the three-axis protocol on an unseen region) demonstrates what happens when these safeguards are absent,
 and we return to its implications for bathymetric UQ evaluation practice in Section 5. Another prospective test, of kriging-
 variance alignment across eight additional regions, is reported in Section 5.2.



5 Discussion

5.1 Why bathymetric UQ needs multi-axis evaluation

These three axes are jointly necessary because no single scalar captures the spatial structure that determines whether uncertainty is useful. The Results demonstrate five instances of the same root cause—collapsing a spatially structured field into a scalar—each producing a misleading verdict: (1) calibration without informativeness (kriging on 010: ECE 4.1%, $r_{\text{pooled}} = +0.16$), (2) significant but operationally negligible within-patch signals, (3) cross-region ranking reversal via Simpson’s paradox, (4) pooled correlation canceling opposing cross-patch and within-patch signals (kriging on 002), and (5) operational verdicts depending on the budget (kriging on 009 capturing 9.1% at 10% vs. 31.3% at 25%, block-level; Table 7). The three-axis protocol, with its within/cross decomposition and budget-resolved capture, is designed to keep this structure visible.

5.2 Kriging-variance alignment: prospective test

In this section, r_{cross} refers to the cross-patch kriging-variance–error correlation, computed with the same n_k -weighted definition as in Section 4.3. Across the three in-domain regions, ordinary-kriging variance displays an alignment spectrum: anti-aligned on the escarpment (002, $r_{\text{cross}} = -0.160$), weakly aligned on the fan (010, $+0.233$), and positively aligned on the ridge (009, $+0.351$, where the variogram range exceeds observation spacing). With $n = 3$, this spectrum is tentative.

To test whether it generalises, we pre-registered predictions for eight additional regions spanning the Atlantic, Pacific, and Indian Oceans. The primary predictor was the error–roughness cross-patch correlation, operationalized through a three-tier threshold rule: Case 2 ($r_{\text{err} \sim \text{rough}} > 0.5$) predicts anti-alignment, Case 1 (< 0.4) predicts weak-positive, and Intermediate (0.4–0.5) carries no pre-locked direction. EPR 15S served as the decisive experiment (Case 2, highest coupling), with its post-training r_{cross} adjudicating between coupling and geometric hypotheses.

The prospective test cleanly falsified the strong (sign-reversal) form of the prediction: all eight new regions produced $r_{\text{cross}} > 0$ (range $+0.08$ to $+0.27$; full table in data supplement). The weak monotonic form—that stronger error–roughness coupling predicts weaker positive alignment—is directionally supported but underpowered ($\rho = -0.54$, $p = 0.21$, $n = 7$ exploratory; see online supplement). The central empirical finding is that positive alignment is the norm: negative r_{cross} remains confined to a single case (002), where extreme error–roughness coupling ($+0.79$) and an inverse distance–roughness relationship create a unique anti-alignment that kriging with external drift (KED) attenuates by 40% but does not reverse (see data supplement for KED results).

The protocol functioned as intended: an exploratory observation was converted into a falsifiable prediction, and the decisive experiment adjudicated. We replace a symmetric three-class description with an asymmetric, empirically grounded picture: predominantly positive alignment modulated by terrain coupling, with anti-alignment as a rare geological exception.



5.3 What the learned uncertainty captures

The learned uncertainty field is not patch-constant: a within-patch variogram analysis places its spatial autocorrelation range at
 540 0.5–2.6 km, well below the 6.4 km patch, so σ does resolve sub-patch spatial structure. This within-patch structure, however, aligns only weakly with within-patch error ($r_{\text{within}} = 0.16\text{--}0.18$); the strong signal is cross-patch ($r_{\text{cross}} = 0.08\text{--}0.50$, locked evaluation protocol; Table 5). In other words, σ varies inside a patch, but that variation does not track which pixels within the patch are actually harder, the difficulty signal it reliably carries is at the patch scale. The operational consequence is that ranking survey blocks by DL σ captures $1.5\text{--}3\times$ the random error expectation at the 10% budget ($1.5\times$ on ridge 009, $2\text{--}3\times$
 545 on slope 002 and fan 010) and beats both kriging variance and a roughness heuristic at the 25% budget in all three regions (Section 4.4).

A natural question is what terrain or data feature drives this learned difficulty signal. A simple re-encoding of per-pixel roughness is unlikely. If σ merely mirrored the per-pixel roughness field (which varies primarily within patches, since the 0.5 km roughness window is far smaller than the 6.4 km patch), then σ 's information would be concentrated at the within-patch
 550 level. The spatial decomposition contradicts this: r_{within} is weak (0.16–0.18) while r_{cross} is strong (0.08–0.50, locked evaluation protocol; Table 5), meaning σ 's error-relevant information is at the patch scale; even though σ itself has sub-patch spatial structure (the variogram range above), that fine-scale variation does not track per-pixel error, which is the scale at which per-pixel roughness operates. A diagnostic regression using σ_{alea} (reported in the data supplement; headline results use σ_{tot}) that includes per-pixel roughness and distance-to-observation still assigns a significant coefficient to σ_{alea} ($\beta_{\sigma} = +10.76 \text{ m/SD}$, $p \approx$
 555 0) with incremental $\Delta R^2 = +0.226$ for $|e|$, confirming that σ carries predictive information beyond what these two covariates capture. The pooled σ_{alea} –roughness correlation is near zero, but we do not rely on this pooled statistic for the claim, since pooled correlations can mask within- or cross-patch structure as the cases in Section 5.1 demonstrate. What specific terrain features (curvature, slope breaks, local relief at scales between the 0.5 km roughness window and the 6.4 km patch) or data features (observation density patterns, prior interpolation artifacts) the learned σ encodes remains an open question. A working
 560 hypothesis, consistent with the patch-scale concentration of the signal, is that σ_{alea} captures meso-scale terrain complexity—the interaction of observation-track geometry with intermediate-wavelength bathymetric features (channel-levee systems, canyon-wall relief, ridge-axis morphology at $\sim 1\text{--}5$ km scales)—that the 0.5 km pixel-level roughness window is too narrow to register. Under this hypothesis, a patch containing a submarine canyon flanked by smooth plateaus would receive a high σ_{alea} not because every pixel is rough, but because the network learns that such juxtaposition of extreme and moderate terrain within a
 565 single context window produces large, spatially correlated residuals. Testing this hypothesis would require a controlled terrain-partition experiment that systematically varies patch composition (homogeneous vs. mixed terrain within a 6.4 km block), which lies beyond the scope of the present study.

We also find that σ is not a context-invariant geographic property. In patch overlap regions, the same pixel receives σ predictions correlated at only 0.40–0.71 across two independent forward passes, indicating a patch-context dependence whose origin
 570 we could not pin down: eight diagnostic experiments (Supplementary S3) ruled out GPU non-determinism, denormalization, mask confounding, input-difference effects, and receptive-field saturation, leaving the source unresolved. Because operational



prioritization uses a fixed patch partition, each location belongs to exactly one patch with a single patch-mean σ , this dependence does not introduce ambiguity into the block-ranking decision, though it does mean σ should not be read as a pure terrain property. Across all three regions, within-patch r_{within} has CV = 7–53% and contributes no detectable gain above pure patch-level ranking (Section 4.4); even in 010, where r_{pooled} and r_{cross} are tightly measured (CV = 3.0% and 4.1% respectively), the within-patch correlation has CV = 42%, consistent with the interpretation that within-patch DL UQ signal is intrinsically weak and poorly measured.

5.4 Operational implications for Seabed 2030

The operational question is whether data-driven uncertainty can guide resurvey prioritization more efficiently than conventional geostatistical proxies. The results support a qualified yes.

DL σ captures $1.5\text{--}3\times$ the random expectation at a 10% budget (block-level; Section 4.4, Fig. 6, Table 7) and never falls below random in any region under MC Dropout. On the continental slope—where prioritization matters most—it concentrates 56–59% of total error into the top 25% of blocks. The standard alternative is unreliable: kriging variance yields below-random block prioritization on the slope ($r_{\text{cross}} = -0.160$) and is uninformative on the fan. The signal is block-scale: σ resolves sub-patch structure (0.5–2.6 km variogram range) but its within-patch variation does not track pixel-level difficulty, so it should rank blocks rather than guide within-block measurement density. This is not limiting for Seabed 2030, which plans at block scale.

Generalization is partially supported. The blind001 prospective test produced a primary-endpoint HIT ($r_{\text{cross}} = 0.358$ in $[+0.35, +0.65]$; Section 4.5.2), but the secondary endpoints were borderline ($\text{capture}_{\text{block}}@10\%$) or invalidated by a lock artifact (Table 9 note), and evidence is limited to a single blind region with one architecture. A second blind test (008, $N = 206$) was inconclusive. Extending the protocol to additional terrains with adequately-powered prospective designs remains necessary.

5.5 Limitations

The empirical evidence rests on three in-domain regions and one prospective blind region—a set that spans contrasting geomorphic regimes but falls well short of global coverage. Trenches, polar shelves, abyssal plains, and densely surveyed areas are absent from the evaluation. Region 009, with only 70 evaluation patches and the largest seed sensitivity among the three in-domain regions, carries particularly high uncertainty and should be interpreted with corresponding caution.

On the method side, the learned-UQ results come from a single probabilistic U-Net backbone with MC Dropout. The deep-ensemble comparison (Section 4.5.1) supports robustness within this specific architecture and training recipe, but it does not speak to other backbones, other epistemic-uncertainty estimators, or alternative training protocols. The geostatistical baseline is ordinary kriging; the results should not be read as a verdict on geostatistical UQ methods in general. Kriging with external drift attenuates but does not reverse the slope anti-alignment (data supplement).

The evaluation design carries its own constraints. GEBCO 2026 is a higher-quality proxy, not direct ground truth. The single degradation pathway tested here (GEBCO 2021→2026) may not represent all operational update scenarios, and all analyses use



605 the GEBCO resolution and 6.4 km block scale. The calibration scalar requires labeled holdout data, which limits deployment to settings where recent validation surveys exist.

Finally, the three-axis protocol itself was developed from the in-domain diagnostic analysis and should be read as a protocol proposal and empirical investigation, not a fully confirmatory study. Four comparisons were pre-registered; the remaining protocol claims are post-hoc, protected by multi-seed replication and block-bootstrap intervals but not by prospective design.
 610 The blind001 ECE lock artifact (Table 9 note) serves as a concrete reminder that pre-registered thresholds must be verified against locked metric definitions before training begins.

5.6 The 008 blind prediction as methodological exhibit

008 is where the five pitfalls of Section 5.1 act together—a pre-registered blind test in which several definition-sensitivity mechanisms operate simultaneously on a single real prediction. The three evaluation rounds (data supplement) illustrate this
 615 concretely: three analysis choices that would be unremarkable in a conventional UQ study (different sample sizes, different r variants, global bias present or removed) produced three contradictory apparent conclusions. When all choices were aligned to a common protocol, the contradictions collapsed to a single finding: $N = 206$ insufficient.

The inconclusive verdict is overdetermined—small sample, analysis-choice sensitivity, and a pre-registered prediction neither confirmed nor falsified—and we report it without retroactive reinterpretation. That the protocol identifies its own boundary
 620 conditions on a held-out region, rather than only rationalizing in-domain successes, is its strongest validation.

6 Conclusions

This paper’s main contribution is TAP-Bathy v1.0, a scale-aware, operationally grounded evaluation protocol for spatial UQ in bathymetry. By jointly assessing calibration, within/cross-patch informativeness, and fixed-budget error capture, the protocol reveals failure modes hidden by single metrics: calibrated-but-uninformative uncertainty, pooled correlations that mask
 625 opposing spatial scales, and aggregation-driven ranking reversals.

The protocol’s necessity is demonstrated by concrete examples. Kriging on the Amazon Fan achieves $ECE = 4.1\%$ but $r_{\text{pooled}} = +0.16$ —well-calibrated yet spatially uninformative. On the Blake Escarpment, kriging variance yields below-random block prioritization ($r_{\text{cross}} = -0.160$). The 008 blind prediction produced three contradictory conclusions from identical data across three analysis rounds, driven entirely by metric-choice sensitivity.

630 Applied to GEBCO 2021→2026 temporal holdouts, learned uncertainty from ProbabilisticMCUNet provides block-scale (6.4 km) information: it captures $1.5\text{--}3\times$ the random error at 10% block-level budgets and remains above-random at 25%, with the cross-patch signal ($r_{\text{cross}} = 0.08\text{--}0.50$, locked evaluation protocol) concentrated at the scale of survey-block difficulty. Pre-registered prospective tests provide partial support: blind001 yielded a primary-endpoint HIT ($r_{\text{cross}} = 0.358$ in $[+0.35, +0.65]$), 008 was inconclusive ($N = 206$), and an eight-region test falsified the strong form of a terrain-coupling pre-
 635 diction. Four comparisons were pre-registered with SHA-256-locked endpoints. Evidence is limited to three in-domain terrains, one blind region, and one architecture; generalization to other terrains or model families remains open.



Appendix A: Exact decomposition of the pooled $|e|$ - σ correlation

Let $\sigma_{k,i}$ and $e_{k,i}$ denote the uncertainty estimate and absolute error at pixel i within patch k ($k = 1, \dots, K$; $i = 1, \dots, n_k$). The pooled Pearson correlation is

$$r_{\text{pooled}} = \frac{\text{Cov}(\sigma, e)}{\sqrt{\text{Var}(\sigma) \text{Var}(e)}} \quad (\text{A1})$$

where Cov and Var are computed over all $\sum_k n_k$ pixels.

Decompose each value into its patch mean and within-patch deviation:

$$\sigma_{k,i} = \bar{\sigma}_k + \delta\sigma_{k,i}, \quad e_{k,i} = \bar{e}_k + \delta e_{k,i}, \quad (\text{A2})$$

where $\bar{\sigma}_k = \frac{1}{n_k} \sum_i \sigma_{k,i}$ and $\sum_i \delta\sigma_{k,i} = 0$ within each patch (identically for e).

The total covariance decomposes as

$$\begin{aligned} \text{Cov}(\sigma, e) &= \frac{1}{N} \sum_k \sum_i (\bar{\sigma}_k + \delta\sigma_{k,i} - \bar{\sigma})(\bar{e}_k + \delta e_{k,i} - \bar{e}) \\ &= \underbrace{\frac{1}{N} \sum_k n_k (\bar{\sigma}_k - \bar{\sigma})(\bar{e}_k - \bar{e})}_{\text{Cov}_{\text{between}}} + \underbrace{\frac{1}{N} \sum_k \sum_i \delta\sigma_{k,i} \delta e_{k,i}}_{\text{Cov}_{\text{within}}} \end{aligned} \quad (\text{A3})$$

where $N = \sum_k n_k$ and $\bar{\sigma}, \bar{e}$ are the global means. The cross-terms $\sum_i \bar{\sigma}_k \delta e_{k,i}$ and $\sum_i \bar{e}_k \delta\sigma_{k,i}$ vanish because $\sum_i \delta\sigma_{k,i} = \sum_i \delta e_{k,i} = 0$ within each patch.

The total variance decomposes analogously:

$$\text{Var}(\sigma) = \text{Var}_{\text{between}}(\sigma) + \text{Var}_{\text{within}}(\sigma), \quad (\text{A4})$$

and identically for e . Define the between-patch variance fractions

$$\eta_\sigma^2 = \frac{\text{Var}_{\text{between}}(\sigma)}{\text{Var}(\sigma)}, \quad \eta_e^2 = \frac{\text{Var}_{\text{between}}(e)}{\text{Var}(e)}. \quad (\text{A5})$$

The cross-patch correlation r_{cross} and within-patch correlation r_{within} are

$$r_{\text{cross}} = \frac{\text{Cov}_{\text{between}}}{\sqrt{\text{Var}_{\text{between}}(\sigma) \text{Var}_{\text{between}}(e)}}, \quad r_{\text{within}} = \frac{\text{Cov}_{\text{within}}}{\sqrt{\text{Var}_{\text{within}}(\sigma) \text{Var}_{\text{within}}(e)}}. \quad (\text{A6})$$

Substituting the variance-decomposed forms into r_{pooled} :

$$\begin{aligned} r_{\text{pooled}} &= \frac{\text{Cov}_{\text{between}} + \text{Cov}_{\text{within}}}{\sqrt{\text{Var}(\sigma) \text{Var}(e)}} \\ &= r_{\text{cross}} \frac{\sqrt{\text{Var}_{\text{between}}(\sigma) \text{Var}_{\text{between}}(e)}}{\sqrt{\text{Var}(\sigma) \text{Var}(e)}} + r_{\text{within}} \frac{\sqrt{\text{Var}_{\text{within}}(\sigma) \text{Var}_{\text{within}}(e)}}{\sqrt{\text{Var}(\sigma) \text{Var}(e)}} \\ &= r_{\text{cross}} \cdot \sqrt{\eta_\sigma^2 \eta_e^2} + r_{\text{within}} \cdot \sqrt{(1 - \eta_\sigma^2)(1 - \eta_e^2)}. \end{aligned} \quad (\text{A7})$$



This identity is exact; it requires only that $\sum_i \delta\sigma_{k,i} = 0$ within each patch, which holds by construction when patch means are subtracted. All r quantities in this paper use unified definitions: r_{within} is the pooled within-patch correlation from concatenated per-patch demeaned residuals, r_{cross} is the n_k -weighted Pearson correlation of patch-mean $\bar{\sigma}_k$ against patch-mean $\bar{e}_k$, and η^2 is the n_k -weighted between-patch variance fraction. Under these definitions the identity holds to numerical precision ($\Delta < 10^{-15}$ for deterministic estimators, $\Delta < 10^{-7}$ for learned estimators), as verified in the accompanying code release.

Appendix B: Heavy-tail sensitivity

Standardized residuals $z_i = e_i / (s \cdot \sigma_i)$ are heavy-tailed (excess kurtosis: +741 on 002, +7 on 009, +17 on 010). Trimming the $\sim 0.1\%$ most extreme $|z_i|$ reduces the 002 kurtosis substantially but changes r_{pooled} by < 0.01 and does not alter any calibration conclusion. The heavy-tail pattern is driven by a small number of pixels where the temporal holdout contains genuinely new, large-amplitude depth information that neither the prior nor the learned correction captures; these are the very locations that a resurvey campaign would target.

Appendix C: Roughness window-size sensitivity

The roughness baseline σ_{rough} (Section 3.6) uses a 5×5 window ($\sim 0.5 \text{ km} \times 0.5 \text{ km}$ at the 100 m pixel spacing). We chose this window empirically: smaller windows (3×3 , $\sim 0.3 \text{ km}$) are dominated by grid-scale noise in flat terrain; larger windows (9×9 , $\sim 0.9 \text{ km}$) begin to smooth out sharp terrain boundaries (canyon walls, ridge axes, channel-levee transitions) that generate the largest residuals. Because locally smoothed standard deviation fields are highly correlated across adjacent window sizes, the operational prioritization conclusions are not sensitive to the specific choice of window; the 5×5 value is used throughout the paper.

Appendix D: Reliability diagrams

Code and data availability. All training, inference, and evaluation scripts, including the kriging baseline, the spatial decomposition, and the multi-seed prioritization pipeline, are archived in the TAP-Bathy v1.0 Zenodo release at <https://doi.org/10.5281/zenodo.23669317> (Cui, 2026) under the MIT license. Model checkpoints, reveal outputs, pre-registration documents with SHA-256 lock manifests, and the data supplement are archived at the same Zenodo record. Pinned software dependencies are provided as an `environment.yml` file. Software stack: PyTorch ≥ 2.0 , CUDA ≥ 11.8 , `gstools` ≥ 1.5 , `pykrige` ≥ 1.7 , `scipy` ≥ 1.10 , `numpy` ≥ 1.24 . Training requires an NVIDIA GPU with $\geq 8 \text{ GB}$ VRAM; CPU inference is supported. GEBCO 2021 and 2026 grids are publicly available at <https://www.gebco.net> (DOIs: 10.5285/c6612cbe-50b3-0cff-e053-6c86abc09f8f, 10.5285/4f68d5c7-45eb-f999-e063-7086abc036fa).

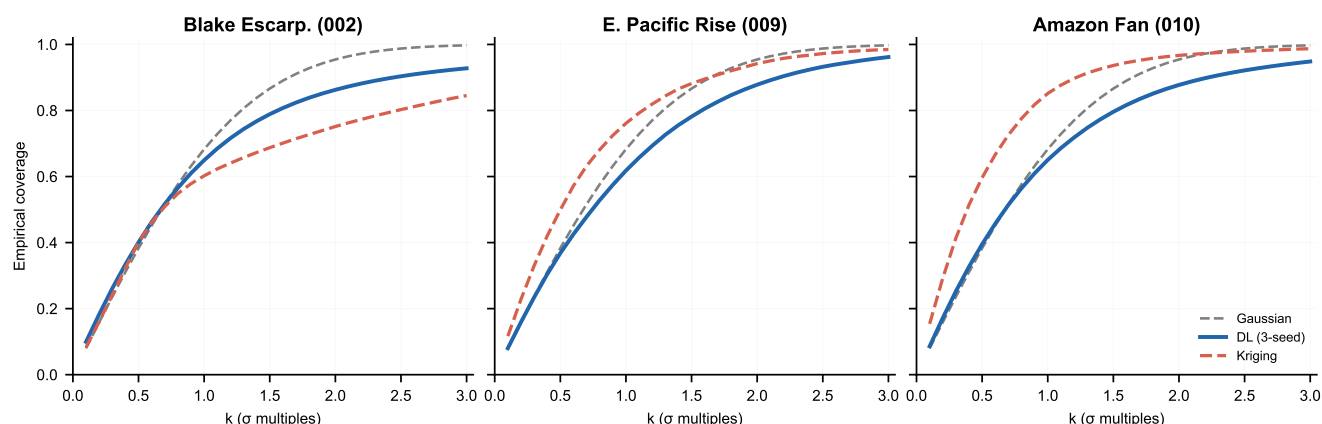


Figure D1. Reliability diagrams for all three in-domain regions. Blue solid line: DL mean coverage across 3 seeds with $\pm 1\sigma$ band; orange dashed: kriging (region-level); black dashed: Gaussian ideal. Annotations show empirical 1σ coverage and excess kurtosis of standardized residuals.

Author contributions. Conceptualization: Xinye Cui, Houpu Li; Methodology: Xinye Cui; Software: Xin Xiong; Validation: Xinye Cui, Xin Xiong, Shaofeng Bian; Formal analysis: Xinye Cui; Investigation: Xinye Cui, Houpu Li; Data curation: Xin Xiong; Writing—original draft preparation: Xinye Cui; Writing—review and editing: all authors; Visualization: Xin Xiong; Supervision: Houpu Li, Guojun Zhai; Project administration: Guojun Zhai; Funding acquisition: Houpu Li, Guojun Zhai. All co-authors have read and approved the submitted version.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank the GEBCO Compilation Group and the Nippon Foundation–GEBCO Seabed 2030 Project for producing and openly distributing the global bathymetric grids used in this study. GEBCO 2021 and 2026 grids were accessed via <https://www.gebco.net>. All model training and evaluation was performed on an NVIDIA RTX 4090 (24 GB GPU). This work was supported by the National Natural Science Foundation of China under Grant 42374050 and Grant 42430101.



008 Blind prediction: evaluation sensitivity

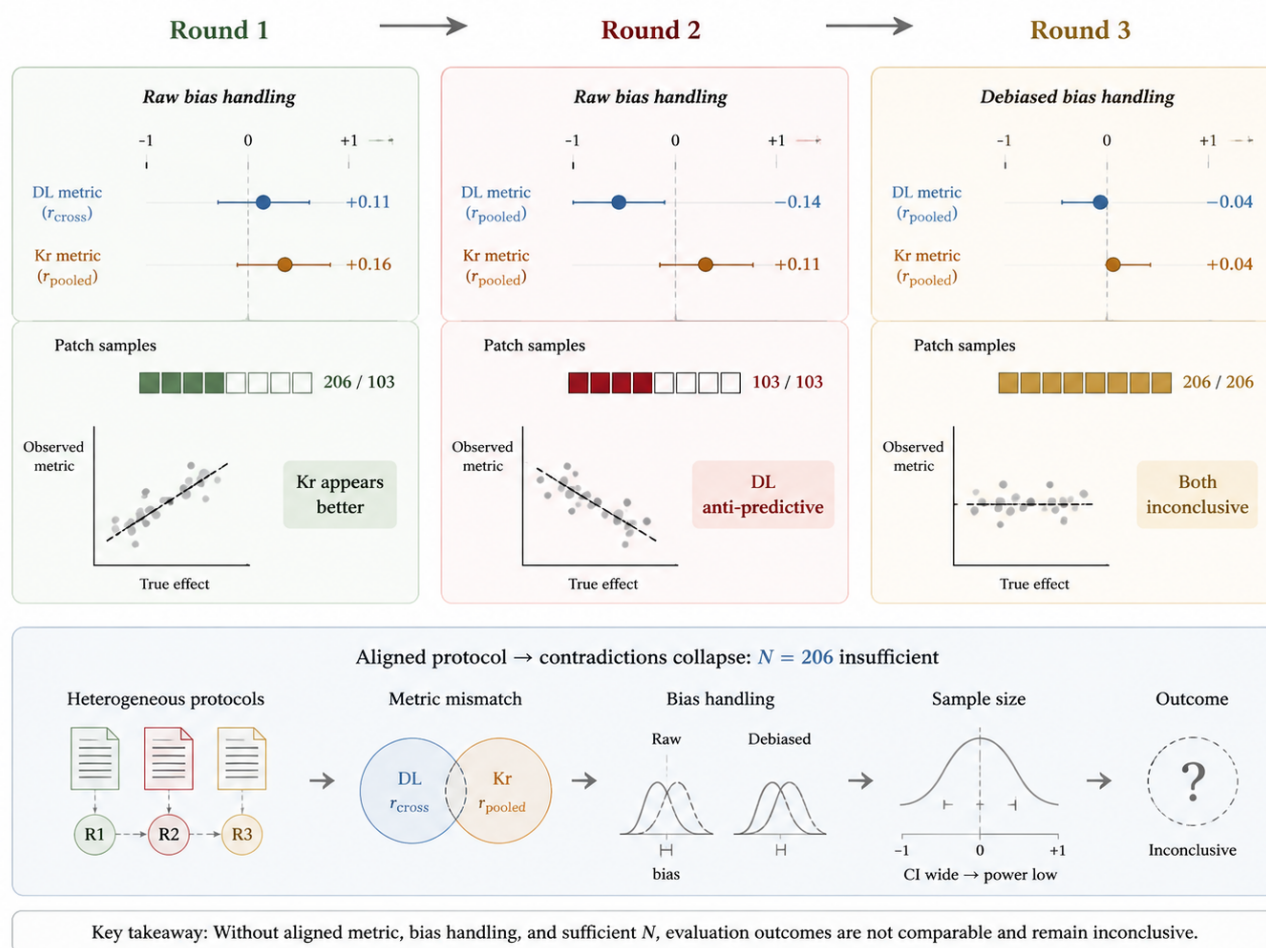


Figure D2. Three rounds of 008 blind-prediction evaluation, three contradictory conclusions from identical underlying data. Round 1 (different sample sizes and metrics) suggested kriging outperforms DL; Round 2 (matched sample size, different metric) reversed the verdict; Round 3 (matched protocol, debiased) collapsed to inconclusive. The three-axis protocol (Section 3.8) is designed to prevent the analysis-choice sensitivity illustrated here. Full round-by-round values are provided in the data supplement.



References

- Calder, B. R. and Mayer, L. A.: Automatic processing of high-rate, high-density multibeam echosounder data, *Geochemistry, Geophysics, Geosystems*, 4, <https://doi.org/10.1029/2002GC000486>, 2003.
- 695 Cressie, N.: *Statistics for Spatial Data*, Wiley, 1993.
- Cui, X.: TAP-Bathy v1.0: Code, Evaluation Outputs, and Manuscript Source for Evaluating Spatial Uncertainty in Bathymetry, <https://doi.org/10.5281/zenodo.23669317>, 2026.
- Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., and Hennig, P.: Laplace Redux – Effortless Bayesian Deep Learning, in: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, <https://doi.org/10.48550/arXiv.2106.14806>, 2021.
- 700 Gal, Y. and Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, in: *Proc. ICML*, pp. 1050–1059, <https://doi.org/10.48550/arXiv.1506.02142>, 2016.
- Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., and Zhu, X. X.: A Survey of Uncertainty in Deep Neural Networks, *Artificial Intelligence Review*, 56, 1513–1589, <https://doi.org/10.1007/s10462-023-10562-9>, 2023.
- 705 GEBCO Bathymetric Compilation Group 2026: The GEBCO_2026 Grid – a continuous terrain model for oceans and land at 15 arc-second intervals, <https://doi.org/10.5285/4f68d5c7-45eb-f999-e063-7086abc036fa>, 2026.
- GEBCO Compilation Group: GEBCO 2021 Grid, <https://doi.org/10.5285/c6612cbe-50b3-0cff-e053-6c86abc09f8f>, 2021.
- Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, *Journal of the American Statistical Association*, 102, 359–378, <https://doi.org/10.1198/016214506000001437>, 2007.
- 710 Goovaerts, P.: *Geostatistics for Natural Resources Evaluation*, Oxford University Press, 1997.
- Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q.: On Calibration of Modern Neural Networks, in: *Proc. ICML*, pp. 1321–1330, <https://doi.org/10.48550/arXiv.1706.04599>, 2017.
- Hüllermeier, E. and Waegeman, W.: Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods, *Machine Learning*, 110, 457–506, <https://doi.org/10.1007/s10994-021-05946-3>, 2021.
- 715 International Hydrographic Organization: IHO Standards for Hydrographic Surveys (S-44), Edition 6.1.0, Tech. rep., https://iho.int/uploads/user/pubs/standards/s-44/S-44_Edition_6.1.0.pdf, international Hydrographic Bureau, Monaco, 2022.
- Kendall, A. and Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, in: *Advances in Neural Information Processing Systems (NeurIPS)*, <https://doi.org/10.48550/arXiv.1703.04977>, 2017.
- Kuleshov, V., Fenner, N., and Ermon, S.: Accurate Uncertainties for Deep Learning Using Calibrated Regression, in: *Proceedings of the 35th International Conference on Machine Learning (ICML)*, vol. 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804, PMLR, <https://doi.org/10.48550/arXiv.1807.00263>, 2018.
- 720 Lakshminarayanan, B., Pritzel, A., and Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, in: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, <https://doi.org/10.48550/arXiv.1612.01474>, 2017.
- Lin, J., Chen, W.-M., Lin, Y., Cohn, J., Gan, C., and Han, S.: MCUNet: Tiny Deep Learning on IoT Devices, in: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, <https://doi.org/10.48550/arXiv.2007.10319>, 2020.
- 725 Liu, Y., Li, S., Yan, W., Suo, Y., Wang, L., Zou, Z., and Wu, L.: High-Resolution Digital Bathymetric Model (DBM) With Uncertainty Reconstructed by Deep Learning With Dropout, *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–17, <https://doi.org/10.1109/TGRS.2024.3396499>, 2024.



- Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, in: Proc. ICLR, <https://doi.org/10.48550/arXiv.1711.05101>, 2019.
- 730 Mayer, L., Jakobsson, M., Allen, G., Dorschel, B., Falconer, R., Ferrini, V., Lamarche, G., Snaith, H., and Weatherall, P.: The Nippon Foundation–GEBCO Seabed 2030 Project: The quest to see the world’s oceans completely mapped by 2030, *Geosciences*, 8, 63, <https://doi.org/10.3390/geosciences8020063>, 2018.
- Naeini, M. P., Cooper, G. F., and Hauskrecht, M.: Obtaining Well Calibrated Probabilities Using Bayesian Binning, in: Proc. AAAI, <https://doi.org/10.1609/aaai.v29i1.9602>, 2015.
- 735 Nix, D. A. and Weigend, A. S.: Estimating the mean and variance of the target probability distribution, in: Proc. IEEE Int. Conf. Neural Networks (ICNN), <https://doi.org/10.1109/ICNN.1994.374138>, 1994.
- Openshaw, S.: The Modifiable Areal Unit Problem, Tech. rep., 1984.
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J.: Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift, in: *Advances in Neural Information Processing Systems* (NeurIPS), vol. 32, <https://doi.org/10.48550/arXiv.1906.02530>, 2019.
- 740 Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guisera-Aroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure, *Ecography*, 40, 913–929, <https://doi.org/10.1111/ecog.02881>, 2017.
- Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Proc. MICCAI, pp. 234–241, https://doi.org/10.1007/978-3-319-24574-4_28, 2015.
- 745 Ryan, W. B. F., Carbotte, S. M., Coplan, J. O., O’Hara, S., Melkonian, A., Arko, R., Weissel, R. A., Ferrini, V., Goodwillie, A., Nitsche, F., Bonczkowski, J., and Zemsky, R.: Global Multi-Resolution Topography synthesis, *Geochem. Geophys. Geosyst.*, 10, <https://doi.org/10.1029/2008GC002332>, 2009.
- Simpson, E. H.: The Interpretation of Interaction in Contingency Tables, *J. R. Stat. Soc. B*, 13, 238–241, 1951.
- 750 Smith, W. H. F. and Sandwell, D. T.: Global sea floor topography from satellite altimetry and ship depth soundings, *Science*, 277, 1956–1962, <https://doi.org/10.1126/science.277.5334.1956>, 1997.
- Tozer, B., Sandwell, D. T., Smith, W. H. F., Olson, C., Beale, J. R., and Wessel, P.: Global bathymetry and topography at 15 arc sec: SRTM15+, *Earth and Space Science*, 6, 1847–1864, <https://doi.org/10.1029/2019EA000658>, 2019.
- Weatherall, P., Marks, K. M., Jakobsson, M., Schmitt, T., Tani, S., Arndt, J. E., Rovere, M., Chayes, D., Ferrini, V., and Wigley, R.: A new digital bathymetric model of the world’s oceans, *Earth and Space Science*, 2, 331–345, <https://doi.org/10.1002/2015EA000107>, 2015.
- 755 Wilson, A. G. and Izmailov, P.: Bayesian Deep Learning and a Probabilistic Perspective of Generalization, in: *Advances in Neural Information Processing Systems* (NeurIPS), vol. 33, <https://doi.org/10.48550/arXiv.2002.08791>, 2020.
- Xie, Y., Bore, N., and Folkesson, J.: Bathymetric Reconstruction From Sidescan Sonar With Deep Neural Networks, *IEEE Journal of Oceanic Engineering*, 48, 372–383, <https://doi.org/10.1109/JOE.2022.3220330>, 2023.