Comparison of Epistemic Uncertainty Quantification Methods for Out-of-Distribution Detection in Autoencoder–RNN Surrogate Model of Molecular-Continuum Flow Simulations

Sabrina Ebert[1], Olga Catalan Aragall[2], Malte Buschmann[1], Philipp Neumann[1,2]

[1] *Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany*
[2] *High Performance Computing and Data Science, Universität Hamburg, Germany*
sabrina.ebert@desy.de

Abstract. Neural network surrogates have been shown to decrease computational costs of simulations, but often at the risk of unreliable predictions. This work integrates and evaluates multiple epistemic uncertainty quantification methods for a reproduced convolutional autoencoder–recurrent neural network surrogate architecture for molecular data in a coupled spatiotemporal molecular-continuum flow prediction. The surrogate is trained on an idealized Kármán vortex street dataset generated using the molecular–continuum simulation framework MaMiCo, and evaluated on three out-of-distribution datasets with progressively increasing shifts from the training distributions. In the Autoencoder model, the Deep Ensemble method sets a strong baseline, but after fine-tuning, both Gaussian Processes and Evidential Deep Learning show promising detection skills and faster inference than Deep Ensemble. This trend continues in the Autoencoder-RNN, which employs a propagation approach for Evidential distributions and an RNN-influenced latent space for Gaussian Processes.

Keywords: uncertainty quantification · epistemic uncertainty · out-of-distribution detection · autoencoder · recurrent neural network · gaussian processes · evidential deep learning · surrogate modeling · molecular-continuum · flow prediction.

1 Introduction

Machine learning models, particularly deep neural networks, are increasingly used for scientific computing to replace or accelerate numerical solvers. In many application areas, data-driven surrogate models have shown the ability to accurately approximate complex dynamical systems while significantly reducing the computational cost compared to traditional physics-based simulations [1,13]. However, standard neural networks only offer point predictions and mostly lack

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

estimates of prediction confidence. This can lead to significant damage and unsafe decisions in critical applications [14,15]. Recently, there has been an increasing emphasis on enhancing the assessment of reliability and trustworthiness in neural network predictions, particularly in regulatory and operational guidance frameworks [20,21]. Uncertainty quantification (UQ) methods in neural networks are a crucial area of research aimed at contributing to this goal, by providing more clarity on the model's uncertainties regarding specific data inputs by integrating methods with different methodological backgrounds, often making use of a variation of Bayesian formalism [5,11]. Generally, epistemic UQ methods for machine learning are well established [5,6,3,4]. However, considering UQ for fluid dynamics, flow data can exhibit strong spacetime coupling, multi-scale structure, or PDE-constrained dynamics, which complicates the direct transfer of insights from standard vision benchmarks. Furthermore, while some studies have investigated UQ for, e.g., turbulence closures, sequence models, and more recently, physics-informed networks [17,18,19], UQ methods for molecular-continuum flow simulations are even scarcer [1,12]. However, further research into UQ methods is especially valuable in this domain, as neural network surrogates are increasingly being explored to reduce their high computational costs by substituting parts or even entire simulations [1]. Clarifying when these surrogates operate outside their training distribution, commonly referred to as out-of-distribution (OOD) detection, can help mitigate incorrect prediction results. For example, this can be achieved by reverting back to the more expensive default simulation for flagged cases of inputs. Furthermore, UQ implementations for architectures such as autoencoder-recurrent networks (AE-RNN), as considered in this work, remain limited as well.

Contributions

This paper describes the main findings from the evaluation of five epistemic UQ methods in a reproduced AE-RNN surrogate architecture [1] for spatiotemporal molecular-continuum flow prediction. These simulations enable a transient molecular zoom-in on flow structures in selected regions, whereas big parts of the fluid dynamics are computed by traditional computational fluid dynamics (CFD) solvers. The contributions are:

- Comparison of key epistemic UQ approaches, including Deep Ensembles, Monte Carlo Dropout, Laplace, Gaussian Processes, and Evidential Deep Learning, along with the spatiotemporal surrogate.
- Out-Of-Distribution-Detection analysis across three increasingly severe shift-datasets: in-domain extrapolation, physics-shifted Couette flow, and physically invalid random inputs.
- Short discussion of uncertainty behavior and limitations in cascade-connected NN surrogate pipelines (AE $\rightarrow$ RNN).

Within the reproduced AE-RNN surrogate of Jarmatz et al. [1], the selected methods for OOD detection cover a wide range of properties described in the methodology section. Useful methods may later be integrated into the surrogate for practical applications. Section 2 summarizes the reproduced AE-RNN

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

surrogate model, the epistemic UQ estimators, the training, and evaluation procedures, as well as the used datasets. Section 3 reports the empirical comparisons of the methods (incl. remarks on the Gaussian Process kernel fine-tuning and AE vs. AE-RNN behavior). Practical implications, limitations, and an outlook are provided in Section 4.

2 Methodology

Molecular-Continuum Simulations and Surrogate Architecture
Data Generation Simulations: The training and evaluation data in this work originate from molecular-continuum flow simulations performed with the open-source MaMiCo (Macro-Micro Coupling) framework [2]. MaMiCo couples a particle-based molecular dynamics (MD) solver, e.g., SimpleMD, with a continuum CFD solver, e.g., based on the lattice Boltzmann method (LBM). The MD simulation is embedded in a larger CFD domain, enabling the simulation of multi-scale flow phenomena where molecular effects interact with continuum dynamics. In MaMiCo, this is achieved through bidirectional coupling of density and momentum samples using overlap layers and boundary conditions [2], indicated by the dashed areas in Figure 1, and nested time stepping, allowing the faster MD dynamics to equilibrate within each continuum update [2]. Molecular-continuum simulations, in particular the nested MD simulations, remain computationally expensive. This motivates the use of neural surrogate models for faster predictions. However, surrogate modeling can be challenging due to sensitivity to initial conditions and settings, which may result in inaccurate behavior with OOD inputs. Thus, integrating methods for quantifying epistemic uncertainty can help detect potential overconfident situations early, before causing incorrect predictions.
Base Surrogate Architecture: We base our analysis on the convolutional AE-RNN surrogate architecture from [1], which is intended to replace expensive molecular simulations in molecular-continuum flow simulations. The model consists of: (i) a convolutional autoencoder that maps the velocity-component field to a low-dimensional latent representation and (ii) a recurrent neural network that predicts auto-regressively the next temporal step in the latent space.

Epistemic Uncertainty Estimators
Given a wide variety of uncertainty estimators [5,6], we representatively compare Deep Ensembles, Monte Carlo (MC) Dropout, Gaussian Processes (GP), Laplace, and Evidential Deep Learning (EDL). Besides Deep Ensembles as a baseline, MC Dropout and GP are used because of their popularity and different mechanisms, while Laplace is implemented as a representative post-hoc method, and EDL as an example of a more novel method with promising computational time reduction and, at the same time, efficient data and model uncertainty information assessment. Further method details and code remain in that repository.[1]

[1] https://github.com/se04ber/flow-uq-thesis

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

Deep Ensembles: Deep Ensembles train multiple models independently, using the variance in their predictions to estimate epistemic uncertainty [4,5].
Diversity in ensembles can be achieved by variations in model architectures, weight initializations, hyperparameters, or data subsets. We focus on weight initialization, which, while having a smaller effect on predictive accuracy than data subsetting, still improves uncertainty estimates and is easier to implement [4,5,6]. Deep Ensembles are considered robust for estimating epistemic uncertainty [5,6] and serve as the baseline for this study. Their effectiveness may arise from exploring multiple modes of the loss landscape instead of relying on a single optimum, as in typical Bayesian approaches [10]. Although performance often stabilizes with a few ensemble members [4], the need for multiple models increases computational costs, highlighting the need for other approaches [5,6]. Six Deep Ensemble members are used as a balance, since convergence is most often observed around five [4].
Monte Carlo Dropout: MC Dropout is a Bayesian approximation method that uses dropout layers during training and inference to estimate epistemic uncertainty. By performing multiple stochastic forward passes, it calculates predictive variance [3,5]. This approach, which can approximate Bayesian neural network inference under certain conditions [3], is valued for its ease of implementation and theoretical basis. However, it requires dropout layers to be active at test time and can reduce predictive accuracy if the dropout rate, which is estimated here through grid search, is too high, often leading to a trade-off between uncertainty estimates and performance [5,6]. Here, three dropout layers are subsequently incorporated into the model with a dropout rate of ($p = 0.001$), where prediction accuracy was not significantly impacted in grid search. The model is then retrained.
Laplace Approximation: models the posterior distribution of network weights locally around a mode by fitting a Gaussian distribution using curvature information (typically based on the Hessian or more sparse approximations) [16]. As a post-hoc method, it requires no changes to the model architecture or training procedure. However, its applicability and interpretability can be limited in high-dimensional neural networks. Here, exact curvature computations are often infeasible, so strong sparsity approximations, e.g. diagonal or low-rank approximations, of the Hessian and dimensionality/parameter reductions are commonly required [5]. We apply a sparse diagonal Laplace with at most 30 last layer parameters, trading posterior quality for at least comparable inference time, here a main criterion for incorporation.
Gaussian Processes: provide a non-parametric Bayesian framework in which epistemic uncertainty naturally arises from predictive variance and increases with distance from observed training data [9]. The prior choices, e.g., kernel function and its hyperparameters, strongly influence GP sensitivity and calibration. Based on preliminary comparisons of common kernels (RBF, Matérn, Rational Quadratic) and subsequent hyperparameter tuning, an RBF kernel is used here, with further fine-tuned hyperparameters: a length scale of 10.0 and a signal noise variance of 1.0, obtained via grid search.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

Evidential Deep Learning: methods estimate predictive uncertainty in a single forward pass by learning to output distributional parameters, including aleatoric and epistemic uncertainty through evidence accumulation. This is achieved by predicting the parameters of a Normal-Inverse-Gamma (NIG) distribution through learned evidence, balanced with an additional error-based regularization term in the loss [7,8]. The implementation is based on [8], the last layer is transformed into a NIG output layer, the training procedure is adjusted, and the model is retrained. Additionally, the offset of some parameters is slightly increased, which has a mildly positive impact on the epistemic uncertainty separation magnitude.

Training Procedure

Training follows the original AE-RNN workflow described by Jarmatz et al. [1] with minor practical adaptations for reproducibility and UQ integration. In brief, the AE is trained first for spatial compression and reconstruction, after which the RNN is trained on sequences in the AE latent space to incorporate temporal behavior. All UQ variants are trained by adjusting the baseline setup to meet the specific requirements of each method. Additional details regarding the models and scripts used for grid search and training, which include information about the hyperparameters and configuration, can be found in the same repository.

Evaluation Metrics

All metrics are derived from the epistemic uncertainty predicted through the epistemic methods for the velocity field prediction results. We obtain a single uncertainty score for each prediction by averaging the uncertainty field across each data array, which results in a scalar value for each run. We then calculate the mean and variance over the data arrays for each dataset. We then evaluate these scores over all samples in each dataset. To assess OOD detection, we combine the in-distribution (ID/Reference) dataset with each OOD dataset and assign binary class labels $y = 0$ for ID and $y = 1$ for OOD. The ROC-AUC then measures how well the raw uncertainty scores rank ID against OOD samples, with values above 0.5 indicating better-than-random discrimination. Further calibration is conducted by mapping uncertainty to a probability-like score using min-max normalization and comparing it to binary threshold targets with the Brier Score. Besides, the Expected Calibration Error (ECE) is another popular metric in the context of epistemic uncertainty quantification, measuring the alignment between predicted confidence and actual frequencies across confidence bins, and also aids in comparability. Lower Brier and ECE values indicate better-calibrated uncertainty.

Datasets

Training Dataset: The surrogate model is trained using data from MD boxes of an idealized vertically symmetric Kármán Vortex Street (KVS) flow dataset, generated with the MaMiCo framework, in which each MD box and solver are embedded and coupled to the KVS running outer domain and CFD solver [2].

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

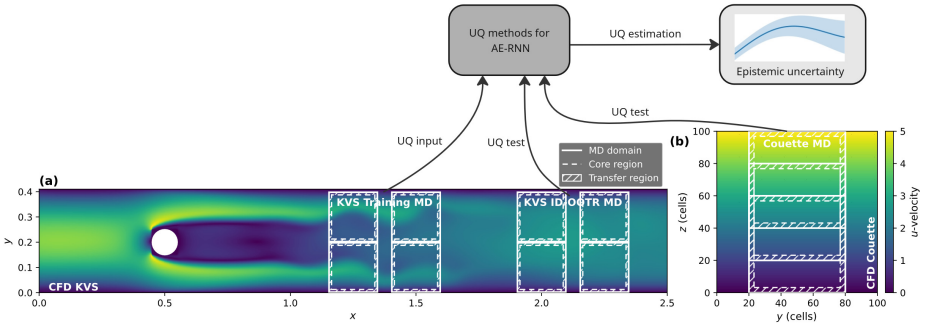

Fig. 1. Sketch of the molecular–continuum simulation setups with molecular dynamics (MD) data generation for model training, testing, and epistemic evaluation. **(a)** Kármán vortex street (KVS) simulation: a z-axis–symmetric cylinder induces vortex shedding. MD boxes are placed in vortex formation regions for training and test ID evaluation, and in regions with decaying vortices to form one out-of-distribution (OOD) test case. Owing to the cylinder and flow symmetry, MD positions are sampled from a horizontal subset along the z-axis. **(b)** Couette flow simulation, used exclusively as an OOD flow case. MD box positions are sampled along the y-axis in accordance with the horizontal flow symmetry. In addition, varying the KVS initialization periods and the Couette wall velocities introduces diversity in vortex-shedding periods and shear strengths, respectively.

Additionally, its split test dataset is used as an ID reference against the increasingly OOD datasets. A 2D slice of the domain setup is shown in Figure 1(a). The CFD KVS exhibits periodic vortex shedding, creating a complex spatiotemporal flow scenario. Simulations run for extended periods to ensure flow convergence. The 2D symmetry of the CFD KVS flow along the z-axis allows efficient data collection along a vertical subset, while multiple initialization durations (20,000 to 28,000 timesteps) enhance temporal data diversity further by capturing different shedding phases [1]. The simulation setup files, including parameter scripts, are available in the public repository. Macroscale CFD would suffice for KVS and the other OOD flow cases below, including Couette flow. We leverage them for the MD-continuum approach here [1] because they present extensively studied benchmarks for laminar flow with well-studied properties. This allows to quantify the quality of the surrogate constructed for the nested MD solver.

OOD Evaluation Datasets: Three datasets are finally considered for analysis. Ordered by increasing severity:

– *In-Domain Out-of-Training Region (ID-OOTR):* Contains spatially extrapolated data within the KVS domain, located far from the original training regions. The dataset exhibits weaker flow gradients while preserving the overall flow structure.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

– *Couette flow:* Represents a domain with a distinctly different flow pattern, exhibiting a horizontally symmetric, steady-state linear velocity profile.
– *TrueRandom:* Consists of randomly sampled values within numerical ranges comparable to the other flow scenarios, but without any underlying physical meaning, besides noise, serving as a stress test for OOD detection.

Additional tested datasets, including linear shifts and rotations, displayed behaviors that fell between these extremes.
Each dataset consists of multiple MD box run results, extracted from runs with different MD spatial locations within the MaMiCo coupled CFD domain and simulations of varying temporal durations. For evaluation, five runs are performed per dataset, and the mean and variance are computed from the averaged results of each dataset. Each run result contains an MD domain volume of $1000 \times 24 \times 24 \times 24$ (time $\times x \times y \times z$) data points. The MD boxes are located within a flow domain measuring $2.5 \times 0.41 \times 0.41$ for KVS, featuring a cylindrical obstacle at $x \in [0.45, 0.55]$ and $y \in [0.15, 0.25]$, which creates a wake region for vortex shedding.
In Couette flow, the cubic domain measures $100 \times 100 \times 100$ between two parallel plates, with MD box locations chosen spanning the channel height to capture the linear velocity profile. Specific positions for the OOD datasets, ID-OOTR and Couette, are illustrated in Figure 1.

3 Results

GP Kernel Hyperparameter Sensitivity
For GPs, we select a Radial Basis Function (RBF) kernel after initially comparing it with common alternatives such as the Matérn and Rational Quadratic kernels. The RBF kernel exhibited the largest uncertainty magnitudes and the clearest separation of OOD test data during the hyperparameter grid search. To aid in interpreting and fine-tuning the resulting distance-based epistemic uncertainty behavior, Figure 2 visualizes the latent space of the AE using Principal Component Analysis (PCA), highlighting the dimensions with the highest variance. Notably, samples from the Couette OOD dataset (shown in purple) are distinctly separated from the training points (shown in grey) from the KVS dataset, both in terms of shape and distance. This distinct separation explains the pronounced OOD response observed in the GP uncertainty for this dataset, especially compared to other methods focused on model generalizability, later.
Grid searches were performed to find optimal prior kernel and kernel hyperparameters (e.g., length scale, signal variance) that maximised separation and saturation of epistemic uncertainty, improving the model's discriminative capability for ID/OOD-data and between OOD scenarios.
Figure 2 illustrates this importance for the capability of two hyperparameter configurations. Notably, in the well-calibrated case shown in the right columns, the distances from OOD datasets to the ID epistemic uncertainty distribution follow the expected order, except for the Couette dataset, which is an outlier due to its distinct, simple, well-organised spatial flow features. The extrapolated IDs

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

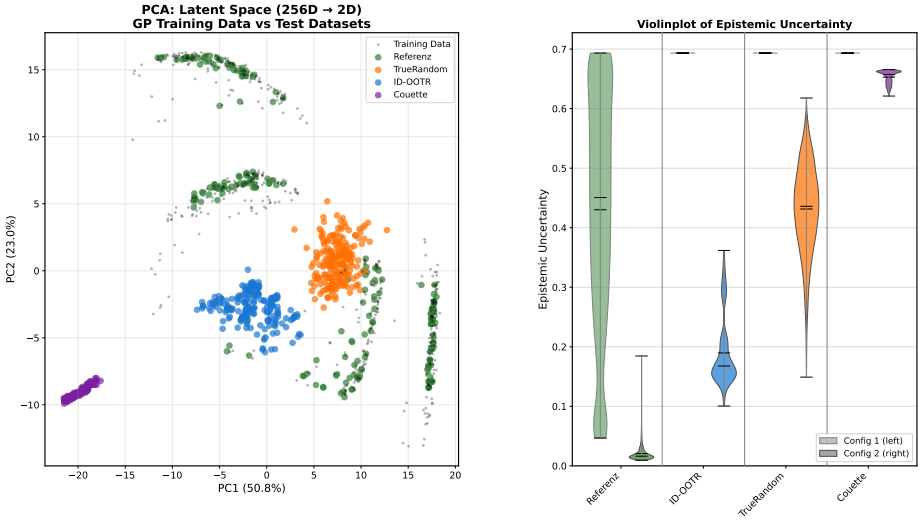

Fig. 2. Left: PCA projection of the AE latent space for all datasets. Right: GP epistemic uncertainty distributions for two RBF kernel hyperparameter configurations obtained from a fine-tuning grid search (initial vs. tuned), shown as violin plots. The PCA visualization indicates a clear separation between the KVS Training and Test Inference datasets (grey and green) and the out-of-distribution (OOD) datasets, most notably the Couette dataset (purple). The GP grid search focuses on RBF kernel hyperparameters (signal variance, noise, and length scale). The violin plot compares the initial configuration (left columns) with the grid-search optimum (right columns), with the primary adjustment in the length scale. The calibration yields a more sensitive and less saturated epistemic uncertainty distribution, improving ID/OOD separability and enabling more robust, case-specific OOD detection.

from the ID-OOTR dataset (in blue) rank nearest in distance, often overlapping in mean with the Reference dataset outliers, followed by the unstructured noise dataset. While the ranking of Couette against the noise dataset may be undesirable under good generalization, further adjustments can be made depending on the distance ordering, such as additional kernel hyperparameter tuning to increase shift sensitivity, trait-informed GP kernel modifications (e.g., less penalizing datasets with high internal density here affecting well-generalizing linear flows such as Couette), or, alternatively, latent-space transformations (e.g., density-aware).

AE Epistemic Results

Due to the time needed to converge and fine-tune a model, the five methods are first only incorporated into the AE for evaluating epistemic uncertainty. The performance and OOD detection results are shown in Figures 3 and 4.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

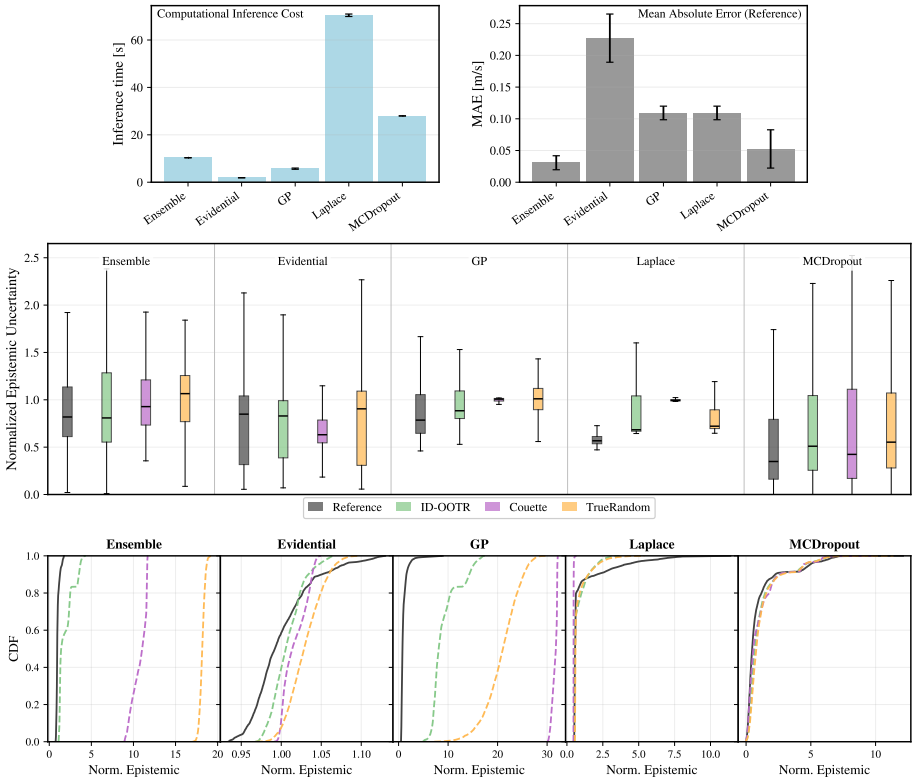

Fig. 3. The upper panel reports computational metrics for inference time and prediction accuracy, while the middle and lower panels present mean-normalized (of their own datas by the means, boxplots and cumulative distribution functions (CDFs) of epistemic uncertainty for the AE-only evaluation. For the boxplots the epistemic results are flattenend spatially and temporally and divided by their own mean per dataset to show the relative change in epistemic distribution. While in the lower panel all CDFs curves are normalized by the reference mean to show separation. The Deep Ensemble reference shows a clear increase in epistemic uncertainty from ID-OOTR to TrueRandom, with the higher ID-OOTR variance reflecting samples that remain closer to the training distribution. All methods consistently identify TrueRandom as out-of-distribution (OOD), whereas Couette behavior is method-dependent. GP and Laplace approaches exhibit comparatively lower variance with near-saturation effects, while Deep Ensembles, MC Dropout, and EDL show higher variance and less consistent OOD separation, highlighting differences in sensitivity to dataset complexity and underlying methodology.

The best-performing methods are integrated into the full AE-RNN pipeline, with results visualized in Figure 4. Results were obtained by running inference for one coupled MD box each through the AE and averaging across multiple runs for

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

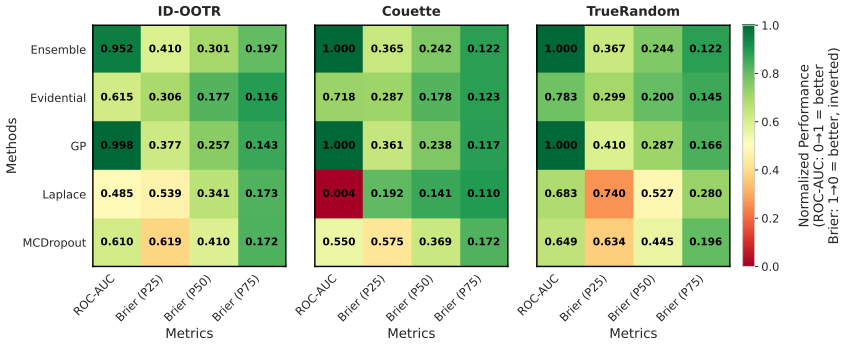

Fig. 4. The AE-only out-of-distribution (OOD) detection heatmaps show ROC-AUC and Brier-type scores at thresholds P25/P50/P75 across OOD datasets ordered by increasing shift severity, culminating in fully unstructured noise. Since OOD detection is typically cast as a binary classification problem via a chosen statistical threshold, this analysis evaluates how effectively each method detects OOD samples under varying shift levels. As discussed in Figure 3, performance depends on the separation between the epistemic uncertainty distributions of OOD and in-distribution/test cases, which can be effectively evaluated using ROC-AUC and calibration using the Brier score. Consistent with the distributional results, Deep Ensembles perform as expected, demonstrating strong results. GP and EDL continue to show competitive performance. In contrast, both MC Dropout and Laplace are less reliable across the scores. The Laplace Couette ROC-AUC is poor despite reasonable exceedance-based Brier results, indicating weak OOD detection reliability without threshold calibration.

each dataset. Figure 3 summarizes operational metrics, including inference time and Mean Absolute Error (MAE) for reconstruction, along with distribution results shown in boxplots and cumulative distribution functions (CDFs).

EDL and GP methods have faster inference speeds than the Deep Ensemble baseline. In contrast, MC Dropout and the Laplace method incur additional computational costs due to repetitions needed for epistemic saturation sampling and curvature calculations, despite the Laplace method's parameter space and Hessian matrix sparsity approximations. Further details on the underlying grid search and fine-tuning plots are available in the GitHub repository. For prediction accuracy, the Deep Ensemble baseline outperforms other methods, as illustrated in Figure 3. While MC Dropout reduces overfitting, it requires careful fine-tuning of the dropout rate to maintain prediction accuracy, leading to a relatively low dropout rate compared to the literature.

EDL has the highest MAE due to its complex loss function but performs adequately after tuning.

The normalized boxplots and CDFs for each method visualize the distribution and shift of epistemic uncertainty estimates. Most methods produce the expected right shift in uncertainty for OOD datasets, with particular variations per method for the Couette and ID-OOTR datasets, compared to a consistent right shift for the randomly sampled data. Figure 3, the lower panel CDFs, shows

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

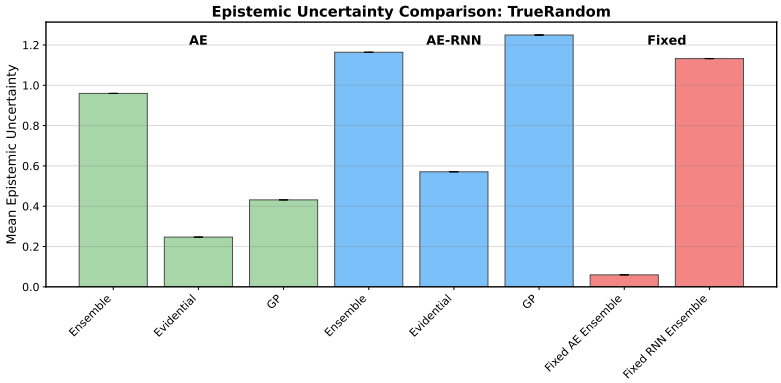

Fig. 5. Comparison of mean epistemic uncertainty magnitudes for AE (green) and AE–RNN (blue) implementations, including fixed partial Deep Ensembles (with either AE or RNN fixed), illustrating each component's contribution to the total epistemic uncertainty on a representative OOD dataset (random noise stress test). All methods exhibit increased epistemic uncertainty when integrated into the full AE–RNN framework. For the fixed partial Deep Ensembles, the AE ensemble contributes substantially more than the RNN ensemble when varied independently. This behavior is likely driven by latent space misalignment, correctly reflected here as a strong OOD response when different AEs are paired with an RNN on which they were not jointly trained.

that Deep Ensembles and GP exhibit strong right shifts and clear separations, particularly in unstructured noise (TrueRandom) and Couette flow scenarios. The EDL method results show some overlap with ID-OOTR and Couette with ID, but still demonstrate a noticeable right shift overall. In contrast, MC Dropout and Laplace methods have less pronounced right-shifts, sometimes even negative in the case of Laplace, and greater overlap in epistemic uncertainty results. The Couette flow behavior is less consistent, with Deep Ensembles and GP showing the strongest separation, followed by EDL, whereas Laplace does not even reliably detect Couette as OOD.

Figure 4 summarizes OOD detection capabilities with ROC-AUC and Brier-type scores across various thresholds in a heatmap. Detailed results and error correlation analyses using Spearman correlation are available in the GitHub repository. Ensembles and GP achieve the best OOD separation, while EDL remains competitive at lower thresholds, despite some reductions in margins. The differences in how various methods increase OOD certainty for Couette are also visible here. The GP is notable for again predicting Couette as the strongest OOD case due to its different flow properties. Also notably, Monte Carlo Dropout, as expected from its epistemic uncertainty distribution overlap, performs poorly, having lower thresholds for the Brier score, indicating bad calibration, likely due to the low dropout rate relative to the literature, which reduces overall sensitivity.

However, it performs adequately at the higher OOD detection percentile threshold P75 while at the same time having a low impact on prediction accuracy,

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

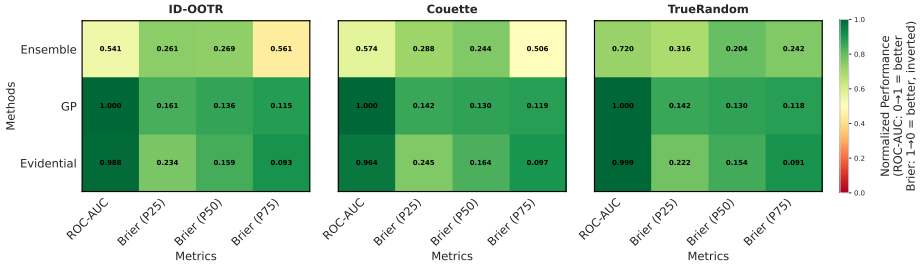

Fig. 6. The AE–RNN OOD detection heatmaps use the same spatial-mean epistemic score per run as the AE-only heatmaps (Figure 4). Similarly to that figure, they illustrate qualitative OOD detection across increasing shift severity via ROC-AUC and Brier at several ID-derived percentiles (Sec. 2). GP and EDL again perform strongly, with AE–RNN EDL slightly outperforming its AE counterpart, highlighting the benefit of capturing interdependent model uncertainty. Divergent ordering of OOD severity for the Couette dataset arises from methodological differences discussed in Figure 3. The Deep Ensemble shows less stable behavior across thresholds due to AE–RNN misalignment when RNNs are trained on a single AE, leading to multimodal epistemic uncertainty distributions from varying member performance, while still contributing to overall OOD detection.

but the trade-off could possibly be optimized further. For Laplace on Couette, the ROC-AUC is about 0.04 (see Fig. 4), indicating extremely poor threshold-independent ID/OOD discrimination based on the raw epistemic scores. This is also reproduced in the full ROC curves, but only for Laplace. The less extreme Brier-type values at P25, P50, and P75 for Couette Laplace reflect performance at fixed threshold targets for the mapped probability. They are less sensitive to overlap and more sensitive to the shape of the uncertainty distribution, rather than to the global ranking measured by ROC-AUC.

AE-RNN Pipeline: OOD Detection and Uncertainty Propagation

Based on method performance and practical considerations at the AE level, a reduced set within the full AE-RNN pipeline has been evaluated, consisting of Deep Ensembles, GP, and EDL.

The OOD metric heatmaps produced for the AE-RNN in Figure 6 are similar to those for the AE, as shown in Figure 4. For detailed results and error correlation metrics, please refer to the GitHub repository.

The Deep Ensemble employs a complete AE-RNN ensemble, while the GP method utilizes the RNN-influenced latent space to capture epistemic uncertainty, focusing on latent space distances calculated post-RNN prediction.

For EDL, a propagation and merging strategy is applied to the Normal-Inverse-Gamma (NIG) distribution outputs from the AE and RNN implementations, accounting for interdependencies of the AE and RNN in the final epistemic uncertainty estimation output of the AE decoder.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

Besides, when the RNN was added, systematic results for epistemic uncertainty were observed. Figure 5 shows that the RNN component (in green) increases epistemic uncertainty compared to the AE alone (in blue).
Additionally, the AE and RNN ensembles showed different contributions to epistemic uncertainty (indicated in red), likely strongly influenced by misalignment between the AE encoder's latent space output and the RNN's expected latent space input, as all RNNs were trained on the same AE. Repeating the experiment with RNNs trained on their corresponding AEs would enhance comparability for the deep ensemble as an AE-RNN baseline and provide greater clarity. The Deep Ensemble, except for the noise stress test, is less competitive. As discussed and shown in Figure 5, its performance likely suffers from missing joint AE–RNN training. The resulting inter–AE–RNN ensemble member variance caused multimodal epistemic uncertainty distributions, making results strongly threshold dependent. Considering full uncertainty distribution shapes, in addition to thresholds, could potentially further refine OOD classification and interpretation.

4 Conclusion

Summary: This work implemented and compared several epistemic UQ methods within a reproduced AE–RNN surrogate trained on KVS simulations and evaluated across datasets with progressively increasing distribution shift. As expected, Deep Ensembles establish a strong baseline, while both GP and EDL show robust OOD detection performance, particularly for the Couette flow case. Notably, these approaches achieve competitive results with generally favorable runtime characteristics. However, EDL requires architectural modifications and careful hyperparameter tuning, and GP performance remains sensitive to kernel and hyperparameter calibration to avoid over- or undersaturated epistemic estimates. GPs may also flag well-generalizing data as OOD, which must be analyzed and adjusted based on specific application needs.
Discussion: EDL is particularly notable for its strong OOD detection while simultaneously modeling both aleatoric and epistemic uncertainty, however, it also produced the highest MAE in this setup. Especially noteworthy is the stronger OOD detection observed in the AE–RNN propagation approach, which potentially highlights the importance of accounting for model interdependence. Beyond strong detection performance, GPs additionally offer easier interpretability through their latent space use and require minimal architectural changes, but they remain sensitive to kernel tuning and may overseparate structurally different yet still predictable flow regimes. Both methods demonstrate robust behavior across the evaluated datasets. In contrast, Laplace and MC Dropout were less reliable in OOD detection across all datasets, particularly for Couette. However, whether this is a disadvantage depends on the specific use case and the characteristics of the data. Laplace is likely limited here by its local and computational approximations around the shape of the minimum, while MC Dropout likely suffers from a conservative dropout configuration that reduces calibration quality and OOD sensitivity.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

For epistemic propagation in the AE–RNN chain, GP and EDL remain robust, whereas Deep Ensemble performance is strongly influenced by AE–RNN pairing. Training each RNN jointly with its corresponding AE would likely reduce latent-space misalignment and improve ensemble consistency, highlighting the importance of coordinated training in dependent surrogate architectures.
Outlook: Future work should further refine MC Dropout calibration and explore additional UQ approaches. A more systematic analysis of the loss landscape could help clarify interactions between model architecture, data characteristics, and epistemic uncertainty behavior. Expanding experiments to more diverse test scenarios and performing cross-architecture studies, including PINNs, CNFs, Transformers, and generative models, would be interesting to better assess robustness, sensitivity, and propagation of epistemic uncertainty in complex flow prediction systems.

Acknowledgments. This work was conducted at the Department of Informatics at DESY Hamburg. The author gratefully acknowledges Philipp Neumann, Juan Pedro Mellado González, Olga Catalan Aragall, and Yannis Schumann for their supervision and guidance, and the rest of the team at DESY for their valuable feedback and insightful discussions.

Disclosure of Interests. The author has no competing interests to declare.

References

1. Jarmatz, P., Lerdo, S., Neumann, P.: Convolutional recurrent autoencoder for molecular-continuum coupling. In: International Conference on Computational Science (ICCS 2023) (2023). https://www.iccs-meeting.org/archive/iccs2023/papers/140760520.pdf
2. Neumann, P., Flohr, H., Arora, R., Bungartz, H.-J., Eckhardt, W.: MaMiCo: Software design for parallel molecular-continuum flow simulations. Comput. Phys. Commun. **200**, 1–11 (2016). https://www.sciencedirect.com/science/article/abs/pii/S0010465515004129
3. Gal, Y.: Uncertainty in Deep Learning. PhD thesis, University of Cambridge (2016). https://www.cs.ox.ac.uk/people/yarin.gal/website/thesis/thesis.pdf
4. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems **30** (2017). https://papers.neurips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf
5. He, W., Jiang, Z., Xiao, T., Xu, Z., Li, Y.: A Survey on Uncertainty Quantification Methods for Deep Learning. arXiv preprint arXiv:2302.13425 (2023). https://arxiv.org/abs/2302.13425
6. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al.: A survey of uncertainty in deep neural networks. Artif. Intell. Rev. **56**, 1513–1589 (2023). https://doi.org/10.1007/s10462-023-10562-9

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31

7. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in Neural Information Processing Systems **31** (2018). https://papers.nips.cc/paper/7580-evidential-deep-learning-to-quantify-classification-uncertainty.pdf
8. Amini, A., Schwarting, W., Soleimany, A., Rus, D.: Deep evidential regression. In: Advances in Neural Information Processing Systems, vol. 33, 14927–14937 (2020). https://proceedings.neurips.cc/paper/2020/hash/aab085461de182608ee9f607f3f7d18f-Abstract.html
9. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press (2006). https://doi.org/10.7551/mitpress/3206.001.0001
10. Fort, S., Dziugaite, G.K., Roy, D.M., Ganguli, S.: Deep ensembles: A loss landscape perspective. arXiv preprint arXiv:1912.02757 (2019). https://arxiv.org/abs/1912.02757
11. MacKay, D.J.C.: Bayesian Interpolation. Neural Computation **4**(3), 415–447 (1992). https://doi.org/10.1162/neco.1992.4.3.415
12. Guo, X., Li, J., Yang, W.: Multiscale modeling and computation of elasto-plastic materials via deep neural networks. Computer Methods in Applied Mechanics and Engineering 344, 1006–1032 (2019). https://doi.org/10.1016/j.cma.2018.10.029
13. Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T.R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., Willson, M.: Probabilistic weather forecasting with machine learning. Nature **637**, 84–90 (2025). https://doi.org/10.1038/s41586-024-08252-9
14. Rappaport, E.N.: Fatalities in the United States from Atlantic tropical cyclones: New data and interpretation. Bull. Am. Meteorol. Soc. **95**(3), 341–346 (2014). https://doi.org/10.1175/BAMS-D-12-00074.1
15. Kompa, B., Snoek, J., Beam, A.L.: Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digit. Med. **4**(1), 4 (2021). https://doi.org/10.1038/s41746-020-00367-3
16. Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., Hennig, P.: Laplace redux—effortless Bayesian deep learning. In: Advances in Neural Information Processing Systems, vol. 34, 20089-20103 (2021). https://papers.neurips.cc/paper_files/paper/2021/hash/a7c9585703d275249f30a088cebba0ad-Abstract.html
17. McDermott, P.L., Wikle, C.K.: Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatio-temporal data. Entropy **21**(9), 905 (2019). https://arxiv.org/abs/1711.00636
18. Pollithy, D., Reith-Braun, M., Pfaff, P., Hanebeck, U.D.: Estimating uncertainties of recurrent neural networks in multi-target tracking. In: Proc. 18th Int. Conf. on Informatics in Control, Automation and Robotics (ICINCO), Workshop on Multisensor Fusion and Integration for Intelligent Systems (2020). https://isas.iar.kit.edu/pdf/MFI20_Pollithy.pdf
19. Sperrer, G.: Uncertainty Quantification of Deep Learning Reduced-Order Models for Fluid Dynamics. PhD thesis, TU Wien (2025). https://repositum.tuwien.at/handle/20.500.12708/221761
20. WMO: Guidelines on Ensemble Prediction Systems and Forecasting. WMO-No. 1091 (2012).
21. European Commission: Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) (2021). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29918-5_31