Bayesian Ensemble Regression for Uncertainty Prediction in Data-Sparse Sensor Calibration

1st Felix Wittich
Measurement and Control Research Group
University of Kassel
Kassel, Germany
felix.wittich@mrt.uni-kassel.de

2nd Andreas Kroll
Measurement and Control Research Group
University of Kassel
Kassel, Germany
andreas.kroll@mrt.uni-kassel.de

***Abstract*—This paper proposes an ensemble-based uncertainty quantification framework for Gaussian Process Regression in small-data regression problems, where reliable uncertainty estimates are essential. The method enables a systematic decomposition of predictive uncertainty by capturing variability induced by model hyperparameters while retaining the data efficiency of standard GPR. By interpreting the ensemble as a finite mixture of Gaussian Processes, the approach provides a principled and computationally tractable alternative to fully Bayesian inference. The framework is demonstrated on a micromagnetic sensor calibration task, highlighting its ability to deliver robust predictions together with meaningful uncertainty information. The results show that the proposed uncertainty-aware modeling approach improves reliability in data-scarce settings and supports informed decision-making in material property prediction.**

***Index Terms*—Uncertainty quantification, Soft sensor, Materials science, Gaussian Process Regression**

I. Introduction and Related Work

In small-data calibration problems, uncertainty quantification is essential for reliable use of data-driven predictive models. This is particularly relevant in micromagnetic sensor calibration, where obtaining reference measurements is costly and only a limited number of training data are typically available. In such settings, regression models must not only achieve low prediction error but also provide calibrated confidence intervals indicating the reliability of individual predictions.

Gaussian Process Regression (GPR) is a widely used probabilistic regression framework due to its strong theoretical foundations and its ability to provide predictive uncertainty alongside point estimates [1]. By defining a prior over functions, GPR yields closed-form posterior predictive distributions under Gaussian likelihoods. However, in practical applications, kernel hyperparameters are commonly determined by maximizing the log marginal likelihood (LML), resulting in a single LML estimate. This point-estimation approach ignores uncertainty in the hyperparameters and often leads to underestimated predictive uncertainty, especially in regions with sparse data or weakly identified kernel parameters [2]. The issue is exacerbated in small-data regimes, where epistemic uncertainty dominates and overconfident predictions can be particularly misleading.

Fully Bayesian Gaussian Process models address this limitation by marginalizing over hyperparameters instead of relying on a single MAP solution. Such approaches, typically implemented using Markov Chain Monte Carlo (MCMC) or variational inference, have been shown to improve uncertainty calibration [3], [4]. Nevertheless, fully Bayesian GPs incur substantial computational cost and implementation complexity, limiting their applicability in many practical calibration scenarios. As a result, most real-world GPR applications continue to rely on marginal likelihood optimization, accepting the risk of overconfident uncertainty estimates.

Ensemble-based approaches offer a practical alternative for capturing model uncertainty and good predictive accuracy [5]. In deep learning, deep ensembles have demonstrated strong empirical performance in uncertainty estimation by averaging predictions across independently trained models [6]. This strategy can be interpreted as an approximate Bayesian model averaging procedure, where diversity among ensemble members reflects epistemic uncertainty. Similar ideas have been applied to Gaussian Processes, for example through bagging or bootstrap aggregation, where multiple GP models are trained on resampled subsets of the data [7]. While such GP ensembles can improve robustness and predictive performance, data resampling reduces the effective training dataset size for each model, which is undesirable in data-limited settings. Moreover, bagging primarily addresses sampling variability and does not explicitly model uncertainty in the GP hyperparameters.

Despite the extensive literature on Gaussian Process regression, ensemble-based formulations of GPR remain comparatively underexplored. Existing work on GP ensembles typically introduces model diversity through data-level perturbations or artificial noise mechanisms rather than through principled sampling of hyperparameter uncertainty.

For instance, [8] generate ensemble diversity by injecting stochastic noise into the training labels of sparse GPR models, yielding efficient uncertainty estimates without sampling kernel hyperparameters. Similarly, ensemble strategies discussed in materials and molecular modeling primarily rely on kernel variation, inducing-point selection, or training-set perturbations to improve robustness [9]. While effective in practice,

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 532921704. The authors would like to thank the Metallic Materials Research Group at the University of Kassel for conducting the hard turning tests as well as the X-ray diffraction and micromagnetic measurements.

these approaches do not approximate Bayesian marginalization over GP hyperparameters and lack a clear probabilistic interpretation in this regard.

In this work, an ensemble Gaussian Process Regression framework is proposed that explicitly targets hyperparameter uncertainty while remaining computationally tractable. The key idea is to generate ensemble members by sampling hyperparameters from an approximate posterior distribution obtained via a Laplace approximation around the marginal likelihood optimum, from which multiple plausible hyperparameter configurations are drawn. Each ensemble member is trained on the full data set, thereby avoiding retraining on data subsets and preserving the data efficiency of standard GPR.

The resulting ensemble can be interpreted as a finite mixture of Gaussian predictive distributions, approximating Bayesian model averaging over hyperparameters. This ensemble formulation yields improved uncertainty calibration compared to single-model GPR, particularly in nonlinear, small-data calibration problems such as micromagnetic sensor modeling.

Laplace approximation of the GP marginal likelihood, hyperparameter posterior sampling, kernel diversification, and Bayesian model averaging are individually established techniques. Their integration into a single ensemble GPR workflow, with Laplace-sampled hyperparameters as the primary source of ensemble diversity, has to our knowledge not been studied. We propose such a framework and demonstrate its calibration benefits on a real small-data sensor calibration task.

The remainder of this paper is structured as follows. Section II introduces the Gaussian Process regression framework and discusses uncertainty quantification in small-data settings. Section III presents the proposed ensemble-based approach and details the decomposition of predictive uncertainty. In Section IV, the method is applied to a micromagnetic sensor calibration problem and evaluated against reference models. Finally, Section V concludes the paper and outlines directions for future work.

II. Methods

This work proposes a probabilistic ensemble Gaussian Process Regression framework designed to address a key limitation of standard GPR models: the systematic underestimation of epistemic uncertainty arising from point-estimated kernel hyperparameters and fixed kernel structures.

While fully Bayesian Gaussian Processes marginalize both latent functions and hyperparameters, exact inference is generally intractable, and scalable approximations such as variational inference or Markov Chain Monte Carlo remain computationally demanding and difficult to integrate into standard regression pipelines. Conversely, practical Gaussian Process implementations typically rely on maximum marginal likelihood estimates, which do not account for hyperparameter uncertainty and yield overly confident predictions.

The proposed method retains the analytical tractability and interpretability of classical Gaussian Processes while approximating Bayesian marginalization through an ensemble approach. Specifically, the contributions of this work are:

Algorithm 1 Ensemble GPR via Laplace Hyperparameter Sampling

Require: Data $\mathcal{D} = (X, y)$, kernel set $\{k_1, \dots, k_K\}$, members per kernel M_k, scale s, temperature T
Ensure: Ensemble $\{(\mathcal{M}_m, w_m)\}$
$\mathcal{E} \leftarrow \emptyset$
for each kernel family k **do**
  $\hat{\theta} \leftarrow \arg\min_\theta \mathcal{L}(\theta)$ ▷ LML maximization, Eq. (7)
  $H \leftarrow \nabla^2_\theta \mathcal{L}(\theta)\big|_{\hat{\theta}}$ ▷ finite-diff Hessian, Eq. (9)
  $\Sigma \leftarrow \left(\frac{1}{2}(H + H^\top) + \lambda I\right)^{-1}$
  for $m = 1$ **to** M_k **do**
    $\theta_m \sim \mathcal{N}(\hat{\theta},\, s^2\Sigma)$, clip to bounds ▷ Eq. (11)
    fit $\mathcal{M}_m$ on full $\mathcal{D}$ with θ_m fixed
    $\ell_m \leftarrow \log p(\mathcal{D} \mid \mathcal{M}_m)$
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{(\mathcal{M}_m, \ell_m)\}$
  end for
end for
$w_m \leftarrow \mathrm{softmax}\big(\ell_m/T + \log \pi(\mathcal{M}_m)\big)$ ▷ BMA weights, Eq. (13)
Predict: mixture $p(y^* \mid x^*) = \sum_m w_m \mathcal{N}\big(\mu_m, \sigma_m^2\big)$ ▷ Eq. (12)
Decompose: $\sigma^2_{\text{tot}} = \underbrace{\sum_m w_m \sigma_m^2}_{\text{aleatoric}} + \underbrace{\sum_m w_m (\mu_m - \mu)^2}_{\text{epistemic}}$ ▷ Eq. (15)

- A Laplace-based hyperparameter posterior sampling strategy is proposed that generates diverse Gaussian Process models by explicitly approximating uncertainty in kernel hyperparameters, rather than relying on heuristic perturbations or data resampling alone.
- Ensemble predictions are formulated as probabilistic model averaging over Gaussian predictive distributions.
- The ensemble predictive distribution is treated as a mixture of Gaussians, allowing uncertainty quantification and predictive intervals to be derived from the full mixture rather than from moment-matched Gaussian approximations.

Together, these elements yield an ensemble Gaussian Process framework that provides calibrated uncertainty estimates suitable for small to medium-sized datasets. The methodology can be summarized as shown in Algorithm 1 and is described in more detail in the following.

A. Problem Setting and Objectives

Let

$$\mathcal{D} = (X, y) \tag{1}$$

denote a supervised regression dataset with inputs $X \in \mathbb{R}^{N \times d}$ and targets $y \in \mathbb{R}^N$. The goal is to learn a predictive distribution

$$p(y^* \mid x^*, \mathcal{D}) \tag{2}$$

that is both accurate in its point estimate and reliable in its uncertainty quantification.

Gaussian Process Regression (GPR) offers a Bayesian framework for regression with uncertainty estimates. However, in most practical implementations, kernel hyperparameters are determined via point estimation (maximum marginal likelihood), which neglects hyperparameter uncertainty and can lead to systematically under-dispersed predictive distributions. This limitation becomes particularly severe in small- to medium-sized datasets, where epistemic uncertainty dominates.

To address these shortcomings while retaining the favorable properties of Gaussian Processes, an ensemble Gaussian Process Regression framework is proposed that explicitly incorporates model uncertainty through kernel diversity, approximate hyperparameter posterior sampling, and probabilistic model averaging.

B. Base Gaussian Process Models

Each ensemble member $m \in \{1, \ldots, M\}$ is a Gaussian Process defined as

$$f_m(x) \sim \mathcal{GP}\big(0, k_m(x, x'; \theta_m)\big), \tag{3}$$

with covariance function

$$k_m(x, x') = \sigma_{f,m}^2 \, \tilde{k}_m(x, x') + \sigma_{n,m}^2 \delta(x, x'). \tag{4}$$

The kernel $\tilde{k}_m$ is drawn from a predefined set of kernel families, listed below with $r = \|x - x'\|$:

- **RBF (squared exponential):** $\exp\big(-r^2/(2\ell^2)\big)$
- **Matérn ($\nu = 1.5$):** $\big(1 + \sqrt{3}\, r/\ell\big) \exp\big(-\sqrt{3}\, r/\ell\big)$
- **Rational Quadratic:** $\big(1 + r^2/(2\alpha\ell^2)\big)^{-\alpha}$
- **Composite RBF + ExpSineSquared:** $\exp\big(-r^2/(2\ell_1^2)\big) + \exp\big(-2\sin^2(\pi r/p)/\ell_2^2\big)$

Each kernel is multiplied by a constant signal-variance factor σ_f^2 and combined with an additive WhiteKernel $\sigma_n^2 \delta(x, x')$, as in (4). Length scales ℓ, ℓ_1, ℓ_2, periodicity p, mixture parameter α, signal variance σ_f^2, and noise variance σ_n^2 are determined per kernel from the marginal likelihood maximum and subsequently sampled via the Laplace approximation described in Section II-D3.

This kernel family selection is motivated by the need to cover a broad spectrum of functional behaviors, ranging from smooth and stationary processes to rougher, multi-scale, and weakly periodic structures. This set reflects the standard taxonomy of stationary GP kernels and is not tailored to any particular target. Kernels poorly suited to a given problem automatically receive low BMA weight rather than being manually excluded. Kernel diversity serves as a form of structural uncertainty modeling, allowing the ensemble to represent uncertainty over function classes rather than committing to a single inductive bias.

For each model, the predictive distribution is Gaussian:

$$p_m(y^* \mid x^*, \mathcal{D}) = \mathcal{N}\big(y^* \mid \mu_m(x^*), \sigma_m^2(x^*)\big) \tag{5}$$

with closed-form expressions for $\mu_m(\cdot)$ and $\sigma_m^2(\cdot)$.

C. Model Diversity: Design Choices and Alternatives

Ensemble performance critically depends on the diversity of its members. Several common strategies exist to induce diversity in GP ensembles:

- **Data resampling (bagging or bootstrap)**, which induces variability through different training subsets;
- **Kernel structure variation**, which induces variability in function classes and
- **Hyperparameter perturbation or sampling**, which captures uncertainty within a fixed kernel family.

While data resampling is effective in large-data regimes, it does not directly address hyperparameter uncertainty and can be suboptimal for moderate dataset sizes. Similarly, ad-hoc hyperparameter jittering lacks a probabilistic interpretation. In this work, hyperparameter posterior uncertainty is prioritized and therefore a Laplace-based hyperparameter sampling strategy is adopted.

D. Hyperparameter Uncertainty via Laplace Approximation

1) Motivation: In a fully Bayesian GP, uncertainty over kernel hyperparameters θ is marginalized by

$$p(y^* \mid x^*, \mathcal{D}) = \int p(y^* \mid x^*, \theta, \mathcal{D}) p(\theta \mid \mathcal{D}) \, d\theta \tag{6}$$

Exact marginalization is intractable for most kernels. Variational or Markov Chain Monte Carlo (MCMC) approaches provide solutions but are computationally expensive and difficult to integrate into standard GP toolchains. Therefore Laplace approximation is adopted as a compromise between Bayesian rigor and computational tractability.

2) LML Hyperparameter Estimation: For each kernel family, hyperparameters are estimated by minimizing the negative log marginal likelihood:

$$\mathcal{L}(\theta) = -\log p(y \mid X, \theta) \tag{7}$$

Optimization yields a LML estimate $\hat{\theta}$.

3) Laplace Approximation of the Hyperparameter Posterior: The posterior distribution around $\hat{\theta}$ is approximated using a second-order Taylor series expansion:

$$p(\theta \mid \mathcal{D}) \approx \mathcal{N}\big(\hat{\theta}, H^{-1}\big), \tag{8}$$

where

$$H = \nabla_\theta^2 \mathcal{L}(\theta)\big|_{\theta=\hat{\theta}} \tag{9}$$

is the Hessian of the negative log marginal likelihood. The Hessian H is computed numerically via central finite differences in the kernel's log-parameter space, using a three-point stencil for the diagonal entries,

$$H_{ii} \approx \frac{L(\hat{\theta} + \epsilon e_i) - 2\, L(\hat{\theta}) + L(\hat{\theta} - \epsilon e_i)}{\epsilon^2}, \tag{10}$$

and the standard four-point stencil for the off-diagonal entries, with a step size of $\epsilon = 10^{-2}$. To ensure that the resulting covariance H^{-1} in (11) is symmetric positive definite, H is first symmetrized, a diagonal damping λI with $\lambda = 10^{-4}$ is added, and an eigendecomposition is performed with an

eigenvalue floor of 10^{-10} prior to inversion. In the subsequent sampling step, drawn hyperparameter vectors θ_m are clipped to the kernel's admissible bounds, and up to 30 resamples are attempted should a particular draw lead to a numerically ill-conditioned Gram matrix during refitting.

4) Hyperparameter Posterior Sampling: Ensemble members are generated by sampling

$$\theta_m \sim \mathcal{N}\big(\hat{\theta}, s^2 H^{-1}\big), \tag{11}$$

where the scaling factor s controls the effective posterior spread and it was fixed to $s = 0.8$ in all experiments. It was chosen empirically to balance ensemble diversity against numerical stability of the refitted models, and acts as an upper bound on the posterior spread, since it is reduced locally if a draw leads to an ill-conditioned Gram matrix. Models are refitted with hyperparameters fixed, thereby preserving the sampled posterior variability. This avoids collapsing back to the LML solution and ensures meaningful ensemble diversity. Compared to heuristic length-scale perturbation, this approach yields a probabilistically grounded approximation to hyperparameter marginalization.

E. Ensemble Predictive Distribution

The ensemble predictive distribution is modeled as a finite mixture of Gaussian distributions:

$$p(y^* \mid x^*, \mathcal{D}) = \sum_{m=1}^{M} w_m \mathcal{N}\big(y^* \mid \mu_m(x^*), \sigma_m^2(x^*)\big) \tag{12}$$

with non-negative weights w_m satisfying $\sum_m w_m = 1$.

This formulation generalizes classical GP prediction by allowing the predictive distribution to deviate from Gaussianity when ensemble members disagree.

F. Model Weighting and Probabilistic Model Averaging

Assigning appropriate weights to ensemble members is critical for both predictive accuracy and uncertainty calibration.

1) Bayesian Model Averaging via Marginal Likelihood: Each GP provides a log marginal likelihood estimate $\log p(\mathcal{D} \mid M_m)$. We compute Bayesian model averaging weights as

$$w_m \propto \exp\left(\frac{1}{T} \log p(\mathcal{D} \mid M_m)\right) \pi(M_m), \tag{13}$$

where T is a parameter and $\pi(M_m)$ is a kernel-family prior. The parameter T mitigates the tendency of marginal likelihood to over-select overly confident models. We selected $T = 10$ on the synthetic benchmark by inspecting calibration and predictive log-likelihood over $T \in [1, 20]$; results were insensitive to the exact value within $T \in [5, 15]$. The same value was reused unchanged on the real data without target-specific tuning.

2) Alternative Weighting Strategies: Alternative ensemble weighting strategies include predictive stacking, which optimizes the predictive log likelihood on a validation data set, and cross-validated predictive weighting, based on the expected log predictive density. These approaches require additional validation data or nested cross-validation. In this work, evidence-based weighting is used due to its simplicity and direct Bayesian interpretation.

G. Uncertainty Decomposition

The ensemble structure enables a decomposition of predictive uncertainty into aleatoric and epistemic components.

Let

$$\mu(x^*) = \sum_m w_m \mu_m(x^*) \tag{14}$$

denote the ensemble predictive mean. The predictive variance decomposes as

$$\sigma^2_{\text{tot}}(x^*) = \underbrace{\sum_m w_m \sigma_m^2(x^*)}_{\text{Aleatoric}} + \underbrace{\sum_m w_m \big(\mu_m(x^*) - \mu(x^*)\big)^2}_{\text{Epistemic}}. \tag{15}$$

where $\mu(x^*)$ is the ensemble mean defined in (14). Aleatoric uncertainty captures irreducible noise, while epistemic uncertainty reflects uncertainty due to limited data, kernel choice, and hyperparameter variability.

H. Predictive Intervals

1) Moment-Matched Gaussian Approximation: A Gaussian approximation using $\mu(x^*)$ and $\sigma^2_{\text{tot}}(x^*)$ yields symmetric predictive intervals. While computationally efficient, this approximation may be inaccurate when the ensemble predictive distribution is skewed or multimodal.

2) Mixture-Based Predictive Intervals: To preserve the full mixture structure, predictive intervals are estimated via Monte Carlo sampling from the mixture distribution. This approach naturally captures non-Gaussian effects and provides more reliable coverage in regions of high epistemic uncertainty.

I. Summary

The proposed ensemble Gaussian Process framework integrates kernel diversity, Laplace-based hyperparameter posterior sampling, and probabilistic model averaging into a unified predictive model. By approximating Bayesian marginalization over hyperparameters and treating the ensemble as a mixture of predictive distributions, the method provides robust uncertainty estimates that significantly extend standard single-model Gaussian Process regression.

III. Case Studies

A. Synthetic Regression Dataset

A one-dimensional synthetic regression dataset with a smooth oscillatory ground-truth function and heteroscedastic Gaussian noise is used. Inputs are sampled from a bounded interval $x \in [x_{\min}, x_{\max}]$ with $x_{\min} = -3$ and $x_{\max} = 3$. The latent signal is

$$f(x) = \sin(1.5x) + 0.35\cos(4x) + 0.15x, \tag{16}$$

and observations are generated as

$$y = f(x) + \epsilon, \qquad \epsilon \sim \mathcal{N}\big(0, \sigma^2(x)\big), \tag{17}$$

with heteroscedastic noise standard deviation

$$\sigma(x) = 0.08 + 0.12 \cdot (1 + e^{-(x-1.5)})^{-1} \tag{18}$$

This noise model yields near-constant low noise for $x \ll 1.5$ and increased noise for $x \gtrsim 1.5$, enabling evaluation of aleatoric uncertainty behavior.

To avoid purely uniform coverage while keeping the dataset size fixed, samples are drawn from the full interval with a lightly under-sampled region around $x \approx -2$, containing only a small fraction of points.

For each experiment, a dataset of $N = 100$ noisy observations is generated and randomly split into training (75%) and test (25%) sets. A single GPR model trained on the training data serves as a baseline, while an ensemble of Gaussian process regressors is trained using the same data split.

We used an ensemble of 60 GP submodels, generated as 15 hyperparameter perturbations per kernel family. This provides sufficient model diversity while keeping the computational cost moderate. A systematic analysis of how calibration quality depends on ensemble size M was not performed and remains for future work.

Predictive intervals at nominal level α are constructed for each test input, for the ensemble using the predictive mixture (12). Empirical coverage is computed as the fraction of test targets falling within the corresponding intervals.

Calibration plots visualize empirical coverage as a function of the nominal interval level and compare it to the ideal calibration line $y = x$.

Results: Figure 1a (Single GPR) shows relatively narrow predictive intervals, even in regions where the model must interpolate or extrapolate. This indicates that the single GP tends to underestimate epistemic uncertainty and remains overconfident due to its fixed kernel and hyperparameter assumptions.

In contrast, Figure 1b demonstrates that the ensemble produces substantially wider uncertainty bands. This behavior is expected: combining multiple GP models increases model diversity and explicitly accounts for uncertainty in the latent function, leading to more conservative but more realistic predictive intervals.

The ensemble achieved a test RMSE of 0.1280, compared to 0.1337 for the single model, indicating a slight improvement in predictive performance through the ensemble approach.

Calibration results in Figure 1c confirm this qualitative observation. The single GP consistently underestimates coverage across all nominal levels, reflecting overly narrow intervals. The ensemble slightly overestimates coverage but is close to ideal calibration at practically relevant levels (around 80 % and 95 %), indicating improved calibration. As an aggregated calibration measure, we used the Mean Absolute Calibration Error (MACE). MACE is defined as the average absolute deviation between the nominal coverage level c_i and the empirically observed coverage $\hat{c}_i$ across K evaluated coverage levels:

$$\mathrm{MACE} = \frac{1}{K} \sum_{i=1}^{K} \left| \hat{c}_i - c_i \right|. \tag{19}$$

For the evaluated coverage levels $c_i \in \{0.5, 0.7, 0.8, 0.9, 0.95\}$, the resulting MACE was 0.102 for the single GPR and 0.042 for the ensemble.

As illustrated in Figure 1e, aleatoric uncertainty rises toward the domain edges; this is attributed to increased predictive variance within the individual ensemble members in data-sparse regions. Importantly, the ensemble exhibits elevated epistemic uncertainty around $x \approx -1$ and $x \approx 1.5$, highlighting regions of model disagreement and illustrating the intended benefit of the ensemble approach. Figure 1d shows the single-GP decomposition for comparison.

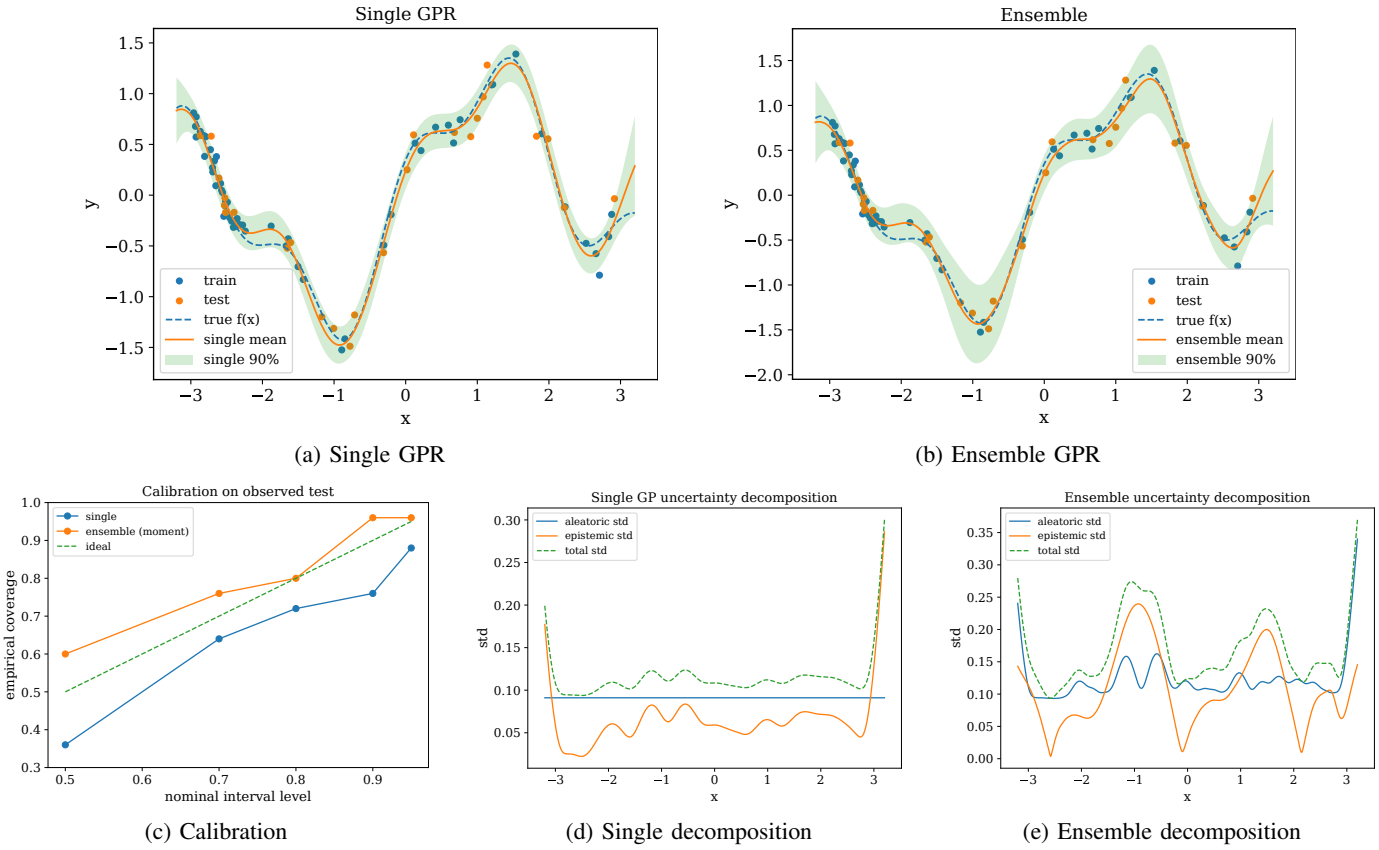
B. Micromagnetic Sensor

The case study is based on micromagnetic material characterization, a nondestructive measurement approach for in-process estimation of mechanical properties in ferromagnetic materials. Micromagnetic measurements exploit the interaction between magnetic domain behavior and microstructural or stress-related material states, enabling the indirect assessment of quantities such as residual stress and hardness. A comprehensive description of the measurement principles, experimental setup, and data acquisition is provided in [10].

The dataset was acquired using a 3MA-II micromagnetic measurement system (Fraunhofer IZFP) and consists of measurements performed on quenched and tempered steel 51CrV4. Different material states were generated through controlled heat treatment, resulting in three distinct initial hardness levels of approximately 400 HV, 500 HV, and 600 HV. Subsequent manufacturing steps were used solely to induce a range of surface and near-surface material conditions relevant for micromagnetic characterization. The data were subject to normalization prior to model training.

In total, the dataset consists of 81 samples. From the available micromagnetic descriptors, the ten most informative features were selected using a regression tree-based feature selection method following [11]. Figure 2 illustrates the micromagnetic sensor setup used for all measurements. These selected features form the input space for all regression models considered in this study. The maximum residual compressive stress and surface hardness were investigated as target parameters for the surface layer condition. The dataset is publicly available and documented in [12]. As before, 60 submodels were generated using 15 hyperparameter perturbations for each of 4 kernel types.

Results: Figure 3 and Figure 4 show the calibration curves obtained on real experimental data for the prediction of the *maximum residual stress* and the *surface hardness*, respectively. The resulting MACE values for maximum residual stress were 0.062 for the single GPR model and 0.022 for the ensemble model. For hardness, the MACE values were 0.070 for the single GPR model and 0.026 for the ensemble model. In both cases, the Gaussian process ensemble exhibits empirical coverage that is consistently closer to the nominal confidence levels than the single Gaussian process regressor. This indicates a better calibrated predictive uncertainty and



(a) Single GPR (b) Ensemble GPR

(c) Calibration (d) Single decomposition (e) Ensemble decomposition

Fig. 1: Predictive mean and uncertainty (top row), calibration and uncertainty decomposition (bottom row).

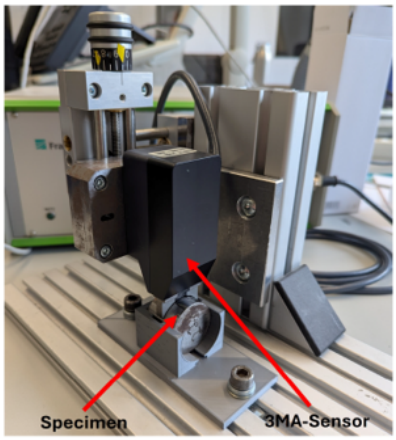


Fig. 2: 3MA-II micromagnetic sensor and the measured specimen for nondestructive material characterization.

demonstrates that the proposed ensemble approach remains effective when applied to real-world data.

Beyond improved calibration, the ensemble also achieves a small but consistent gain in predictive accuracy. Table I summarizes the results of a 10-fold cross-validation in terms of the coefficient of determination (R^2). For both target quantities, the ensemble slightly outperforms the single GPR model, confirming that the improved uncertainty quantification does not come at the cost of reduced accuracy. Note that improved predictive accuracy is a secondary benefit; the main contribution lies in uncertainty calibration. On this real-data task (N=81, 10 features), combined training and inference took 0.28 s for the single GPR and 2.33 s for the full ensemble (M=60), i.e. a 8.3x overhead, confirming that the ensemble remains computationally tractable in the small-data regime (All experiments were run on a 14-core Intel Core Ultra 5 125H, 32 GB RAM using Python 3.13.3 and scikit-learn).

TABLE I: 10-fold cross-validation R^2 scores for single GPR and ensemble models.

Target	Single GPR	Ensemble
Surface hardness	0.8717	0.8779
Maximum residual stress	0.8904	0.9121

Overall, these results show that the ensemble approach yields better-calibrated uncertainty estimates and slightly improved predictive performance compared to a single Gaussian process, demonstrating its suitability for real-data applications.

IV. Conclusions and Outlook

This paper presented an ensemble-based uncertainty quantification framework for Gaussian Process regression tailored to small-data calibration problems. By modeling hyperparameter-induced variability through an ensemble of

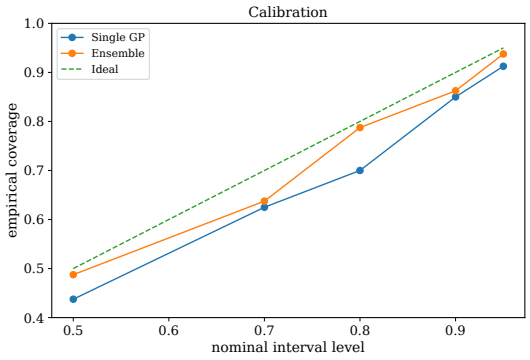


Fig. 3: Calibration curve for the prediction of the maximum residual stress (real data, 10-fold CV).

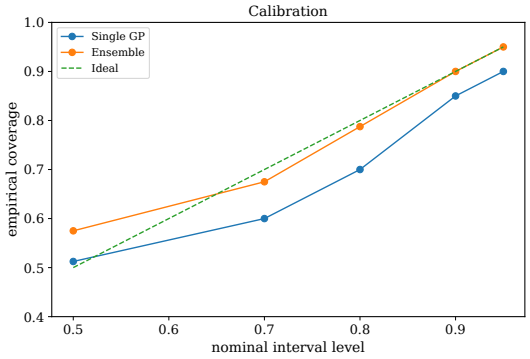


Fig. 4: Calibration curve for the prediction of surface hardness (real data, 10-fold CV).

Gaussian Processes, the proposed approach enables a structured decomposition of predictive uncertainty while preserving the data efficiency and interpretability of standard GPR. The formulation as a finite mixture of Gaussian Processes offers a computationally tractable alternative to fully Bayesian inference and yields robust predictions with meaningful uncertainty estimates. The application to a micromagnetic sensor calibration task demonstrated the practical relevance of the approach and its ability to support reliable decision-making in data-scarce settings.

Some aspects of the proposed framework may pose limitations in certain settings. The Laplace approximation is a local quadratic expansion and could be inaccurate for multimodal or strongly skewed hyperparameter posteriors, where sampling-based alternatives such as HMC would be more faithful at higher cost. In addition, the cubic per-model cost of GPR scales with the ensemble size M, which is tractable in the small-data regime but may preclude direct application to large datasets without sparse-GP approximations.

Future work will focus on systematic study of the ensemble construction, including the use of different kernel structures, as well as strategies for selecting kernel types, ensemble size, and perturbation strength. Although fixed values of s, T, and M were used throughout the experiments, a systematic sensitivity analysis of these design parameters remains an important direction for future work. In addition, the ensemble-induced variability could be exploited for feature importance and feature selection by analyzing the sensitivity of predictive uncertainty to input dimensions or by comparing marginal likelihood contributions across ensemble members. Beyond small-data settings, evaluating the approach on larger datasets, additional benchmarks, and further application domains will help to assess scalability and generalization. Furthermore, where sufficient data is available, a comparison against deep ensembles would help to position the proposed approach relative to neural-network-based uncertainty quantification methods.

A systematic comparison against fully Bayesian GP baselines — using MCMC or variational inference for hyperparameter marginalization — as well as against a bootstrapped GP ensemble was beyond the scope of this work due to the associated computational cost. Such benchmarks, evaluating both calibration quality and computational efficiency, are an important direction for future work.

References

[1] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[2] V. Lalchand and C. E. Rasmussen, "Approximate inference for fully bayesian gaussian process regression," in *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*, ser. Proceedings of Machine Learning Research, C. Zhang, F. Ruiz, T. Bui, A. B. Dieng, and D. Liang, Eds., vol. 118. PMLR, dec 2020, pp. 1–12.

[3] R. M. Neal, "Monte carlo implementation of gaussian process models for bayesian regression and classification," *arXiv: Data Analysis, Statistics and Probability*, 1997.

[4] J. Hensman, N. Fusi, and N. D. Lawrence, "Gaussian processes for big data," in *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'13. Arlington, Virginia, USA: AUAI Press, 2013, p. 282–290.

[5] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information fusion*, vol. 76, pp. 243–297, 2021.

[6] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, Red Hook, NY, USA, 2017, p. 6405–6416.

[7] T. G. Dietterich, "Ensemble methods in machine learning," Berlin, Heidelberg, p. 1–15, 2000.

[8] M.-P. V. Christiansen, N. Rønne, and B. Hammer, "Efficient ensemble uncertainty estimation in gaussian processes regression," *Machine Learning: Science and Technology*, vol. 5, no. 4, p. 045029, oct 2024.

[9] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," *Science Advances*, vol. 7, no. 10, 2021.

[10] T. Wegener, A. Liehr, A. Bolender, S. Degener, F. Wittich, A. Kroll, and T. Niendorf, "Calibration and validation of micromagnetic data for non-destructive analysis of near-surface properties after hard turning," *HTM Journal of Heat Treatment and Materials*, vol. 77, no. 2, pp. 156–172, 2022.

[11] F. Wittich and A. Kroll, "Calibration modeling and feature selection for micromagnetic material characterization using machine learning [Kalibriermodellerstellung und Merkmalsselektion für die mikromagnetische Materialcharakterisierung mittels maschineller Lernverfahren]," *at - Automatisierungstechnik*, vol. 73, no. 10, pp. 778–790, 2025.

[12] F. Wittich, T. Wegener, A. Liehr, W. Zinn, T. Niendorf, A. Kroll, A. Bolender, and S. Degener, "Micromagnetic measurement of the surface layer state in hard turned 51crv4 steel," 2025. [Online]. Available: https://daks.uni-kassel.de/handle/123456789/614