Uncertainty Quantification for Regression: A Unified Framework based on Kernel Scores

Christopher Bülte[1,2] **Yusuf Sale**[1,2] **Gitta Kutyniok**[1,2,3,4] **Eyke Hüllermeier**[1,2,5]

[1] LMU Munich
[2]Munich Center for Machine Learning (MCML)
[3]DLR-German Aerospace Center
[4]University of Tromsø
[5]German Research Center for Artificial Intelligence (DFKI, DSA)

Abstract

Regression tasks, notably in safety-critical domains, require reliable uncertainty quantification, yet the literature remains largely classification-focused. To address this, we introduce a family of measures for total, aleatoric, and epistemic uncertainty in multivariate regression based on strictly proper kernel scores. The framework provides a principled recipe for designing new uncertainty measures whose behavior, such as tail sensitivity or out-of-distribution responsiveness, is governed by the choice of the underlying kernel, while also encompassing existing measures under a joint analysis. We prove explicit correspondences between properties of the kernel and behavior of resulting uncertainty measures, yielding concrete design guidelines for practitioners. Extensive experiments across structured regression tasks, including spatial and functional domains, demonstrate effectiveness on downstream tasks such as out-of-distribution detection and active learning, and reveal that different kernel choices lead to distinct trade-offs, offering practitioners guidance for task-specific selection.

1 INTRODUCTION

Predictive models now drive decision-making in safety-critical domains such as weather forecasting [Price et al., 2025, Alet et al., 2025], autonomous driving [Michelmore et al., 2018] or healthcare [Löhr et al., 2024, Edupuganti et al., 2020]; tasks where careful analysis of the model predictions and accurate uncertainty quantification are indispensable. Many studies have analyzed different approaches to quantify predictive uncertainty, often distinguishing between different sources of uncertainty. In particular, one usually considers two sources of uncertainty: *aleatoric uncertainty* and *epistemic uncertainty* [Hüllermeier and Waegeman, 2021]. Broadly speaking, aleatoric uncertainty describes the inherent randomness in the data-generating process, for example, due to measurement errors and, as it describes variability that is independent of the amount of data, is often referred to as *irreducible* uncertainty. Epistemic uncertainty, on the other hand, arises from a lack of knowledge about the data-generating process and can be reduced by improving the model or acquiring more data; therefore, it is also referred to as *reducible* uncertainty.

While aleatoric uncertainty is well captured in predictive models, epistemic uncertainty is more difficult to represent and requires higher-order formalisms, such as second-order distributions (distributions of distributions), which is referred to as *uncertainty representation* [Hüllermeier and Waegeman, 2021]. Given such a representation, the key question is how to measure or quantify the total, aleatoric, and epistemic uncertainty (*uncertainty quantification*). While the representation mainly determines predictive performance, the choice of uncertainty measure plays a vital role in decision making and can have an additional impact on the performance of downstream tasks, with numerous works developing and analyzing new measures for uncertainty quantification [Sale et al., 2024, Malinin and Gales, 2021, Kotelevskii et al., 2022, Berry and Meger, 2025]. In addition, recent work focuses on steps towards more unified approaches that incorporate many existing measures and give guidance on how to construct new ones [Hofman et al., 2024b, Kotelevskii et al., 2025]. However, research has focused either on uncertainty quantification in classification or on parametric univariate regression tasks, neglecting the increasing amount of structured domains where generative models and other nonparametric methods show great performance [Alet et al., 2025, Ke et al., 2024].

In *regression* tasks, a practitioner is generally interested in predictive uncertainty, which describes the uncertainty of the target $\boldsymbol{y} \in \mathcal{Y}$ given some covariates $\boldsymbol{x} \in \mathcal{X}$. While the notions of total, aleatoric, and epistemic uncertainty remain the same [Hüllermeier and Waegeman, 2021], the corresponding uncertainty measures fundamentally differ from

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:839–879.

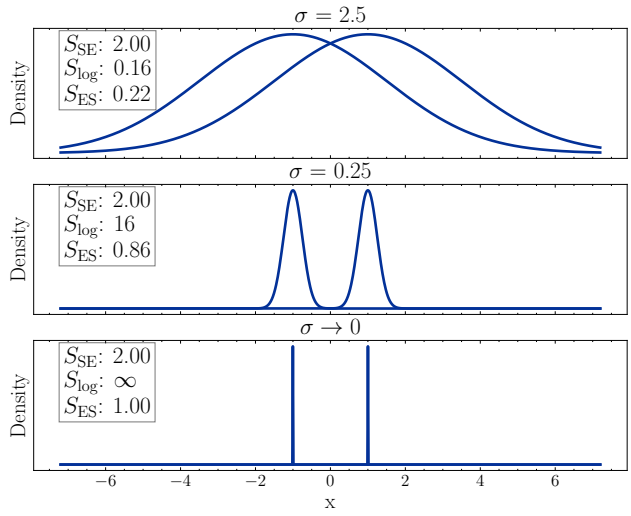

Figure 1: Illustration of epistemic uncertainty for a two-member Gaussian ensemble with shared variance. As σ shrinks, the variance-based measure (S_{SE}) stays constant, the entropy-based measure ($S_{\log}$) diverges, while our proposed energy-score-based measure (S_{ES}) converges to half the Euclidean distance between component means.

the classification case. Unlike classification, where the label space is discrete and bounded, regression targets lie in an (often) unbounded, continuous, and possibly high-dimensional domain, which renders existing measures unsuitable. While many regression methods focus on uncertainty *representation* [Amini et al., 2020, Lakshminarayanan et al., 2017, Kelen et al., 2025], only a few works study the underlying uncertainty measures from a theoretical standpoint [Berry and Meger, 2025, Bülte et al., 2025]. Score-divergence-based decompositions of uncertainty have recently been formalized for *classification* [Kotelevskii et al., 2025, Hofman et al., 2024b], yet no analogous, theoretically grounded measures exist for the multivariate regression setting.

Contributions We address this gap in two steps. *(i)* We transfer the score-divergence formulation of total, aleatoric, and epistemic uncertainty [Kotelevskii et al., 2025, Hofman et al., 2024b] from classification to the multivariate regression setting, yielding a well-defined framework for designing uncertainty measures in continuous target spaces. *(ii)* Within this setting, we identify strictly proper *kernel scores* [Gneiting and Raftery, 2007] as a particularly well-suited family for instantiating these measures: they carry a metric structure, and come with an unbiased, sample-based estimator, which keeps them applicable to complex predictive distributions where density-based measures break down. The choice of kernel then acts as a design lever: we prove explicit connections between properties of the kernel and desirable behavior of the associated uncertainty measure, with regards to the assessment of uncertainties, translation invariance, and robustness. These properties target concrete failure modes of existing measures, as illustrated in Figure 1. Finally, we validate the proposed measures empirically, demonstrating the derived theoretical properties in practice and showcasing their strong performance across a wide range of complex structured regression tasks, including depth estimation and the prediction of dynamical systems.

2 UNCERTAINTY IN REGRESSION

In the following, we denote by $\mathcal{X} \subseteq \mathbb{R}^k$ and $\mathcal{Y} \subseteq \mathbb{R}^d$ the (real-valued) feature and target space, respectively. Furthermore, let $\mathcal{P}(\mathcal{Y})$ denote a convex set of probability measures on the measure space $(\mathcal{Y}, \sigma(\mathcal{Y}))$, where $\sigma(\mathcal{Y})$ is a suitable σ-algebra, and let $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. In addition, we write $\mathcal{D} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ for the training data. For $i \in \{1, \dots, n\}$, each pair $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is a realization of the random variables (X_i, Y_i), which are assumed to be independent and identically distributed (i.i.d) according to some probability measure $\mathbb{P}$. Therefore, each feature vector $\boldsymbol{x} \in \mathcal{X}$ induces a conditional probability distribution $\mathbb{P}(\cdot \mid \boldsymbol{x}) \in \mathcal{P}(\mathcal{Y})$ over the outcome space $\mathcal{Y}$.

Uncertainty representation Regarding second-order uncertainty quantification, we similarly define by $\mathcal{P}(\mathcal{P}(\mathcal{Y}))$ the set of all probability measures on $(\mathcal{P}(\mathcal{Y}), \sigma(\mathcal{P}(\mathcal{Y})))$, with a suitable σ-algebra. We refer to $Q \in \mathcal{P}(\mathcal{P}(\mathcal{Y}))$ as a *second-order distribution*. In contrast to the classification setting, the probability measures $\mathbb{P} \in \mathcal{P}(\mathcal{Y})$ are not necessarily defined on a bounded domain. While we keep the setup as general as possible and this article mainly revolves around uncertainty quantification rather than uncertainty representation, the following examples illustrate how a second-order distribution could be specified within our framework:

Parametric distributions: Given absolute continuity with respect to the Lebesgue measure and a (fixed) parametric distribution $p(\cdot \mid \boldsymbol{\theta}(\boldsymbol{x}))$ with $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$, we can consider the second-order distribution to be on the (measurable) parameter space $(\Theta, \sigma(\Theta))$, e.g. $Q \in \mathcal{P}(\Theta)$. In particular, this includes many uncertainty quantification methods, such as deep ensembles [Lakshminarayanan et al., 2017], deep evidential regression [Amini et al., 2020], or distributional regression [Kneib et al., 2023].

Ensemble approaches: Given an empirical measure, i.e. $Q = Q_m := \frac{1}{M} \sum_{m=1}^M \delta_{\mathbb{P}_m}$ for first-order distributions $\mathbb{P}_m \sim Q$, the setting includes ensembles of general first-order methods such as normalizing flows [Berry and Meger, 2023], mixture density networks [Bishop, 1994], nonparametric ensembles [Kelen et al., 2025] or diffusion models [Wolleb et al., 2022].

Unless noted otherwise, we will consider arbitrary first- and second-order distributions, where we assume that we have a first-order distribution $\mathbb{P} \sim Q$, distributed to some second-order distribution Q and $Y \sim \mathbb{P}$. In addition, we define the (first-order) predictive mixture distribution $\overline{\mathbb{P}} := \mathbb{E}_Q[\mathbb{P}]$,

which can be interpreted as the Bayesian model average (BMA) predictive distribution [Schweighofer et al., 2023].

3 UNCERTAINTY QUANTIFICATION BASED ON PROPER SCORING RULES

In this section, we recall how proper scoring rules can be established to define uncertainty measures. A *scoring rule* [Gneiting and Raftery, 2007] is a function $S : \mathcal{P}(\mathcal{Y}) \times \mathcal{Y} \to \overline{\mathbb{R}}$, such that $S(\mathbb{P}, \mathbb{Q}) \coloneqq \int S(\mathbb{P}, \boldsymbol{y})\, d\mathbb{Q}(\boldsymbol{y})$ is well-defined for all $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{Y})$. S is called *proper*, if $S(\mathbb{Q}, \mathbb{Q}) \leq S(\mathbb{P}, \mathbb{Q})$, for all $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ and *strictly proper* if equality holds only when $\mathbb{P} = \mathbb{Q}$. Intuitively, proper scoring rules quantify the discrepancy between a predictive distribution and an observed outcome. Following Dawid [2007], every scoring rule S can be associated with a *(generalized) entropy* $H : \mathcal{P}(\mathcal{Y}) \to \overline{\mathbb{R}}$ and a *divergence* $D : \mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{Y}) \to \overline{\mathbb{R}}$, via

$$H : \mathbb{P} \mapsto H(\mathbb{P}) \coloneqq S(\mathbb{P}, \mathbb{P}) \tag{1}$$

$$D : (\mathbb{P}, \mathbb{Q}) \mapsto D(\mathbb{P}, \mathbb{Q}) \coloneqq S(\mathbb{P}, \mathbb{Q}) - H(\mathbb{Q}). \tag{2}$$

For (strictly) proper scoring rules, H is (strictly) concave on $\mathcal{P}(\mathcal{Y})$, while the divergence satisfies $D(\mathbb{P}, \mathbb{Q}) \geq 0$ for $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ with equality if and only if $\mathbb{P} = \mathbb{Q}$ [compare Dawid, 2007]. These quantities generalize the familiar notions of Shannon entropy and Kullback-Leibler divergence: H captures the average surprisal under a distribution, and D measures the discrepancy between two distributions. Under mild assumptions, proper scoring rules can be characterized in terms of their entropy function [Gneiting and Raftery, 2007], so either can be used to construct the other.

Scoring rules have been utilized to construct uncertainty measures [Kotelevskii et al., 2025, Hofman et al., 2024b], which can be adapted to our second-order distribution Q in the following way:

$$\begin{aligned} \mathrm{TU_B}(Q) &\coloneqq \mathbb{E}_{\mathbb{P}\sim Q}[S(\overline{\mathbb{P}}, \mathbb{P})], \\ \mathrm{EU_B}(Q) &\coloneqq \mathbb{E}_{\mathbb{P}\sim Q}[D(\overline{\mathbb{P}}, \mathbb{P})], \\ \mathrm{AU_B}(Q) &\coloneqq \mathbb{E}_{\mathbb{P}\sim Q}[H(\mathbb{P})], \end{aligned} \tag{3}$$

Here, epistemic uncertainty (EU) measures the spread of the predictive distributions around their mixture, while aleatoric uncertainty (AU) captures average irreducible noise. Total uncertainty (TU) is the sum thereof.

Alternatively, Kotelevskii et al. [2025], Schweighofer et al. [2023], Berry and Meger [2025] proposes *pairwise estimators* that replace the mixture $\overline{\mathbb{P}}$ with expectations over independent draws from Q:

$$\begin{aligned} \mathrm{TU_P}(Q) &\coloneqq \mathbb{E}_{\mathbb{P},\mathbb{P}'\sim Q}[S(\mathbb{P}', \mathbb{P})], \\ \mathrm{EU_P}(Q) &\coloneqq \mathbb{E}_{\mathbb{P},\mathbb{P}'\sim Q}[D(\mathbb{P}', \mathbb{P})], \end{aligned} \tag{4}$$

with AU unchanged. Both variants satisfy the additive decomposition $\mathrm{TU} = \mathrm{EU} + \mathrm{AU}$. The pairwise estimator avoids computing or sampling from the typically intractable mixture distribution, and—crucially—admits closed-form expressions for many parametric families. The BMA estimator is the less expensive alternative ($\mathcal{O}(M)$ vs. $\mathcal{O}(M^2)$ for an ensemble of size M), but usually requires approximation of $\overline{\mathbb{P}}$. When S is convex in its first argument, Jensen's inequality gives $\mathrm{TU}_P \geq \mathrm{TU}_B$, so that the pairwise estimator provides an upper bound [Schweighofer et al., 2023].

While this decomposition is general, its application to regression has been limited: the log-score, which leads to the familiar entropy-based measure [Fishkov et al., 2025], requires absolute continuity and therefore density estimation, which is intractable for high-dimensional data. In the following section, we propose kernel scores as a principled and practically advantageous instantiation of this framework for general regression settings.

4 KERNEL SCORES

We now introduce kernel scores as the central tool of our framework. The key insight is that kernel scores inherit all the structural properties required for the decomposition in (3)–(4), while additionally providing closed-form expressions for a broad class of distributions, unbiased nonparametric estimators, and applicability to structured domains such as graphs or functional data. Here, we draw mainly on the notation of Waghmare and Ziegel [2026].

Definition 4.1 (Kernel score)**.** *Let $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a continuous, conditionally negative definite kernel*[1]*, and $\mathcal{P}_k = \{\mathbb{P} \in \mathcal{P}(\mathcal{Y}) : \iint k(\boldsymbol{x}, \boldsymbol{x}')\, d\mathbb{P}(\boldsymbol{x})\, d\mathbb{P}(\boldsymbol{x}') < \infty\}$. Then, the associated kernel score is*

$$\begin{aligned} S_k(\mathbb{P}, \boldsymbol{y}) = &\int k(\boldsymbol{x}, \boldsymbol{y})\, d\mathbb{P}(x) \\ &- \frac{1}{2} \iint k(\boldsymbol{x}, \boldsymbol{x}')\, d\mathbb{P}(\boldsymbol{x})\, d\mathbb{P}(\boldsymbol{x}') - \frac{1}{2} k(\boldsymbol{y}, \boldsymbol{y}), \end{aligned} \tag{5}$$

for $\mathbb{P} \in \mathcal{P}_k, \boldsymbol{y} \in \mathcal{Y}$, with induced entropy and divergence

$$\begin{aligned} H_k(\mathbb{P}) = &\frac{1}{2} \iint k(\boldsymbol{x}, \boldsymbol{x}')\, d\mathbb{P}(\boldsymbol{x})\, d\mathbb{P}(\boldsymbol{x}') \\ &- \frac{1}{2} \int k(\boldsymbol{x}, \boldsymbol{x})\, d\mathbb{P}(\boldsymbol{x}), \end{aligned} \tag{6}$$

$$D_k(\mathbb{P}, \mathbb{Q}) = -\frac{1}{2} \iint k(\boldsymbol{y}, \boldsymbol{y}')\, d(\mathbb{P} - \mathbb{Q})(\boldsymbol{y})\, d(\mathbb{P} - \mathbb{Q})(\boldsymbol{y}'), \tag{7}$$

for $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_k$.

S_k is nonnegative and (strictly) proper for a (strongly) conditionally negative definite kernel [Waghmare and Ziegel, 2026].

[1] A kernel $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is conditionally negative definite if $\sum_{i,j=1}^{n} a_i a_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq 0$, $\forall n \in \mathbb{N}$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{Y}$, and $a_1, \ldots, a_n \in \mathbb{R}$ with $\sum_{j=1}^{n} a_j = 0$. [Waghmare and Ziegel, 2026].

Instantiating the pairwise estimator (4) with S_k directly yields tractable uncertainty measures, whose closed-form expressions for Gaussian and mixture distributions are derived in Appendix B. Kernel scores have been increasingly applied in forecast evaluation and machine learning [Gneiting and Raftery, 2007, Allen et al., 2023, Shen and Meinshausen, 2024], including complex regression settings such as weather forecasting [Chen et al., 2024, Alet et al., 2025] or solving PDEs [Bülte et al., 2025].

Crucially, D_k is essentially the squared distance between the kernel mean embeddings of the probability distributions into some Hilbert space [Steinwart and Ziegel, 2021] and is closely related to the Maximum Mean Discrepancy (MMD^2), a well-studied divergence in statistics and machine learning [Gretton et al., 2012, Sejdinovic et al., 2013]. This connection provides both a theoretical basis and practical advantages that distinguish our framework from alternatives such as the log- or quadratic score.

Metric structure: Under mild conditions, kernel scores are the only scoring rules that induce a valid metric on $\mathcal{P}_k$ [Theorem 19, Waghmare and Ziegel, 2026]. Furthermore, existence only requires $H_k(\mathbb{P}) < \infty$, which allows for measuring the divergence between continuous, discrete, or degenerate distributions, as opposed to other scoring rules that require absolute continuity with respect to the Lebesgue measure (compare Figure 1).

Sample-based estimation: The MMD^2 (and therefore also S_k and H_k) admits an unbiased empirical estimator [Gretton et al., 2012]; therefore, the uncertainty measures can be estimated consistently from samples alone. This makes the framework applicable to implicit or sample-based models, such as diffusion, or flow-based models, where likelihood evaluation is intractable in high dimensions.

Structured domains: The kernel k can be adapted to the underlying output domain: stationary kernels for Euclidean regression, variogram-based kernels for spatial outputs [Scheuerer and Hamill, 2015], graph kernels for molecular data [Vishwanathan et al., 2010], or functional kernels for PDE solution spaces [Wynne and Duncan, 2022]. This flexibility is unique among common scoring rules and is central to providing domain-independent uncertainty measures.

Translation invariance and homogeneity: When $k(\boldsymbol{x}, \boldsymbol{y}) \equiv \kappa(\boldsymbol{x} - \boldsymbol{y}),\ \boldsymbol{x}, \boldsymbol{y} \in \mathcal{Y}$, for some conditionally negative definite function $\kappa : \mathcal{Y} \to \mathbb{R}$, the score is *translation invariant*, i.e., $S_k(\mathbb{P}, \boldsymbol{y}) = S_k(\mathbb{P}_{\boldsymbol{h}}, \boldsymbol{y} + \boldsymbol{h})$ for $\boldsymbol{y}, \boldsymbol{h} \in \mathcal{Y}$. A scoring rule is *homogeneous of degree* α if $S(\mathbb{P}_c, c\boldsymbol{y}) = c^\alpha S(\mathbb{P}, \boldsymbol{y})$ for every $c > 0, \boldsymbol{y} \in \mathcal{Y}, \mathbb{P} \in \mathcal{P}$ [Waghmare and Ziegel, 2026]. This ensures that affine rescalings of the data do not change the relative performance assessment—a desirable invariance for regression tasks spanning different output scales.

5 PROPERTIES OF KERNEL SCORES AS AN UNCERTAINTY MEASURE

The properties of kernel scores described above carry over directly to the induced uncertainty measures. For instance, the ability to compare arbitrary distributions—including degenerate ones—via a sample-based estimator is particularly relevant when first-order distributions are combined via a linear pool [Hall and Mitchell, 2007], as is common in forecast ensembles [Krüger and Nolte, 2016]. Beyond these inherited characteristics, we now show that principled choices of k lead to uncertainty measures satisfying additional desirable properties, extending previous studies on axiomatic frameworks [Wimmer et al., 2023, Hüllermeier et al., 2022, Bülte et al., 2025]. One trivial aspect of the corresponding measures is that they are all nonnegative, which follows directly from the kernel score being nonnegative.

Let $\mathbb{P} \sim Q, \mathbb{P}' \sim Q'$ be random first-order distributions with $Q, Q' \in \mathcal{P}(\mathcal{P}(\mathcal{Y}))$ and let $\delta_{\mathbb{P}} \in \mathcal{P}(\mathcal{P}(\mathcal{Y}))$ denote the Dirac measure at $\mathbb{P} \in \mathcal{P}(\mathcal{Y})$. For $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\mathcal{Y})$ let $\leq_{\text{cx}}$ denote the convex order [Shaked and Shanthikumar, 2007], meaning that $\mathbb{P}_1 \leq_{\text{cx}} \mathbb{P}_2 \iff \mathbb{E}_{X\sim\mathbb{P}_1}[\phi(X)] \leq \mathbb{E}_{Y\sim\mathbb{P}_2}[\phi(Y)]$ for all convex $\phi : \mathcal{Y} \to \mathbb{R}$. Similarly, for $Q_1, Q_2 \in \mathcal{P}(\mathcal{P}(\mathcal{Y}))$, let $\leq^2_{\text{cx}}$ denote the convex order with respect to all convex functionals $\Phi : \mathcal{P}(\mathcal{Y}) \to \mathbb{R}$. In particular for $\mathbb{P}_1 \leq_{\text{cx}} \mathbb{P}_2$ it holds that $\mathbb{E}_{X\sim\mathbb{P}_1}[X] = \mathbb{E}_{Y\sim\mathbb{P}_2}[Y]$ and $\mathbb{V}_{X\sim\mathbb{P}_1}[X] \leq \mathbb{V}_{Y\sim\mathbb{P}_2}[Y]$, since the stochastic order is a measure of variability [Shaked and Shanthikumar, 2007]. We propose the following properties of the corresponding uncertainty measures, which are proved for both types of estimators in Appendix A.

First, we formalize the intuition that a fully concentrated second-order distribution, i.e., no model disagreement, should yield zero epistemic uncertainty. In addition, epistemic uncertainty should be monotone with respect to the convex order: a second-order distribution with greater spread over models implies greater epistemic uncertainty. This leads to the following proposition:

Proposition 5.1 (Epistemic uncertainty)**.** *For any proper scoring rule S, for which the map $\mathbb{P} \mapsto S(\mathbb{P}, \mathbb{Q})$ is convex for fixed $\mathbb{Q}$, it holds that*

1. $Q = \delta_{\mathbb{P}} \implies \text{EU}(Q) = 0$*, while for a strictly proper scoring rule the converse holds as well,*
2. $\text{EU}(\delta_{\mathbb{P}}) \leq \text{EU}(Q_1) \leq \text{EU}(Q_2), \quad \forall Q_1 \leq^2_{\text{cx}} Q_2.$

Similarly, if a *first-order* predictive distribution has more variability, it should be assigned a higher value of *aleatoric* uncertainty, as formalized in the following proposition:

Proposition 5.2 (Aleatoric uncertainty)**.** *Any kernel score S_k with a translation invariant kernel $k(\boldsymbol{x}, \boldsymbol{x}')$ that is convex in one of its arguments fulfills* $\text{AU}(\delta_{\mathbb{P}_1}) \leq \text{AU}(\delta_{\mathbb{P}_2}),\ \forall \mathbb{P}_1 \leq_{\text{cx}} \mathbb{P}_2$.

Finally, we want to analyze how robust an uncertainty measure is to deviations in the second-order distribution. We consider robustness in terms of the influence function [Hampel et al., 1986, Chapter 2], which analyzes the limiting behavior if the underlying (second-order) distribution is perturbed by a single point diverging to infinity. If the influence function is bounded, any outlier in Q can only have a finite impact on the estimation of the uncertainty measure M, making it robust against such outliers. This is formalized in the following proposition:

Proposition 5.3 (Robustness)**.** *Consider a parametric first-order distribution $\mathbb{P}_{\boldsymbol{\theta}} \in \mathcal{P}(\mathcal{Y})$ with $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$, a second-order distribution $Q \in \mathcal{P}(\Theta)$, and $\boldsymbol{\vartheta} \sim Q$. Let S_k be a kernel score with bounded kernel, i.e., $\|k\|_\infty := \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{Y}} |k(\boldsymbol{x},\boldsymbol{y})| < \infty$ and let $Q_\varepsilon := (1-\varepsilon)Q + \varepsilon\delta_{\boldsymbol{\theta}_0}$, $\boldsymbol{\theta}_0 \in \Theta$, with influence function*

$$\mathrm{IF}(\boldsymbol{\theta}_0; M, Q) := \lim_{\varepsilon\to 0} \frac{M(Q_\varepsilon) - M(Q)}{\varepsilon}.$$

Then, for each uncertainty measure $M \in \{\mathrm{AU}, \mathrm{EU}\}$, we have $M(Q) < \infty$, and

$$\sup_{\boldsymbol{\theta}_0\in\Theta} |\mathrm{IF}(\boldsymbol{\theta}_0; M, Q)| \;\le\; C_M \,\|k\|_\infty \;<\; \infty,$$

for some constant C_M so M is robust in terms of the influence function.

Together, Propositions 5.1–5.3 characterize certain desirable behavior of uncertainty measures: epistemic uncertainty vanishes if and only if all models agree, aleatoric uncertainty increases with the variability of the predictive distribution, and neither measure can be destabilized by outlying ensemble members when k is bounded. Crucially, these properties are not guaranteed by properness alone—they depend on the specific choice of kernel. The propositions, therefore, serve as a principled guide for selecting k in dependence on the underlying task and corresponding requirements. In particular, we propose the following instantiations, with closed-form expressions derived in Appendix B.

Energy score: The energy score (S_{ES}), with kernel $k(\boldsymbol{x},\boldsymbol{x}') = \|\boldsymbol{x}-\boldsymbol{x}'\|$ is strictly proper, translation invariant, and homogeneous, satisfying both Propositions 5.1 and 5.2. In fact, it is the unique homogeneous translation invariant kernel score on $\mathbb{R}^d$. Its univariate special case $d = 1$ recovers the continuous ranked probability score, arguably one of the most widely used proper scoring rules in regression settings [Gneiting and Raftery, 2007]. Any univariate strictly proper score, such as the CRPS, can be extended to a multivariate strictly proper rule via marginal averaging (see Appendix B), which we denote by S_{CRPS}. However, in that case, the dependence structure is not accounted for.

Gaussian kernel score: The (negative) Gaussian kernel score (S_{k_γ}) corresponding to the kernel $k(\boldsymbol{x},\boldsymbol{x}') = -\exp\left(-\|\boldsymbol{x}-\boldsymbol{x}'\|^2/\gamma^2\right)$ and bandwidth $\gamma > 0$ is strictly proper, satisfies Propositions 5.1, but fails 5.2 due to the kernel not being convex. As the only bounded kernel among our proposals, it is however, the only one satisfying the robustness condition of Proposition 5.3, making it the most conservative choice when outlying ensemble members are anticipated.

Squared-error: Finally, the squared-error (S_{SE}) with kernel $k(\boldsymbol{x},\boldsymbol{x}') = \|\boldsymbol{x}-\boldsymbol{x}'\|^2$ falls within the kernel score framework and recovers the commonly-used variance-based uncertainty measure in the univariate case [Amini et al., 2020], providing a natural link to already existing measures. However, since it is not strictly proper, it fails the converse in Proposition 5.1 and has stronger assumptions (existence of second moments) than the other scoring rules. We therefore include it primarily as a baseline for comparison rather than as a recommended instantiation.

For completeness, while the strictly proper *log-score* ($S_{\log}$) is not a kernel score, it satisfies Propositions 5.1 and 5.2 under standard regularity conditions (see Appendix A) and leads to the well-known entropy-based uncertainty measure [Kendall and Gal, 2017]. However, it requires absolute continuity with respect to the Lebesgue measure, density evaluation and can assign negative uncertainty values, limiting its applicability as a general uncertainty measure. In summary, we propose the kernel-based uncertainty decomposition (4) instantiated with $S_{\mathrm{ES}}, S_{\mathrm{CRPS}}, S_{k_\gamma}$ as principled uncertainty measures for regression, supported by the theoretical properties of the measures itself, as well as the guarantees established in Propositions 5.1–5.3. The log-score and squared-error remain valid instantiations within the scoring rule framework, serving as a natural ground for comparison.

6 NUMERICAL EXPERIMENTS

We evaluate our kernel-based uncertainty measures across four experimental protocols: robustness evaluation, selective prediction, out-of-distribution detection, and active learning, probing complementary aspects of uncertainty quality from calibration under shift to data-efficient acquisition.

Table 1: Predictive uncertainty representations used across the experiments.

Method	Predictive form	Second-order
NG	$\mathcal{N}(\mu, \sigma^2)$	Ensemble
DER	$t_{2\alpha}(\cdot;\gamma, \frac{\beta(1+\upsilon)}{\upsilon\alpha}, 2\alpha)$	Conjugate prior
LoRa	$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{U}\boldsymbol{U}^\top + \boldsymbol{D})$	Ensemble
MDN	$\sum_k w_k \mathcal{N}(\mu_k, \sigma_k^2)$	Ensemble
SB	$\frac{1}{N}\sum_n \delta_{\boldsymbol{y}_n}$	Ensemble

To demonstrate that the proposed measures are agnostic to the choice of predictive model, we evaluate them across five

uncertainty representation methods, summarized in Table 1. The natural Gaussian (NG) method [Immer et al., 2023] uses a predictive univariate normal distribution with a second-order ensemble, while the deep evidential regression (DER) approach [Amini et al., 2020] uses the corresponding conjugate prior. Further, we consider a univariate mixture density network (MDN) [Bishop, 1994] and a multivariate Gaussian with a low-rank covariance matrix (LoRa) [Rezende et al., 2014], both using second-order ensembling. Finally, we utilize a nonparametric sampling-based generative model (SB) [Pacchiardi et al., 2024]. These models cover a variety of predictive representations, including closed-form, sampling-based, multimodal, or multivariate.

Further, we consider a variety of benchmark datasets, covering different predictive tasks, dimensions, and data modalities. In particular, we consider the UCI dataset for univariate regression [Hernández-Lobato and Adams, 2015], two one-dimensional PDE prediction tasks [Takamoto et al., 2022], and two two-dimensional vision tasks, namely depth regression [Amini et al., 2020] and surface temperature prediction [Rasp et al., 2024]. Note that for the PDEs, we use the probabilistic neural operator [Bülte et al., 2025] as the sample-based method, which generates solution samples in the corresponding function space. However, the properties of our selected kernels also hold in the corresponding Hilbert spaces [Ziegel et al., 2024], highlighting the broad applicability of our framework. A sample prediction and corresponding uncertainty estimates are shown in Figure 2.

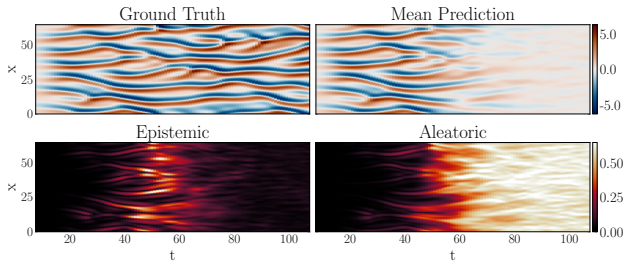

Figure 2: Predictions for the (chaotic) Kuramoto-Sivashinsky equation using the sampling-based method and corresponding estimates of epistemic and aleatoric uncertainty using the S_{CRPS} measure. Both components show structural consistency with the underlying task; aleatoric uncertainty increases with time, as physical predictability decreases, while epistemic uncertainty peaks around the predictability limit of the system.

As uncertainty quantification baselines, we include the log-score $S_{\log}$ and the squared-error S_{SE}, which are commonly used in practice but either lack the theoretical guarantees established in the previous section or do not fit into the kernel framework at all. The bandwidth for the Gaussian kernel score S_{k_γ} is chosen via the median heuristic [Garreau et al., 2018] on each dataset, which we found to perform reliably across tasks.

In the following, we use the pairwise estimator throughout, as it admits closed-form expressions for all considered first-order distributions (compare Appendix B). Detailed descriptions of the experimental setup, as well as additional results and visualizations are provided in Appendix D. For completeness, we also provide an analysis of computational complexity and approximation error of the different instantiations, as well as a corresponding empirical runtime analysis in Appendix C.

6.1 ROBUSTNESS ANALYSIS

Table 2: MAPE (↓) of AU and EU estimates on the concrete dataset, comparing a base ensemble of $M = 25$ members against an augmented ensemble with one additional member trained on targets distorted by $\tilde{y} = y + \mathcal{N}(0, \delta^2)$.

	Aleatoric			
S/δ	0.0	1.0	2.5	5.0
$S_{\log}$	0.2	3.2	4.5	4.8
S_{SE}	1.1	4.6×10^3	6.8×10^4	4.8×10^5
S_{ES}	0.6	6.7×10^1	2.2×10^2	5.0×10^2
S_{k_γ}	0.0	0.2	0.2	0.2
	Epistemic			
$S_{\log}$	3.5	1.5×10^4	5.3×10^4	2.4×10^5
S_{SE}	3.7	3.0×10^4	6.2×10^4	4.7×10^5
S_{ES}	2.7	8.1×10^2	1.2×10^3	4.1×10^3
S_{k_γ}	1.6	1.3×10^1	1.1×10^1	1.3×10^1

To empirically validate the robustness (in terms of the influence function) of different measures, we use three datasets from the UCI benchmark [Hernández-Lobato and Adams, 2015] and train a deep ensemble [Lakshminarayanan et al., 2017] on each task. Then, we train one additional ensemble member using a target variable with added noise, i.e. $\tilde{y} = y + \mathcal{N}(0, \delta^2)$ with gradually increasing noise. While this is a synthetic outlier creation, it allows for comparing the robustness of each uncertainty measure and the corresponding scoring rule to a single corrupted ensemble member. To measure the deviation, we use the mean absolute percentage error (MAPE) with respect to the uncertainty in the base ensemble, i.e.,

$$\mathrm{MAPE} \coloneqq \frac{100}{n} \sum_{i=1}^{n} \left| \frac{M_i(Q^\delta) - M_i(Q)}{M_i(Q)} \right|,$$

where $M_i \in \{\mathrm{AU}, \mathrm{EU}\}$ denotes the uncertainty estimate at input $\boldsymbol{x}_i$, $i = 1, \ldots, n$, Q denotes the base ensemble and Q^δ denotes the corrupted ensemble. Table 2 shows the results for the concrete dataset. The Gaussian kernel score S_{k_γ} remains stable across all distortion levels for both aleatoric and epistemic uncertainty. In contrast, the other measures, most notably the squared-error degrade by several

orders of magnitude even at moderate δ, consistent with their unbounded influence functions.

6.2 SELECTIVE PREDICTION

In selective prediction, the model is evaluated only on parts of the (test-) dataset, typically a specific subset with low uncertainty. Therefore, this task assesses the ability of the uncertainty measure to indicate whether a prediction is correct or not. Here, one typically uses total uncertainty [Kotelevskii et al., 2025, Hofman et al., 2026], as neither component, aleatoric or epistemic, determines the prediction correctness alone. The performance for selective prediction is measured using prediction-reject-ratios [PRRs, Malinin and Gales, 2021], which are negatively oriented (lower is better) of inaccurate predictions using the corresponding uncertainty measure. We evaluate this experiment for all datasets and uncertainty representation methods; Table 3 shows the corresponding average ranks, while the full results table and selected visualizations are available in Appendix D.5.

Table 3: Average rank (↓) of the PRR aggregated over all datasets, with the best measure in bold. The total average is calculated per task, i.e. Univariate, 1D, and 2D.

Method	$S_{\log}$	S_{SE}	S_{CRPS}	S_{ES}	S_{k_γ}
NG	**2.56**	2.61	2.61	2.61	2.67
DER	3.39	**1.94**	3.00	2.56	2.44
MDN	-	1.78	**1.67**	1.89	2.67
LoRa	5.00	2.00	4.00	3.00	**1.00**
SB	-	2.45	2.41	2.05	**1.77**
Task-weighted	3.51	2.34	2.69	2.43	**2.29**

Overall, S_{k_γ} achieves the best aggregated performance across all tasks, with S_{ES} and S_{CRPS} obtaining slightly worse but comparable ranks. For more complex multimodal or multivariate representation methods these strictly proper scoring rules consistently outperform alternatives, which follows directly from Proposition 5.1: scoring rules sensitive to distributional shape beyond the first two moments are better equipped to reflect uncertainty in more expressive predictive distributions. The comparatively weaker performance of S_{SE} for the multimodal methods is therefore expected: reducing uncertainty to a point estimate of the predictive mean cannot distinguish between, e.g., a peaked unimodal and a diffuse multimodal distribution. Nevertheless, S_{SE} remains competitive, however, only for methods assuming a unimodal or multivariate Gaussian output, where the mean is a sufficient summary.

6.3 OUT-OF-DISTRIBUTION DETECTION

Out-of-distribution (OOD) detection is a commonly used task to assess and compare the quality of uncertainty measures and uncertainty quantification methods. In essence, a model is trained on in-distribution (ID) data and its predictions, as well as corresponding uncertainty estimates, are compared between ID and OOD data. Since the model has not seen the OOD data before, it should assign higher epistemic uncertainty to those inputs.

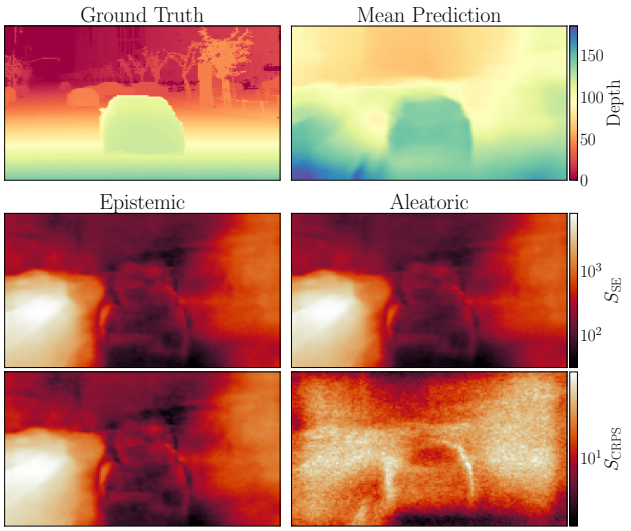

Figure 3: Predictions of the sampling-based method for the depth regression OOD dataset, as well as uncertainty estimates of the (pointwise) measures S_{SE} and S_{CRPS}.

Table 4: Average rank (↓) of the AUROC for out-of-distribution detection for each method aggregated over all datasets used. The best measure is highlighted in bold.

Method	$S_{\log}$	S_{SE}	S_{CRPS}	S_{ES}	S_{k_γ}
NG	**2.00**	4.00	2.50	3.00	3.50
DER	**1.00**	3.50	3.00	4.50	3.00
MDN	-	3.25	2.25	**1.75**	2.75
LoRa	5.00	3.25	**1.75**	2.25	2.75
SB	-	3.25	**2.00**	**2.00**	2.75
Average	2.67	3.42	**2.25**	2.58	2.92

We generate OOD data for the 1D PDE tasks by changing the underlying coefficients of the PDE (i.e., viscosity and length scale) and for the 2D tasks using a domain shift (different scene for the depth regression and a different geographical domain for the weather prediction task). To evaluate the performance of the different measures, we evaluate the AUROC of the uncertainty scores between OOD and ID samples. Table 4 shows the corresponding rank of the uncertainty measures, averaged across the uncertainty representation methods.

Overall, the marginal score S_{CRPS} shows the best performance, followed by its multivariate version S_{ES}. Here, the log-score $S_{\log}$ performs quite well for a first-order univariate

Gaussian, but does not lead to a good OOD recognition for the multivariate Gaussian. In this experiment, S_{SE} performs worst across all measures. Figure 3 provides an exemplary visualization of the out-of-distribution prediction for the ApolloScape dataset and the sampling-based method. It is evident that the S_{SE} measure leads to almost identical predictions for AU and EU, while the S_{CRPS} measure shows better disentangled uncertainty estimates.

It is worth noting that OOD detection benchmarks are inherently difficult to construct in a way that is both realistic and discriminative [Hofman et al., 2026, Li et al., 2025], as detection difficulty is inseparable from how the distribution shift is defined. This is reflected in Table 11, where AUROC values are near one across almost all datasets and methods—a result that should be interpreted carefully, since not all distribution shifts meaningfully increase predictive difficulty. Crucially, this near-ceiling performance is nonetheless consistent across diverse datasets, shift types, and uncertainty representation methods, providing broader evidence for the reliability of kernel scoring rules as uncertainty measures beyond what any single dataset could establish.

6.4 ACTIVE LEARNING

Table 5: Average rank (↓) of the final test loss in the active learning setting aggregated over all datasets, with the best measure in bold. Here, $\mathcal{U}$ denotes the random baseline.

Method	$S_{\log}$	S_{SE}	S_{CRPS}	S_{ES}	S_{k_γ}	$\mathcal{U}$
NG	3.57	4.00	4.00	4.00	**1.43**	2.00
DER	3.57	4.00	**1.57**	**1.57**	2.86	3.00
MDN	-	2.57	3.29	3.29	**1.57**	2.57
SB	-	2.22	2.00	**1.67**	2.44	4.11
Average	3.57	3.13	2.67	2.57	**2.10**	3.00

As a final task, we consider an active learning experiment, which is frequently used to evaluate probabilistic predictions and uncertainty measures. Here, the objective is to select new training instances under a computational budget, using epistemic uncertainty as the selection criterion [Nguyen et al., 2022, Kirsch et al., 2019]. Starting from a small training set, the learner iteratively selects new datapoints based on the corresponding (predictive) uncertainty to minimize the corresponding loss function with as few labels as possible. We use this task as a comparison ground for our different estimators of epistemic uncertainty.

Due to the high computational load of the 2D tasks, we focus on the univariate and 1D PDE datasets, which still offer a diverse ground for comparison. We split the data into train, validation, and test datasets and use 5% of the training data size for a random initialization. In each of 20 rounds, the learner acquires 1% of the dataset size as new datapoints. In each round, $M = 5$ ensemble members are trained for 50 epochs from scratch, and the whole experiment is repeated across three independent seeds. Table 5 shows the performance ranks of the uncertainty measures per representation method, while Table 12 provides the full results. Here, the three strictly proper scoring rules (S_{CRPS}, S_{ES}, S_{k_γ}) lead to the best performance. In particular, S_{k_γ} obtains a significantly lower rank than the comparison methods. As opposed to the other experiments, the methods $S_{\log}$ and S_{SE} do not perform well, even worse than the random baseline $\mathcal{U}$, also for the first-order Gaussian predictions.

Findings & Insights Across four evaluation protocols spanning various data domains, the kernel-based (strictly proper) scoring rules demonstrate consistently strong performance. S_{k_γ} achieves the best overall rank in selective prediction and active learning, while S_{CRPS} and S_{ES} lead in out-of-distribution detection. The strictly proper kernel scores particularly outperform $S_{\log}$ and S_{SE} for expressive predictive distributions beyond unimodal Gaussians, where sensitivity to distributional shape beyond the first two moments is critical. Additionally, S_{k_γ} remains stable under ensemble corruption, whereas other measures degrade by orders of magnitude. Overall, our proposed measures consistently outperform the baselines: on every task, at least one kernel-based measure outperforms both baselines, and each kernel-based measure outperforms the baselines on the majority of tasks. Although no single uncertainty measure dominates uniformly across all tasks and representation methods, this is expected given the broad framework and the fact that different kernels lead to different uncertainty assessments. However, in general, the kernel scores that fulfill the posed theoretical properties collectively offer the most reliable uncertainty quantification.

7 RELATED WORK

Novel uncertainty measures. Many studies focus on quantifying uncertainty for predictive models, especially for classification. While the most commonly used measures are based on the Shannon entropy [Houlsby et al., 2011], those have been criticized for having undesirable properties [Wimmer et al., 2023]. Several generalizations have been proposed, such as variance-based [Sale et al., 2023], distance-based [Sale et al., 2024] or pairwise [Schweighofer et al., 2023, Berry and Meger, 2025] estimators. Closest to our work are recent developments in deriving uncertainty measures based on proper scoring rules and divergences. Gruber and Buettner [2023], Adlam et al. [2022] derive a bias-variance decomposition based on Bregman divergences, which was extended to kernel scores by Gruber and Buettner [2024], where the corresponding uncertainty measures are conceptually similar to our proposed ones. However, their study focuses on generative models and on assessing predictive performance. Recently, Kotelevskii et al. [2025], Hofman et al. [2024a] introduced a framework for decomposing

and quantifying uncertainty based on proper scoring rules, which was extended to the univariate (Gaussian) regression case [Fishkov et al., 2025]. While similar in nature, our work specifically considers kernel scores with advantageous properties and works in more general regression domains, moving away from the univariate Gaussian assumptions to more complex uncertainty representation.

Uncertainty quantification in regression. While many works focus on uncertainty representation in regression, for example, via second-order distributions [Amini et al., 2020, Meinert and Lavin, 2022, Malinin et al., 2020] or ensembles [Berry and Meger, 2023, Lakshminarayanan et al., 2017, Kelen et al., 2025], little is usually done in the direction of analyzing the underlying uncertainty measures. The studies usually employ either the variance-based measure [Amini et al., 2020, Meinert and Lavin, 2022, Valdenegro-Toro and Mori, 2022] or (a variant of) the entropy-based measure [Malinin et al., 2020, Berry and Meger, 2025, Postels et al., 2021]. While Bülte et al. [2025] compare both measures with respect to a given set of preferable properties, they do not consider other measures or the pairwise variants thereof. In contrast, our work proposes a general way to construct uncertainty measures that can be used with many different instantiations, leading to different properties.

8 CONCLUSION

We propose a general framework for uncertainty quantification in regression, based on strictly proper kernel scores, which encompasses different uncertainty measures in a single principled construction. In particular, it turns the design of uncertainty measures into the selection of an underlying kernel k, providing a systematic way for domain or task-specific constructions. Our analysis shows how structural properties of kernels, such as strictly properness or boundedness, directly translate into distinct characteristics of the induced uncertainty measures, allowing practitioners to design measures aligned with specific requirements. Beyond theoretical contributions, our empirical results demonstrate the validity of the proposed measures, yielding consistently strong performance across diverse data modalities, uncertainty representations, and benchmark tasks.

Limitations and future work While our framework provides a principled foundation, it is not unique; alternative measures may satisfy the same properties. Developing principled selection procedures for choosing or learning suitable kernels remains an important direction for future research. In particular, exploring alternative score constructions, such as weighted kernel scores [Allen et al., 2023], different kernel families (e.g., Laplace or inverse-multiquadratic), or learnable kernels could enable more targeted sensitivity to specific phenomena, such as extreme events. Similarly, improving the bandwidth selection procedure within the kernel, for example, by mixing different scales, could further improve empirical performance. In general, our work focused on a specific set of theoretical properties, and extending the analysis to other aspects such as computational efficiency, scalability, or interpretability could broaden the framework and extend its applicability. Exciting opportunities also lie in exploring other data domains, such as graph-structured data, where the usage of kernel scores could open up new possibilities of uncertainty quantification in, for example, molecule design. Empirically, extending the evaluation to a broader range of generative models, including diffusion and flow-based architectures, may provide further insight into the practical behavior of kernel-based scores. Finally, further theoretical investigation of the relationship between kernel scores and maximum mean discrepancy could yield new insights into their geometric and statistical properties, potentially guiding the principled design of uncertainty measures that are optimal for a specific downstream task.

Acknowledgements

The authors acknowledge support by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Research, Technology and Space.

C. Bülte and G. Kutyniok acknowledge support by the German Research Foundation under the grant DFG-SPP-2298.

E. Hüllermeier acknowledges support by the German Research Foundation under the grant GRK 3081 (project number 534429653).

G. Kutyniok also acknowledges support by the gAIn project, which is funded by the Bavarian Ministry of Science and the Arts (StMWK Bayern) and the Saxon Ministry for Science, Culture and Tourism (SMWK Sachsen). Furthermore, G. Kutyniok is supported by LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder as well as by the Hightech Agenda Bavaria.

References

Ben Adlam, Neha Gupta, Zelda Mariet, and Jamie Smith. Understanding the bias-variance tradeoff of Bregman divergences, 2022. arXiv:2202.04167.

Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R. Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Skillful joint probabilistic weather forecasting from marginals, 2025. arXiv:2506.10772.

Sam Allen, David Ginsbourger, and Johanna Ziegel. Evaluating forecasts for high-impact events using transformed kernel scores. *SIAM/ASA Journal*

on Uncertainty Quantification, 11(3):906–940, 2023. doi:10.1137/22M1532184.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14927–14937. Curran Associates, Inc., 2020.

Lucas Berry and David Meger. Normalizing Flow Ensembles for Rich Aleatoric and Epistemic Uncertainty Modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6806–6814, June 2023. doi:10.1609/aaai.v37i6.25834.

Lucas Berry and David Meger. Epistemic uncertainty estimation in regression ensemble models with pairwise epistemic estimators. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, volume 39, pages 14927–14937. Curran Associates, Inc., 2025.

Christopher M. Bishop. Mixture density networks. Working paper, Aston University, 1994.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Christopher Bülte, Philipp Scholl, and Gitta Kutyniok. Probabilistic neural operators for functional uncertainty quantification. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Christopher Bülte, Yusuf Sale, Timo Löhr, Paul Hofman, Gitta Kutyniok, and Eyke Hüllermeier. An Axiomatic Assessment of Entropy- and Variance-based Uncertainty Quantification in Regression, 2025. arXiv:2504.18433.

Jieyu Chen, Tim Janke, Florian Steinke, and Sebastian Lerch. Generative machine learning methods for multivariate ensemble postprocessing. *The Annals of Applied Statistics*, 18(1):159–183, March 2024. doi:10.1214/23-AOAS1784.

A. P. Dawid. The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, 59(1):77–93, February 2007. doi:10.1007/s10463-006-0099-8.

A. D´Isanto and K. L. Polsterer. Photometric redshift estimation via deep learning - generalized and pre-classification-less, image based, fully probabilistic redshifts. *Astronomy&Astrophysics*, 609:A111, 2018. doi:10.1051/0004-6361/201731326.

Vineet Edupuganti, Morteza Mardani, and Shreyas Vasanawala. Uncertainty Quantification in Deep MRI Reconstruction. *IEEE transactions on medical imaging*, PP, 09 2020. doi:10.1109/TMI.2020.3025065.

Alexander Fishkov, Kajetan Schweighofer, Mykyta Ielanskyi, Nikita Kotelevskii, Mohsen Guizani, and Maxim Panov. Uncertainty Quantification for Regression using Proper Scoring Rules, 2025. arXiv:2509.26610.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic, 2018. arXiv:1707.07269.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi:10.1198/016214506000001437.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13:723–773, March 2012. ISSN 1532-4435.

Sebastian Gruber and Florian Buettner. Uncertainty estimates of predictions via a general bias-variance decomposition. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 11331–11354. PMLR, 2023.

Sebastian Gregor Gruber and Florian Buettner. A Bias-Variance-Covariance Decomposition of Kernel Scores for Generative Models. In *Forty-first International Conference on Machine Learning*, 2024.

Stephen Hall and James Mitchell. Combining density forecasts. *International Journal of Forecasting*, 23:1–13, 03 2007. doi:10.1016/j.ijforecast.2006.08.001.

Frank Hampel, Elvezio Ronchetti, Peter Rousseeuw, and Werner Stahel. *Robust Statistics: The Approach Based on Influence Functions*. 03 1986. doi:10.1002/9781118186435.

José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks, 2015. arXiv:1502.05336.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia De Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, July 2020. doi:10.1002/qj.3803.

Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty: A credal approach. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024a.

Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty with proper scoring rules, 2024b. arXiv:2404.12215.

Paul Hofman, Yusuf Sale, and Eyke Hüllermeier. Uncertainty Quantification for Machine Learning: One Size Does Not Fit All. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(26):21726–21734, 2026. 10.1609/aaai.v40i26.39323.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning, 2011. arXiv:1112.5745.

Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The ApolloScape dataset for autonomous driving. 03 2018. doi:10.1109/CVPRW.2018.00141.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3): 457–506, 2021.

Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In *Uncertainty in Artificial Intelligence*, pages 548–557. PMLR, 2022.

Alexander Immer, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Effective Bayesian heteroscedastic regression with deep neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Domokos M. Kelen, Ádám Jung, Péter Kersch, and Andras A Benczur. Distribution-free data uncertainty for neural network regression. In *The Thirteenth International Conference on Learning Representations*, 2025.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. arXiv:1412.6980.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep Bayesian active learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Thomas Kneib, Alexander Silbersdorff, and Benjamin Säfken. Rage Against the Mean – A Review of Distributional Regression Approaches. *Econometrics and Statistics*, 26:99–123, April 2023. doi:10.1016/j.ecosta.2021.07.006.

Nikita Kotelevskii, Aleksandr Artemenkov, Kirill Fedyanin, Fedor Noskov, Alexander Fishkov, Artem Shelmanov, Artem Vazhentsev, Aleksandr Petiushko, and Maxim Panov. Nonparametric uncertainty quantification for single deterministic neural network. In *Advances in Neural Information Processing Systems*, volume 35, pages 36308–36323. Curran Associates, Inc., 2022.

Nikita Kotelevskii, Vladimir Kondratyev, Martin Takáč, Eric Moulines, and Maxim Panov. From risk to uncertainty: Generating predictive uncertainty measures via bayesian estimation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Fabian Krüger and Ingmar Nolte. Disagreement versus uncertainty: Evidence from distribution forecasts. *Journal of Banking & Finance*, 72(S):172–186, 2016. doi:10.1016/j.jbankfin.2015.05.007.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim G. J. Rudner, C. Bayan Bruss, and Andrew Gordon Wilson. Position: Supervised classifiers answer the wrong questions for OOD detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 81689–81713. PMLR, 2025.

Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Timo Löhr, Michael Ingrisch, and Eyke Hüllermeier. Towards aleatoric and epistemic uncertainty in medical image classification. In *International Conference on Artificial Intelligence in Medicine*, pages 145–155. Springer, 2024.

Osama Makansi, Eddy Ilg, Ozgun Cicek, and Thomas Brox. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021.

Andrey Malinin, Sergey Chervontsev, Ivan Provilkov, and Mark Gales. Regression prior networks, 2020. arXiv:2006.11590.

Nis Meinert and Alexander Lavin. Multivariate deep evidential regression, 2022. arXiv:2104.06135.

Rhiannon Michelmore, Marta Kwiatkowska, and Yarin Gal. Evaluating uncertainty quantification in end-to-end autonomous driving control, 2018. arXiv:1811.06817.

Thomas Muschinski, Georg J. Mayr, Thorsten Simon, Nikolaus Umlauf, and Achim Zeileis. Cholesky-based multivariate Gaussian regression. *Econometrics and Statistics*, 29:261–281, January 2024. doi:10.1016/j.ecosta.2022.03.001.

Vu-Linh Nguyen, Mohammad Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111, 01 2022. doi:10.1007/s10994-021-06003-9.

D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60 vol.1, 1994. doi:10.1109/ICNN.1994.374138.

Lorenzo Pacchiardi, Rilwan A. Adewoyin, Peter Dueben, and Ritabrata Dutta. Probabilistic forecasting with generative networks via scoring rule minimization. *Journal of Machine Learning Research*, 25(45):1–64, 2024.

P. B. Patnaik. The non-central χ2- and F-distribution and their applications. *Biometrika*, 36(1/2):202–232, 1949. ISSN 00063444, 14643510. URL `http://www.jstor.org/stable/2332542`.

R. Pic, C. Dombry, P. Naveau, and M. Taillardat. Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation. *Advances in Statistical Climatology, Meteorology and Oceanography*, 11(1): 23–58, 2025. doi:10.5194/ascmo-11-23-2025.

Janis Postels, Hermann Blum, Yannick Strümpler, Cesar Cadena, Roland Siegwart, Luc Van Gool, and Federico Tombari. The hidden uncertainty in a neural networks activations, 2021. arXiv:2012.03082.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044): 84–90, January 2025. doi:10.1038/s41586-024-08252-9.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 11 2005. doi:10.7551/mitpress/3206.001.0001.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6), 2024. doi:10.1029/2023MS004019.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

Yusuf Sale, Paul Hofman, Lisa Wimmer, Eyke Hüllermeier, and Thomas Nagler. Second-order uncertainty quantification: Variance-based measures, 2023. arXiv:2401.00276.

Yusuf Sale, Viktor Bengs, Michele Caprio, and Eyke Hüllermeier. Second-order uncertainty quantification: A distance-based approach. In *Forty-first International Conference on Machine Learning*, 2024.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4): 1321 – 1334, 2015. doi:10.1175/MWR-D-14-00269.1.

Kajetan Schweighofer, Lukas Aichberger, Mykyta Ielanskyi, and Sepp Hochreiter. Introducing an improved information-theoretic measure of predictive uncertainty, 2023. arXiv:2311.08309.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, October 2013. doi:10.1214/13-AOS1140.

Moshe Shaked and J. Shanthikumar. *Stochastic Orders*. 01 2007. doi:10.1007/978-0-387-34675-5.

Xinwei Shen and Nicolai Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):653–677, 11 2024. doi:10.1093/jrsssb/qkae108.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. volume 7576, pages 746–760, 10 2012. doi:10.1007/978-3-642-33715-4_54.

Ingo Steinwart and Johanna F. Ziegel. Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis*, 51:510–542, March 2021. doi:10.1016/j.acha.2019.11.005.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1596–1611. Curran Associates, Inc., 2022.

Matias Valdenegro-Toro and Daniel Saromo Mori. A Deeper Look into Aleatoric and Epistemic Uncertainty Disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1508–1516, Los Alamitos, CA, USA, June 2022. IEEE Computer Society. doi:10.1109/CVPRW56347.2022.00157.

S.V.N. Vishwanathan, Nicol N. Schraudolph, Risi Kondor, and Karsten M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11(40):1201–1242, 2010.

Kartik Waghmare and Johanna Ziegel. Proper scoring rules for estimation and forecast evaluation. *Annual Review of Statistics and Its Application*, 13:271–296, 2026. doi:10.1146/annurev-statistics-042424-050626.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2282–2292. PMLR, 2023.

Andreas Winkelbauer. Moments and absolute moments of the normal distribution, 2014. arXiv:1209.4340.

Julia Wolleb, Robin Sandkuehler, Florentin Bieder, Philippe Valmaggia, and Philippe C. Cattin. Diffusion models for implicit image segmentation ensembles. In *Medical Imaging with Deep Learning*, 2022.

George Wynne and Andrew B. Duncan. A kernel two-sample test for functional data. *Journal of Machine Learning Research*, 23(73):1–51, 2022.

Johanna Ziegel, David Ginsbourger, and Lutz Dümbgen. Characteristic kernels on Hilbert spaces, Banach spaces, and on sets of measures. *Bernoulli*, 30(2):1441–1457, May 2024. doi:10.3150/23-BEJ1639.

David Zwicker. py-pde: A Python package for solving partial differential equations. *Journal of Open Source Software*, 5(48):2158, April 2020. doi:10.21105/joss.02158.

Uncertainty Quantification for Regression: A Unified Framework based on kernel scores (Supplementary Material)

Christopher Bülte[1,2] **Yusuf Sale**[1,2] **Gitta Kutyniok**[1,2,3,4] **Eyke Hüllermeier**[1,2,5]

[1] LMU Munich
[2]Munich Center for Machine Learning (MCML)
[3]DLR-German Aerospace Center
[4]University of Tromsø
[5]German Research Center for Artificial Intelligence (DFKI, DSA)

A PROOFS

A.1 PROOFS OF PROPOSITIONS 5.1 - 5.3

Proof of Proposition 5.1. Here we prove that for any proper scoring rule S, it holds that

1. $Q = \delta_{\mathbb{P}} \implies \mathrm{EU}(Q) = 0$, while for a *strictly* proper scoring rule the converse holds as well,
2. $\mathrm{EU}(\delta_{\mathbb{P}}) \leq \mathrm{EU}(Q_1) \leq \mathrm{EU}(Q_2)$.

1. Consider the BMA estimator. For $Q = \delta_{\mathbb{P}}$ we have $\overline{\mathbb{P}} = \mathbb{P}$ and $\mathrm{EU}(Q) = \mathbb{E}_{\mathbb{P}\sim Q}[D(\overline{\mathbb{P}}, \mathbb{P})] = D(\mathbb{P}, \mathbb{P}) = 0$, since D is a divergence. For a strictly proper scoring rule, we obtain

$$\mathrm{EU}(Q) = \mathbb{E}_{\mathbb{P}\sim Q}[D(\overline{\mathbb{P}}, \mathbb{P})] = 0 \implies \overline{\mathbb{P}} = \mathbb{E}_Q[\mathbb{P}] = \mathbb{P} \implies Q = \delta_{\mathbb{P}}.$$

For the pairwise estimator, the proof works in an analogous way.

2 (BMA). The lower bound follows immediately from the nonnegativity of the divergence D and the first part of the proposition being fulfilled for a proper scoring rule. Furthermore, we are given $Q_1 \leq_{\mathrm{cx}}^2 Q_2$ and $\mathrm{EU}(Q) = \mathbb{E}_{\mathbb{P}\sim Q}[D(\overline{\mathbb{P}}, \mathbb{P})]$. Recall that for a scoring rule with $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{Y})$, the divergence is given as $D(\mathbb{P}, \mathbb{Q}) = S(\mathbb{P}, \mathbb{Q}) - S(\mathbb{Q}, \mathbb{Q})$. We want to show that

$$\mathrm{EU}(Q_1) = \mathbb{E}_{\mathbb{P}\sim Q_1}[D(\overline{\mathbb{P}}, \mathbb{P})] \leq \mathbb{E}_{\mathbb{P}\sim Q_2}[D(\overline{\mathbb{P}}, \mathbb{P})] = \mathrm{EU}(Q_2).$$

We will show that $D(\overline{\mathbb{P}}, \mathbb{P})$ is a convex functional in $\mathbb{P}$. Then, by definition of the convex order, it follows that $\mathrm{EU}(Q_1) \leq \mathrm{EU}(Q_2)$.

First, note that by definition of the convex order we have a fixed $\overline{\mathbb{P}} = \mathbb{E}_{\mathbb{P}\sim Q_1}[\mathbb{P}] = \mathbb{E}_{\mathbb{P}\sim Q_2}[\mathbb{P}]$. By definition of proper scoring rules, the term $S(\mathbb{P}, \mathbb{Q})$ is affine in $\mathbb{Q}$ [Dawid, 2007] and therefore convex. Furthermore, we know that $H(\mathbb{Q}) = S(\mathbb{Q}, \mathbb{Q})$ is a concave function in $\mathbb{Q}$ [Waghmare and Ziegel, 2026] and therefore $-H(\mathbb{Q})$ is convex. In total, $D(\overline{\mathbb{P}}, \mathbb{P})$ consists of an affine function plus a convex function in $\mathbb{P}$ and is therefore also convex in $\mathbb{P}$ [Boyd and Vandenberghe, 2004].

2 (Pairwise). For the pairwise estimator, we require the additional assumption that for a fixed $\mathbb{Q}$, the map $\mathbb{P} \mapsto S(\mathbb{P}, \mathbb{Q})$ is convex, which is fulfilled by kernel scores or scoring rules of Bregman type. First, write $F(\mathbb{P}, \mathbb{P}') := D(\mathbb{P}, \mathbb{P}') = S(\mathbb{P}, \mathbb{P}') - H(\mathbb{P}')$.

By the convexity assumption, for fixed $\mathbb{P}'$, the map $\mathbb{P} \mapsto F(\mathbb{P}, \mathbb{P}')$ is convex. Furthermore, since $S(\mathbb{P}, \mathbb{Q})$ is affine in $\mathbb{Q}$ and $H(\mathbb{P})$ is concave [Dawid, 2007], for fixed $\mathbb{P}$, the map $\mathbb{P}' \mapsto F(\mathbb{P}, \mathbb{P}')$ is affine + convex, hence convex.

For every fixed $\mathbb{P}'$, we obtain the following via the convex order

$$\mathbb{E}_{\mathbb{P}\sim Q_1} F(\mathbb{P}, \mathbb{P}') \;\leq\; \mathbb{E}_{\mathbb{P}\sim Q_2} F(\mathbb{P}, \mathbb{P}').$$

Integrating over $\mathbb{P}' \sim Q_1$ gives

$$\mathbb{E}_{\mathbb{P},\mathbb{P}'\sim Q_1} F(\mathbb{P},\mathbb{P}') \;\leq\; \mathbb{E}_{\mathbb{P}'\sim Q_1}\mathbb{E}_{\mathbb{P}\sim Q_2} F(\mathbb{P},\mathbb{P}').$$

Similarly, for every fixed $\mathbb{P}$, we obtain

$$\mathbb{E}_{\mathbb{P}'\sim Q_1} F(\mathbb{P},\mathbb{P}') \;\leq\; \mathbb{E}_{\mathbb{P}'\sim Q_2} F(\mathbb{P},\mathbb{P}').$$

Integrating over $\mathbb{P} \sim Q_2$ gives

$$\mathbb{E}_{\mathbb{P}\sim Q_2}\mathbb{E}_{\mathbb{P}'\sim Q_1} F(\mathbb{P},\mathbb{P}') \;\leq\; \mathbb{E}_{\mathbb{P},\mathbb{P}'\sim Q_2} F(\mathbb{P},\mathbb{P}').$$

Since both sides coincide (by Fubini's theorem), we ultimately get

$$\mathbb{E}_{\mathbb{P},\mathbb{P}'\sim Q_1} F(\mathbb{P},\mathbb{P}') \;\leq\; \mathbb{E}_{\mathbb{P},\mathbb{P}'\sim Q_2} F(\mathbb{P},\mathbb{P}'),$$

i.e.

$$\mathrm{EU}_P(Q_1) \;\leq\; \mathrm{EU}_P(Q_2).$$

□

Proof of Proposition 5.2. Here we prove that any kernel score S_k with a translation invariant kernel $k(x, x')$ that is convex in one of its arguments fulfills $\mathrm{AU}(\delta_{\mathbb{P}_1}) \leq \mathrm{AU}(\delta_{\mathbb{P}_2})$.

We know by assumption that $\mathbb{P}_1 \leq_{\mathrm{cx}} \mathbb{P}_2$ and $\mathrm{AU}(\delta_{\mathbb{P}}) = H(\mathbb{P})$. Therefore, we need to show that $H(\mathbb{P}_1) \leq H(\mathbb{P}_2)$. Recall that for any translation invariant kernel score we have $k(x, x') = \psi(x - x')$ for some $\psi : \mathcal{Y} \to \mathbb{R}$ and the corresponding entropy is given as

$$H(\mathbb{P}) = \frac{1}{2}\mathbb{E}_{X,X'\sim\mathbb{P}}[k(X - X')] - \frac{1}{2}\mathbb{E}_{X\sim\mathbb{P}}[\underbrace{k(X - X)}_{\equiv k(0)}],$$

where the last part is a constant, due to the translation invariance, and therefore does not affect the inequality. Now define $\phi_P(x) \coloneqq \mathbb{E}_{X'\sim\mathbb{P}}[\psi(x - X')]$, which is convex in x, since ψ is convex and linearity in expectation preserves convexity.

Now, using convex order, we have

$$\mathbb{E}_{X\sim\mathbb{P}_1}[\phi_{\mathbb{P}_1}(X)] \leq \mathbb{E}_{Y\sim\mathbb{P}_2}[\phi_{\mathbb{P}_1}(Y)].$$

Similarly, we can also obtain an order for the convex function $\phi_{\mathbb{P}_2}$ as

$$\mathbb{E}_{X\sim\mathbb{P}_1}[\phi_{\mathbb{P}_2}(X)] \leq \mathbb{E}_{Y\sim\mathbb{P}_2}[\phi_{\mathbb{P}_2}(Y)].$$

Now, note that using Fubini's theorem, we obtain

$$\mathbb{E}_{X\sim\mathbb{P}_1}[\phi_{\mathbb{P}_2}(X)] = \mathbb{E}_{Y\sim\mathbb{P}_2}[\phi_{\mathbb{P}_1}(Y)] = \mathbb{E}_{U\sim\mathbb{P}_1,V\sim\mathbb{P}_2}[\psi(U - V)].$$

Therefore, we obtain

$$\mathbb{E}_{X\sim\mathbb{P}_1}[\phi_{\mathbb{P}_1}(X)] \leq \mathbb{E}_{Y\sim\mathbb{P}_2}[\phi_{\mathbb{P}_2}(Y)],$$

and therefore

$$\mathrm{AU}(\delta_{\mathbb{P}_1}) = H(\mathbb{P}_1) \leq H(\mathbb{P}_2) = \mathrm{AU}(\delta_{\mathbb{P}_2}).$$

□

Proof of Proposition 5.3. Write $\kappa \coloneqq \|k\|_\infty$ and, for $\mathbb{P}, \mathbb{P}' \in \mathcal{P}(\mathcal{Y})$, set $a(\mathbb{P}, \mathbb{P}') \coloneqq \mathbb{E}_{X\sim\mathbb{P},\, X'\sim\mathbb{P}'}[k(X, X')]$. The form a is symmetric, satisfies $|a(\mathbb{P}, \mathbb{P}')| \leq \kappa$, and is linear in each argument under a mixture representation. Since sampling from a mixture means sampling its component first, we have $a\big(\int \mathbb{P}_{\boldsymbol{\theta}}\, dQ(\boldsymbol{\theta}),\, \mathbb{P}'\big) = \int a(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{P}')\, dQ(\boldsymbol{\theta})$. The entropy and divergence read

$$H_k(\mathbb{P}) = \tfrac{1}{2}a(\mathbb{P}, \mathbb{P}) - \tfrac{1}{2}\mathbb{E}_{X\sim\mathbb{P}}[k(X, X)], \qquad D_k(\mathbb{P}, \mathbb{Q}) = a(\mathbb{P}, \mathbb{Q}) - \frac{1}{2}a(\mathbb{P}, \mathbb{P}) - \frac{1}{2}a(\mathbb{Q}, \mathbb{Q}),$$

where $D_k \geq 0$ as the divergence of a (strictly) proper score, and $D_k(\mathbb{P}, \mathbb{P}) = 0$. From $|a| \leq \kappa$ we get $|H_k(\mathbb{P})| \leq \kappa$ and $0 \leq D_k(\mathbb{P}, \mathbb{P}') \leq 2\kappa$, hence $\mathrm{AU}(Q), \mathrm{EU}(Q) < \infty$.

Aleatoric. As $\mathrm{AU}(Q) = \mathbb{E}_{\boldsymbol{\vartheta}\sim Q}[H_k(\mathbb{P}_{\boldsymbol{\vartheta}})]$ is linear in Q, we have $\mathrm{AU}(Q_\varepsilon) = (1-\varepsilon)\,\mathrm{AU}(Q) + \varepsilon H_k(\mathbb{P}_{\boldsymbol{\theta}_0})$, so

$$\mathrm{IF}(\boldsymbol{\theta}_0; \mathrm{AU}, Q) = H_k(\mathbb{P}_{\boldsymbol{\theta}_0}) - \mathbb{E}_{\boldsymbol{\vartheta}\sim Q}[H_k(\mathbb{P}_{\boldsymbol{\vartheta}})], \qquad \sup_{\boldsymbol{\theta}_0\in\Theta} |\mathrm{IF}(\boldsymbol{\theta}_0; \mathrm{AU}, Q)| \le 2\kappa.$$

Epistemic, pairwise. With $\phi(\boldsymbol{\theta}, \boldsymbol{\theta}') \coloneqq D_k(\mathbb{P}_{\boldsymbol{\theta}'}, \mathbb{P}_{\boldsymbol{\theta}})$ we have $\mathrm{EU}_\mathrm{P}(Q) = \mathbb{E}_{Q\otimes Q}[\phi]$ and $\phi(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) = 0$. Substituting $Q_\varepsilon \otimes Q_\varepsilon$,

$$\mathrm{EU}_\mathrm{P}(Q_\varepsilon) = (1-\varepsilon)^2\,\mathrm{EU}_\mathrm{P}(Q) + 2\varepsilon(1-\varepsilon)\mathbb{E}_{\boldsymbol{\vartheta}\sim Q}[D_k(\mathbb{P}_{\boldsymbol{\theta}_0}, \mathbb{P}_{\boldsymbol{\vartheta}})],$$

so using $0 \le \mathbb{E}_{\boldsymbol{\vartheta}\sim Q}[D_k(\mathbb{P}_{\boldsymbol{\theta}_0}, \mathbb{P}_{\boldsymbol{\vartheta}})], \mathrm{EU}_\mathrm{P}(Q) \le 2\kappa$, we obtain

$$\mathrm{IF}(\boldsymbol{\theta}_0; \mathrm{EU}_\mathrm{P}, Q) = 2\big(\mathbb{E}_{\boldsymbol{\vartheta}\sim Q}[D_k(\mathbb{P}_{\boldsymbol{\theta}_0}, \mathbb{P}_{\boldsymbol{\vartheta}})], -\mathrm{EU}_\mathrm{P}(Q)\big), \qquad \sup_{\boldsymbol{\theta}_0\in\Theta} |\mathrm{IF}(\boldsymbol{\theta}_0; \mathrm{EU}_\mathrm{P}, Q)| \le 4\kappa.$$

Epistemic, BMA. Set $A \coloneqq \mathbb{E}_{\boldsymbol{\vartheta}\sim Q}[a(\mathbb{P}_{\boldsymbol{\vartheta}}, \mathbb{P}_{\boldsymbol{\vartheta}})]$ and $\bar{a} \coloneqq \mathbb{E}_{\boldsymbol{\vartheta},\boldsymbol{\vartheta}'\sim Q}[a(\mathbb{P}_{\boldsymbol{\vartheta}}, \mathbb{P}_{\boldsymbol{\vartheta}'})]$. Expanding the divergence form,

$$\mathrm{EU}_\mathrm{P}(Q) = \mathbb{E}_{\boldsymbol{\vartheta},\boldsymbol{\vartheta}'}\big[a(\mathbb{P}_{\boldsymbol{\vartheta}}, \mathbb{P}_{\boldsymbol{\vartheta}'}) - \frac{1}{2}(\mathbb{P}_{\boldsymbol{\vartheta}}, \mathbb{P}_{\boldsymbol{\vartheta}}) - \frac{1}{2}a(\mathbb{P}_{\boldsymbol{\vartheta}'}, \mathbb{P}_{\boldsymbol{\vartheta}'})\big] = \bar{a} - A.$$

For the mixture $\overline{\mathbb{P}} = \int \mathbb{P}_{\boldsymbol{\theta}}\, dQ(\boldsymbol{\theta})$, linearity of a gives $a(\overline{\mathbb{P}}, \mathbb{P}_{\boldsymbol{\vartheta}}) = \mathbb{E}_{\boldsymbol{\vartheta}'}[a(\mathbb{P}_{\boldsymbol{\vartheta}'}, \mathbb{P}_{\boldsymbol{\vartheta}})]$ and $a(\overline{\mathbb{P}}, \overline{\mathbb{P}}) = \bar{a}$, hence

$$\mathrm{EU}_\mathrm{B}(Q) = \mathbb{E}_{\boldsymbol{\vartheta}}\big[D_k(\overline{\mathbb{P}}, \mathbb{P}_{\boldsymbol{\vartheta}})\big] = \mathbb{E}_{\boldsymbol{\vartheta}}\big[\mathbb{E}_{\boldsymbol{\vartheta}'}[a(\mathbb{P}_{\boldsymbol{\vartheta}'}, \mathbb{P}_{\boldsymbol{\vartheta}})] - \frac{1}{2}\bar{-}\frac{1}{2}a(\mathbb{P}_{\boldsymbol{\vartheta}}, \mathbb{P}_{\boldsymbol{\vartheta}})\big] = \bar{a} - \frac{1}{2}\bar{a} + \frac{1}{2}A = \frac{1}{2}(A - \bar{a}) = \tfrac{1}{2}\,\mathrm{EU}_\mathrm{P}(Q).$$

Thus $\mathrm{IF}(\boldsymbol{\theta}_0; \mathrm{EU}_\mathrm{B}, Q) = \frac{1}{2}\,\mathrm{IF}(\boldsymbol{\theta}_0; \mathrm{EU}_\mathrm{P}, Q)$, bounded by 2κ.

In all cases $\sup_{\boldsymbol{\theta}_0\in\Theta} |\mathrm{IF}(\boldsymbol{\theta}_0; M, Q)| \le C_M\,\kappa < \infty$ for a finite constant C_M, so M is robust in terms of the influence function. □

A.2 ADDITIONAL PROPOSITIONS FOR EXISTING MEASURES

Here, we introduce and prove two more propositions regarding the variance- and entropy-based measures.

Proposition A.1. *The variance-based measure (squared error) does not fulfill point 1 of Proposition 5.1.*

Proof. Consider the BMA estimator, two first-order Gaussian distribution, e.g. $\mathbb{P}_1 = \mathcal{N}(0, \sigma_1^2), \mathbb{P}_2 = \mathcal{N}(0, \sigma_2^2)$ with $\sigma_1^2 \neq \sigma_2^2$ and a second-order distribution, specified as a Dirac mixture, i.e. $Q = \frac{1}{2}\delta_{\mathbb{P}_1} + \frac{1}{2}\delta_{\mathbb{P}_2}$. Recall that for the variance-based measure, we have $D(\mathbb{P}, \mathbb{Q}) = (\mathbb{E}_{Y\sim\mathbb{P}}[Y] - \mathbb{E}_{Y'\sim\mathbb{Q}}[Y'])^2$. In addition, we obtain $\overline{\mathbb{P}} = \frac{1}{2}\mathbb{P}_1 + \frac{1}{2}\mathbb{P}_2$ and $\mathbb{E}_{Y'\sim\overline{P}}[Y'] = 0$. Then we obtain

$$\begin{aligned}\mathrm{EU}(Q) &= \mathbb{E}_{\mathbb{P}\sim Q}[D(\overline{\mathbb{P}}, \mathbb{P})] = \mathbb{E}_{\mathbb{P}\sim Q}[(\underbrace{\mathbb{E}_{Y'\sim\overline{\mathbb{P}}}[Y']}_{=0} - \mathbb{E}_{Y\sim\mathbb{P}}[Y])^2] = \mathbb{E}_{\mathbb{P}\sim Q}[(\mathbb{E}_{Y\sim\mathbb{P}}[Y])^2] \\ &= \frac{1}{2}\mathbb{E}_{\mathbb{P}_1\sim Q}[(\underbrace{\mathbb{E}_{Y\sim\mathbb{P}_1}[Y]}_{=0})^2] + \frac{1}{2}\mathbb{E}_{\mathbb{P}_2\sim Q}[(\underbrace{\mathbb{E}_{Y\sim\mathbb{P}_2}[Y]}_{=0})^2] = 0.\end{aligned}$$

Therefore, we obtain $\mathrm{EU}(Q) = 0$ although $Q \neq \delta_{\mathbb{P}}$. The same argument also works for the pairwise estimator. □

Proposition A.2. *Assume that $\mathbb{P}_1, \mathbb{P}_2$ are absolutely continuous with respect to the Lebesgue measure and therefore admit a probability density. Further assume that both densities are log-concave Then, the entropy-based measure (log-score) fulfills* $\mathrm{AU}(\delta_{\mathbb{P}_1}) \le \mathrm{AU}(\delta_{\mathbb{P}_2})$.

Proof. A probability distribution has log-concave density if the density can be expressed as $p(x) \equiv \exp(\varphi(x))$ for a concave function $\varphi(x)$. Recall that the log-score corresponds to the differential entropy, which can be expressed as

$$H(\mathbb{P}) = -\int p(x)\log p(x) d\mu(x) = \mathbb{E}_{\mathbb{P}}[-\log p(X)].$$

Then, for a log-concave density, we have that $\phi(x) \coloneqq -\log p_2(x)$ is a convex function in x. By convex order, we then have

$$\mathbb{E}_{X\sim\mathbb{P}_1}[-\log p_2(X)] = \mathbb{E}_{X\sim\mathbb{P}_1}[\phi(X)] \le \mathbb{E}_{Y\sim\mathbb{P}_2}[\phi(Y)] = H(\mathbb{P}_2).$$

The left-hand side is the cross-entropy of $\mathbb{P}_1, \mathbb{P}_2$, which, by definition, can be decomposed into

$$\mathbb{E}_{X\sim\mathbb{P}_1}[-\log p_2(X)] = H(\mathbb{P}_1) + D_{\mathrm{KL}}(\mathbb{P}_1\|\mathbb{P}_2) \geq H(\mathbb{P}_1),$$

where the inequality follows from the KL-divergence being nonnegative. Combining the above gives

$$H(\mathbb{P}_1) \leq \mathbb{E}_{X\sim\mathbb{P}_1}[-\log p_2(X)] = \mathbb{E}_{X\sim\mathbb{P}_1}[\phi(X)] \leq \mathbb{E}_{Y\sim\mathbb{P}_2}[\phi(Y)] = H(\mathbb{P}_2),$$

and therefore

$$\mathrm{AU}(\delta_{\mathbb{P}_1}) = H(\mathbb{P}_1) \leq H(\mathbb{P}_2) = \mathrm{AU}(\delta_{\mathbb{P}_2}).$$

□

B DERIVATION OF MEASURES FOR SPECIFIC CHOICES OF SCORING RULES

In this section, we derive expressions for the (generalized) entropy- and divergence term of the uncertainty measures introduced in this article. Recall that in order to assess EU, AU and TU, one requires expressions for the entropy, divergence and expected scoring rule. This is regardless whether one chooses the pairwise or the BMA estimator. Therefore, for $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ and $X, X' \sim \mathbb{P},\ Y, Y' \sim \mathbb{Q}$, and $\mathbb{P}, \mathbb{P}' \sim Q,\ \overline{\mathbb{P}} = \mathbb{E}_Q[\mathbb{P}]$, we will derive the quantities $H(\mathbb{P}), D(\mathbb{P}, \mathbb{Q})$, as well as the gap between the BMA and pairwise estimation Δ, for different scoring rules.

Log-score Let $\mathcal{P}$ be the set of distributions on $\mathcal{Y}$ that are absolutely continuous with respect to the Lebesgue measure μ and $\mathbb{P}, \mathbb{Q} \in \mathcal{P}$ with corresponding densities p, q. The *logarithmic score* $S_{\log} : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$, given by

$$S_{\log}(\mathbb{P}, \boldsymbol{y}) = -\log p(\boldsymbol{y})$$

is a strictly proper scoring rule. The associated entropy and divergence are given as

$$\begin{aligned} H_{\log}(\mathbb{P}) &= -\int p(\boldsymbol{x}) \log p(\boldsymbol{x})\, d\mu(\boldsymbol{x}), \\ D_{\log}(\mathbb{P}, \mathbb{Q}) &= \int q(\boldsymbol{y}) \log\left(\frac{q(\boldsymbol{y})}{p(\boldsymbol{y})}\right) d\mu(\boldsymbol{y}) = D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}), \end{aligned}$$

which are the Shannon entropy and Kullback-Leibler divergence, respectively. Utilizing the BMA estimator, we obtain the entropy-based measure, while for the pairwise estimator we obtain the pairwise KL-divergence, as shown by Schweighofer et al. [2023]. For their difference, we obtain the so-called reverse mutual information

$$\Delta = \mathbb{E}_Q\left[D_{\mathrm{KL}}\left(\overline{\mathbb{P}}\|\mathbb{P}\right)\right].$$

Kernel score Consider the kernel score $S_k : \mathcal{P}_k \times \mathcal{Y}$ associated with a negative definite kernel k. We obtain the following expressions for the pairwise estimator:

$$\begin{aligned} H(\mathbb{P}) &= \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[k(X, X')\right] - \frac{1}{2}\mathbb{E}_{\mathbb{P}}[k(X, X)], \\ D(\mathbb{P}, \mathbb{Q}) &= \mathbb{E}_{\mathbb{P},\mathbb{Q}}\left[k(X, Y)\right] - \frac{1}{2}\mathbb{E}_{\mathbb{P}}\left[k(X, X')\right] - \frac{1}{2}\mathbb{E}_{\mathbb{Q}}\left[k(Y, Y')\right]. \end{aligned}$$

The corresponding uncertainty measures are obtained by plugging the selected kernel into the above quantities.

Squared error Let $\mathcal{P}$ be the set of distributions on $\mathcal{Y} \subseteq \mathbb{R}^p$ such that $\int \|\boldsymbol{x}\|^2\, d\mathbb{P}(\boldsymbol{x}) < \infty$ and $Y \sim \mathbb{P} \in \mathcal{P}(\mathcal{Y})$. The squared error $S_{\mathrm{SE}} : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ given by

$$S_{\mathrm{SE}}(\mathbb{P}, \boldsymbol{y}) = (\boldsymbol{y} - \mathbb{E}_{\mathbb{P}}[Y])^2,$$

is a proper (but not strictly proper) kernel rule, with $k(\boldsymbol{x}, \boldsymbol{x}') = \|\boldsymbol{x} - \boldsymbol{x}'\|^2$. The associated entropy and divergence are given as

$$H_{\mathrm{SE}}(\mathbb{P}) = \mathrm{tr}(\mathrm{Cov}_{\mathbb{P}}[Y]), \qquad D_{\mathrm{SE}}(\mathbb{P}, \mathbb{Q}) = \|\boldsymbol{\mu}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{Q}}\|^2.$$

In the case of the squared error, the corresponding uncertainty measures can be expressed in terms of moments of the first-order distribution, leading to the following measures for the BMA estimator:

$$\begin{aligned}
\mathrm{AU}_B(Q) &= \mathbb{E}_Q\left[\mathrm{tr}(\mathrm{Cov}_{\mathbb{P}}[Y])\right], \\
\mathrm{EU}_B(Q) &= \mathbb{E}_Q\left[\left\|\boldsymbol{\mu}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}'}\right\|^2\right] = \mathrm{tr}\left(\mathrm{Cov}_Q[\boldsymbol{\mu}_{\mathbb{P}}]\right), \\
\mathrm{TU}_B(Q) &= \mathbb{E}_Q\left[\left\|Y - \mathbb{E}_Q[\boldsymbol{\mu}_{\mathbb{P}}]\right\|^2\right],
\end{aligned}$$

which reduces to the variance-based decomposition in the univariate case $\mathcal{Y} \subseteq \mathbb{R}$. For the pairwise estimator, we obtain

$$\begin{aligned}
\mathrm{AU}_P(Q) &= \mathbb{E}_Q\left[\mathrm{tr}(\mathrm{Cov}_{\mathbb{P}}[Y])\right], \\
\mathrm{EU}_P(Q) &= 2\mathbb{E}_Q\left[\left\|\boldsymbol{\mu}_{\mathbb{P}} - \boldsymbol{\mu}_{\mathbb{P}'}\right\|^2\right] = 2\mathrm{tr}\left(\mathrm{Cov}_Q[\boldsymbol{\mu}_{\mathbb{P}}]\right), \\
\mathrm{TU}_P(Q) &= \mathbb{E}_Q\left[\left\|Y - \mathbb{E}_Q[\boldsymbol{\mu}_{\mathbb{P}}]\right\|^2\right] + \mathrm{tr}\left(\mathrm{Cov}_Q[\boldsymbol{\mu}_{\mathbb{P}}]\right),
\end{aligned}$$

which shows that both estimators only differ by a factor of two for the epistemic uncertainty. The gap between both estimators is

$$\Delta = \mathrm{tr}\left(\mathrm{Cov}_Q[\boldsymbol{\mu}_{\mathbb{P}}]\right) = \mathbb{E}_Q[D_{\mathrm{SE}}(\overline{\mathbb{P}}, \mathbb{P})].$$

This quantity measures the expected (score-) divergence between the BMA against all possible models.

B.1 CLOSED-FORM EXPRESSIONS FOR GAUSSIANS

Here, we derive closed-form expressions for the entropy and divergence term of different scoring rules for first-order (univariate) Gaussian and mixture of Gaussian distributions. Recall that for kernel scores S_k with a conditionally negative definite kernel k, the entropy and divergence of two probability measures $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ are given as

$$H_k(\mathbb{P}) = \frac{1}{2}\mathbb{E}_{X,X'\sim\mathbb{P}}[k(X, X')] - \frac{1}{2}\mathbb{E}_{X\sim\mathbb{P}}[k(X, X)] \tag{8}$$

$$D_k(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{X\sim\mathbb{P},Y\sim\mathbb{Q}}[k(X, Y)] - \frac{1}{2}\mathbb{E}_{X,X'\sim\mathbb{P}}[k(X, X')] - \frac{1}{2}\mathbb{E}_{Y,Y'\sim\mathbb{Q}}[k(Y, Y')]. \tag{9}$$

Consider two first-order Gaussian distributions $X \sim \mathbb{P} = \mathcal{N}(\mu, \sigma^2),\ Y \sim \mathbb{Q} = \mathcal{N}(\nu, \tau^2)$. Then we obtain the following expressions:

Log-score

$$H(\mathbb{P}) = \frac{1}{2}\log(2\pi e\sigma^2), \tag{10}$$

$$D(\mathbb{P}, \mathbb{Q}) = \log\left(\frac{\sigma}{\tau}\right) + \frac{\tau^2 + (\mu - \mu)^2}{2\sigma^2} - \frac{1}{2}. \tag{11}$$

These expressions are obtained via well-known results from the differential entropy and KL-divergence for Gaussian distributions (compare for example Rasmussen and Williams [2005]).

Squared error

$$H(\mathbb{P}) = \sigma^2, \tag{12}$$

$$D(\mathbb{P}, \mathbb{Q}) = (\mu - \nu)^2. \tag{13}$$

Proof. For the entropy, we obtain

$$H(\mathbb{P}) = \frac{1}{2}\mathbb{E}_{X,X'\sim\mathbb{P}}[(X - X')^2)] = \frac{1}{2}\left(\mathbb{E}_{\mathbb{P}}[X^2] - 2\mathbb{E}_{\mathbb{P}}[X]\mathbb{E}_{\mathbb{P}}[X'] + \mathbb{E}_{\mathbb{P}}[X'^2]\right) = \mathbb{V}_{\mathbb{P}}[X] = \sigma^2.$$

In addition, we have that $\mathbb{E}_{X\sim\mathbb{P},Y\sim\mathbb{Q}}[(X-Y)^2] = \mathbb{E}_{\mathbb{P}}[X^2] - 2\mathbb{E}_{\mathbb{P}}[X]\mathbb{E}_{\mathbb{Q}}[Y] + \mathbb{E}_{\mathbb{Q}}[Y^2]$ such that for the divergence we obtain

$$\begin{aligned}
D(\mathbb{P},\mathbb{Q}) &= \mathbb{E}_{\mathbb{P}}[X^2] - 2\mathbb{E}_{\mathbb{P}}[X]\mathbb{E}_{\mathbb{Q}}[Y] + \mathbb{E}_{\mathbb{Q}}[Y^2] - \mathbb{V}_{\mathbb{P}}[X] - \mathbb{V}_{\mathbb{Q}}[Y] \\
&= \mathbb{E}_{\mathbb{P}}[X^2] - 2\mathbb{E}_{\mathbb{P}}[X]\mathbb{E}_{\mathbb{Q}}[Y] + \mathbb{E}_{\mathbb{Q}}[Y^2] - \mathbb{E}_{\mathbb{P}}[X^2] + \mathbb{E}_{\mathbb{P}}[X]^2 - \mathbb{E}_{\mathbb{Q}}[Y^2] + \mathbb{E}_{\mathbb{Q}}[Y]^2 \\
&= \mathbb{E}_{\mathbb{P}}[X]^2 - 2\mathbb{E}_{\mathbb{P}}[X]\mathbb{E}_{\mathbb{Q}}[Y] + \mathbb{E}_{\mathbb{Q}}[Y]^2 = (\mathbb{E}_{\mathbb{P}}[X] - \mathbb{E}_{\mathbb{Q}}[Y])^2 \\
&= (\mu-\nu)^2.
\end{aligned}$$

□

CRPS

$$H(\mathbb{P}) = \frac{\sigma}{\sqrt{\pi}}, \tag{14}$$

$$D(\mathbb{P},\mathbb{Q}) = \left(\sqrt{\sigma^2+\tau^2}\right)\frac{\sqrt{2}}{\sqrt{\pi}}{}_1F_1\left(-\frac{1}{2},\frac{1}{2};-\frac{1}{2}\frac{(\mu-\nu)^2}{\sigma^2+\tau^2}\right) - \left(\frac{\sigma+\tau}{\sqrt{\pi}}\right). \tag{15}$$

Proof. Winkelbauer [2014] show that for the raw absolute moment of a Gaussian we have

$$\mathbb{E}[|X|^p] = \sigma^p 2^{p/2}\frac{\Gamma(\frac{p+1}{2})}{\sqrt{\pi}}{}_1F_1\left(-\frac{p}{2},\frac{1}{2};-\frac{\mu^2}{2\sigma^2}\right),$$

where ${}_1F_1$ denotes Kummer's confluent hypergeometric function. Furthermore, we know that $X-Y\sim\mathcal{N}(\mu-\nu,\sigma^2+\tau^2)$, $X-X'\sim\mathcal{N}(0,2\sigma^2)$ and $Y-Y'\sim\mathcal{N}(0,2\tau^2)$. Therefore, we obtain

$$H(\mathbb{P}) = \frac{1}{2}\mathbb{E}_{X,X'\sim\mathbb{P}}[|X-X'|] = \frac{1}{2}\sqrt{2\sigma^2}\sqrt{2}\frac{\Gamma(1)}{\sqrt{\pi}}{}_1F_1\left(-\frac{1}{2},\frac{1}{2};0\right) = \frac{\sigma}{\sqrt{\pi}}.$$

With $\mathbb{E}_{X\sim\mathbb{P},Y\sim\mathbb{Q}}[|X-Y|] = \sqrt{\sigma^2+\tau^2}\frac{\sqrt{2}}{\sqrt{\pi}}{}_1F_1\left(-\frac{1}{2},\frac{1}{2};-\frac{1}{2}\frac{(\mu-\nu)^2}{\sigma^2+\tau^2}\right)$ we obtain the divergence $D(\mathbb{P},\mathbb{Q})$ by plugging in the corresponding expectations. □

Gaussian kernel score Given the (negative) Gaussian kernel $k(x,y) = -\exp(-(x-y)^2/\gamma^2)$ with scalar bandwidth γ, we obtain

$$H(\mathbb{P}) = \frac{1}{2}\left(1 - \frac{\gamma}{\sqrt{\gamma^2+4\sigma^2}}\right) \tag{16}$$

$$D(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\frac{\gamma}{\sqrt{\gamma^2+4\sigma^2}} + \frac{1}{2}\frac{\gamma}{\sqrt{\gamma^2+4\tau^2}} - \frac{\gamma}{\sqrt{\gamma^2+2(\sigma^2+\tau^2)}}\exp\left(-\frac{(\mu-\nu)^2}{\gamma^2+2(\sigma^2+\tau^2)}\right) \tag{17}$$

Proof. Let $Z \coloneqq X-Y\sim\mathbb{P}_Z \coloneqq \mathcal{N}(\delta,\upsilon)$ with $\delta \coloneqq \mu-\nu, \upsilon \coloneqq \sigma^2+\tau^2$. Then $\frac{Z^2}{\delta}$ follows a noncentral chi-squared distribution, i.e. $\frac{Z^2}{\delta}\sim\chi^2(1,\lambda)$ with noncentrality parameter $\lambda = \frac{\delta^2}{\upsilon}$. Furthermore, we have

$$\mathbb{E}_{X\sim\mathbb{P},Y\sim\mathbb{Q}}[k(X,Y)] = -\mathbb{E}_{\mathbb{P}_Z}\left[\exp\left(-\frac{\frac{Z^2}{\upsilon}\upsilon}{\gamma^2}\right)\right] = -M_{\chi^2(1,\lambda)}\left(-\frac{\upsilon}{\gamma^2}\right).$$

Here, $M_{\chi^2(k,\lambda)}(t)$ is the moment-generating function of $\chi^2(k,\lambda)$, with $t = -\frac{\upsilon}{\gamma^2}$, which can be expressed analytically (compare, for example, Patnaik [1949]) as $M_{\chi^2(k,\lambda)}(t) = \frac{\exp\left(\frac{\lambda t}{1-2t}\right)}{(1-2t)^{k/2}}$. Therefore, we obtain

$$\mathbb{E}_{X\sim\mathbb{P},Y\sim\mathbb{Q}}[k(X,Y)] = -\frac{\gamma}{\sqrt{\gamma^2+2(\sigma^2+\tau^2)}}\exp\left(-\frac{(\mu-\nu)^2}{\gamma^2+2(\sigma^2+\tau^2)}\right)$$

and

$$H(\mathbb{P}) = \frac{1}{2}\mathbb{E}_{X,X'\sim\mathbb{P}}[k(X,X')] - \frac{1}{2}\mathbb{E}_{X\sim\mathbb{P}}[k(X,X)]$$
$$= \frac{1}{2}\left(1 - \frac{\gamma}{\sqrt{\gamma^2+4\sigma^2}}\right).$$

By plugging these expressions into the definition of the divergence $D(\mathbb{P},\mathbb{Q})$, we obtain the corresponding closed form. □

Gaussian mixtures Here, we consider a mixture of Gaussians, i.e. $X \sim \mathbb{P} = \sum_{i=1}^{M} w_i \mathcal{N}(\mu_i, \sigma_i^2), Y \sim \mathbb{Q} = \sum_{j=1}^{N} v_j \mathcal{N}(\mu_j, \sigma_j^2)$ with nonnegative weights w_i, v_j that sum to one. For a mixture of Gaussians, closed-form expressions are not necessarily available, as is the case for the log-score. However, for specific cases, closed-form expressions are available via the corresponding marginals. For a translation-invariant kernel score, the expressions for the mixture density network can be derived in terms of the kernel score of the individual components. By linearity of the expectation, we obtain

$$\mathbb{E}[k(X,Y)] = \sum_{i=1}^{M}\sum_{j=1}^{N} w_i v_j \mathbb{E}_{X\sim\mathcal{N}(\mu_i,\sigma_i^2),Y\sim\mathcal{N}(\mu_j,\sigma_j^2)}[k(X,Y)].$$

In the case of a translation invariant kernel, i.e. $k(X,Y) \equiv k(X-Y)$, this reduces to a weighted sum of the corresponding Gaussian score, as we have $X - Y \sim \mathcal{N}(\mu_i - \mu_j, \sigma_i^2 + \sigma_j^2)$. Therefore, we can use the results from the previous section to derive the scores for the Gaussian mixtures analytically.

Marginal scores In the multivariate setting $\mathcal{Y} \subseteq \mathbb{R}^d$ for $d > 1$, closed-form expressions are more difficult to obtain than in the univariate setting. For instance, for a Gaussian distribution, the energy score admits an analytic solution for $d = 1$ but not for $d > 1$. However, one can always define a multivariate proper scoring rule from a univariate one. Let $\{Y_j\}_{j=1}^{d}$ be a collection of marginal distributions from the multivariate random variable $\boldsymbol{Y}$. Then one can construct a *marginal score* for $\boldsymbol{Y}$ as

$$S_M(\mathbb{P}, \boldsymbol{y}) = \sum_{j=1}^{d} S(\mathbb{P}_j, y_j),$$

where $Y_j \sim \mathbb{P}_j$ when $\boldsymbol{Y} \sim \mathbb{P}$ and S is a (strictly) proper scoring rule for the marginal Y_j [Pic et al., 2025]. Then, the scoring rule S_M is proper, but not strictly proper. This is especially interesting if the main interest is in the marginals, for example, if the dependence structure across the marginals is of little interest.

C COMPUTATIONAL COMPLEXITY AND APPROXIMATION ERROR

Computational complexity. The cost of the pairwise estimator separates into a second-order level over the M ensemble members and a first-order level for evaluating the divergence/entropy per member. The second-order level requires $\mathcal{O}(M^2)$ divergence evaluations for EU and $\mathcal{O}(M)$ entropy evaluations for AU. This is shared by *all* pairwise measures, including $S_{\log}$ and S_{SE}. The first-order cost depends on the scoring rule and the uncertainty representation; if it is available in closed-form, the cost is negligible, otherwise the cost depends on the additional number of first-order samples N (compare Table 6):

- S_{SE} is cheapest at $\mathcal{O}(MNd)$ and scales linearly in M, N, and d: since $\mathrm{EU}_{\mathrm{SE}} = \mathrm{tr}(\mathrm{Cov}_Q[\boldsymbol{\mu}_{\mathbb{P}}])$, it suffices to compute M empirical means and the trace of their covariance, with no pairwise sample comparisons. This efficiency comes at the price of discarding all distributional information beyond the first moment.
- With the closed-form expressions of NG, DER, MDN, LoRa, the kernel scores attain $\mathcal{O}(M^2 d)$, adding negligible overhead for typical ensemble sizes ($M \leq 10$).
- Without closed forms, sample-based kernel scores cost $\mathcal{O}(M^2N^2d)$, i.e. $\mathcal{O}(N^2)$ per pair, but require no density evaluation.
- $S_{\log}$ requires a density: sample-based evaluation therefore falls back to kernel density estimation (KDE), and the BMA log-score admits no closed form when $\bar{\mathbb{P}}$ is a mixture, whereas kernel scores decompose linearly over mixture components and remain exact.

Table 6: First-order computational cost and estimation error of the pairwise estimator per scoring rule. The second-order level adds a shared $\mathcal{O}(M^2)$ (EU) / $\mathcal{O}(M)$ (AU) term. M: ensemble size, N: first-order samples, d: output dimension. The error column refers to sample-based evaluation.

Measure	Closed-form cost	Sample-based cost	First-order error
S_{SE}	$\mathcal{O}(MNd)$	$\mathcal{O}(MNd)$	$\mathcal{O}(1/\sqrt{N})$
S_{ES}, S_{k_γ}	$\mathcal{O}(M^2d)$	$\mathcal{O}(M^2N^2d)$	$\mathcal{O}(1/\sqrt{N})$
S_{CRPS}	$\mathcal{O}(M^2d)$	$\mathcal{O}(M^2N^2d)$	$\mathcal{O}(1/\sqrt{N})$
$S_{\log}$	— (no closed form for mixtures)	KDE	$\mathcal{O}(N^{-4/(d+4)})$

Empirical runtime. We validate these costs on NYU Depth v2 (100 test images, $d = 55 \times 74$; Figure 4). Varying $M \in \{2, \ldots, 10\}$ with closed-form (NG) and sample-based (SB, $N = 10$) representations, all kernel measures evaluate in well under one second at $M = 10$, while S_{SE} is near-constant owing to its linear scaling. Varying $N \in \{5, \ldots, 50\}$ at fixed $M = 10$, the sample-based kernel measures follow the $\mathcal{O}(N^2)$ reference curve and S_{SE} stays flat. (The marginal S_{CRPS} appears slower under the closed form only because its ${}_1F_1$ evaluation falls back to a CPU `scipy` routine.) At our 2D experimental setting ($N = 10$), all runtimes are well below one second per 100 images.

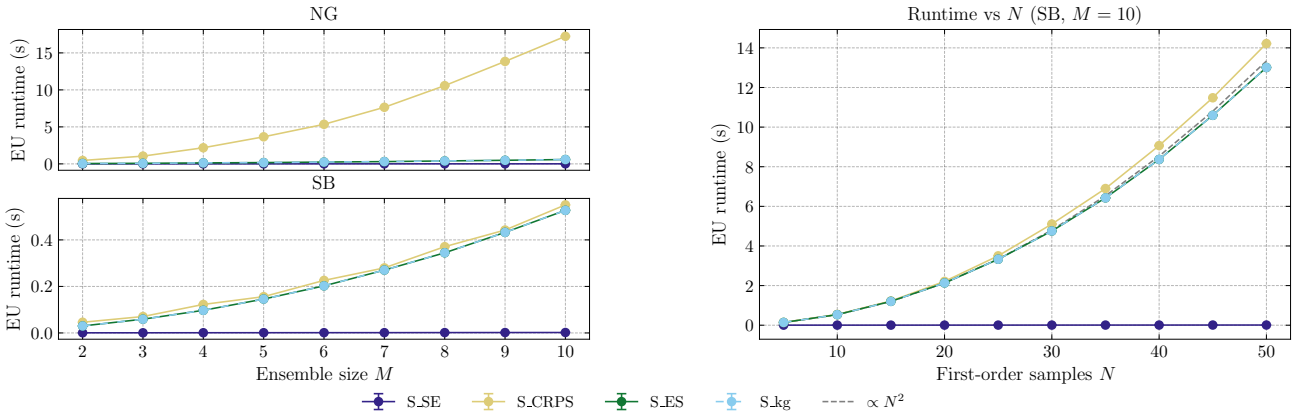

Figure 4: Wall-clock runtime of the uncertainty measures on NYU Depth v2 (100 images). Left: varying ensemble size M. Right: varying sample count N (sample-based, $M = 10$).

Approximation error. Two error sources arise: finite ensemble size M (second-order) and, for sample-based evaluation, finite sample count N (first-order). The second-order estimators both converge at $\mathcal{O}(1/\sqrt{M})$; AU, because it is estimated as a sample mean and EU, as it is estimated via a one-sample U-statistic of order two. Crucially, this rate is identical for every scoring rule (including $S_{\log}$ and S_{SE}): the finite-ensemble error is a property of the uncertainty representation, not of the measure. For the first-order error, if a closed-form expression is available, the estimation error reduces to zero. When sampling is required, the U-statistic estimator of the kernel divergence D_k is unbiased and converges at $\mathcal{O}(1/\sqrt{N})$ independently of the output dimension d [Gretton et al., 2012], in contrast to KDE-based evaluation of $S_{\log}$, which converges at $\mathcal{O}(N^{-4/(d+4)})$ and is impractical for large d. The total estimation error of the sample-based estimator thus decomposes as

$$\mathcal{O}\left(1/\sqrt{M}\right) + \mathcal{O}\left(1/\sqrt{N}\right),$$

with the second term vanishing whenever closed-form first-order expressions are available.

D EXPERIMENT DETAILS

Our model implementations and reproducible experiments are available at `https://github.com/cbuelt/kernel-uq`.

D.1 DATASETS

In this section, we describe all the datasets used, as well as their generation procedure and out-of-distribution version, if applicable. All datasets and corresponding tasks are some sort of regression problem, where the quality of a prediction is evaluated using the mean-squared-error (MSE). An overview of the datasets is available in Table 7.

Table 7: Overview of datasets and evaluation protocols used in experiments.

Task	**Domain**	**Datasets**	**OOD shift**
Univariate	Tabular regression	UCI	-
1D PDEs	PDE	Burgers'	Change in viscosity ν
	PDE	Kuramoto-Sivashinsky	Change in length scale L
2D	Climate	ERA5	Geographic domain shift
	Computer vision	NYU Depth / ApolloScape	Scene shift

Univariate regression For the univariate regression task, we utilize the UCI benchmark [Hernández-Lobato and Adams, 2015], from which we use all datasets, except *boston*, due to raised ethical concerns[1], and *wine-quality* due to its categorical prediction objective.

1D PDE tasks To accommodate more complex tasks, we use two time-dependent one-dimensional partial differential equations (PDEs). While both PDEs could be analyzed in an autoregressive manner, we focus on a single-step prediction, where at each step the probabilistic model samples from a conditional distribution $u_s \sim f_\theta(\cdot \mid u_{s-1}, u_{s-2})$, predicting the dynamics $u_s - u_{s-1}$, given the last two timesteps. In particular, we consider the following two dynamical systems:

Burgers' equation. The Burgers' equation is given as

$$\begin{aligned} \partial_t u(t,x) + \partial_x (u^2(t,x)/2) &= \nu/\pi \partial_{xx} u(t,x), \quad x \in (0,1),\ t \in (0,2] \\ u(0,x) &= u_0(x), \quad x \in (0,1) \end{aligned}$$

where $u \in C([0,T]; H^r_{\text{per}}((0,1);\mathbb{R}))$ for any $r > 0$, $u_0 \in L^2_{\text{per}}((0,1);\mathbb{R})$ is the initial condition and $\nu \in \mathbb{R}_+$ is the diffusion coefficient[2]. We utilize data from the PDEBENCH repository [Takamoto et al., 2022], which assumes a constant diffusion coefficient, which we choose as $\nu = 0.1$. The data is generated with a periodic boundary condition from a superposition of sinusoidal waves with the temporally and spatially 2nd-order upwind difference scheme for the advection term, and the central difference scheme for the diffusion term. As an out-of-distribution dataset, we use the same equation but with diffusion coefficient $\nu = 0.01$, which leads to the corresponding solutions being less smooth and having steeper gradients. Figure 5 shows a sample rollout of the two different systems.

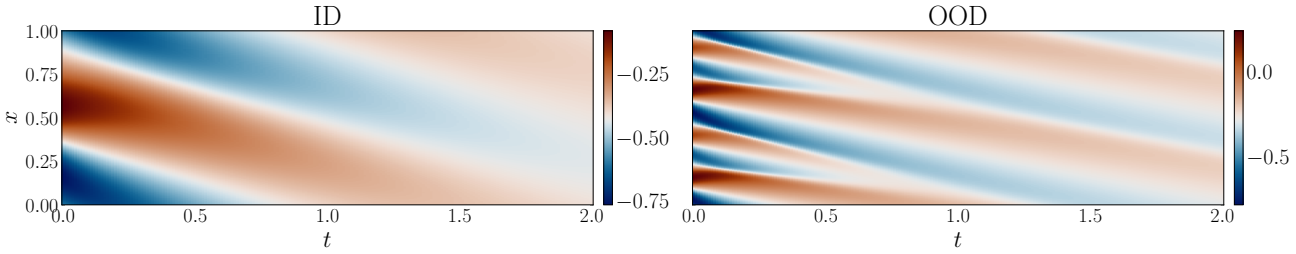

Figure 5: Selected samples of the in-distribution ($\nu = 0.1$) and out-of-distribution dataset ($\nu = 0.01$) for the Burgers' equation.

Kuramoto–Sivashinsky equation. The Kuramoto–Sivashinsky (KS-) equation in one spatial dimension is given as:

$$\begin{aligned} \partial_t u(x,t) + u\partial_x u(x,t) + \partial_x^2 u(x,t) + \partial_x^4 u(x,t) &= 0, \qquad x \in \mathcal{D}, t \in (0,T] \\ u(x,0) &= u_0(x), \qquad x \in \mathcal{D} \end{aligned}$$

[1] https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8

[2] $H^r_{\text{per}}(\mathcal{D};\mathbb{R})$, $L^2_{\text{per}}(\mathcal{D};\mathbb{R})$ denote the periodic Sobolev and L^2 spaces, respectively.

where $\mathcal{D} \subseteq \mathbb{R}$, $u \in C([0,T]; H^4_{\text{per}}(\mathcal{D};\mathbb{R}))$, and $u_0 \in L^2_{\text{per}}(\mathcal{D};\mathbb{R})$. We follow the setup in Bülte et al. [2025] and simulate the KS-equation from random uniform noise $\mathcal{U}(-1,1)$ on a periodic domain $\mathcal{D} = [0, L],\ L = 64$ using the py-pde package [Zwicker, 2020]. We generate 10000 samples with a resolution of 256×50 and $\Delta t = 2$.
As an out-of-distribution dataset, we again simulate from the KS-equation, but with an adjusted length scale of $L = 80$, which leads to an increasing number of unstable spatial modes. Figure 6 shows a sample rollout of the two different systems.

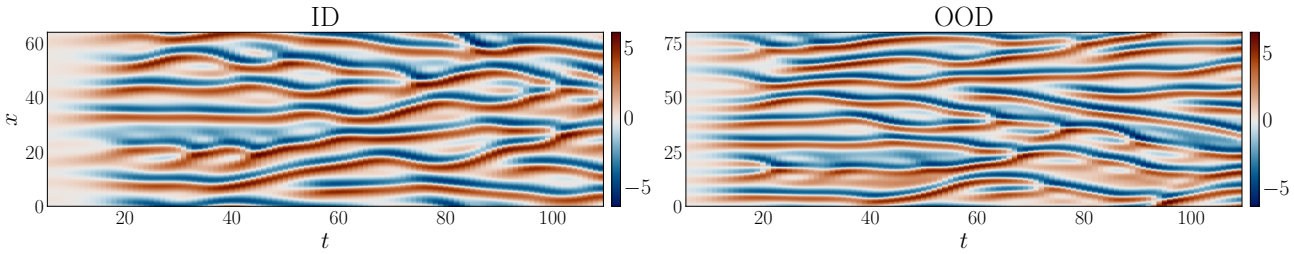

Figure 6: Selected samples of the in-distribution ($L = 64$) and out-of-distribution dataset ($L = 80$) for the Kuramoto-Sivashinsky equation.

2D tasks Finally, we use the following two high-dimensional imaging tasks:

Depth regression. As a typical vision task, we utilize the NYU Depth v2 dataset [Silberman et al., 2012], similar to Amini et al. [2020], which consists of image-depth pairs of indoor scenes. As an out-of-distribution dataset, we utilize ApolloScape [Huang et al., 2018], a dataset of outdoor driving scenes. Figure 7 shows the depth-image pairs for the two different datasets.

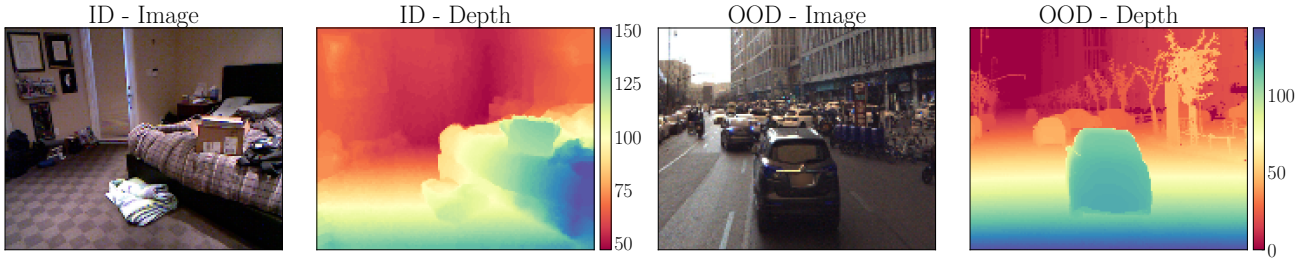

Figure 7: Selected samples of the NYU (in-distribution) and ApolloScape (out-of-distribution) dataset with corresponding ground truth image and depth target.

Surface temperature prediction (T2M). Finally, we use a grid-based surface temperature prediction task, where we utilize the ERA5 dataset [Hersbach et al., 2020] provided via the WeatherBench2 benchmark [Rasp et al., 2024]. We fix a 6-hour forecast horizon and use training data from 2011 to 2018, validation data from 2019 to 2020, and test data from 2022. Similar to Bülte et al. [2025], we use data with a spatial resolution of $0.25° \times 0.25°$ and a time resolution of $6h$. For computational reasons, we restrict the data to a European domain, covering an area from 35°N – 75°N and 12.5°W – 42.5°E with selected user-relevant weather variables (u-component and v-component of 10-m wind speed (U10 and V10), temperature at 2m and 850 hPa (T2M and T850), geopotential height at 500 hPa (Z500), as well as land-sea mask and orography) that serve as input to the model, while the prediction target is only T2M. The total number of input channels is 12. As an out-of-distribution dataset, we use a completely different domain that is still roughly similar to the in-distribution data regarding topography and climate regime, namely a domain over North America with similar latitudes. In particular, we use a domain of the same size and spatial resolution, but ranging from 30°N - 69.75°N and 125°W - 70°E. Due to the size of the data and corresponding computational restraints, the resolution of the dataset is halved for the evaluation of the uncertainty measures in the selective prediction and OOD experiments. Figure 8 shows a sample of the two-meter surface temperature for the two different domains.

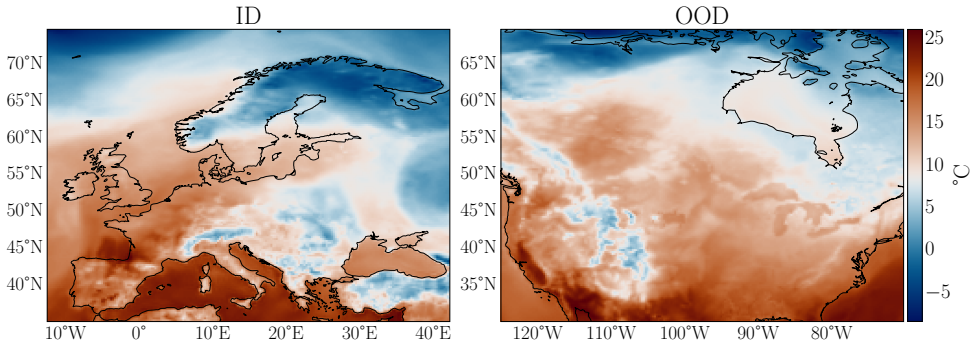

Figure 8: Selected sample of the two-meter surface temperature (28 October 2022, 00:00 UTC) for the in-distribution and out-of-distribution domain.

D.2 BACKBONE MODELS

For all models, we use the Adam optimizer [Kingma and Ba, 2017] with early stopping after 10 epochs and a learning rate schedule that halves the learning rate if no improvement in validation loss has been recorded for more than 5 epochs. For all datasets we use a 10% train-test split and an additional 10% split into training and validation.

Univariate regression As a model backbone, we use a MLP with two hidden layers with 50 neurons each and GELU activation function, where the final activation depends on the chosen uncertainty representation method. Here we use a learning rate of 1e-4, an early stopping patience of 25, and train for a maximum number of 5000 epochs. The batch size is chosen as 32, 64, or 128, depending on the size of the dataset.

1D PDE tasks For the two PDE tasks, we use a Fourier neural operator [Li et al., 2021], which is a neural network architecture that directly learns solutions in the corresponding function space and has shown great success in modeling and solving partial differential equations. Our implementation is based on the publicly available *neuraloperator* library[3]. The models for both tasks are specified identically, namely with 20 Fourier modes, 64 hidden channels, and 256 projection and lifting channels. In total, both networks have roughly 463k parameters. The neural operators are trained with a learning rate of 1e-3, an early stopping patience of 10, and a batch size of 128 for a maximum of 500 epochs.

2D tasks For both two-dimensional tasks, we use a ConvNeXt architecture [Liu et al., 2022], which is based on a simple ResNet, but adapted to be more similar to the architecture of a Vision Transformer. It has shown comparable performance to different Transformer versions across a variety of tasks, while maintaining the simplicity and efficiency of regular convolutional neural networks. In particular, we can make use of the pre-trained version, available via PyTorch[4], where we use the tiny variant for the depth regression and the small variant for the temperature prediction task. The models are pre-trained on ImageNet1K, and we only adapt the very first and last layer to accommodate for the different number of input and output channels. In total, the models have 34M and 56M parameters for the depth regression and temperature prediction tasks, respectively. We use a learning rate of 1e-3, an early stopping patience of 10, and a batch size of 64 trained for a maximum of 250 epochs.

D.3 UNCERTAINTY REPRESENTATION METHODS

In this section, we describe the different uncertainty representation methods used in our experiments. Except for deep evidential regression, which directly learns a second-order distribution Q, all methods are first-order predictors and the corresponding second-order distribution is created via ensembling [Lakshminarayanan et al., 2017] with $M = 10$ ensembles. The different uncertainty representation methods can be applied to any architecture; they just require an adjustment in the final layer processing. An exception is the neural operator, since the corresponding output is in function space, where probability densities are not properly defined; only the sampling-based approach is theoretically valid [Bülte et al., 2025]. For the remaining methods, we share the model-specific parameters across the different backbones. For the multivariate

[3] `https://github.com/neuraloperator/neuraloperator`

[4] `https://docs.pytorch.org/vision/main/models/convnext.html`

tasks, the univariate methods (Deep evidential regression, natural Gaussian, and mixture density network) are interpreted as pointwise predictions.

Natural Gaussian Using a predictive normal distribution $p(\cdot \mid \boldsymbol{\theta}(\boldsymbol{x})) = \mathcal{N}(\mu(\boldsymbol{x}), \sigma^2(\boldsymbol{x}))$ to approximate the probability distribution over $\boldsymbol{y}$ given $\boldsymbol{x}$ for a given model has been common practice in many machine learning tasks [Nix and Weigend, 1994, Lakshminarayanan et al., 2017]. However, when training with the log-likelihood, direct optimization of μ and σ^2 usually leads to training instabilities. To prevent this, we follow the approach of Immer et al. [2023], which use the natural parametrization of the normal distribution, given by $\eta_1 = \mu/\sigma^2,\ \eta_2 = -1/2\sigma^2$ with $\eta_2 < 0$, to obtain more stable optimization. The corresponding log-likelihood loss can be expressed in a closed form; for more details, compare Immer et al. [2023]. To fulfill the parameter restrictions, we use a softplus activation function on the parameter η_2 and reverse the sign. Figure 9 shows a generated sample, the mean prediction, and the corresponding standard deviation of this method for different datasets.

Deep evidential regression Moving beyond the first-order Gaussian, Amini et al. [2020] propose to directly model second-order uncertainty by predicting the parameters of a Normal Inverse-Gamma (NIG) distribution, which is the conjugate prior of a Gaussian. In particular, one obtains $p(\cdot \mid \boldsymbol{\theta}(\boldsymbol{x})) = \mathcal{N}(\mu, \sigma^2)$, with $\mu \sim \mathcal{N}(\gamma(\boldsymbol{x}), \sigma^2 \upsilon(\boldsymbol{x})^{-1})$ and $\sigma^2 \sim \Gamma^{-1}(\alpha(\boldsymbol{x}), \beta(\boldsymbol{x}))$. Here, the outputs of our neural networks are the four parameters $\boldsymbol{\theta} = (\gamma, \upsilon, \alpha, \beta)^\top$, with $\gamma \in \mathbb{R}, \upsilon > 0, \alpha > 1, \beta > 0$, where the individual constraints are realized using a softplus activation. We use the log-likelihood loss function specified in Amini et al. [2020] with a regularization factor $\lambda = 0.01$, as specified by the authors. Figure 10 shows a generated sample, the mean prediction, and the corresponding standard deviation of this method for different datasets.

Mixture density network To accommodate a potential (univariate) multimodal data distribution, we also employ a mixture density network, where the predictive density is specified as $p(\cdot \mid \boldsymbol{\theta}(\boldsymbol{x})) = \sum_{k=1}^{K} w_k \mathcal{N}(\mu_k(\boldsymbol{x}), \sigma_k^2(\boldsymbol{x}))$, which is a weighted mixture of several Gaussian distributions. Mixture density networks, as proposed by Bishop [1994] have been employed in machine learning for a long time, but recent work has focused on optimization and performance improvements [Makansi et al., 2019, Kelen et al., 2025], showing competitive performance across several benchmark tasks [Kelen et al., 2025]. To stabilize training, we move beyond the typical log-likelihood loss and train the model using the continuous ranked probability score (CRPS), as proposed by D´Isanto and Polsterer [2018]. We use a softplus activation for the variance and a softmax for the weights in order to satisfy the corresponding constraints. We choose $K = 10$ across all experiments to allow for a higher level of multimodality in the distribution, but omit detailed hyperparameter tuning, as this is not the focus of these experiments. Figure 11 shows a generated sample, the mean prediction, and the corresponding standard deviation of this method for different datasets.

Low-rank multivariate normal Now, we move to a multivariate method, by again specifying a predictive Gaussian, but this time in a multivariate setting, via $p(\cdot \mid \boldsymbol{\theta}(\boldsymbol{x})) = \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}), \Sigma(\boldsymbol{x}))$, where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean vector and $\Sigma \in \mathbb{R}^{d \times d}$ is the (positive definite and symmetric) covariance matrix. While the covariance matrix can be modeled using its Cholesky decomposition, which also ensures positive definiteness and symmetry Muschinski et al. [2024], for large covariance matrices, this method becomes numerically unstable. Instead, we use a low-rank + diagonal decomposition [Rezende et al., 2014], where the covariance matrix is approximated via $\Sigma = UU^\top + D$ with a low-rank matrix $U \in \mathbb{R}^{d \times r},\ r \ll d$ and a positive diagonal matrix $D \in \mathbb{R}^{d \times d}$. This allows for computationally efficient modeling of the covariance matrix, while still enabling the model to learn the underlying covariance structure. Throughout the experiments, we use the softplus activation for the diagonal, a rank of $r = 10$, and train the model using the Gaussian kernel score, which admits a closed-form expression and offers more stable training as compared to the log-likelihood. Figure 12 shows a generated sample, the mean prediction, and the corresponding standard deviation of this method for different datasets.

Sampling-based Finally, we also use a nonparametric, sampling-based method, which is based on the idea of incorporating noise into the neural network and training it with a (strictly) proper scoring rule. This generative method has recently gained popularity, both regarding theoretical analysis [Shen and Meinshausen, 2024, Pacchiardi et al., 2024], as well as applications [Chen et al., 2024, Alet et al., 2025, Bülte et al., 2025]. In particular, this method has been used both for the univariate setting, using the CRPS as a loss functionKelen et al. [2025], as well as the multivariate setting, using the energy score [Chen et al., 2024]. Furthermore, this method can also be employed together with neural operators, leading to an empirical distribution over the output function space [Bülte et al., 2025]. For the univariate tasks, we adopt the setup and loss from Kelen et al. [2025], while the multivariate methods are similar to Shen and Meinshausen [2024], where random noise is concatenated to the input channel of the model and the energy score is used as a loss function. Figure 13 shows a generated sample, the mean prediction and the corresponding standard deviation of this method for different datasets. For this model, we use $N = 50$ samples for the univariate, $N = 10$ samples for the 1D PDE, and $N = 5$ samples for the 2D tasks.

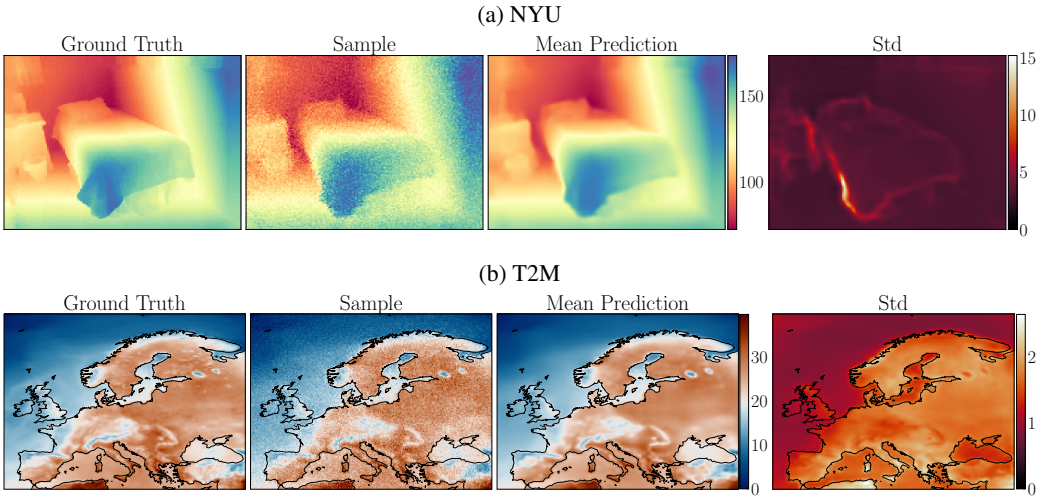

Figure 9: Visualizations of the predictions of the natural Gaussian method for the depth regression and surface temperature prediction tasks.

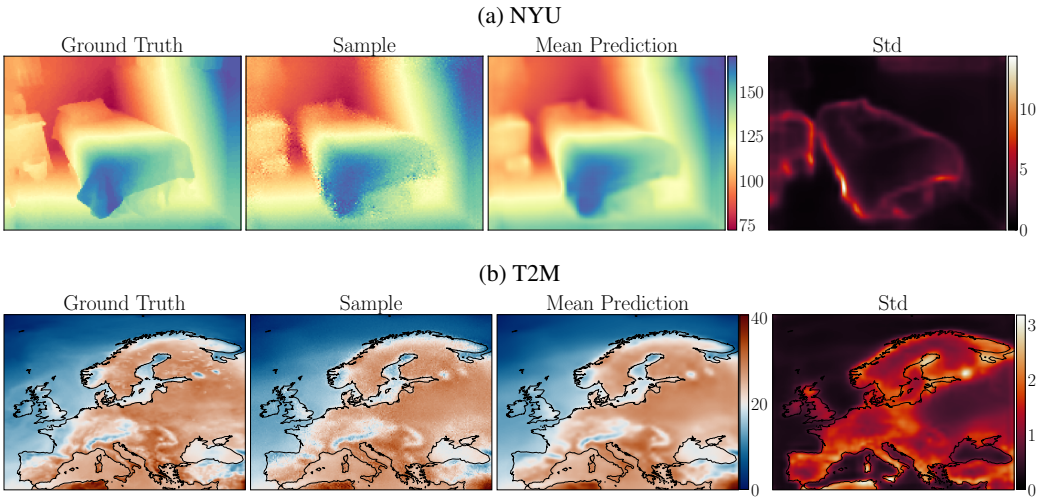

Figure 10: Visualizations of the predictions of the deep evidential regression method for the depth regression and surface temperature prediction tasks.

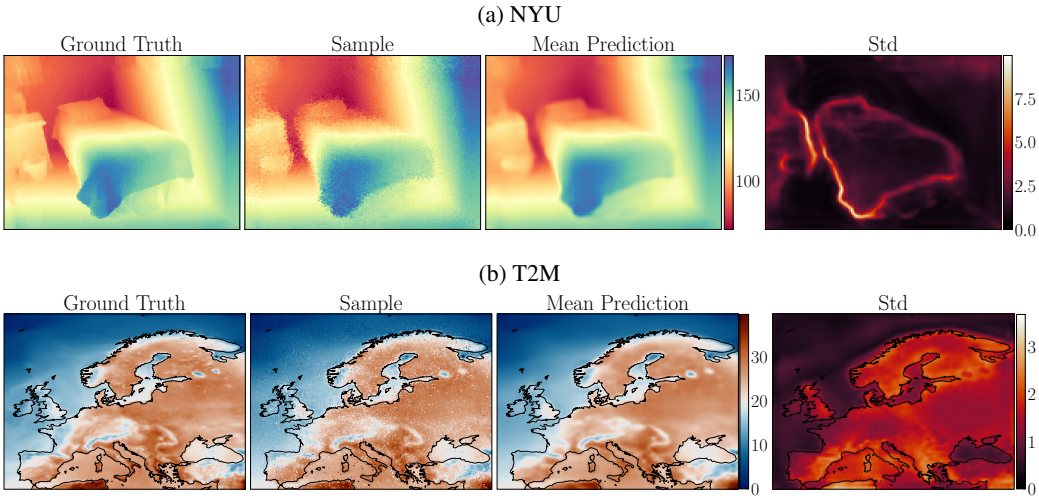

Figure 11: Visualizations of the predictions of the mixture density network for the depth regression and surface temperature prediction tasks.

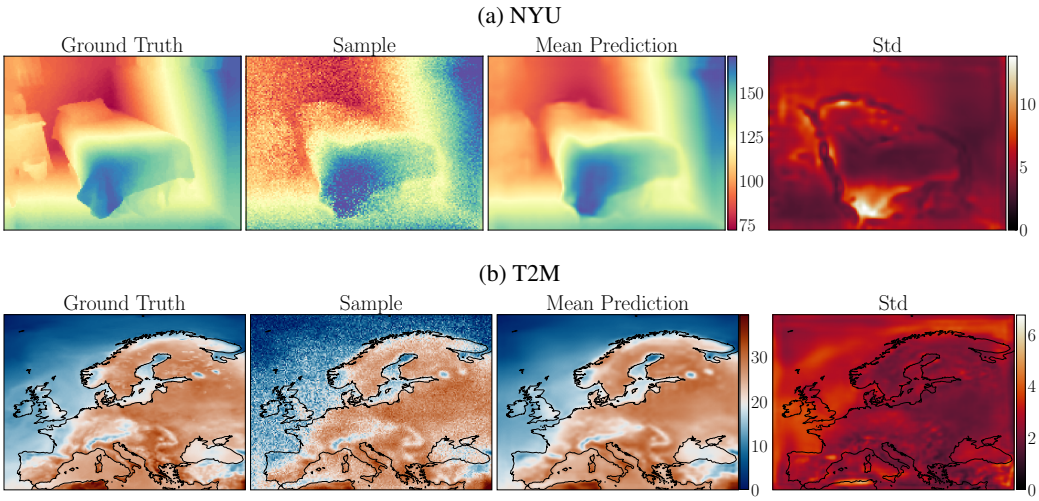

Figure 12: Visualizations of the predictions of the low-rank multivariate normal method for the depth regression and surface temperature prediction tasks.

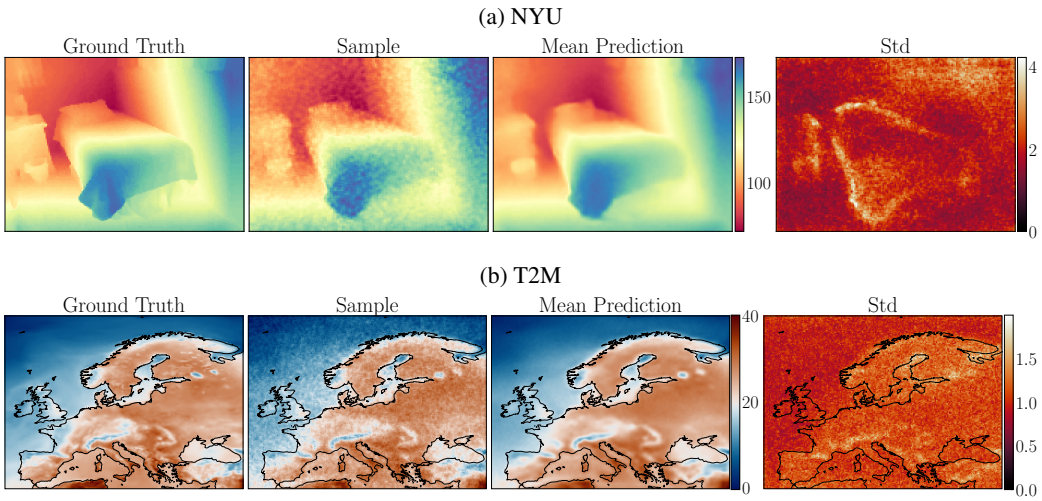

Figure 13: Visualizations of the predictions of the sampling-based method for the depth regression and surface temperature prediction tasks.

D.4 ROBUSTNESS

Here, we use a regular deep ensemble [Lakshminarayanan et al., 2017] on the concrete, energy, and yacht dataset from the UCI regression benchmark [Hernández-Lobato and Adams, 2015]. We train a base ensemble of $M = 25$ and $M = 5$ members and one additional member that is trained on a distorted target $\hat{y} = y + \mathcal{N}(0, \delta^2)$. This allows us to analyze the robustness of the different uncertainty measures with respect to an outlier in the ensemble prediction.

First, we provide a theoretical analysis of the robustness in the case of this deep ensemble, which admits a first-order predictive Gaussian distribution $p(y \mid \boldsymbol{\theta}) = \mathcal{N}(\mu, \sigma^2), \boldsymbol{\theta} = (\mu, \sigma^2)^\top$. Assume that the second-order distribution fulfills $\|\mathbb{E}_Q[H(P_{\boldsymbol{\vartheta}})]\| < \infty$, meaning that the aleatoric uncertainty of the sample distribution Q is well defined, which always holds for a finite ensemble. In that case, we can analyze the influence function $\mathrm{IF}(\boldsymbol{\theta}_0; \mathrm{AU}, Q)$ directly by analyzing the limit $\lim_{\boldsymbol{\theta}_0 \to \infty} H(P_{\boldsymbol{\theta}_0})$, since $\mathbb{E}_Q[H(P_{\boldsymbol{\vartheta}})]$ is a finite constant. Table 8 shows the closed-form expressions for $H(\boldsymbol{\theta}_0)$, as well as the corresponding growth rates in the contamination $\boldsymbol{\theta}_0$. While the Gaussian kernel score is the only scoring rule that is robust, since it admits a bounded influence function, the log-score and CRPS have a notably slower growth rate in $\boldsymbol{\theta}_0$ as the variance-based measure, which grows linearly with σ_0^2.

Table 8: Limit and corresponding growth rates for the influence function $\mathrm{IF}(\boldsymbol{\theta}_0; \mathrm{AU}, Q)$ in the limit $\boldsymbol{\theta}_0 \to \infty$.

S	$H(P_{\boldsymbol{\theta}_0})$	$\lim_{\boldsymbol{\theta}_0 \to \infty} H(P_{\boldsymbol{\theta}_0})$	Growth
$S_{\log}$	$\frac{1}{2}\log(2\pi e\sigma_0^2)$	∞	$\mathcal{O}(\log(\sigma_0^2))$
S_{SE}	σ_0^2	∞	$\mathcal{O}(\sigma_0^2)$
S_{ES}	$\frac{\sigma_0}{\sqrt{\pi}}$	∞	$\mathcal{O}(\sqrt{\sigma_0^2})$
S_{k_γ}	$\frac{1}{2}\left(1 - \frac{\gamma}{\sqrt{\gamma^2+4\sigma_0^2}}\right)$	0.5	$\mathcal{O}(1/\sqrt{\sigma_0^2})$

This theoretical analysis is also supported by the numerical results on the UCI benchmark. Table 9 shows the mean absolute percentage error of the epistemic and aleatoric uncertainty from the base ensemble for different values of δ. Figure 14 shows corresponding visualizations of the MAPE vs δ for the different datasets.

Table 9: Effect of the added noise δ on the different epistemic and aleatoric uncertainty measures across all three datasets for $M = 25$ ensemble members. The reported values are the mean absolute percentage error from the corresponding measure for the base ensemble.

Experiment	Type	S	0.0	0.2	0.5	1.5	2.5	5.0
Concrete	Aleatoric	$S_{\log}$	0.25	0.90	1.56	3.34	4.47	4.82
		S_{SE}	1.11	2.53×10^1	3.24×10^2	7.21×10^3	6.77×10^4	4.78×10^5
		S_{ES}	0.55	4.65	1.47×10^1	7.83×10^1	2.24×10^2	5.04×10^2
		S_{k_γ}	0.03	0.10	0.13	0.18	0.19	0.20
	Epistemic	$S_{\log}$	3.49	4.99×10^2	3.10×10^3	3.32×10^4	5.29×10^4	2.43×10^5
		S_{SE}	3.71	9.49×10^2	4.08×10^3	6.17×10^4	6.22×10^4	4.74×10^5
		S_{ES}	2.68	1.16×10^2	3.35×10^2	1.09×10^3	1.16×10^3	4.05×10^3
		S_{k_γ}	1.57	1.13×10^1	1.47×10^1	1.21×10^1	1.17×10^1	1.33×10^1
Energy	Aleatoric	$S_{\log}$	0.11	0.57	1.17	2.28	2.62	3.79
		S_{SE}	1.12	1.07×10^1	6.41×10^1	2.86×10^3	6.57×10^3	1.09×10^7
		S_{ES}	0.52	3.52	1.13×10^1	6.24×10^1	9.52×10^1	1.93×10^3
		S_{k_γ}	0.34	1.41	2.31	2.97	3.07	3.17
	Epistemic	$S_{\log}$	3.33	3.86×10^2	2.94×10^3	1.84×10^4	7.74×10^4	6.45×10^5
		S_{SE}	3.57	6.16×10^2	5.47×10^3	3.62×10^4	1.52×10^5	3.12×10^5
		S_{ES}	2.38	7.49×10^1	2.26×10^2	6.05×10^2	1.26×10^3	1.84×10^3
		S_{k_γ}	0.88	1.12	1.19	1.98	2.08	2.24
Yacht	Aleatoric	$S_{\log}$	0.09	1.17	2.08	2.75	3.61	4.3
		S_{SE}	0.60	2.25×10^1	2.88×10^3	2.1×10^5	6.42×10^5	6.1×10^6
		S_{ES}	0.31	1.59×10^1	5.75×10^1	2.99×10^2	7×10^2	2.36×10^3
		S_{k_γ}	0.18	2.65	3.18	3.36	3.52	3.54
	Epistemic	$S_{\log}$	6.04	2.43×10^4	4.39×10^4	3.89×10^5	2.34×10^6	3.77×10^6
		S_{SE}	6.58	4.95×10^4	8.69×10^4	5.53×10^5	4.06×10^6	3.31×10^6
		S_{ES}	4.87	1.25×10^3	1.71×10^3	4.13×10^3	1.12×10^4	1.02×10^4
		S_{k_γ}	2.75	1.03×10^1	9.13	8.49	8.32	8.31

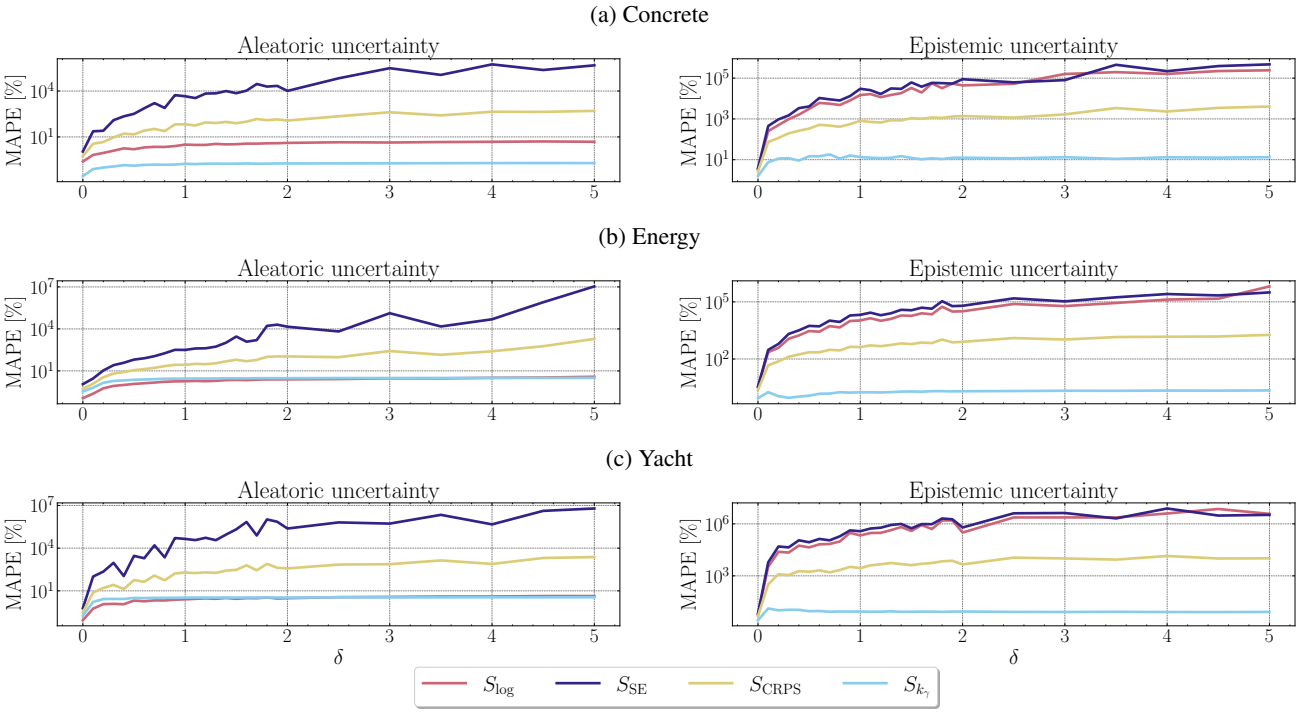

Figure 14: Effect of the added noise δ on the different uncertainty measures for an ensemble of size $M = 25$ across all three datasets. The reported values are the mean absolute percentage error from the corresponding measure for the base ensemble.

D.5 SELECTIVE PREDICTION

We evaluate selective prediction on all datasets using the corresponding and uncertainty representation methods described in this section. As an evaluation criterion, we evaluate the prediction-reject ratio [PRR, Malinin and Gales, 2021], which can be extended to the regression setting using the mean squared error (MSE) as a performance metric [Fishkov et al., 2025]. We evaluate using retention rates, i.e. 1−rejection, from 0.5 to 1. The full results are available in Table 10 and corresponding visualizations of the retention curves for the different methods in Figures 15, 16, and 17.

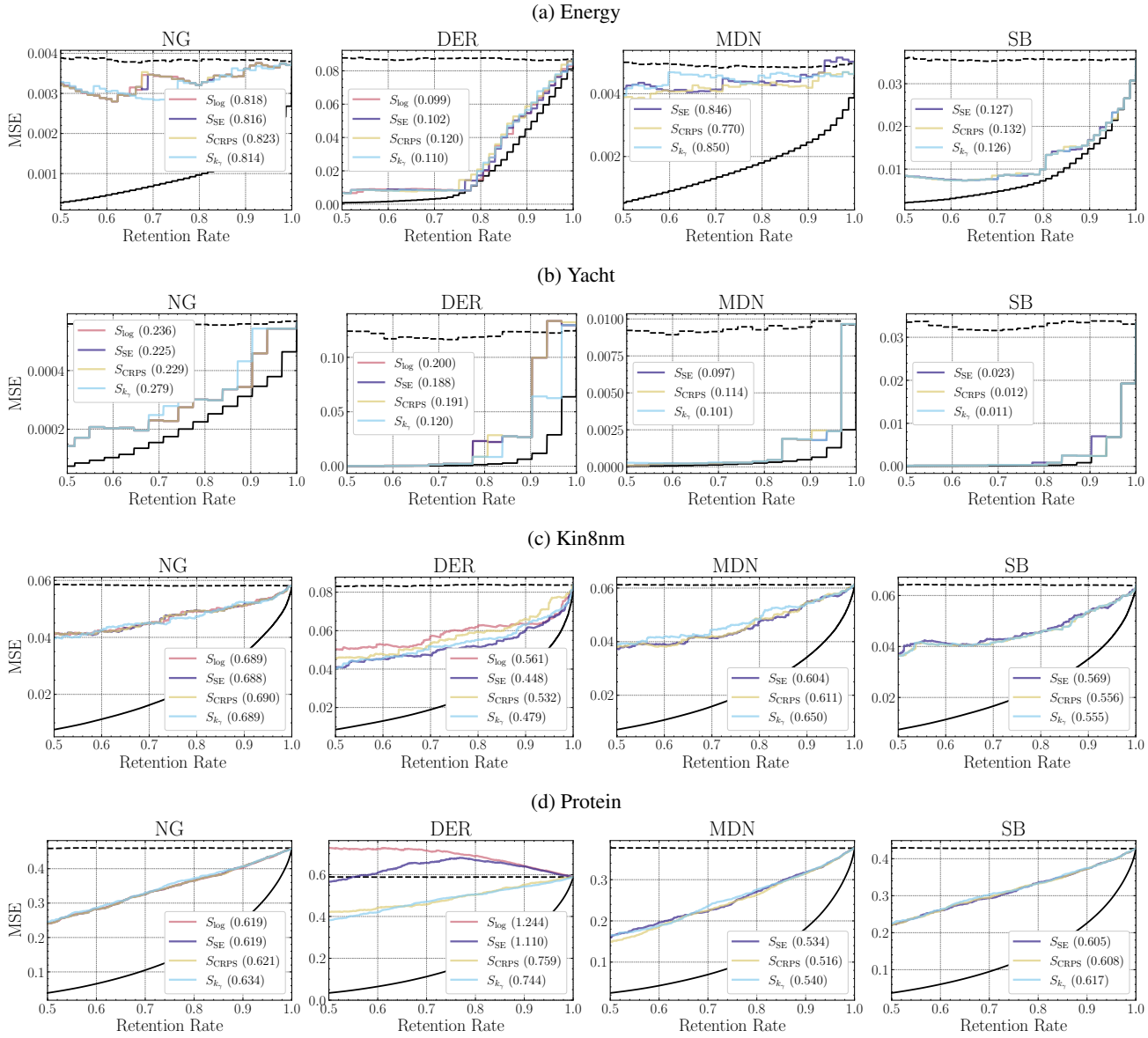

Figure 15: Retention curves with the reported PRR per uncertainty measure for four of the different UCI datasets. The black solid line is the optimal retention, where the curve is sorted by the actual MSE per prediction. The black dashed line is the random baseline.

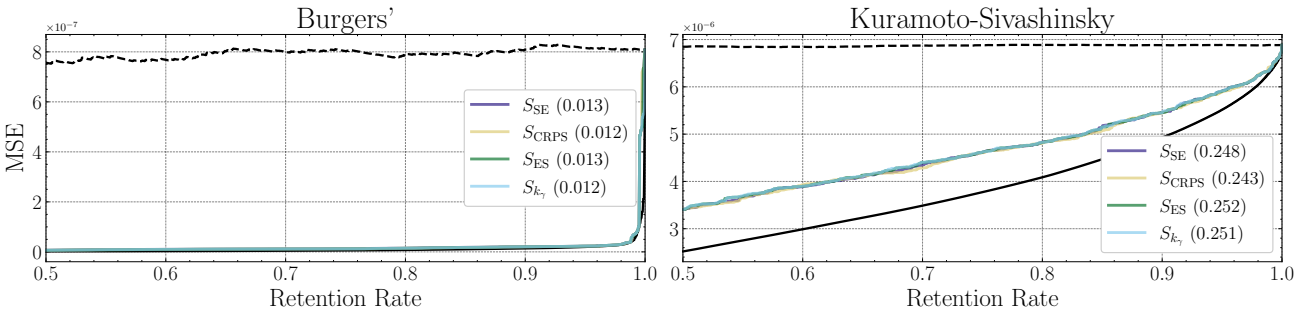

Figure 16: Retention curves with the reported PRR per uncertainty measure for the two one-dimensional PDE datasets. The black solid line is the optimal retention, where the curve is sorted by the actual MSE per prediction. The black dashed line is the random baseline.

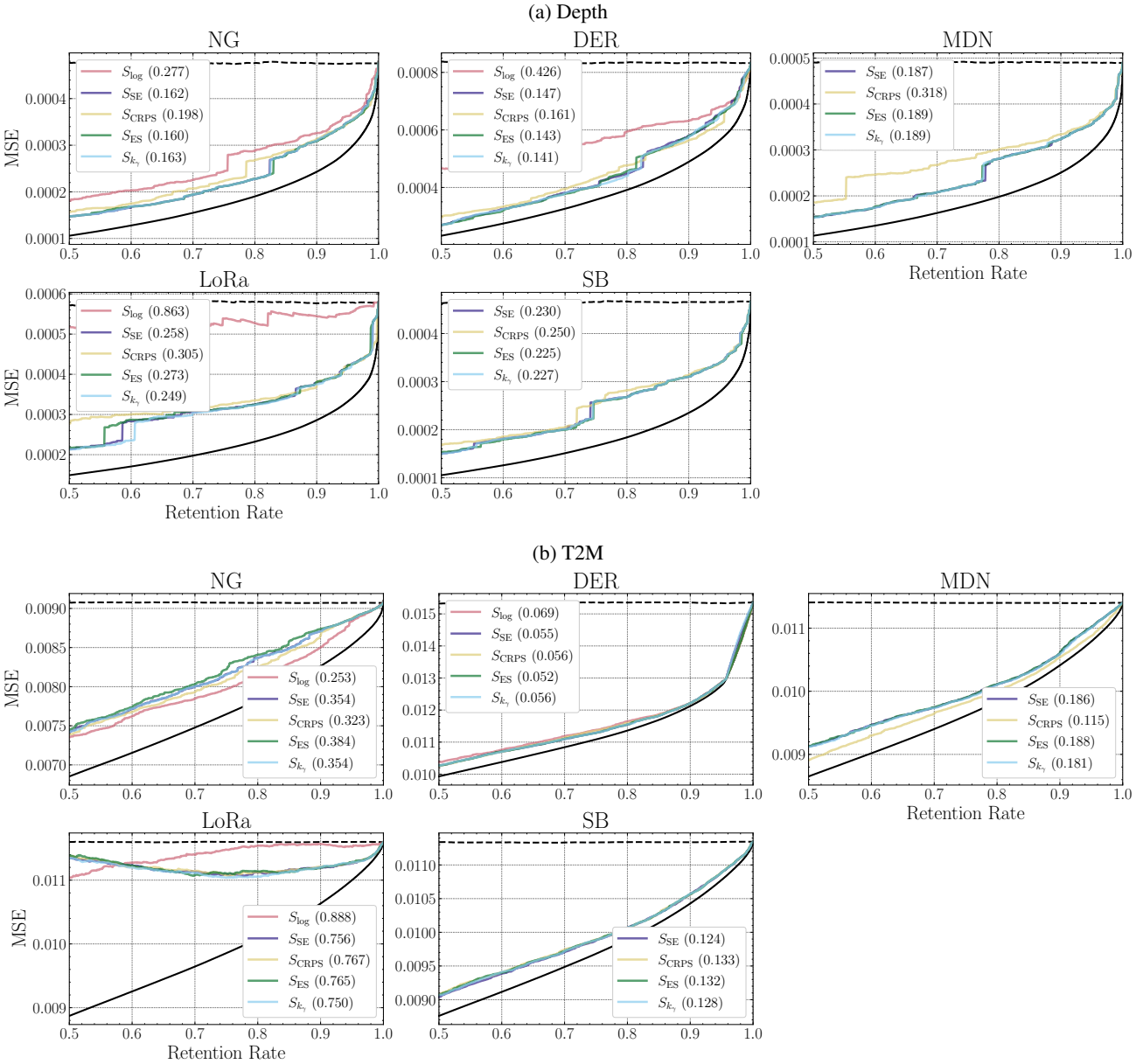

Figure 17: Retention curves with the reported PRR per uncertainty measure for the two 2D tasks. The black solid line is the optimal retention, where the curve is sorted by the actual MSE per prediction. The black dashed line is the random baseline.

Table 10: Prediction-reject-ratios (PRR ↓) for all different datasets, uncertainty representation methods and uncertainty measures, where the best measure for each configuration is highlighted in bold. Note that for the univariate tasks, the S_{ES} coincides with S_{CRPS} and the values are omitted for readability.

Dataset	Method	$S_{\log}$	S_{SE}	S_{CRPS}	S_{ES}	S_{k_γ}
Burgers'	SB	-	0.013	0.013	0.013	**0.012**
KS	SB	-	0.247	**0.242**	0.251	0.252
Depth	NG	0.278	0.162	0.197	**0.161**	0.162
	DER	0.426	0.147	0.160	0.143	**0.142**
	MDN	-	**0.188**	0.314	0.189	0.189
	LoRa	0.859	0.262	0.310	0.272	**0.251**
	SB	-	0.229	0.251	**0.225**	0.227
T2M	NG	**0.253**	0.356	0.322	0.385	0.356
	DER	0.068	0.054	0.055	**0.052**	0.056
	MDN	-	0.186	**0.116**	0.188	0.181
	LoRa	0.889	0.759	0.766	0.764	**0.750**
	SB	-	**0.124**	0.134	0.132	0.129
Concrete	NG	0.648	**0.644**	**0.644**	-	0.742
	DER	0.591	**0.566**	0.606	-	0.709
	MDN	-	0.503	**0.497**	-	0.531
	SB	-	0.612	0.609	-	**0.608**
Energy	NG	0.824	0.818	0.823	-	**0.798**
	DER	**0.101**	0.103	0.121	-	0.110
	MDN	-	0.820	**0.770**	-	0.848
	SB	-	0.125	0.125	-	**0.124**
Kin8nm	NG	0.690	0.687	0.688	-	**0.682**
	DER	0.560	**0.448**	0.535	-	0.479
	MDN	-	**0.602**	0.610	-	0.646
	SB	-	0.573	**0.556**	-	0.558
Naval	DER	0.564	**0.177**	0.743	-	0.187
	MDN	-	0.278	**0.231**	-	0.422
	NG	**0.530**	0.551	0.545	-	0.608
	SB	-	0.305	0.300	-	**0.299**
Power	NG	0.936	0.936	0.936	-	**0.891**
	DER	0.931	**0.851**	0.891	-	0.958
	MDN	-	0.745	**0.735**	-	0.765
	SB	-	0.848	0.842	-	**0.841**
Protein	NG	**0.620**	**0.620**	**0.620**	-	0.633
	DER	1.243	1.110	0.759	-	**0.743**
	MDN	-	0.535	**0.518**	-	0.540
	SB	-	**0.607**	0.609	-	0.616
Yacht	NG	**0.241**	**0.241**	**0.241**	-	0.296
	DER	0.201	0.201	0.198	-	**0.124**
	MDN	-	**0.103**	0.107	-	0.105
	SB	-	0.024	**0.012**	-	**0.012**

D.6 OUT-OF-DISTRIBUTION DETECTION

For evaluation, we use ID and OOD datasets of the same size, which we achieve by subsampling, if necessary, and compute the AUROC across all datasets and uncertainty representation methods. For the out-of-distribution generation procedure for each dataset, compare Appendix D.1. The full results are listed in Table 11. Selected visualizations for the sampling-based method over the different datasets can be found in Figure 18, 19, 20, and 21.

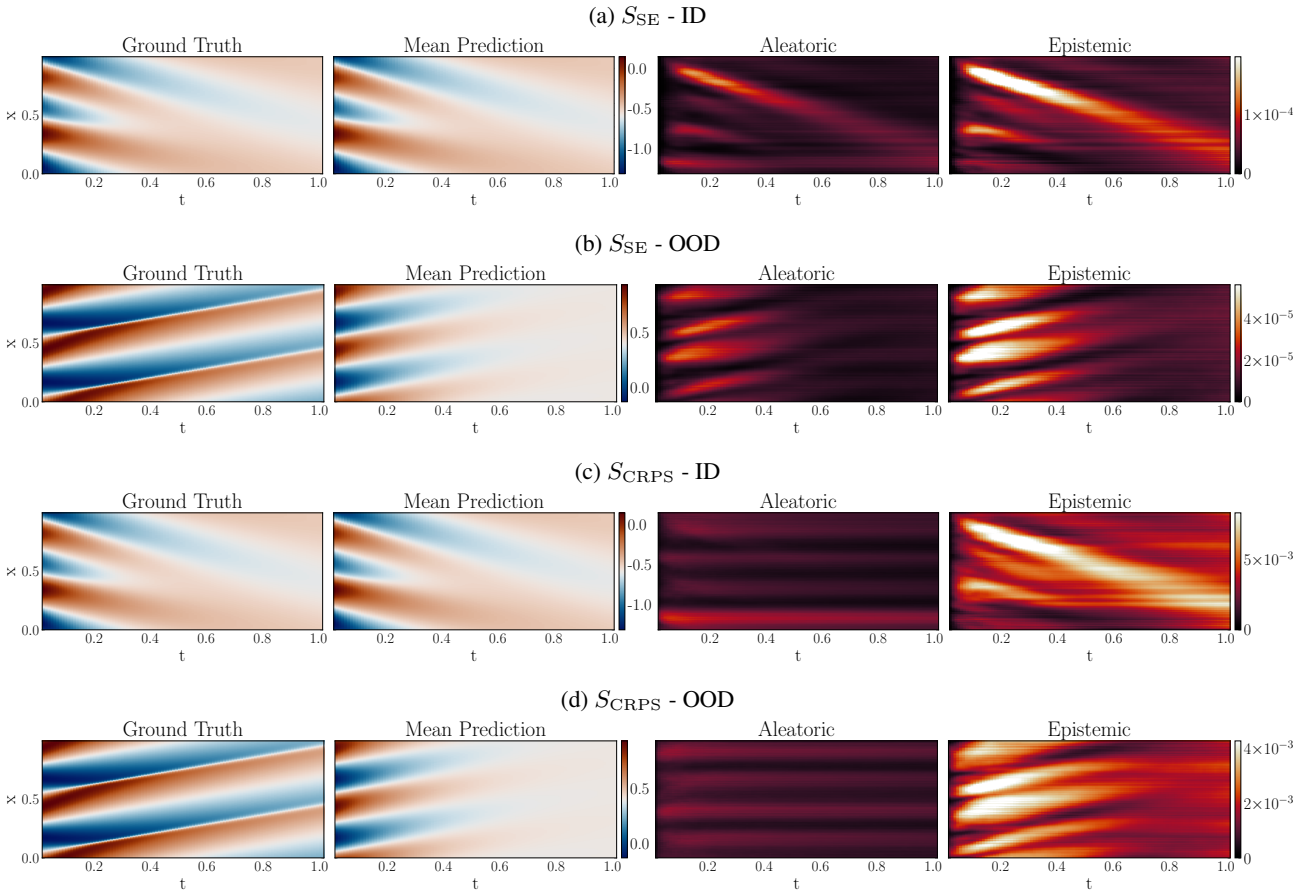

Figure 18: Predictions and uncertainty estimates for a selected sample of the Burgers' equation. Shown are the in-distribution (ID) and out-of-distribution (OOD) predictions for the S_{SE} and S_{CRPS} measures, as those can be visualized pointwise.

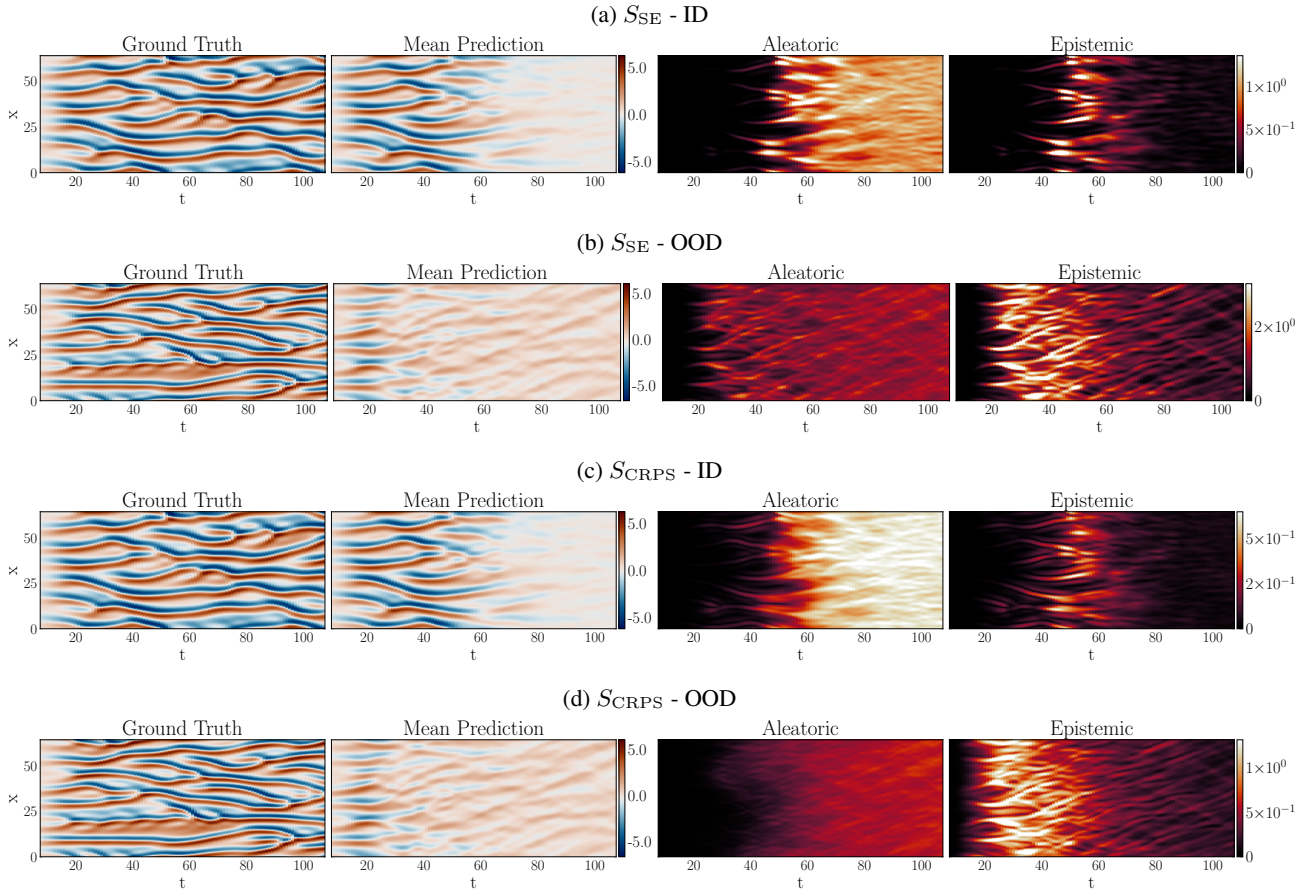

Figure 19: Predictions and uncertainty estimates for a selected sample of the Kuramoto-Sivashinsky equation. Shown are the in-distribution (ID) and out-of-distribution (OOD) predictions for the S_{SE} and S_{CRPS} measures, as those can be visualized pointwise.

Table 11: Out-of-distribution detection (AUROC ↑) for all different datasets, uncertainty representation methods and uncertainty measures, where the best measure for each configuration is highlighted in bold.

Dataset	Method	$S_{\log}$	S_{SE}	S_{CRPS}	S_{ES}	S_{k_γ}
Burgers'	SB	-	0.9307	0.9371	**0.9397**	0.9327
KS	SB	-	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Depth	NG	**0.9999**	0.9995	0.9998	0.9998	0.9997
	DER	**0.9923**	0.9542	0.9886	0.9549	0.9590
	MDN	-	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	LoRa	0.2951	0.9991	**0.9997**	0.9994	0.9993
	SB	-	0.9994	**0.9999**	**0.9999**	0.9996
T2M	NG	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	DER	**1.0000**	0.9986	0.9559	0.9547	0.9747
	MDN	-	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	LoRa	0.8066	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	SB	-	**1.0000**	**1.0000**	**1.0000**	**1.0000**

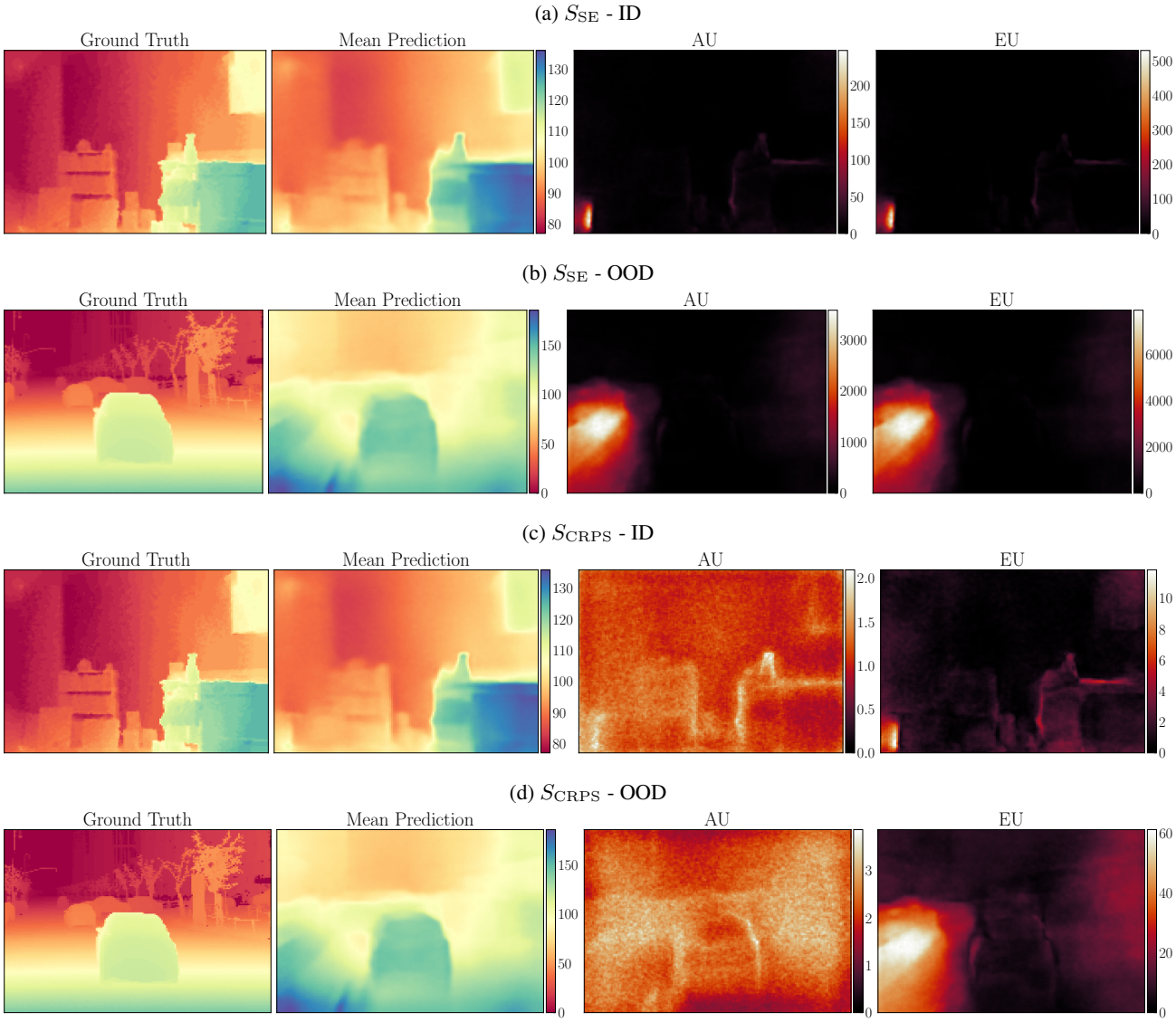

Figure 20: Predictions and uncertainty estimates for a selected sample of the NYU (ID) and ApolloScape (OOD) datasets. Shown are the in-distribution (ID) and out-of-distribution (OOD) predictions for the S_{SE} and S_{CRPS} measures, as those can be visualized pointwise.

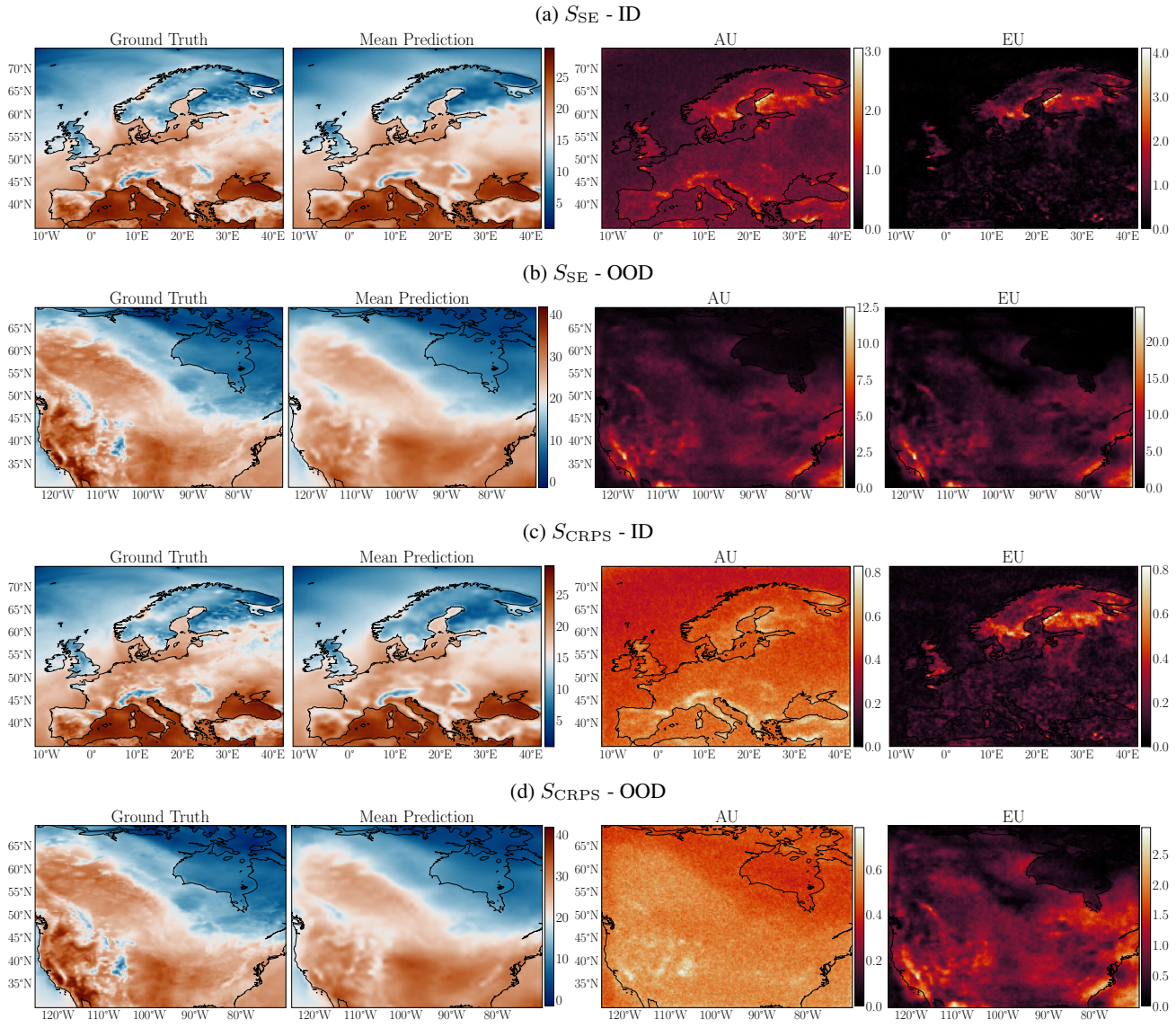

Figure 21: Predictions and uncertainty estimates for a selected sample of surface temperature prediction task across the European (ID) and North American (OOD) domains. Shown are the in-distribution (ID) and out-of-distribution (OOD) predictions for the S_{SE} and S_{CRPS} measures, as those can be visualized pointwise.

D.7 ACTIVE LEARNING

To make the active learning task consistent across the different data modalities, we base the configuration on the total size of the data. In particular, all datasets are split into 90/10 train and test datasets, and the former further into a 90/10 train validation split. We choose an initial fraction of 5% of the training data size to randomly select the starting pool. The active learning loop is then run for 20 rounds, after which a total of 25% of the data has been selected, meaning that the amount of new instances in each round is 1% of the training data size. In each round, $M = 5$ ensemble members are trained for a total of 50 epochs from scratch, and the whole process is repeated across three independent seeds. As a fundamental comparison, we also employ a random baseline, which selects the next samples using a uniform distribution. The full results are listed in Table 12. Selected visualizations of the MSE against the acquisition steps can be found in Figure 22, and 23.

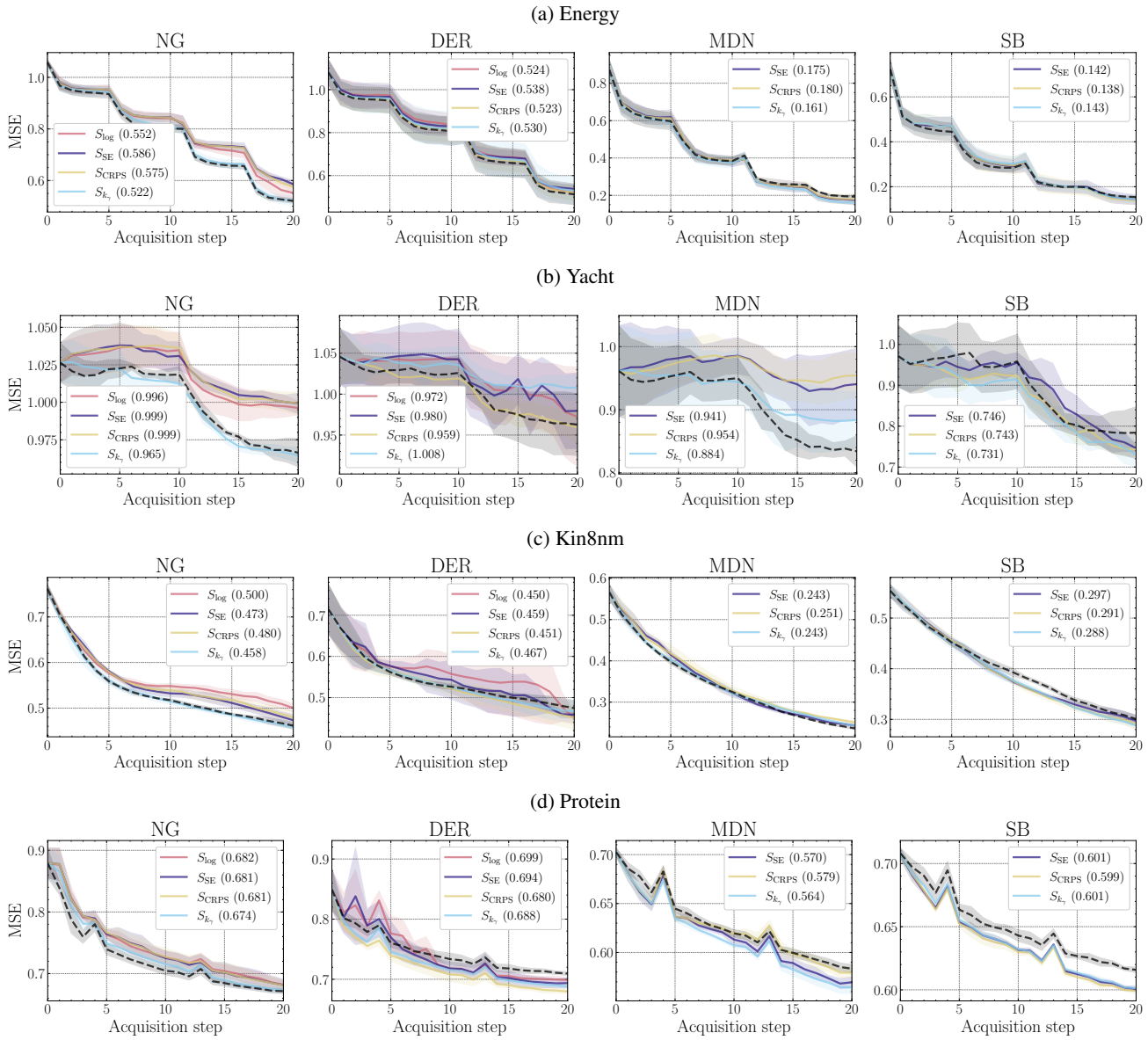

Figure 22: Detailed results of the active learning runs for selected UCI datasets. Shown are the mean values and standard deviation (shaded regions) of the test MSE across the three different runs, with the final loss in brackets. The black dotted line is the random baseline.

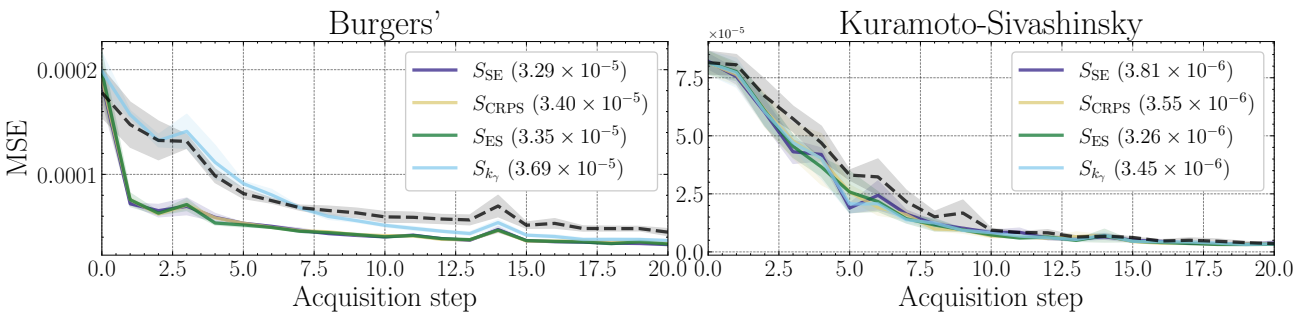

Figure 23: Detailed results of the active learning runs for PDE datasets. Shown are the mean values and standard deviation (shaded regions) of the test MSE across the three different runs, with the final loss in brackets. The black dotted line is the random baseline.

Table 12: Final test loss (MSE ↓) for the active learning task for all different datasets, uncertainty representation methods, and uncertainty measures. The best measure for each configuration is highlighted in bold, while the standard deviation across the different runs is given in brackets. Note that for the univariate tasks, the S_{ES} coincides with S_{CRPS} and the values are omitted for readability.

Dataset	Method	$S_{\log}$	S_{SE}	S_{CRPS}	S_{ES}	S_{k_γ}	$\mathcal{U}$
Burgers'	SB	–	$\mathbf{3.29 \times 10^{-5}}$ (8.05×10^{-7})	3.40×10^{-5} (9.12×10^{-7})	3.35×10^{-5} (4.93×10^{-7})	3.69×10^{-5} (4.52×10^{-7})	4.48×10^{-5} (3.36×10^{-6})
KS'	SB	–	3.81×10^{-6} (9.25×10^{-7})	3.55×10^{-6} (4.42×10^{-7})	$\mathbf{3.26 \times 10^{-6}}$ (3.19×10^{-7})	3.45×10^{-6} (2.63×10^{-7})	3.58×10^{-6} (2.62×10^{-7})
Concrete	NG	0.654 (0.041)	0.657 (0.050)	0.654 (0.049)	-	**0.635** (0.009)	**0.635** (0.010)
	DER	0.685 (0.036)	0.690 (0.018)	**0.640** (0.020)	-	0.653 (0.014)	0.654 (0.016)
	MDN	–	0.391 (0.020)	**0.387** (0.014)	-	0.394 (0.018)	0.402 (0.004)
	SB	–	**0.346** (0.025)	0.360 (0.026)	-	0.356 (0.023)	0.373 (0.007)
Energy	NG	0.552 (0.034)	0.586 (0.014)	0.575 (0.018)	-	0.522 (0.016)	**0.521** (0.009)
	DER	0.524 (0.009)	0.538 (0.017)	0.523 (0.039)	-	0.530 (0.063)	**0.513** (0.048)
	MDN	–	0.175 (0.025)	0.180 (0.027)	-	**0.161** (0.006)	0.193 (0.002)
	SB	–	0.142 (0.021)	**0.138** (0.018)	-	0.143 (0.017)	0.154 (0.001)
Kin8nm	NG	0.500 (0.009)	0.473 (0.012)	0.480 (0.013)	-	**0.458** (0.005)	0.461 (0.007)
	DER	**0.450** (0.014)	0.459 (0.008)	0.451 (0.028)	-	0.467 (0.025)	0.475 (0.019)
	MDN	–	0.243 (0.002)	0.251 (0.004)	-	0.243 (0.002)	**0.236** (0.003)
	SB	–	0.297 (0.010)	0.291 (0.011)	-	**0.288** (0.005)	0.301 (0.008)
Naval	NG	1.143 (0.077)	1.067 (0.028)	1.148 (0.088)	-	0.931 (0.010)	**0.924** (0.012)
	DER	1.141 (0.027)	1.054 (0.182)	0.918 (0.042)	-	**0.911** (0.062)	0.913 (0.048)
	MDN	–	0.287 (0.049)	0.309 (0.147)	-	**0.141** (0.039)	0.167 (0.038)
	SB	–	**0.284** (0.037)	0.295 (0.050)	-	0.300 (0.058)	0.309 (0.033)
Power	NG	**0.060** (0.001)	**0.060** (0.001)	**0.060** (0.001)	-	**0.060** (0.000)	0.061 (0.000)
	DER	0.058 (0.001)	0.058 (0.002)	**0.057** (0.001)	-	**0.057** (0.001)	**0.057** (0.001)
	MDN	–	**0.055** (0.000)	0.056 (0.000)	-	**0.055** (0.000)	**0.055** (0.000)
	SB	–	**0.056** (0.001)	**0.056** (0.001)	-	**0.056** (0.000)	0.057 (0.001)
Protein	NG	0.682 (0.007)	0.681 (0.009)	0.681 (0.007)	-	0.674 (0.005)	**0.671** (0.004)
	DER	0.699 (0.005)	0.694 (0.008)	**0.680** (0.004)	-	0.688 (0.002)	0.709 (0.003)
	MDN	–	0.570 (0.005)	0.579 (0.006)	-	**0.564** (0.006)	0.583 (0.004)
	SB	–	0.601 (0.002)	**0.599** (0.002)	-	0.601 (0.003)	0.616 (0.002)
Yacht	NG	0.996 (0.005)	0.999 (0.007)	0.999 (0.008)	-	**0.965** (0.005)	0.966 (0.009)
	DER	0.972 (0.061)	0.980 (0.046)	**0.959** (0.020)		1.008 (0.016)	0.963 (0.038)
	MDN	–	0.941 (0.057)	0.954 (0.036)	-	0.884 (0.045)	**0.834** (0.025)
	SB	–	0.746 (0.009)	0.743 (0.018)	-	**0.731** (0.031)	0.784 (0.064)