Uncertainty Estimation for Deep Reconstruction in Acuatic Disaster Scenarios with Autonomous Vehicles

Samuel Yanes Luis[1 [0000−0002−7796−3599]], Alejandro Casado Pérez[2 [0009−0007−5572−5913]], Alejandro Mendoza Barrionuevo[2 [0000−0002−6931−7124]], Dame Seck Diop[2 [0009−0004−0752−1041]], Sergio Toral Marín[2 [0000−0003−2612−0388]], and Daniel Gutiérrez Reina[2 [0000−0002−9017−7349]]

[1] *Corresponding author.* Department of Electronic Technology, University of Sevilla, Av. de la Reina Mercedes S/N, 41012, Sevilla, Spain. syanes@us.es
[2] Department of Electronic Engineering, University of Sevilla, Av. de los Descubrimientos S/N, 41005, Sevilla, Spain.

Abstract. Accurate reconstruction of environmental scalar fields from sparse onboard observations is essential for autonomous vehicles engaged in aquatic monitoring. Beyond point estimates, principled uncertainty quantification is critical for active sensing strategies such as Informative Path Planning, where epistemic uncertainty drives data collection decisions. This paper compares Gaussian Processes, Monte Carlo Dropout, Deep Ensembles, and Evidential Deep Learning for simultaneous scalar field reconstruction and uncertainty decomposition under three perceptual models representative of real sensor modalities. Results show that Evidential Deep Learning achieves the best reconstruction accuracy and uncertainty calibration across all sensor configurations at the lowest inference cost, while Gaussian Processes are fundamentally limited by their stationary kernel assumption and become intractable as observation density grows. These findings support Evidential Deep Learning as the preferred method for uncertainty-aware field reconstruction in real-time autonomous vehicle deployments.

Keywords: Uncertainty estimation · Informative Path Planning · Evidential Deep Learning · Autonomous Surface Vehicles · Environmental Monitoring

1 Introduction

Deploying autonomous vehicles for environmental monitoring is, at its core, a decision-making problem under uncertainty. A surface vessel or aerial drone tasked with mapping a marine hydrocarbon spill must continuously answer two intertwined questions: *what does the contaminant field look like right now?* and *where is that estimate least trustworthy?* The first question drives reconstruction;

the second drives action. Without a principled answer to both, any adaptive monitoring strategy is blind [6].

This coupling between estimation and uncertainty is especially acute for phenomena that are chaotic in nature. Oil-spill dynamics, algal blooms, or chemical plumes can be simulated via advection-diffusion equations, yet their real-world realisation is sensitive to unobserved boundary conditions, such that reconstruction algorithms must express calibrated ignorance over unvisited regions and distinguish between *epistemic uncertainty* arising from lack of data and *aleatoric uncertainty* from irreducible sensor noise [1]. Only the former is reducible by collecting additional measurements and therefore actionable for path planning. This challenge is amplified by platform heterogeneity: autonomous surface vehicles with contact sensors deliver scalar readings at isolated points, while aerial vehicles with multispectral or thermal cameras return spatially dense but indirect observations over wide areas [13]. Each observation model induces a different geometry of information and uncertainty structure, and a practical framework must accommodate all of them.

Gaussian processes (GP) have long been the standard tool for this setting precisely because uncertainty is native to their formulation: posterior variance is available in closed form and well-calibrated by construction [6]. Their limitations are equally well-known: cubic scaling with the number of observations and an inability to exploit the rich spatial structure encoded in image-like measurements. Deep learning addresses both shortcomings, as networks trained on physics-based synthetic data generalise across observation modalities and run in milliseconds at inference time, but standard deterministic architectures produce point estimates with no notion of confidence [5]. Recovering calibrated uncertainty from deep models has therefore become a central research question. Monte Carlo Dropout [4] approximates a Bayesian posterior by averaging stochastic forward passes; Deep Ensembles [5] derive uncertainty from the disagreement of independently trained networks; and Evidential Deep Learning [1] places a higher-order evidential prior over the likelihood, providing both uncertainty types in a single forward pass without sampling.

To the best of our knowledge, this is the first systematic comparison of these four families under heterogeneous observation models in an environmental-monitoring setting. This gap matters because the utility of uncertainty estimates in active monitoring depends not only on their average calibration, but on whether epistemic uncertainty specifically correlates with prediction error, so that it can serve as a reliable criterion for directing the vehicle towards maximally informative locations [3, 6]. We present a comparative study of GP regression, Monte Carlo Dropout, Deep Ensembles, and Evidential Deep Learning (EDL) for simultaneous field reconstruction and uncertainty estimation in autonomous environmental monitoring. Concretely: *(i)* we formulate the problem under a unified observation-model framework covering real-world sensor modalities typical of aerial and surface platforms; *(ii)* we evaluate reconstruction accuracy and uncertainty calibration, with explicit separation of aleatoric and epistemic components, on physics-based oil-spill simulations; and *(iii)* we analyse the correlation between

epistemic uncertainty and prediction error as an indicator of active-learning suitability.

The paper is organised as follows. Section 2 reviews related work. Section 3 formalises the problem. Section 4 describes estimation and observation models. Section 5 presents results, and Section 6 concludes.

2 Related Work

GPs have been widely adopted for scalar field estimation in autonomous vehicle missions due to their principled probabilistic formulation [10]. In the context of aquatic monitoring, GPs provide a natural mechanism for uncertainty quantification: the posterior variance directly reflects the confidence of the model at unobserved locations, making them a natural fit for IPP strategies that seek to reduce uncertainty along the planned path [11]. However, classical GP formulations rely on a fixed, stationary kernel, typically the Radial Basis Function (RBF) , which imposes isotropy and stationarity on the estimated field. This implies a homoscedastic noise assumption, where the aleatoric uncertainty is treated as a global hyperparameter fitted by maximizing the marginal log-likelihood, rather than being learned as a spatially varying quantity. As a result, GPs struggle to capture complex, non-stationary phenomena or sensor noise patterns that vary across the environment [12].

Deep generative models, particularly Variational Autoencoders (VAE), have been explored for reconstructing environmental phenomena such as oil spills and floating debris [2]. Trained on physics-based simulators, these models capture complex non-stationary spatial patterns that GPs struggle to represent. However, predictive uncertainty is not a native output of these architectures, which restricts their use in active sensing frameworks. This motivates equipping similarly flexible deep architectures with principled uncertainty estimation, as pursued in this work.

Several methods have been proposed to equip deep neural networks with uncertainty estimation capabilities. Monte Carlo Dropout [4] interprets dropout at inference time as approximate Bayesian inference, yielding a simple yet effective way to obtain epistemic uncertainty estimates. Deep Ensembles [5] train multiple independent networks and exploit their predictive disagreement as a proxy for uncertainty, achieving strong empirical calibration at the cost of increased computation. Evidential Deep Learning [1] takes a different approach by placing a conjugate prior over the output distribution, enabling closed-form decomposition of epistemic and aleatoric uncertainty from a single forward pass.

Despite their demonstrated effectiveness, these methods have been evaluated primarily in the context of image classification tasks, with comparatively little attention paid to regression problems. More importantly, their behavior under the perceptual models characteristic of autonomous vehicles , including cone-shaped fields of view, point-wise sensors, and nadir-looking cameras, has not been systematically studied. The feasibility and calibration of these methods under such observation models, and their suitability for integration into IPP pipelines

for Unmanned Surface Vehicles (USV) and Unmanned Aerial Vehicles (UAV), remains an open question that this work aims to address.

3 Problem Statement

3.1 Scalar Field Reconstruction

Let $\mathcal{X} \subset \mathbb{R}^2$ denote the spatial domain of interest, and let $f : \mathcal{X} \to \mathbb{R}$ be an unknown scalar field representing the environmental phenomenon to be monitored (e.g., pollutant concentration or debris density). We assume that f is static or quasi-static over the duration of the mission.

A vehicle collects a set of N observations $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathcal{X}$ denotes the measurement location and $y_i \in \mathbb{R}$ is the corresponding observation. The observation is related to the true field through a perceptual model $g : \mathcal{X} \to \mathcal{Y}$, which encodes the sensor's measurement process:

$$y_i = g(f, \mathbf{x}_i) + \varepsilon_i, \tag{1}$$

where ε_i represents sensor noise. The perceptual model g depends on the sensor modality: a point-wise measurement (e.g., a contact water quality sensor), a cone-shaped field of view (e.g., a lateral camera for debris detection), or a nadir-looking footprint (e.g., an aerial camera). In all cases, g defines how the true field is integrated or aggregated into a scalar observation at each vehicle pose.

Rather than seeking a point estimate $\hat{f}$, the primary objective is to learn a predictive distribution $p(f(\mathbf{x}) \mid \mathbf{x}, \mathcal{D})$ for each location $\mathbf{x} \in \mathcal{X}$. We model this distribution as a Gaussian,

$$p\big(f(\mathbf{x}) \mid \mathbf{x}, \mathcal{D}\big) = \mathcal{N}\Big(\hat{f}(\mathbf{x}),\, \hat{\sigma}^2(\mathbf{x})\Big)\,, \tag{2}$$

where the mean $\hat{f}(\mathbf{x})$ serves as the point reconstruction and $\hat{\sigma}^2(\mathbf{x})$ captures the total predictive uncertainty. This Gaussian assumption is standard in heteroscedastic regression [7] and is well-suited to the present setting: the sensor noise ε_i in Eq. 1 is physically reasonable to model as Gaussian, and the epistemic component reflects parameter uncertainty that, under ensemble averaging, also converges to a Gaussian marginal.

Under this assumption, the model is trained by minimising the Negative Log-Likelihood (NLL) over the collected observations:

$$\mathcal{L}_{\mathrm{NLL}} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{\big(y_i - \hat{f}(\mathbf{x}_i)\big)^2}{2\,\hat{\sigma}^2(\mathbf{x}_i)} + \frac{1}{2} \log \hat{\sigma}^2(\mathbf{x}_i) \right]. \tag{3}$$

This objective jointly incentivises accurate reconstruction and well-calibrated predictive variance: overestimating $\hat{\sigma}^2$ is penalised by the log term, while underestimating it inflates the squared residual. This contrasts with reconstruction-only approaches trained with mean squared error (MSE), where any inferred variance is disconnected from the training signal and collapses to near-zero values, rendering uncertainty estimates unreliable [2].

3.2 Uncertainty Decomposition

Beyond the total predictive variance, we seek to characterise the separate sources of uncertainty associated with $\hat{f}$. Following the standard decomposition [1], the total predictive uncertainty at a location $\mathbf{x}$ can be written as:

$$\sigma^2_{\text{total}}(\mathbf{x}) = \underbrace{\sigma^2_{\text{ale}}(\mathbf{x})}_{\text{aleatoric}} + \underbrace{\sigma^2_{\text{epi}}(\mathbf{x})}_{\text{epistemic}}, \tag{4}$$

where aleatoric uncertainty σ^2_{ale} reflects irreducible noise inherent to the measurement process, and epistemic uncertainty σ^2_{epi} reflects the model's lack of knowledge due to limited observations, which can in principle be reduced by collecting additional data. How $\hat{\sigma}^2(\mathbf{x})$ is decomposed into these two components depends on the estimation method and is detailed in Section 4.

This decomposition is central to the active sensing use case: only epistemic uncertainty is actionable for path planning, as aleatoric uncertainty remains invariant to further sampling. A method that conflates both sources provides a misleading signal for guiding data collection.

4 Methods

4.1 Reconstruction Backbone: U-Net Architecture

All deep learning methods share a common U-Net backbone [9] that takes as input two spatial matrices of size $H \times W$: an observation mask $\mathbf{M} \in \{0,1\}^{H\times W}$ and a value matrix $\mathbf{V} \in \mathbb{R}^{H\times W}$, concatenated along the channel dimension to form $\mathbf{X} = [\mathbf{M}, \mathbf{V}] \in \mathbb{R}^{2\times H\times W}$. The encoder comprises 4 blocks of two convolutional layers with batch normalization and ReLU activations followed by max-pooling. The decoder mirrors this structure via transposed convolutions with skip connections, and a final 1×1 convolution produces the reconstructed field $\hat{\mathbf{F}} \in \mathbb{R}^{H\times W}$.

4.2 Monte Carlo Dropout

Monte Carlo Dropout [4] inserts dropout layers with rate p after each convolutional block. At inference, dropout remains active and T stochastic forward passes yield predictions $\{\hat{\mathbf{F}}^{(t)}\}_{t=1}^{T}$, from which epistemic uncertainty and the predictive mean are estimated as:

$$\hat{f}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} \hat{f}^{(t)}(\mathbf{x}), \qquad \sigma^2_{\text{epi}}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{f}^{(t)}(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2. \tag{5}$$

Aleatoric uncertainty is obtained from an auxiliary log-variance head trained with the heteroscedastic NLL loss:

$$\mathcal{L}_{\text{NLL}} = \frac{1}{|\mathcal{X}|}\sum_{\mathbf{x}}\left[\frac{\left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2}{2\hat{\sigma}^2_{\text{ale}}(\mathbf{x})} + \frac{1}{2}\log\hat{\sigma}^2_{\text{ale}}(\mathbf{x})\right], \tag{6}$$

4.3 Deep Ensembles

Deep Ensembles [5] train M independent backbone instances from different random seeds, each with an auxiliary log-variance head optimized with Eq. (6). At inference, epistemic uncertainty is derived from inter-member disagreement and aleatoric uncertainty is averaged across members:

$$\hat{f}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \hat{f}^{(m)}(\mathbf{x}) \tag{7}$$

$$\sigma^2_{\text{epi}}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{f}^{(m)}(\mathbf{x}) - \hat{f}(\mathbf{x}) \right)^2, \quad \hat{\sigma}^2_{\text{ale}}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} \hat{\sigma}^{2(m)}_{\text{ale}}(\mathbf{x}) \tag{8}$$

4.4 Evidential Deep Learning

Evidential Deep Learning [1] places a Normal-Inverse-Gamma prior over the likelihood, modifying the backbone to output four quantities $(\gamma, \nu, \alpha, \beta)$ per spatial location, with constraints $\nu > 0$, $\alpha > 1$, $\beta > 0$. Uncertainty estimates are available in closed form:

$$\hat{f}(\mathbf{x}) = \gamma, \quad \sigma^2_{\text{epi}}(\mathbf{x}) = \frac{\beta}{\nu(\alpha - 1)}, \quad \sigma^2_{\text{ale}}(\mathbf{x}) = \frac{\beta}{\alpha - 1}, \tag{9}$$

where we omit the explicit dependence on $\mathbf{x}$ for brevity. Training minimizes the NIG log-likelihood plus an evidence regularizer that penalizes high evidence on incorrect predictions:

$$\mathcal{L}_{\text{EDL}} = \sum_{\mathbf{x}} \left[\mathcal{L}_{\text{NIG}}(\mathbf{x}) + \lambda \cdot |f(\mathbf{x}) - \gamma(\mathbf{x})| \cdot (2\nu(\mathbf{x}) + \alpha(\mathbf{x})) \right], \tag{10}$$

with $\lambda = 10^{-3}$ and gradient clipping at max norm 1.0.

4.5 Gaussian Process Baseline

As a probabilistic baseline, we employ a GP [8] with composite kernel:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2} \right) + \sigma_n^2 \, \delta(\mathbf{x}, \mathbf{x}'), \tag{11}$$

with hyperparameters $\{\sigma_f^2, \ell, \sigma_n^2\}$ optimized by maximizing the marginal log-likelihood per test sample. The posterior variance captures epistemic uncertainty through contraction at observed locations, while σ_n^2 provides a spatially uniform aleatoric estimate — a homoscedastic limitation that, as shown in Section 5, proves critical for fields with heterogeneous spatial regularity.

4.6 Perceptual Models

Three perceptual models define the observation operator g in Eq. (1), each representing a real sensor modality used in aquatic monitoring (see Fig. 1).

Point-wise sensor. A contact sensor returns the field value at a single location: $y_i = f(\mathbf{x}_i) + \varepsilon_i$.

Conical field of view (FOV). A lateral USV camera integrates field values over a cone-shaped footprint:

$$y_i = \frac{1}{|\mathcal{R}_i|} \sum_{\mathbf{x} \in \mathcal{R}_i} f(\mathbf{x}) + \varepsilon_i, \tag{12}$$

where $\mathcal{R}_i$ is the set of cells within the conical footprint of half-angle ϕ and range $d_{\max}$ centered on the vehicle heading θ_i.

Nadir-looking camera. A downward-facing UAV camera returns an $s \times s$ image, contributing s^2 spatially distributed observations per step:

$$y_i^{(p,q)} = f\left(\mathbf{x}_i^{(p,q)}\right) + \varepsilon_i^{(p,q)}, \qquad p, q \in \{1, \ldots, s\}. \tag{13}$$

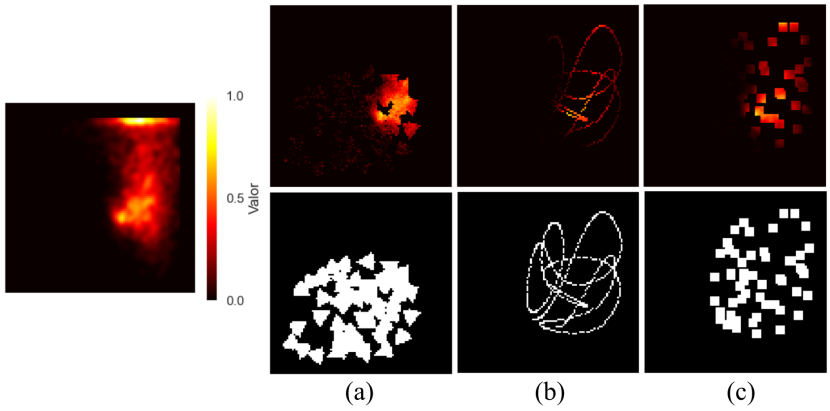


Fig. 1: Example field and observation footprints for the three perceptual models for a particular scenario (left): (a) Conic FOV, (b) Point-wise sensor, (c) Nadir-looking camera. Top row shows the measured values, while the bottom row shows the corresponding observation masks. Both matrices are fed into the U-Net.

5 Results

5.1 Experimental Setup

Three synthetic datasets of 2000 oil spill simulations each are generated on a 100×100 grid using a physics-based particle simulator from [2], differing only in

their observation model. Base simulation parameters are fixed across datasets: number of particles, spill radius, wind and tidal speed 1.0, etc. although every simulation generates a distinct field realization. The measurement noise ε_i is fixed for Nadir and Point-wise models as $\varepsilon_i \sim \mathcal{N}(0, 0.1^2)$. For the Conical model, the noise follows a linear relationship with the pixel distance, with $\sigma_n \in [0.01, 0.08]$. The Nadir dataset simulates a UAV with $s = 5$ and 10 m altitude, while the Conical dataset simulates a USV with $\phi = 30°$ and $d_{\max} = 20$ m. Every ground truth field is normalized to $[0, 1]$ with respect to the sensor range. For every field, a random smooth trajectory of random length between 50 and 200 steps is generated, always within the maximum trajectory length of each vehicle. Every trajectory is converted into a sequence of observations according to the corresponding perceptual model. The dataset is split 80%/20%, yielding 400 held-out test samples per dataset.

All deep learning models share the U-Net backbone with channel progression $32 \to 512$ and a bottleneck of 1024 channels, trained for 50 epochs with batch size 16 using Adam. Monte Carlo Dropout uses $p = 0.2$ and $T = 30$ inference passes; the Ensemble comprises $M = 5$ members with learning rate 3×10^{-4}; EDL uses $\lambda = 10^{-3}$, gradient clipping at max norm 1.0, and learning rate 5×10^{-4}. The GP baseline requires no offline training and is fitted per test sample via L-BFGS. All experiments were conducted on a MacBook Pro with Apple M3 Pro and 18 GB unified RAM. Full implementation details and hyperparameters are available at https://gitlab.ratatosk.cc/syanes/uncertainty-estimation-caepia-2026.

5.2 Evaluation Metrics

Reconstruction accuracy is assessed using the Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{X}_{\text{test}}|} \sum_{\mathbf{x} \in \mathcal{X}_{\text{test}}} \left(f(\mathbf{x}) - \hat{f}(\mathbf{x}) \right)^2}. \tag{14}$$

Normalized Uncertainty calibration is evaluated using the Uncertainty Calibration Error (UCE), which measures the consistency between predicted minmax-normalized uncertainty $\hat{\sigma}_{\text{total}}$ magnitudes and normalized empirical error $\hat{e}(\mathbf{x})$ variance. The normalized uncertainty range is discretized into B equally spaced bins, and for each bin U_b the normalized mean predicted uncertainty is compared against $\hat{e}(\mathbf{x})$:

$$\text{UCE} = \sum_{b=1}^{B} \frac{|U_b|}{|\mathcal{X}_{\text{test}}|} \left| \frac{1}{|U_b|} \sum_{\mathbf{x} \in U_b} \hat{\sigma}^2_{\text{total}}(\mathbf{x}) - \frac{1}{|U_b|} \sum_{\mathbf{x} \in U_b} \hat{e}(\mathbf{x})^2 \right|, \tag{15}$$

The normalization imposes a common scale across methods and datasets, enabling direct comparison of the uncertainty as a surrogate for error magnitude. A lower UCE indicates better calibration, with 0 representing perfect alignment between predicted uncertainty and empirical error.

5.3 Experiment results

Table 1 reports the RMSE across the 400 test samples for each method and perceptual model. EDL achieves the best reconstruction accuracy across all three sensor models by a considerable margin, with a mean RMSE roughly half that of the Ensemble and three to four times lower than GP. The Deep Ensemble ranks second, while MC Dropout and GP trail significantly. Notably, reconstruction performance is largely consistent across perceptual models for all deep learning methods, suggesting that the U-Net backbone adapts well to different observation footprints. The GP baseline shows the highest variance, particularly under the Nadir model, reflecting its sensitivity to observation sparsity.

Table 1: RMSE (mean ± std) across 400 test samples per dataset. Best result per column in **bold**.

Method	Conic FOV	Nadir	Pointwise
EDL	$\mathbf{0.0274 \pm 0.0234}$	$\mathbf{0.0293 \pm 0.0209}$	$\mathbf{0.0272 \pm 0.0204}$
Ensemble	0.0506 ± 0.0238	0.0570 ± 0.0200	0.0586 ± 0.0188
MC Dropout	0.0878 ± 0.0222	0.0915 ± 0.0160	0.0992 ± 0.0167
GP	0.1146 ± 0.0348	0.1193 ± 0.0648	0.1146 ± 0.0321

Table 2 reports the UCE for each method and perceptual model. EDL achieves the lowest UCE across all sensor configurations, indicating that its predicted uncertainty magnitudes are most consistent with the actual reconstruction errors. The Ensemble ranks second with stable and low UCE values. MC Dropout shows moderate calibration, while the GP baseline presents the highest UCE and largest variance across all methods. This behavior is a direct consequence of the isotropic RBF kernel assumption: oil spill fields exhibit spatially heterogeneous Lipschitz continuity — sharp gradients at spill boundaries coexist with smooth interior regions — which a stationary, isotropic kernel cannot capture. As a result, the GP is forced to commit to a single global length-scale ℓ that is simultaneously too smooth in high-gradient regions and too rough in homogeneous ones. Combined with the homoscedastic noise term σ_n^2, which imposes a spatially uniform aleatoric estimate regardless of local observation density, the GP loses the key black-box advantage of principled uncertainty quantification precisely where it is most needed.

Table 2: UCE (mean ± std) across 400 test samples per dataset. Best result per column in **bold**.

Method	Conic FOV	Nadir	Pointwise
EDL	$\mathbf{0.0018 \pm 0.0031}$	$\mathbf{0.0016 \pm 0.0025}$	$\mathbf{0.0013 \pm 0.0022}$
Ensemble	0.0026 ± 0.0016	0.0040 ± 0.0016	0.0040 ± 0.0017
MC Dropout	0.0084 ± 0.0026	0.0063 ± 0.0018	0.0110 ± 0.0029
GP	0.0110 ± 0.0158	0.0241 ± 0.2443	0.0096 ± 0.0093

Figure 2 shows the calibration curves for each method across the three perceptual models. EDL achieves the best calibration across all sensor models (deviation

≈ 0.038–0.040), followed closely by the Ensemble (deviation ≈ 0.043–0.072). MC Dropout presents moderate calibration errors with notable improvement under the Pointwise model. The GP baseline shows severe miscalibration across all configurations (deviation > 0.52), confirming that the combination of a stationary kernel and a global homoscedastic noise term is fundamentally misspecified for fields with non-homogeneous spatial regularity.

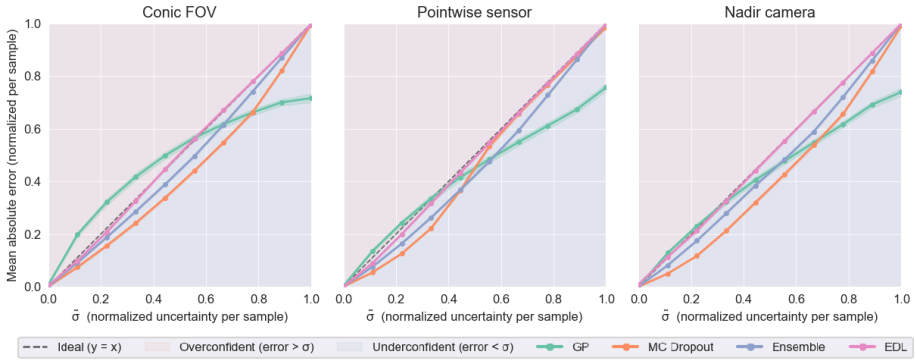


Fig. 2: Calibration curves for each method and perceptual model. The x-axis represents the normalized predicted uncertainty bins, while the y-axis shows the corresponding normalized empirical error. The closer the curve is to the diagonal, the better the calibration.

Table 3 reports mean inference time per sample. Beyond raw speed, inference time has direct implications for real-time onboard deployment. The GP baseline requires per-sample kernel fitting via L-BFGS, making its computational cost cubic in the number of observations: as the vehicle collects more data during a mission, GP inference becomes increasingly intractable, ranging from 81 ms under the sparse Pointwise model to over 2200 ms under Conic FOV, where the wider sensor footprint yields denser observation sets. This scaling behavior fundamentally limits the suitability of GPs for online deployment on resource-constrained platforms such as USVs or UAVs. Among deep learning methods, EDL is the fastest at $\sim$7 ms per sample, as it requires only a single deterministic forward pass through the network.

Table 3: Mean inference time per sample (ms). Best result per column in **bold**.

Method	**Conic FOV**	**Nadir**	**Pointwise**
EDL	$\mathbf{7.1 \pm 1.5}$	$\mathbf{7.2 \pm 1.4}$	$\mathbf{7.4 \pm 1.3}$
Ensemble	20.3 ± 2.5	20.6 ± 0.5	20.9 ± 0.6
MC Dropout	202.4 ± 12.4	202.0 ± 2.7	207.1 ± 8.3
GP	2247.1 ± 0.0	531.8 ± 0.0	81.2 ± 0.0

Taken together, the results highlight two methods as particularly well-suited for uncertainty-aware field reconstruction in autonomous vehicle deployments: Deep Ensembles and EDL. Both achieve strong reconstruction accuracy and

well-calibrated uncertainty estimates across all perceptual models, clearly outperforming MC Dropout and the GP baseline.

Between the two, EDL presents a compelling additional advantage: it yields calibration performance comparable to the Ensemble at a fraction of the inference cost, while intrinsically modeling the predictive distribution as a *multivariate Gaussian* through the Normal-Inverse-Gamma formulation. This means that epistemic and aleatoric uncertainty are not post-hoc approximations derived from sampling, but closed-form quantities that emerge directly from the evidential prior. This makes EDL particularly attractive for integration into probabilistic IPP frameworks where the uncertainty map must be queried repeatedly and efficiently during online mission execution.

6 Conclusion

This paper has presented a comparative evaluation of four uncertainty estimation methods for scalar field reconstruction under heterogeneous observation models, motivated by the requirements of uncertainty-aware Informative Path Planning with autonomous surface and acuatic vehicles.

Evidential Deep Learning and Deep Ensembles consistently outperform Monte Carlo Dropout and the Gaussian Process baseline across all perceptual models. Evidential Deep Learning achieves the lowest RMSE and UCE with a single deterministic forward pass, while Deep Ensembles offer competitive calibration at a modest increase in inference cost. Crucially, both methods produce uncertainty estimates that are well aligned with the actual reconstruction error, making predicted uncertainty a reliable surrogate for the true error surface. This calibration property is directly exploitable in path planning strategies that optimize reconstruction quality without access to ground truth, and opens the door to online model selection criteria grounded in uncertainty rather than held-out performance.

The Gaussian Process baseline, despite its principled formulation, suffers from severe miscalibration across all configurations due to its stationary isotropic kernel, which cannot adapt to fields with spatially heterogeneous regularity. Its cubic scaling with observation count further limits its suitability for real-time onboard deployment.

Future work will integrate these methods into a closed-loop Informative Path Planning system to evaluate whether well-calibrated epistemic uncertainty translates into improved path efficiency under budget-constrained missions.

Acknowledgments. Proyecto PID2024-158365OB-C21 financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Amini, A., Schwarting, W., Soleimany, A., Rus, D.: Deep evidential regression. In: Advances in Neural Information Processing Systems 33 (NeurIPS) (2020)
2. Casado-Pérez, A., Yanes, S., Toral, S.L., Perales-Esteve, M., Gutiérrez-Reina, D.: Variational autoencoder for the prediction of oil contamination temporal evolution in water environments. Sensors **25**(6) (2025). https://doi.org/10.3390/s25061654
3. Chen, W., Khardon, R., Liu, L.: Robotic active information gathering for spatial field reconstruction with rapidly-exploring random trees and online learning of Gaussian processes. Sensors **19**(5), 1016 (2019)
4. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML). pp. 1050–1059 (2016)
5. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems 30 (NeurIPS). pp. 6405–6416 (2017)
6. Mansfield, S., Montazeri, A.: A survey on autonomous environmental monitoring approaches: towards unifying active sensing and reinforcement learning. Frontiers in Robotics and AI **11**, 1336612 (2024)
7. Nix, D.A., Weigend, A.S.: Estimating the mean and variance of the target probability distribution. In: Proceedings of the 1994 IEEE International Conference on Neural Networks (ICNN). vol. 1, pp. 55–60. IEEE (1994)
8. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. Adaptive Computation and Machine Learning, The MIT Press (2006). https://doi.org/10.7551/mitpress/3206.001.0001, https://doi.org/10.7551/mitpress/3206.001.0001
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015), https://arxiv.org/abs/1505.04597
10. Samaniego, F.P., Reina, D.G., Marín, S.L.T., Arzamendia, M., Gregor, D.O.: A bayesian optimization approach for water resources monitoring through an autonomous surface vehicle: The ypacarai lake case study. IEEE Access **9**, 9163–9179 (2021)
11. Yanes Luis, S., Gutiérrez-Reina, D., Toral Marín, S.: Censored deep reinforcement patrolling with information criterion for monitoring large water resources using autonomous surface vehicles. Applied Soft Computing **132**, 109874 (2023). https://doi.org/10.1016/j.asoc.2022.109874
12. Yanes Luis, S., Shutin, D., Marchal Gómez, J., Gutiérrez Reina, D., Toral Marín, S.: Deep reinforcement multiagent learning framework for information gathering with local gaussian processes for water monitoring. Advanced Intelligent Systems **6**(8), 2300850 (2024). https://doi.org/10.1002/aisy.202300850
13. Zakaria, N.A., et al.: UAV-based remote sensing for the petroleum industry and environmental monitoring: State-of-the-art and perspectives. Journal of Petroleum Science and Engineering **208**, 109633 (2022)