



Original article

Goal-driven adaptive sampling strategies for machine learning models predicting fields

Jigar Parekh^{a,b}, Philipp Bekemeyer^{a,b,*}^a Cluster of Excellence SE²A - Sustainable and Energy-Efficient Aviation, TU Braunschweig, Braunschweig, Germany^b Institute of Aerodynamics and Flow Technology, German Aerospace Center (DLR), Braunschweig/Göttingen, Germany

ARTICLE INFO

Editor: Dr Mehdi Ghoreyshi

Keywords:

Adaptive sampling
Active learning
Field-accurate surrogate modeling
Graph neural network
Surrogate based uncertainty quantification
Transition modeling

ABSTRACT

Machine learning models are widely regarded as a way forward to tackle multi-query challenges that arise once expensive black-box simulations such as computational fluid dynamics are investigated. However, ensuring the desired level of accuracy for a certain task at minimal computational cost, e.g. as few black-box samples as possible, remains a challenge. Active learning strategies are used for scalar quantities to overcome this challenge and different so-called infill criteria exist and are commonly employed in several scenarios. Even though needed in various fields, an extension of active learning strategies towards field predictions is still lacking or limited to very specific scenarios and/or model types. In this paper we propose an active learning strategy for machine learning models that are capable of predicting fields, which is agnostic to the model architecture itself. For doing so, we combine a well-established Gaussian process model for a scalar reference value and simultaneously aim at reducing the epistemic model error and the difference between scalar and field predictions. Different specific forms of the above-mentioned approach are introduced and compared to each other as well as only scalar-valued based infill. Results are presented for the NASA common research model for an uncertainty propagation task showcasing high level of accuracy at significantly smaller cost compared to an approach without active learning.

1. Introduction

High-fidelity aerodynamic simulations (e.g. CFD solutions of the flow around an aircraft) provide detailed flow field information and accurate performance metrics, but they remain computationally expensive especially if multiple parameter combination are to be investigated. Surrogate modeling has become a cornerstone in aerodynamic design to at least partially alleviate this cost. Data-driven models such as Gaussian process regressions, reduced-order models (ROMs), or machine learning and deep learning models, such as graph neural networks (GNNs) are trained to approximate the simulation outputs, trading a high upfront simulation cost for a fast predictive model. A critical challenge is to ensure that the surrogate is sufficiently accurate with a limited number of training samples (simulation runs). The design and analysis of computer experiments is well established in the literature [1–3] providing a solid basis with samples spread of the design space. However, domain specifics like the difference between subsonic and transonic flow are unknown to such design of experiment techniques and, hence, can not be accounted for. Adaptive sampling, or active learning, addresses this by iteratively selecting new sample points in the design or parameter space that are expected to maximally improve the surrogate. For scalar objec-

tives - for example, predicting a single quantity like a drag coefficient or lift-to-drag ratio - there are well-established infill (acquisition) criteria that guide this selection. A classic approach is the Efficient Global Optimization (EGO) algorithm using the expected improvement (EI) criterion [4,5], which balances exploration and exploitation by favoring points likely to yield a better optimum of the objective. Alternatively, when the goal is accurate surrogate modeling for uncertainty quantification (UQ) rather than purely optimization, one can target reductions in the surrogate's prediction error across the input distribution. In Gaussian process (GP) surrogates, for instance, methods like Active Learning Cohn (ALC) choose new samples that would maximally reduce the integrated predictive variance of the model [6]. Such variance-based criteria (related to integrated mean squared error) seek to minimize overall surrogate uncertainty with respect to the input probability density function, ensuring the model is accurate on average for UQ analysis [7–9]. In summary, for single-output surrogates, techniques like EI and variance reduction provide principled ways to improve surrogate fidelity with minimal samples [6,7]. These methods have been successfully applied in aerodynamic optimization and analysis, where adaptively sampling the design space yields faster convergence to optimal designs and more reliable surrogate predictions than space-filling designs alone.

* Corresponding author.

E-mail addresses: jigarparekh279@gmail.com (J. Parekh), philipp.bekemeyer@dlr.de (P. Bekemeyer).<https://doi.org/10.1016/j.ast.2026.112407>

Received 29 January 2026; Received in revised form 17 April 2026; Accepted 17 April 2026

Available online 20 April 2026

1270-9638/© 2026 The Author(s). Published by Elsevier Masson SAS. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

However, many problems in aerodynamic design and analysis demand field predictions rather than just scalar outputs. For example, predicting the pressure or velocity field around an airfoil or across a wing's surface is crucial for understanding shock locations, flow separation, and structural loads — information that a single scalar, e.g. the drag coefficient, cannot capture. Building surrogates for such distributed outputs which can be thought of as high-dimensional vectors or functions defined on a spatial domain is far more complex. First, the dimensionality of the output — potentially thousands of values describing a field — often necessitates some form of dimensionality reduction. Common approaches include reduced-order modeling techniques such as proper orthogonal decomposition (POD) to represent the field in terms of a few dominant modes or the mapping of a latent variable to the field [10,11]. Alternatively, one may use deep learning models such as autoencoders that learn a compressed latent representation of the field. Second, even if a surrogate can predict a field, adaptively sampling to improve a field prediction surrogate is not straightforward. In contrast to a scalar objective, a field has many features and possible sources of error, e.g. a surrogate might accurately predict overall trends but miss a local shock entirely. Hence, there is a need for field-based infill criteria that guide the selection of new simulation points by considering the surrogate's performance on the entire field output. Some studies have begun to explore this; for instance, adaptive sampling strategies for POD-based surrogates have been proposed to iteratively fill the parametric space and improve the surrogate accuracy [12]. Furthermore, vector-valued Gaussian processes, which extend standard GP models to directly handle multi-output fields and quantify uncertainty over them, have been developed to drive max-variance based infill strategies [13,14]. However, proposed techniques are either limited in scope or tailored towards a certain model type.

Extending adaptive sampling to field prediction tasks raises several key challenges.

- i) **Scalarizing field outputs:** Unlike a single objective, a field output must be summarized or reduced to a one-dimensional criterion for decision-making. Determining an appropriate scalarization (for example, an integrated error norm over the field, a worst-case error, or a weighted sum of errors at critical locations) is non-trivial and can bias the sampling toward certain regions of the input space. Multi-objective optimization research offers many scalarizing functions to combine multiple objectives into one [15], but those methods typically handle a handful of objectives, not the thousands of degrees of freedom in a field. Applying a naive scalarization to a field (e.g. simply using the average error) might overlook localized phenomena, whereas focusing only on the maximum error might overweight outliers. Thus, defining a scalar criterion that truly reflects surrogate field accuracy is an open question.
- ii) **Managing epistemic uncertainty:** Surrogate models for fields must capture uncertainty in high-dimensional outputs. For example, Gaussian process surrogates require multi-output covariance structures to estimate uncertainties and their correlations [13]. In practice, reduced bases with probabilistic models or neural network ensembles are used. However, epistemic uncertainty arises not only from sparse data but also from uncertain model parameters such as correlation lengths or basis truncation. Graham and Cortés [16] demonstrate that accounting for covariance uncertainty can significantly impact predictive variance, driving adaptive sampling via both model and conditional uncertainties. Ignoring these factors may lead to misinformed infill decisions; hence, rigorous statistical modeling is essential to correctly identify where the surrogate is truly uncertain [17,18].
- iii) **Consistency between scalars and fields:** Design often requires both global coefficients, e.g. lift and drag, and detailed field information, e.g. pressure or shear distribution. Training separate surrogates for scalars and fields raises the challenge of allocating new samples optimally. For example, improving drag prediction might

not reduce pressure distribution errors, and vice versa. This leads to a multi-output active learning problem where one must balance competing objectives. Focusing solely on one aspect may waste samples or leave the other under-trained, and existing multiobjective infill methods still require reducing field error to a single metric, which does not scale well with multiple outputs [6]. Achieving an optimal training set that accurately captures both scalar and field outputs is challenging and underexplored in current literature. Traditional multiobjective infill methods can treat, for example, scalar and field error as competing objectives; however, they require reducing field error to a single number, which does not scale well when multiple scalar metrics or field features are of interest.

Given these challenges, current approaches to surrogate modeling for aerodynamic fields show significant gaps. Bayesian multiobjective optimization techniques can scalarize a few objectives [15], but they do not extend well to high-dimensional field outputs. Meanwhile, spatial sensor placement studies [16] focus on field estimation under specific covariance assumptions and do not incorporate concurrent optimization objectives or engineering constraints. A separate line of research focuses on improving the surrogate model itself by embedding physical constraints into the training process, for example through physics-informed loss functions or architectures that enforce boundary conditions and monotonicity [19,20]. While these physics-embedding approaches enhance surrogate fidelity, they address a complementary problem: *how* to build a better model from a given dataset, rather than *where* to sample next to most efficiently grow that dataset. Our work targets the latter question and is agnostic to whether the underlying surrogate is purely data-driven or physics-informed. There is a clear need for a unified, adaptive sampling strategy that is both field-aware and objective-aware, while rigorously accounting for surrogate uncertainty [21] and at the same time being agnostic to the specific machine learning model employed. To the best of our knowledge, no current methodology fully addresses the challenge of efficiently learning both scalar performance metrics and high-dimensional field outputs.

In this work, we propose field-based adaptive sampling strategies for surrogate modeling that extend to UQ, optimization, reliability analysis and general data-base generation. The specific objectives of this work are:

1. To develop model-agnostic infill criteria that extend classical scalar-based acquisition functions to field outputs by combining a well-established Gaussian process error estimate with a misfit term that enforces consistency between scalar and field predictions.
2. To introduce a unified active learning framework that jointly trains scalar and field surrogates, ensuring that new samples reduce overall epistemic uncertainty in both scalar and field outputs while being applicable to any field prediction model independent of its architecture.
3. To demonstrate the effectiveness of the proposed strategies on the NASA Common Research Model, showing that the coupled infill criteria achieve higher accuracy at significantly reduced computational cost compared to non-adaptive sampling and scalar-only infill, and to apply the best-performing strategy for forward uncertainty propagation.

These infill criteria effectively balance exploration, i.e. improving the surrogate where it is most uncertain in the flow field, and exploitation, i.e. refining predictions of important global quantities. The proposed approach enables efficient and robust training of surrogates that can predict full flow fields with quantified uncertainty, reducing the number of expensive simulations needed. This leads to surrogates that remain accurate across the input distribution and reliable for downstream tasks such as global sensitivity analysis, probabilistic design optimization, and risk assessment in aerodynamic applications.

The paper is structured as follows: [Section 2](#) introduces the goal-driven adaptive sampling strategies for surrogate models predicting

fields. Section 3 details the setup for NASA CRM simulations and surrogate modeling; including problem formulation, flight conditions, transition modeling, environmental and operational parameters, and the numerical model for computing aerodynamic quantities of interest (QoIs). Section 4 presents the results from different infill approaches and the application of best-of-all model for the UQ analysis. Finally, conclusions are offered in Section 6.

2. Methodology

This section presents the proposed adaptive sampling framework in a general form, independent of any specific application, surrogate architecture, or engineering discipline. In this study, surrogate models are developed for both scalar outputs and spatially distributed field outputs. The adaptive sampling proposed accommodates both the surrogates directly estimate quantity of interest (QoI). The following sections detail the surrogate modeling methods and adaptive infill criteria employed to ensure high predictive accuracy in both scalar and field contexts. Even though results are later on presented for an aerodynamic use-case the methodology is applicable to any discipline or setting as long as the field prediction can be integrated into an engineering-feasible scalar quantity that than be modelled using a Gaussian Process.

2.1. Surrogate modeling without infill

We first construct baseline surrogate models using an initial design of experiments (DoE) before any adaptive refinement. Three surrogate modeling techniques are employed, reflecting the diversity of scalar vs. field outputs.

2.1.1. Gaussian process regression for scalars

Gaussian Process (GP) models are employed to surrogate scalar quantities of interest (QoIs), due to their flexibility and intrinsic uncertainty estimation capability [22]. The true response $y(\mathbf{x})$ is assumed to be a realization of a random function $Y(\mathbf{x})$, composed of a deterministic trend and a zero-mean GP noise term. Formally, this can be expressed as:

$$Y(\mathbf{x}) = g(\mathbf{x})\beta + \epsilon(\mathbf{x}), \quad (1)$$

where $g(\mathbf{x})\beta$ represents a known regression model (trend) and $\epsilon(\mathbf{x})$ is a GP with covariance given by

$$\text{Cov}[\epsilon(\mathbf{x}), \epsilon(\mathbf{x}')] = \sigma^2 R_\theta(\mathbf{x}, \mathbf{x}'), \quad (2)$$

using stationary covariance kernels, e.g., Gaussian-exponential family, with hyperparameters θ estimated via maximum likelihood [23].

The GP predictor provides interpolation of training data and posterior variance estimates, yielding the best linear unbiased estimator:

$$\hat{y}(\mathbf{x}) = g(\mathbf{x})\beta + \mathbf{r}(\mathbf{x})^T R^{-1}(Y - F\beta), \quad (3)$$

where R is the correlation matrix of training points, $\mathbf{r}(\mathbf{x})$ is the correlation vector between $\mathbf{x}$ and training points, Y are observed outputs, and F is the design matrix. The GP further estimates the mean-squared error (MSE) at any input $\mathbf{x}$ [24]:

$$s^2(\mathbf{x}) = \sigma^2 (1 - \mathbf{r}(\mathbf{x})^T R^{-1} \mathbf{r}(\mathbf{x}) + \mathbf{u}^T (F^T R^{-1} F)^{-1} \mathbf{u}), \quad (4)$$

with $\mathbf{u} = F^T R^{-1} \mathbf{r}(\mathbf{x}) - g(\mathbf{x})$. This $s(\mathbf{x})$ acts as a local surrogate error indicator, making GPs particularly suited for uncertainty quantification.

2.1.2. Proper orthogonal decomposition with interpolation for fields

For high-dimensional field outputs, Proper Orthogonal Decomposition with Interpolation (PODI) can be utilized [25]. POD generates an efficient representation of output fields by expressing the field as a linear combination of orthonormal modes. Given an initial set of N_0 field snapshots $\{u(\mathbf{x}_k)\}_{k=1}^{N_0}$, each of dimension $p \gg 1$, singular value decomposition (SVD) yields a reduced basis $\{\phi_i\}_{i=1}^M$ capturing the most significant variations [26]. Any field can be approximated as:

$$u(\mathbf{x}) \approx u_0 + \sum_{i=1}^M \alpha_i(\mathbf{x}) \phi_i, \quad (5)$$

where u_0 is the mean field and $\alpha_i(\mathbf{x})$ are the mode amplitudes dependent on inputs $\mathbf{x}$. The problem reduces to modeling the mapping from inputs to mode amplitudes, typically via multi-dimensional regression techniques like GPs or Radial Basis Functions (RBFs) [25]. This PODI approach significantly reduces output dimensionality, providing efficient and accurate field predictions, as demonstrated by recent studies.

2.1.3. Graph neural networks for fields

Graph Neural Networks (GNNs) offer a powerful data-driven alternative for modeling complex spatial fields, particularly suited for outputs defined over unstructured meshes common in aerodynamic and fluid dynamics problems [27,28]. GNNs interpret mesh points as graph nodes and mesh connections as graph edges, utilizing iterative message-passing algorithms to learn intricate spatial interactions directly from data [29,30].

The message-passing mechanism in GNNs generally follows an iterative update process. For each node i , the latent feature representation $\mathbf{h}_i^{(l)}$ at the l th iteration is updated according to:

$$\mathbf{m}_i^{(l+1)} = \sum_{j \in \mathcal{N}(i)} \phi(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}, \mathbf{e}_{ij}), \quad (6)$$

$$\mathbf{h}_i^{(l+1)} = \psi(\mathbf{h}_i^{(l)}, \mathbf{m}_i^{(l+1)}), \quad (7)$$

where $\mathcal{N}(i)$ denotes the set of neighboring nodes to node i , $\mathbf{e}_{ij}$ represents edge features, and ϕ and ψ are neural network functions typically realized as multi-layer perceptrons (MLPs). For further details on the employed GNN methodology the interested reader is referred to [28].

The main advantages of GNN-based approaches include inherent flexibility in mesh topology representation, the ability to directly incorporate geometric and topological information into learning, and enhanced accuracy in capturing spatial correlations compared to traditional surrogate modeling techniques. These capabilities make GNNs a compelling choice for surrogate modeling tasks in aerodynamics and fluid mechanics.

2.2. Adaptive sampling (infill) approaches

The adaptive sampling (infill) step is solely governed by the acquisition function which decides how the next sample is determined. For applications with scalar output the next sample point ξ^* is selected by maximizing the criteria value $I(\xi)$:

$$\xi^* = \underset{\xi}{\operatorname{argmin}} \{-I(\xi)\} \quad (8)$$

A typical and simple choice for the criteria is the surrogate error, i.e. $I(\xi) = \hat{s}(\xi)$. Such error-based adaptive sampling has been shown to greatly improve efficiency over unguided MC sampling [31,32]. The optimization of the acquisition function is performed with a global optimizer (here, differential evolution) which is effective in navigating complex, multimodal objective landscapes.

It should be noted that, for applications with non-uniformly distributed inputs the next infill point is typically selected by maximizing the product of the joint input probability density function of the inputs PDF_ξ and the respective criteria value, i.e. $\xi^* = \underset{\xi}{\operatorname{argmin}} \{-\text{PDF}(\xi)I(\xi)\}$. This criterion prioritizes sampling in high-probability regions while simultaneously targeting regions with high criteria value.

2.2.1. Adaptive sampling for scalars

The surrogate-error criterion works well for single-output (scalar) GP surrogates as these have an inherent estimation of model uncertainty i.e. $\hat{s}(\xi) = \sigma_{GP}(\xi)$. We will refer to this approach as *GP Infill:SE*. Note that this approach still needs a surrogate model for field predictions. We employ a PODI or GNN using together the DoE and the infill samples of the GP Infill:SE approach to construct a field surrogate enabling inference at new inputs.

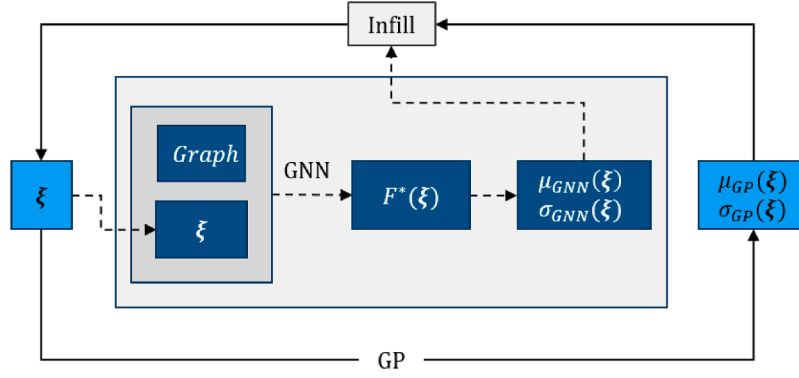


Fig. 1. Coupled goal-driven scalar-field adaptive sampling framework. A Gaussian process (GP) surrogate maps the inputs ξ to scalar coefficients resulting in a mean and variance approximation denoted by μ_{GP} and σ_{GP} , while a field model receives the same inputs together with other additional information such as the surface-mesh graph and predicts field quantities F^* from which the scalars are integrated out with thereby estimated scalar mean and variance denoted by μ_{GNN} and σ_{GNN} . The misfit and epistemic uncertainties of both surrogates feed a coupled infill criterion that selects the next design point. The new snapshot closes the loop and updates both surrogates.

While using a scalar GP's error is convenient, it may not fully capture where the field surrogate itself has high errors. For instance, the pressure distribution might be poorly predicted in some region of input parameter space that does not greatly affect the integrated scalars C_D and/or C_L . Infill purely based on σ_{GP} may not explore such regions.

2.2.2. Adaptive sampling for fields

For deep learning field surrogates, we must extend the concept of SE-based infill. A straightforward approach is to derive a scalar metric from the field and then apply a similar criterion. However, unlike GPs, GNNs lack inherent uncertainty estimates. To circumvent this issue, Monte Carlo dropout or ensemble methods can be used to approximate predictive uncertainty [33,34]. In this study we employ the former, the details of which are deferred to [Appendix A](#). Using the approximated model variance as the SE indicator i.e. $\hat{s}(\xi) = \sigma_{GNN}$, we can apply the SE-based infill just like we did with GPs effectively reducing the field infill problem to the scalar case. We refer to this approach as *GNN Infill:SE*.

Just like GP, the GNN surrogate is retrained after every infill iteration. Therefore, the GNN surrogate keeps changing in terms of its prediction and its variance at each training sample after every infill iteration. One potential problem with this is that it may not guarantee a significant reduction in the model's epistemic uncertainty even after a large number of infill steps.

2.2.3. Coupled adaptive sampling

To simultaneously improve the accuracy of scalar and field surrogates, we introduce two acquisition strategies that couple the scalar surrogate and field surrogate predictions. Crucially, these criteria are model-agnostic: they interact with the field surrogate only through its integrated scalar QoI and, optionally, its predictive variance, without accessing internal model parameters, gradients, or any other part of the architecture. Any field model that can produce an integrated scalar and, optionally, an uncertainty estimate can therefore be used.

The first criterion augments the GP's uncertainty with the misfit between the GP and field surrogate predictions. At a given $\mathbf{x}$, let $\mu_{GP}(\mathbf{x})$ be the GP-predicted QoI (and $\mu_{GNN}(\mathbf{x})$ be the corresponding QoI computed from the GNN-predicted field $F^*(\xi)$). We define the misfit $\Delta(\xi) = |\mu_{GP}(\xi) - \mu_{GNN}(\xi)|$ as a measure of disagreement between the two surrogates. The acquisition function is then given by:

$$\xi^* = \underset{\xi}{\operatorname{argmin}} \{ -[\sigma_{GP}(\xi) + \lambda \Delta(\xi)] \} \quad (9)$$

Here λ is a weighting factor (we take $\lambda = 1$ for simplicity). We refer to this approach as surrogate error with misfit; *GP-GNN Infill:SEwMisfit*. This method essentially prioritizes points that either have a large GP

predictive uncertainty or a large discrepancy between the GP and field surrogate. In applications with non-uniform inputs, as discussed earlier, the criteria value is weighted by the input PDF to remain focused on probable regions. This can accelerate learning for the field surrogate - by adding points where the field prediction disagrees with the well-trained GP, we specifically target improvements in the field model. SEwMisfit is in general related to model discrepancy sampling [35] or multi-fidelity infill [36], where one model guides sampling for another.

The second coupled strategy is more statistically rigorous. We consider the predictive probability distributions of the GP and the field surrogate for a chosen QoI at an input sample. The GP yields a normal distribution $P_\xi = \mathcal{N}(\mu_{GP}(\xi), \sigma_{GP}^2(\xi))$. For the field surrogate, as mentioned earlier, we can obtain an approximate distribution Q_ξ e.g. via Monte Carlo dropout ensembles, yielding mean $\mu_{GNN}(\xi)$ and variance $\sigma_{GNN}^2(\xi)$. We then use the Jensen-Shannon divergence [37] between these two distributions as the acquisition metric. The Jensen-Shannon divergence is a symmetrized measure of difference between P and Q , defined by:

$$JSD(P||Q) = \frac{1}{2} KL(P||M) + \frac{1}{2} KL(Q||M) \quad (10)$$

where $M = \frac{1}{2}(P + Q)$ and KL denotes Kullback-Leibler divergence [38]. We evaluate $JSD(P||Q)$ between the two model predictive distributions at each ξ . The acquisition function is then given by:

$$\xi^* = \underset{\xi}{\operatorname{argmin}} \{ -\operatorname{PDF}_\xi(\xi) JSD(P_\xi||Q_\xi) \} \quad (11)$$

This criterion targets points where the entire predictive distributions disagree maximally. We refer to this approach simply as *GP-GNN Infill:JSD*. Unlike SEwMisfit which only uses the mean discrepancy (and GP variance), JSD accounts for cases where, for example, the GNN is very uncertain (broad Q_ξ) while the GP is confident (narrow P_ξ), or vice versa. By sampling such locations, we expect to reduce the epistemic uncertainty of both models. In effect, JSD infill tries to make the GP and field surrogate agree, by adding data until their predictive distributions align in the input space. This strategy has parallels to recent active learning methods in reliability engineering that use JSD to update Kriging models [39], demonstrating improved efficiency and convergence in failure probability estimation. However, even though mathematically rigorous with respect to matching distributions it is affected from the same drawbacks with respect to Monte-Carlo dropout discussed before.

Both SEwMisfit and JSD reduce to standard SE when the field surrogate aligns perfectly with the GP. When field surrogate lags in accuracy, they prioritize reducing this discrepancy, while still weighting samples by the input PDF. This results in infill samples that simultaneously improve scalar and field predictions.

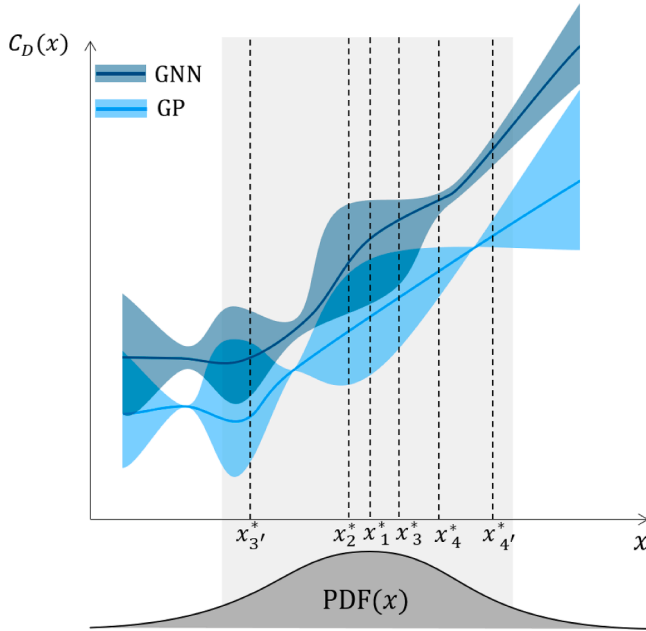


Fig. 2. Conceptual comparison of the four adaptive-sampling criteria used for the coupled GP-GNN surrogate system. The solid lines represent the GP (light-blue) and GNN (dark-blue) predictions of the drag coefficient $C_D(x)$; the shaded bands indicate their epistemic uncertainty. The gray bell curve is the joint input probability density $PDF(x)$. Vertical dashed lines mark candidate infill locations x_i^* proposed by each strategy: x_1^* - Surrogate Error (SE) criterion applied to the GP; x_2^* - SE applied to the GNN; x_3^* - SE with GP-GNN misfit (SEwMisfit); x_4^* - Jensen-Shannon divergence (JSD) coupling the two surrogates. The possibility of almost equally probable next infill sample is represented as $x_{4'}^*$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 2 depicts all adaptive sampling approaches discussed. In the schematic example, the location of the points x_1^* and x_2^* is with respect to the Infill:SE criterion for GP and GNN, respectively - sampling where the product of model uncertainty and input distribution is highest. Points x_3^* would typically be chosen based on SEwMisfit criterion. However, point $x_{3'}^*$ may also be chosen based on this criterion - indicating that a sample may be chosen even in the lower probability region if the sum of the GP uncertainty and the GP-GNN misfit is significant. Similarly, using JSD criterion, sample points x_4^* and $x_{4'}^*$ may be chosen in support of large mismatch between the two surrogates - even in the lower probability regions of the input distribution.

3. Numerical setup

Having introduced the general methodology, this section describes the specific application used as a proof of concept. The aerodynamic configuration considered is the NASA Common Research Model with a Natural Laminar Flow wing (CRM-NLF). This wing was originally designed by Lynde et al. [40] using a Crossflow Attenuated Natural Laminar Flow (CATNLF) method to delay both Tollmien-Schlichting (TS) and crossflow (CF) instabilities. A semi-span wind tunnel model of the CRM-NLF was later tested at NASA's National Transonic Facility, confirming the design's ability to achieve extensive laminar flow at transonic cruise conditions [41]. For the present study, we utilize the same geometry and focus on similar flow conditions.

3.1. Numerical model

The flow around the wing and fuselage NASA CRM configuration is simulated using the TAU CFD solver from the German Aerospace Center (DLR) [42]. The aerodynamic properties of interest are determined by

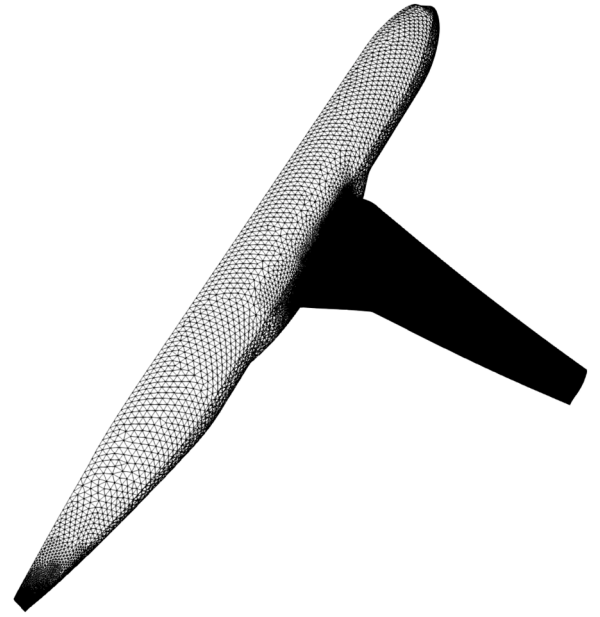


Fig. 3. Computational mesh for the NASA CRM.

solving the Reynolds-averaged Navier-Stokes (RANS) equations along with the $k - \omega$ SST turbulence model [43]. The solver's configuration includes a 4w multigrid cycle, a backward Euler solver for pseudo-time integration and a central flux discretization scheme. The employed unstructured mesh comprises of about 50 million points and about 0.3 million surface points, as shown in **Fig. 3**. Laminarity (transition prediction) is considered for both the upper and the lower surface of the wing plus fuselage configuration.

To predict the transition front over the wing, a recently developed DLR γ - model [44] is used. The class of transition transport models has recently gained significant attention, in particular for application in laminar airfoil design as well as in transition front prediction over a wing [45]. The DLR γ model, an enhancement of the $\gamma - Re_{\theta}$ model, for transport aircraft applications is a member of this class and used as transition prediction method in this work. It includes the "Simple-AHD" criterion [46], i.e. an adapted criterion accounting for pressure gradients and considering the effect of compressibility. Furthermore, it includes a crossflow extension [47]. It has been effective for laminar wings at high Reynolds numbers (around 10^7) and is continuously validated and extended [48].

3.2. Environmental and operational parameters

The goal-based adaptive sampling strategies discussed in **Section 2** is equally applicable for uniformly or non-uniformly distributed inputs. We consider training the surrogate models based on joint input distribution, as done in previous studies [31,32], enabling exploitation of regions with high-probability in the parameter space.

To understand the effect of environmental and operational variations on the transition front and pressure distribution over the wing, relevant factors and their effects must be studied. The critical N-factors (N_{TS} , N_{CF}) used in the e^N transition model [49] captures flow quality, significantly influenced by freestream disturbances (e.g., clouds, turbulence intensity), surface imperfections, and acoustic disturbances [50–52]. Environmental uncertainties in these N-factors are modeled as uniformly distributed variables, e.g., $N_{TS} \sim \mathcal{U}[5, 14]$, $N_{CF} \sim \mathcal{U}[4, 11]$ [53]. For the DLR γ model, uncertainties are represented through variations in freestream turbulence intensity Tu , related to N_{TS} by Mack's formula [54]:

$$N_{TS} = -8.43 - 2.4 \ln(Tu). \quad (12)$$

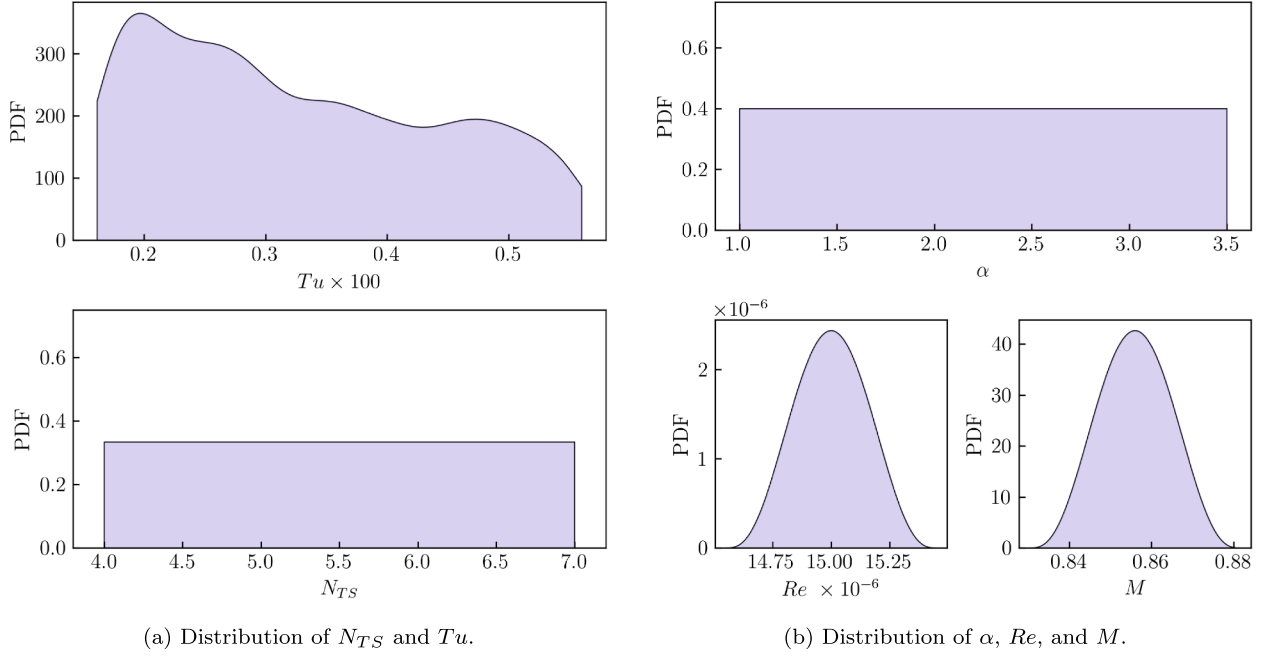


Fig. 4. (Left) N_{TS} and Tu uncertainty distributions. (Right) Uncertainty distributions for angle of attack, Reynolds number, and Mach number.

Sampling N_{TS} from the specified uniform distribution yields the corresponding distribution for Tu . The influence of surface roughness (impacting N_{CF}) is not included here due to implementation limitations. Distributions for the N-factors and turbulent intensity are shown in Fig. 4 (left). Sampling N_{TS} between 4 and 7 captures the range of critical N-factors determined experimentally in the NTF (typical values 5–6) versus calmer flight environments ($N_{TS} \approx 10$) [40]. Mapping to Tu via Mack's formula preserves the physical coupling between modal amplification and disturbance level without committing to a particular analytic PDF.

In many aerodynamic studies the operational conditions such as freestream Mach number, Reynolds number and aircraft angle of attack are prescribed as deterministic inputs. For short-haul transports, however, continuous throttle adjustments, weather-related altitude excursions and mild gust encounters lead to non-negligible variability in these quantities even during nominal cruise, which in turn alters boundary-layer stability and overall performance [55]. Therefore, in the present study the three quantities are therefore modeled as random variables with carefully chosen probability laws. The operational uncertainty space is described by three independent random variables as shown in Fig. 4 (right). A uniform distribution of angle of attack α over $[1.0^\circ, 3.5^\circ]$ reproduces the small trim adjustments that were necessary in the NTF to match target sectional lift and to compensate for support-system offsets [56]. Transition fronts on the CRM-NLF wing shifted by several percent chord within this narrow α band, making the interval critical. To guarantee repeatable flow quality and TSP imaging, the wind tunnel can be assumed to be operated at a Reynolds number which is normally distributed with $\mu = 15 \times 10^6$ and a Coefficient of Variation (CoV) of 1% reflecting those tight operational constraints while still allowing viscous-scale variations that affect crossflow growth. Similarly, Mach number is modeled with $\mu = 0.856$ and a 1% CoV, capturing small variations which may significantly modify shock strength and therefore the pressure-gradients. These distributions are deliberately conservative (uniform or very narrow distributions) yet encompass the reasonable spread around nominal conditions, recorded in the CRM-NLF data set. They therefore provide a physically consistent and statistically unbiased basis e.g. in a surrogate-based uncertainty propagation.

3.3. Surrogate modeling and adaptive sampling setup

All surrogate models utilize the same set of four environmental and operational conditions ξ as the inputs, i.e. freestream turbulence intensity (Tu), angle of attack (α), Reynolds number (Re), and Mach number (M) as inputs. A Gaussian process (GP) surrogate maps the inputs to scalar coefficients (C_L, C_D), while a PODI or GNN receives the inputs and predicts field quantities ($C_p, C_f, C_{f_x}, C_{f_y}, C_{f_z}$) from which the lift and drag coefficients are integrated. In addition the GNN model also has access to the surface-mesh graph.

The initial design of experiments (DoE) is constructed using Sobol quasi-random sequences [31] based on prescribed input distributions, ensuring an even coverage of highly probable samples. We consider two configurations: one with a fixed DoE (no infill) and one with adaptive sampling i.e. DoE with infill. In the no-infill case, the surrogate models are trained using the DoE dataset and validated on a separate holdout set, with a further independent test set for final assessment. In the infill case, the models start from the same DoE dataset but are iteratively refined by adding new samples (infill points) selected via an active learning criterion. After each sequential infill iteration, the training set grows by one sample (with the validation and test sets remaining the same), allowing direct comparison. The infill strategy follows the flowchart illustrated in Fig. 1. Surrogate models are trained on the current dataset, their predictive uncertainties (or error indicators) are evaluated across the input space, and the next sample(s) are chosen where the model confidence is lowest or error highest, then the high-fidelity simulation is run at those points and the results appended to the training dataset. This loop is repeated until a budget of infill samples is exhausted. In cases where both a field surrogate and a scalar surrogate are used together (e.g. a GNN for flow field and a GP for an integrated quantity), a coupled infill strategy is employed such that the selection of new points considers uncertainties of either GP or both models so that the added sample improves the accuracy of both field and global scalar prediction. All DoE and Surrogate Models are constructed and trained using the DLR Surrogate Modeling for Aero Data Toolbox (SMARTy) [57]

For the scalar response the GP uses a squared-exponential kernel with a regularization of $1e-12$ to ensure numerical stability. Kernel

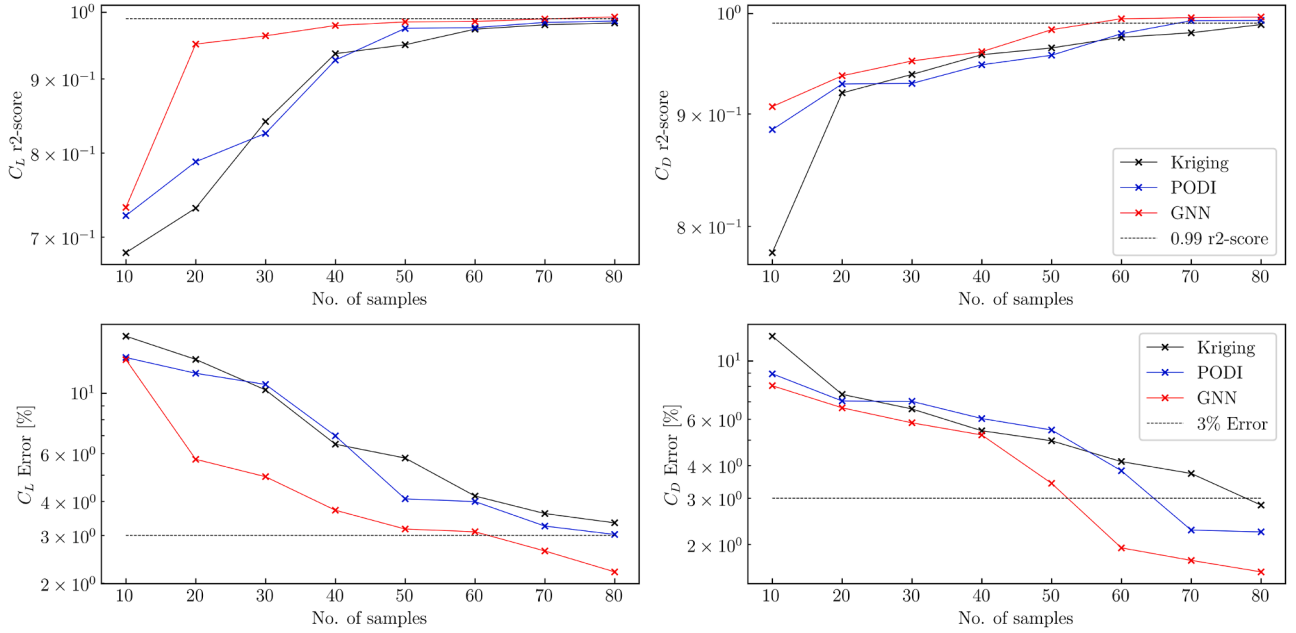


Fig. 5. Convergence of the no-infill surrogates. Top: r2-score for C_L (left) and C_D (right); the dashed line marks the 0.99 target. Bottom: normalized RMSE for C_L (left) and C_D (right); the dashed line marks the 3% target. Kriging, PODI and GNN are trained on DoE dataset with size ranging from 10 to 80 samples.

length-scales and signal variance are determined by maximizing the log-marginal likelihood. The field surrogate GNN uses a ResGatedGraphConv convolution with node feature vector comprising of the normalized spatial co-ordinates appended by the four input variables. Four message-passing layers with 64 nodes each are followed by a linear head that output surface pressure and friction coefficient fields. Inputs are min-max scaled, targets standardized, activations use ReLU with 5% dropout, and training employs Adam (learning rate $lr = 1e-3$ with an exponential decay to $5e-4$). The initial DoE surrogate is trained for 3000 epochs. Each infill retraining typically converges in 100 – 300 epochs, with an early-stopping patience of 20 epochs based on validation RMSE. The acquisition function is optimized at each infill step using the differential evolution algorithm from SciPy with a population size of 15 and a convergence tolerance of 10^{-6} , bounded by the input parameter ranges.

4. Adaptive sampling results

In the upcoming subsections, we will discuss the performance of both no-infill and adaptive sampling based modeling approaches described in the previous sections, followed by the application of the best model for surrogate-based uncertainty propagation.

4.1. Surrogate models without infill

First, no-infill surrogates are built from pure DoE data in order to measure the raw predictive capability of each model. Three models are considered: a GP regressor for scalar quantities, a PODI model for field variables, and a GNN that operates directly on the surface-mesh graph. Separate data sets of 10-80 DoE points are used for training, while 20 samples each are reserved for validation and testing. Fig. 5 shows the prediction quality of the three surrogates as the DoE samples are gradually increased from 10 to 80 samples. Two complementary metrics are used throughout: the coefficient of determination (r2-score), which measures the proportion of output variance captured by the surrogate, and the root-mean-square error normalized by the data range (normalized RMSE), which provides a scale-independent measure of absolute prediction accuracy. Together they ensure that both correlation and magnitude of the predictions are assessed. A performance target of r2-score ≥ 0.99 and a error (normalized RMSE) below 3% was set for scalar outputs

(predicted or integrated from field predictions). The GNN reaches the r2-score threshold for both C_L and C_D with as few as 50 samples, whereas PODI requires roughly 60-70 samples and the GP (Kriging) about 70-80 samples. Similarly, GNN reaches the target error in 50–60 samples, followed by PODI with about 70 samples and GP requiring all 80 samples. While the GNN meets both accuracy goals with about 50 DoE samples, it lacks a rigorous built-in uncertainty estimate. Conversely, the GP provides an analytic variance but needs far more data to reach accuracy thresholds. These findings motivate an adaptive sampling (infill) stage after an initial DoE - (i) to accelerate GP convergence by targeting regions of highest predictive variance, and (ii) to supply GNN with strategically placed snapshots that refine the local flow physics without incurring the cost of a brute-force DoE expansion.

4.2. Surrogate models with adaptive sampling

Next we employ adaptive sampling strategy. A common initial DoE set of 30 points is augmented by 30 infill samples that are acquired sequentially. Thus every adaptive surrogate is trained iteratively and monitored with the same validation and test sets as earlier. Four variants are updated after each infill step: a stand-alone scalar GP, a stand-alone GNN for the field outputs, and two coupled GP-GNN models (misfit and JSD). Here, we only rely on the GNN model for field predictions given its superior performs compared to PODI in the no infill setting. However, the methodologies (at least SEwMisfit) can be applied to PODI models in a straightforward fashion. For uncertainty aware sampling w.r.t. the field model we would suggest to use GP model for latent space regression to also obtain uncertainty estimates for the PODI model. Note that, the total computational cost is dominated by the CFD evaluations (each RANS simulation on the 50-million-point mesh requires several hours on an HPC cluster), while surrogate training and acquisition function optimization are negligible in comparison. In practice, the GP re-optimization takes seconds, the GNN retraining requires roughly 2 to 5 min on a single GPU, and the MC-dropout uncertainty pre-computation adds approximately 10 to 15 min, so that one complete infill iteration finishes in less than 30 min. Consequently, the number of high-fidelity evaluations is the most meaningful and hardware-independent measure of computational effort.

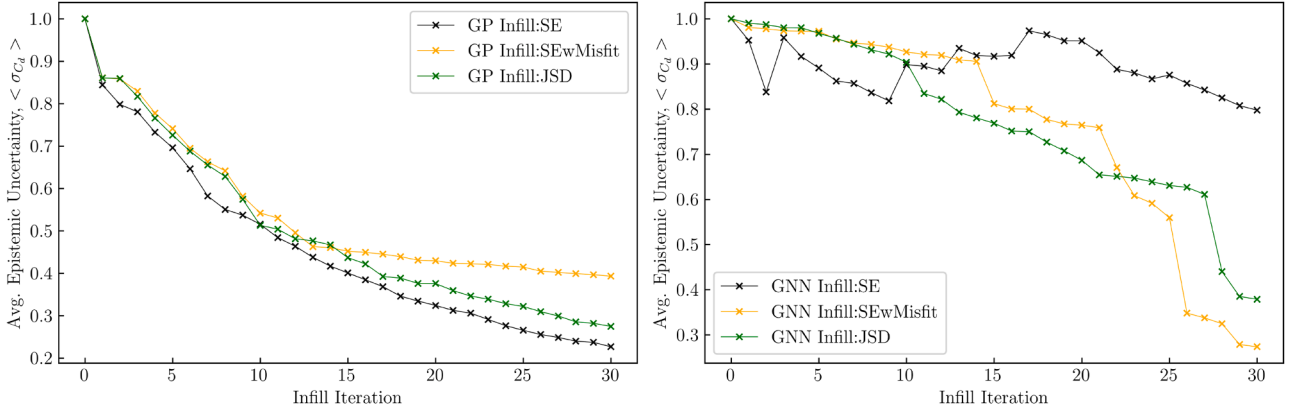


Fig. 6. Reduction of scalar epistemic uncertainty (mean predictive variance $\langle \sigma_{C_d} \rangle$) under different infill approaches. Left: GP variance for Infill:SE, Infill:SEwMisfit and Infill:JSD. Right: GNN variance for Infill:SE and the corresponding GNN variance for the two coupled approaches. Values are averaged over 10,000 MC points in input space. Note that GP Infill:SE and GNN Infill:SE are independent approaches, and use GP and GNN surrogate errors, respectively.

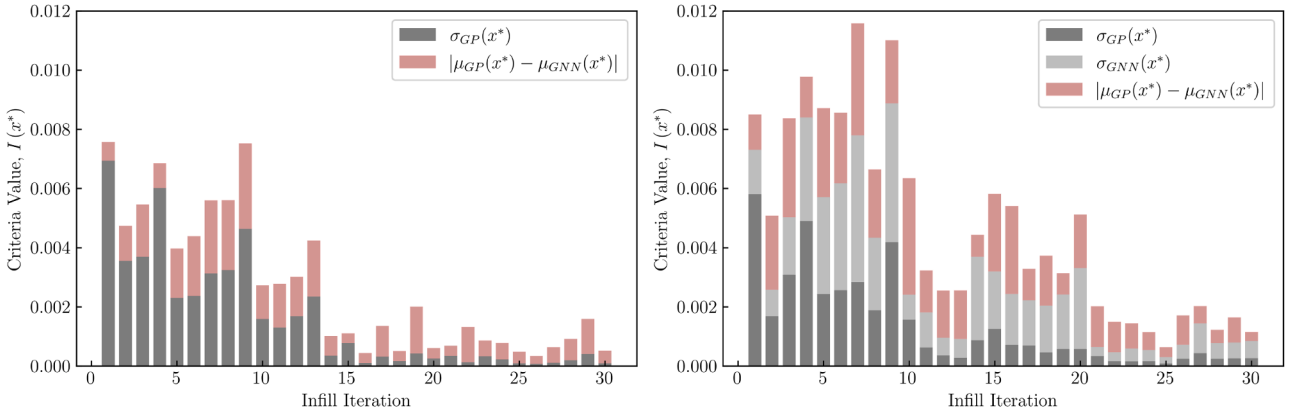


Fig. 7. Evolution of the criteria value over infill iterations for different infill strategies with iteration 0 marking the end of the DoE stage. The criteria value is the term $I(x^*)$ in the expression of the general acquisition function $\xi^* = \arg\min_{\xi} \{-\text{PDF}_{\xi} I(\xi^*)\}$. Stacked bars show contributions from different terms in the criteria value for SEwMisfit (left) and JSD (right) approaches. Such breakdown doesn't directly relate to JSD criteria, but shown here for the sake of comparison with SEwMisfit.

Fig. 6 shows the scalar epistemic variance $\langle \sigma_{C_d} \rangle$ for both GP and GNN, averaged over 10,000 MC points in the input space, as successive infill points are added. For the GP surrogate, Infill:SE reduces epistemic uncertainty fastest to about 25% of the initial spread, followed by JSD coupled infill strategy which reaches about 30% of the initial level. The SEwMisfit infill approach reduces the GP surrogate average uncertainty to about 40% of the initial. When the GNN is updated only using the MC-Dropout based estimate of uncertainty, its own variance remains high and erratic. However, the coupled criteria that combine GP variance with either GP-GNN misfit (SEwMisfit) or their Jensen-Shannon divergence reduce GNN's average uncertainty by roughly one order of magnitude. It is worth noting that beyond the 14th iteration the uncertainty in GNN drastically reduces indicating the dominance of the misfit term in the acquisition function. The average uncertainty in GNN using JSD based infill is slightly higher compared to the SEwMisfit based approach.

Fig. 7 shows the progression of the criteria value of the acquisition function across infill iterations for the two coupled infill strategies. During the first 13 infill iterations the SEwMisfit criterion drops sharply with GP predictive variance dominating the criteria value. Over the remaining iterations the criteria value remain significantly smaller and is mostly dominated by the GP-GNN misfit. For the JSD based approach we stack the predictive variance from both surrogates along with the misfit. The criteria values drop more gradually as compared to SEwMisfit. The GNN predictive variance mostly dominates the value until 20 iterations, indicating high epistemic uncertainty in the neural network. Since the JSD rewards overlap between the two predictive distributions

rather than absolute error, it continues to propose samples that steadily reconcile GP and GNN even after the largest disagreements have been resolved. By iteration 30 SEwMisfit results in a lower overall criteria value while JSD maintains a slightly higher value. Both approaches suggest greater robustness and allow for the surrogate pair to gradually converge as compared to the SE based approach using either GP and GNN alone.

The uncertainty in the surface friction coefficient field is presented in Fig. 8. At the beginning of the infill stage (iteration 0) large uncertainty around the shock and the highly-accelerated leading-edge region is observed. This can be attributed to the low number of DoE samples used for the initial GNN surrogate resulting in high variance predictions. By iteration 15, the coupled infill strategies have significantly reduced the uncertainty as compared to the SE based infill which leaves them virtually unchanged. After 30 infill points SEwMisfit and JSD move the SE based approach still struggles to even mildly diminish the uncertainty. The coupled approach therefore ensures that remaining model epistemic uncertainty is sufficiently small to not contaminate the propagation of other sources of uncertainty.

Fig. 9 summarizes how the GP surrogate predictions for the scalar outputs on the test dataset improves once adaptive sampling is activated. GP with SE-based infill reaches the target r2-score and error within 10 iterations. Whereas, GP with SEwMisfit and JSD infill takes about 30 and 23 iterations, respectively. Note that, in comparison to no-infill surrogates (Fig. 5), the adaptive sampling strategies tends to converge faster towards the target accuracy levels.

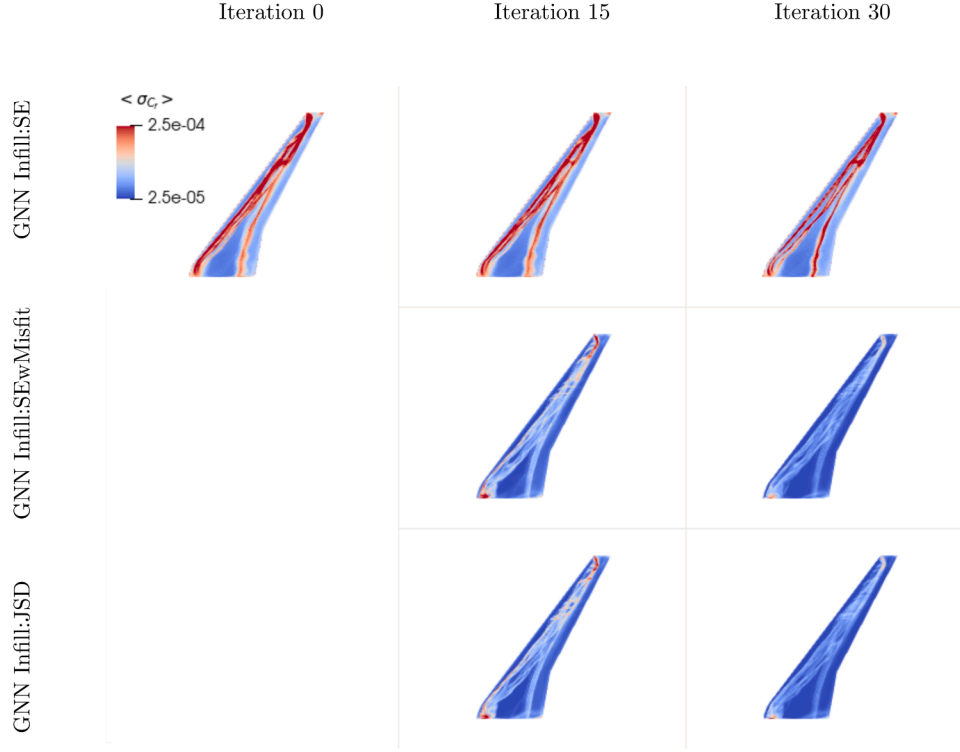


Fig. 8. Spatial distribution of the skin-friction epistemic uncertainty $\sigma_{C_f}(x, y)$ on the CRM wing for the different infill strategies averaged over 100 input samples. Columns correspond to the initial DoE with 30 samples (i.e. infill iteration 0), infill iteration 15, and iteration 30. Rows correspond to the approaches Infill:SE, Infill:SEwMisfit and Infill:JSD.

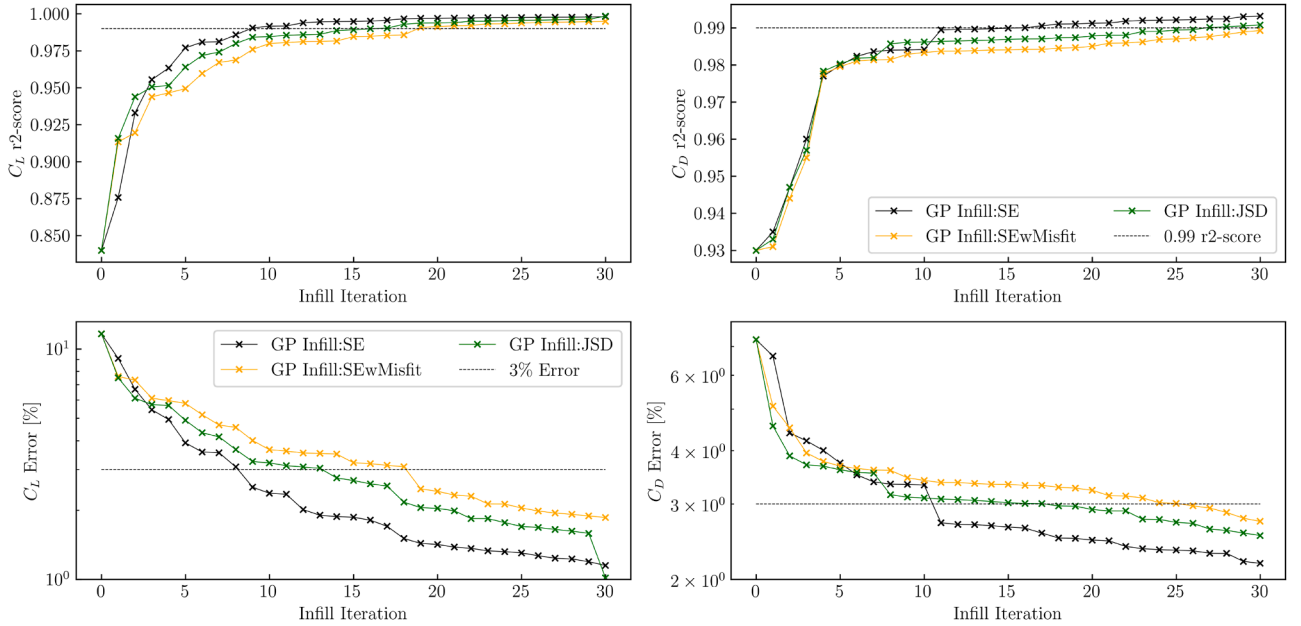


Fig. 9. Evolution of GP accuracy under three acquisition rules. Upper row: coefficient of determination (r2-score) for lift (left) and drag (right); the dashed line marks the 0.99 target. Lower row: relative RMSE normalized by the data range for lift (left) and drag (right); the 3% specification is indicated by the dashed line. Curves start at the 30 sample DoE (iteration 0).

The predicted ($\hat{C}_L, \hat{C}_D$) versus true (C_L, C_D) scalar coefficients for the test dataset after 30 infill iterations are presented in Fig. 10. The GNN (DoE + Infill) is trained on DoE samples plus SE based infill samples used to enrich GP. The GP scalar predictions using all three infill strategies are significantly closer to the true scalar values. Compared to GNN (DoE + Infill), the adaptive sampling based GNN scalar predictions (integrated using the field predictions) are much more accurate

and closer to the GP predictions. This can be attributed to the reduction of the misfit term between GP and GNN during the infill iterations of the coupled strategies.

In Fig. 11, we compare for three test points, the field predictions with the true field outputs for all the field surrogate models discussed previously. The model without infill are trained using 60 DoE samples. The difference between the predicted and the true fields is used to identify

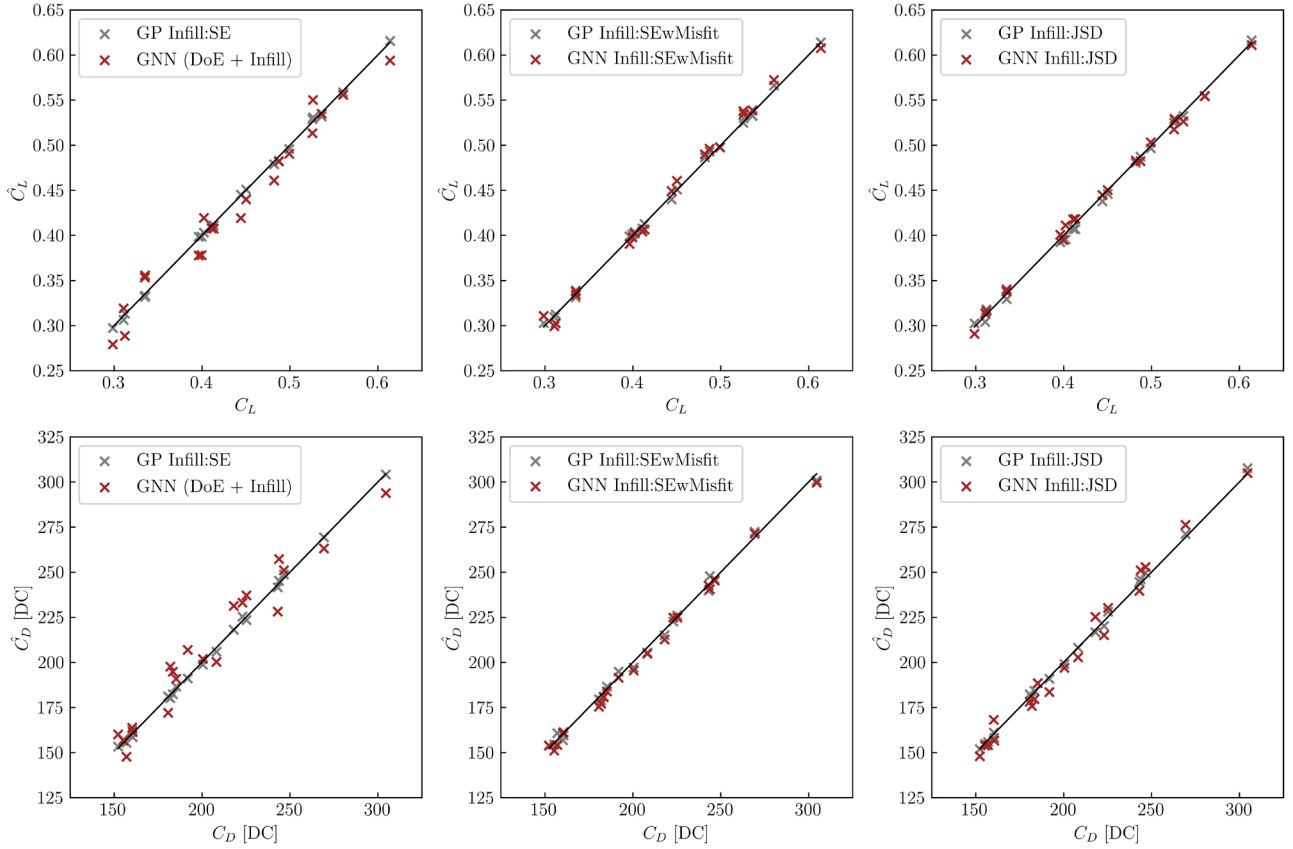


Fig. 10. Parity plots for lift (top row) and drag counts (bottom row) after 30 adaptive sampling iterations. Columns correspond to the three acquisition strategies namely, variance-only (Infill:SE), coupled variance and misfit (Infill:SEwMisfit) and JSD based coupling (Infill:JSD).

Table 1
Comparison of accuracy criterion for different active learning methods.

Method	C_p RMSE	C_f [x, y, z] RMSE in e^{-4}	C_p r2-score	C_f [x, y, z] r2-score
PODI	0.1772	7.3816, 2.6253, 2.8166	0.7740	0.5073, 0.6691, 0.8892
GNN	0.0345	3.0214, 0.9096, 0.7151	0.9914	0.9176 , 0.9613, 0.9920
GNN (Infill:SE)	0.0627	5.2325, 1.5350, 1.1119	0.9717	0.7530, 0.8898, 0.9828
GNN (Infill:SEwMisfit)	0.0225	1.9923, 0.6403, 0.5204	0.9963	0.9641 , 0.9808, 0.9962
GNN (Infill:JSD)	0.0203	1.7977, 0.7124, 0.5535	0.9970	0.9708 , 0.9762, 0.9957

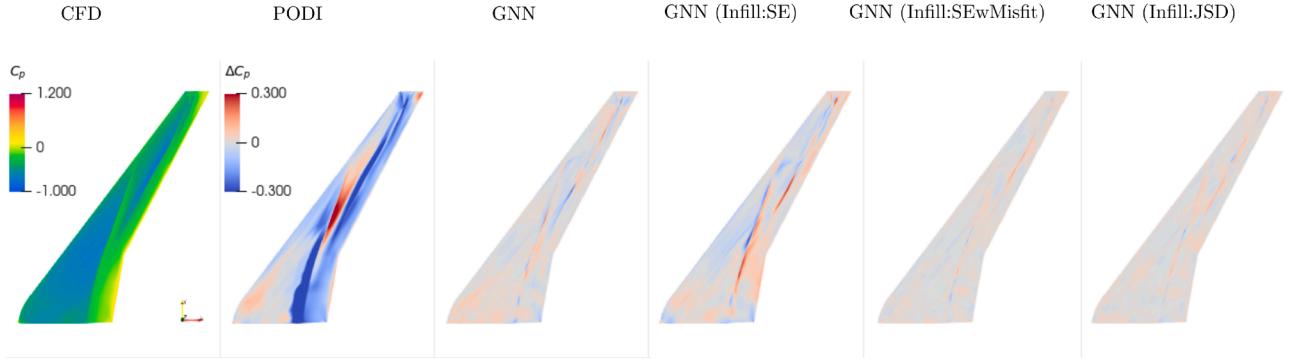
the regions of inaccuracies. The PODI fails to accurately capture the surface pressure coefficients. It shows large predictions errors close to and beyond the transition front, and near the shock locations as well. The no-infill GNN shows significant improvement over the PODI model but still fails to capture the transition front and shock locations accurately. GNN (Infill:SE) model performs better than PODI but still struggles to reach the accuracy of no-infill GNN. The coupled sampling based GNN models (Infill:SEwMisfit and Infill:JSD) have the highest accuracy of all the field surrogates. They have very low error along the transition front and are able to precisely capture the shock locations. Note that these are the only surrogate models which have a reasonable accuracy even for the test samples close to the edge of design space (Fig. 11(c)). However, it should be acknowledged that since the infill criteria weight the acquisition function by the input PDF, fewer samples are placed in low-probability regions. Surrogate predictions near or outside the boundary of the input distribution are therefore generally less reliable, as evident from the larger errors at test point (c) compared to the interior points (a) and (b).

Error metrics for all models with respect to RMSE and r2-score are given in Table 1 highlighting that the adaptive sampling strategies are outperforming the non-adaptive techniques with respect to all metrics consistently. Comparing the adaptive sampling criterions to each other

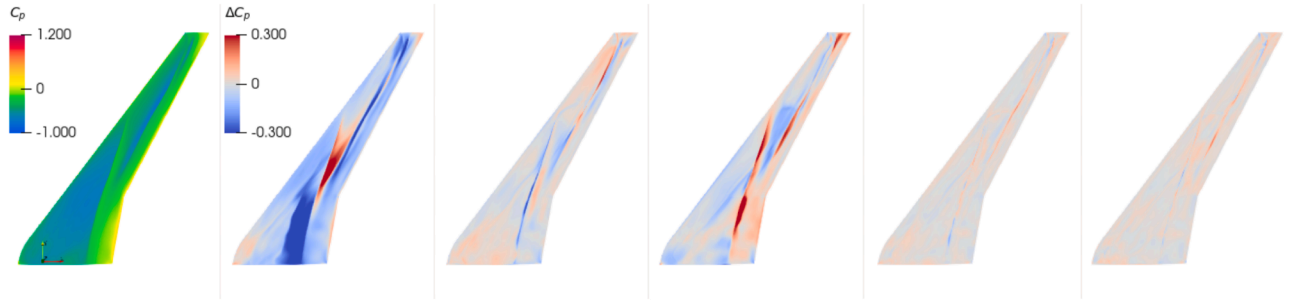
it can be stated that the SEwMisfit as well as the JSD scheme yield very similar results while the SE approach is performing worse. Given that the latter does not account for both models simultaneously this does not come as a surprise.

It is worth noting that overfitting of the GNN after multiple infill steps is a potential concern. In the present setup, this risk is mitigated by several mechanisms: early stopping with a patience of 20 epochs monitors the validation RMSE on a fixed holdout set, dropout regularization (5%) remains active during training, and each infill re-training is warm-started from the previously converged weights (typically converging in 100 to 300 additional epochs). Since the validation and test sets remain unchanged throughout the infill process, any overfitting would be directly visible in the monitored accuracy metrics.

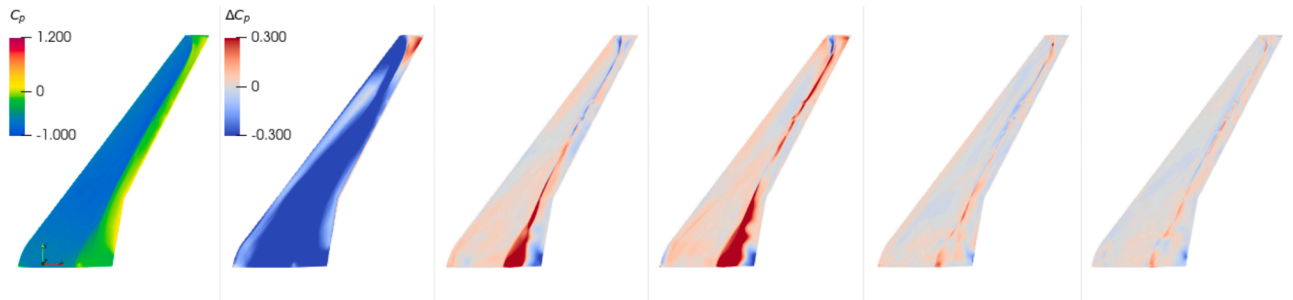
In terms of computational savings, the coupled infill approach achieves field prediction accuracy above $r2 = 0.99$ for C_p with 60 total simulations (30 DoE + 30 infill), whereas the no-infill GP requires 70 to 80 samples to meet the scalar accuracy targets alone and does not yield comparable field accuracy. This corresponds to a reduction of roughly 25 to 30% in the number of expensive CFD evaluations while simultaneously providing accurate field predictions that a purely scalar surrogate cannot deliver.



(a) Prediction output for: $Tu = 4.5551e - 3, \alpha = 1.2539, Re = 1.4949e7, M = 0.8604$



(b) Prediction output for: $Tu = 5.3816e - 3, \alpha = 1.0927, Re = 1.5131e7, M = 0.8599$



(c) Prediction output for: $Tu = 3.5746e - 3, \alpha = 2.5429, Re = 1.4853e7, M = 0.8767$

Fig. 11. Surface pressure coefficient field predictions using different field surrogate models at three different test points. Points (a) and (b) belong to high probability region of the input space, while point (c) lies closer to the boundary of the joint distribution. The difference between surrogate and CFD output is plotted to clearly spot the prediction inaccuracies. Note that the model without infill are trained using 60 DoE samples.

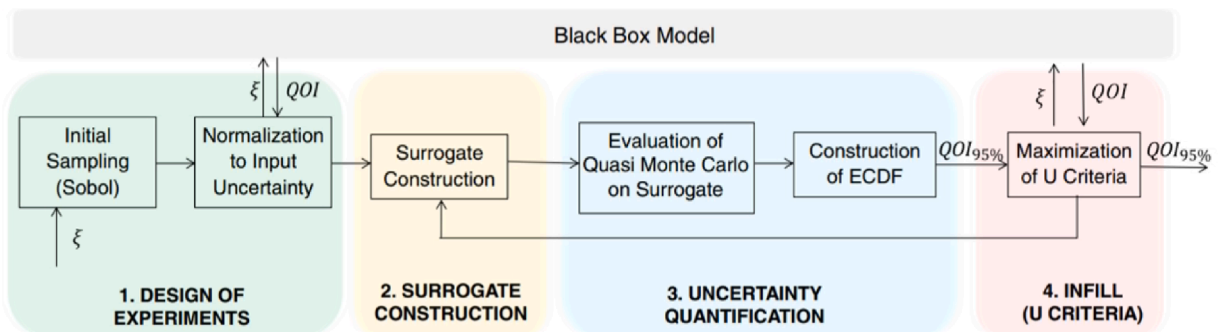


Fig. 12. Steps involved in surrogate based uncertainty propagation approach.

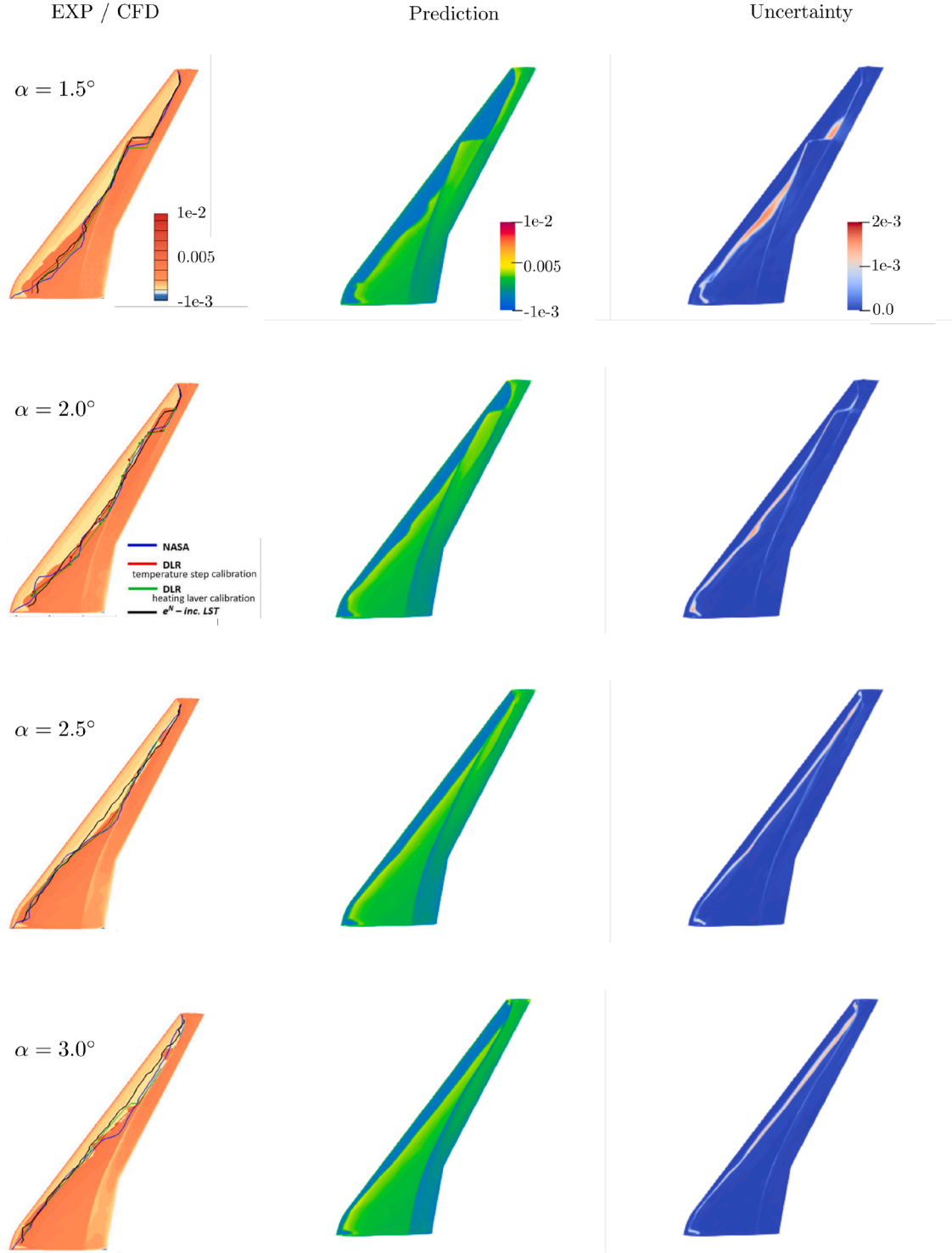


Fig. 13. Surface skin friction coefficient at four different angles of attack. The CFD simulations with experimental data (first column) is adapted from [45]. The nominal prediction and uncertainty reflected due to uncertain turbulent freestream intensity is based on the SBUQ approach using the field surrogate with GP-GNN Infill:SEwMisfit as infill criterion.

5. Uncertainty propagation – an engineering application

We employ the Surrogate-Based Uncertainty Quantification (SBUQ) framework introduced by Sabater et al. [31] and available in SMARTy [57], as illustrated in Fig. 12. In this framework, a surrogate model maps the uncertain input variables ξ to a quantity of interest (QoI), enabling efficient propagation of input uncertainties. Rather than performing computationally intensive direct Monte Carlo simulations on

the high-fidelity model, the framework uses a surrogate model to estimate statistical properties of the QoI. This strategy significantly reduces the computational burden by leveraging a fast-to-evaluate approximation of the true model response. To compute QoI statistics, the Monte Carlo method is employed not on the expensive black-box function, but on the surrogate model, which approximates its behavior. The surrogate model is trained on a set of initial Design of Experiments (DoE) samples in the stochastic input space. The predictive mean and an associated

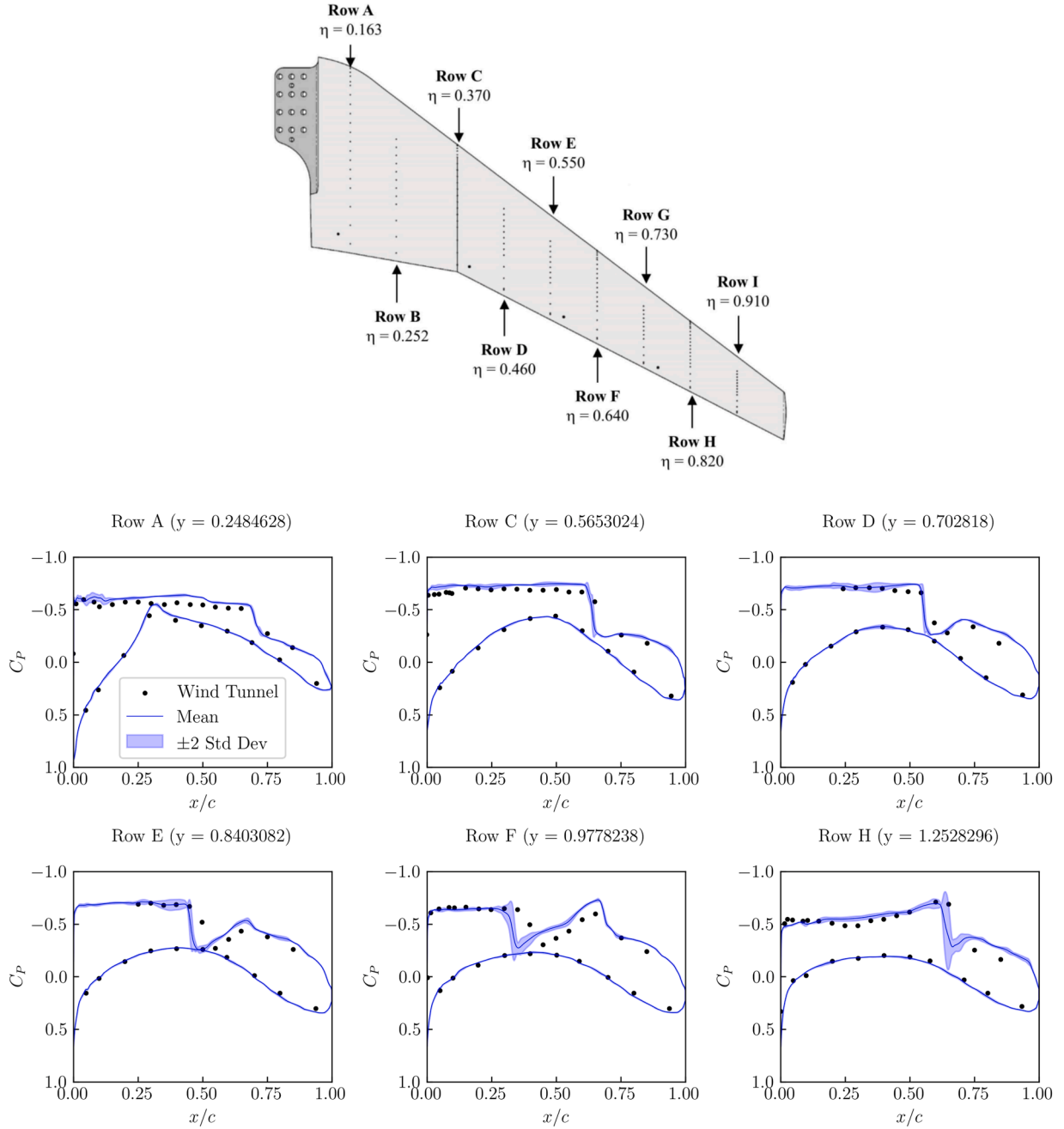


Fig. 14. Surface pressure coefficient over different stations along the wingspan at an angle of attack of 1.5° . The uncertainty propagation results are indicated using the ± 2 standard deviation bounds around the surrogate mean prediction. The mean prediction and uncertainty propagation is based on the SBUQ approach using the field surrogate with GP-GNN Infill:SEwMisfit as infill criterion.

uncertainty (variance) estimate of the surrogate model, makes it well-suited for guiding adaptive sampling strategies. To enhance surrogate accuracy, we adopt an active learning (infill) strategy that sequentially refines the surrogate model. In the baseline version, the infill criterion focuses on improving global prediction of the QoI. Once the surrogate is refined, QoI statistics are computed using a large number of samples generated via Quasi-Monte Carlo (QMC) methods. Compared to conventional Monte Carlo, QMC offers improved convergence rates and better integration accuracy in high-dimensional settings, while maintaining ease of implementation.

In contrast to [31] herein we substitute the scalar-valued based infill criterion with the previously discussed SEwMisfit criterion. This allows

us to simultaneously propagate input uncertainties to the drag coefficient as well as the surface skin friction and pressure coefficients. While the drag coefficient provides insight into the aerodynamic performance, the skin friction and pressure coefficients enable us to investigate the influence of input uncertainties on transition locations and local aerodynamic behavior leading to a further improvement of our aerodynamic understanding. Similar to before we compute 30 DoE samples using a Halton sequence and add 30 infill samples based on the SEwMisfit criterion. Sampling is enhanced by accounting for the input PDFs to concentrate samples in locations of high likelihood. This results in the same model discussed in Section 4 labeled SEwMisfit but in addition we can investigate the influence of input uncertainties on distributed quanti-

ties by drawing a lot of samples from the GNN model using a quasi Monte Carlo evaluation. Note that, very similar results have been obtained when using the JSD criterion for infill.

In Fig. 13 the influence of the uncertain turbulent freestream intensity on the surface skin friction coefficient is shown for four different angles of attack. The left column shows a combination of experimental and numerical results and is adapted from [45]. Note the, changes in transition locations derived based on different criteria/approaches. The section column displays the mean surface skin friction coefficient while its variance is displayed in the third column. A close correlation between mean prediction and the combined EXP/CFD is clearly visible. Interestingly the variation in turbulent freestream intensity results in a variance around the transition location that resembles the differences in transition location derived based on changing criteria. Transition for angles of attack above 2.5 degrees are primarily caused from the occurring shock on the upper surface. Since the influence of the turbulent freestream intensity is minimal, also the predicted variance shrinks compared to smaller angles of attack.

In addition to the skin friction coefficient also the surface pressure coefficient under uncertainties can be analyzed and compared to experimental results. Fig. 14 shows the c_p distribution at six different stations along the wingspan for an angle of attack of 1.5° with experimental results as dots and the mean prediction as solid lines. Variances are indicated by shaded areas. Overall reasonable agreement between the model predictions and experimental results is observed. For the most inboard section Row A the effect of variations in turbulent intensity is occurring towards the leading edge at around $x/c = 0.1$ which relates to variations seen in surface skin friction coefficient in Fig. 13. For Row C and D there is barely any influence of the turbulent freestream intensity visible whereas for more outboard sections variance are primarily present in the proximity of the shock. Interestingly the uncertainty in turbulent freestream intensity seems to cause a variation in shock intensity but seems to have a very limited influence on the shock location.

6. Conclusion

In this paper we present a goal-driven adaptive sampling strategy for models predicting fields such as the surface pressure distribution of an aircraft. The proposed strategy is agnostic to the field prediction model at hand and hence suitable for arbitrary deep-learning models. Effectively it consists of two components that a combined to form a composite infill criterion. First, a Gaussian Process which inherent error estimate is used to determine undersampled areas of the domain and, second, a misfit term between the integrated field model and Gaussian process prediction ensuring consistency between both models. The misfit term can either be computed in a deterministic fashion by comparing mean prediction or in a stochastic fashion by leveraging the Jensen-Shannon Divergence term.

The performance of the newly proposed adaptive sampling strategy is investigated for the NASA Common Research Model comparing the model prediction accuracy for different infill strategies and also a baseline quasi-monte Carlo, model-agnostic space-filling sampling strategy. The different strategies are infill solely-based on the Gaussian Process error and infill based on the composite formulations, namely surrogate error with misfit (SEwMisfit) as well as a stochastic formulation based on the Jensen-Shannon Divergence term (JSD). Graph neural networks are used a field-prediction models while also some initial results are shown for proper orthogonal decomposition with latent-space interpolation. Both composite infill criteria, SEwMisfit and JSD, consistently outperform the GP-only based criterion with respect to field as well as scalar value predictions. It has been observed that first the GP error portion of the criterion is active ensuring that the entire domain is adequately covered while in the later stage of the infill process the misfit term dominates and effectively refines critical flow regions, e.g. high angles of attack and Mach number, where a mismatch between local and global predictions is likely. Finally the SEwMisfit criterion is em-

ployed for forward uncertainty propagation analysis investigating the influence of uncertainties in turbulent intensity on the surface pressure and skin friction coefficient of the natural laminar flow version of the NASA CRM. Results coincide with experimental findings with respect to pressure distributions and derived laminar to turbulent transition location.

The proposed methodology is formulated independently of the application domain and could in principle be applied to other engineering fields where a distributed field output can be integrated into a meaningful scalar quantity, e.g. structural stress fields and peak stress, thermal fields and heat flux integrals, or acoustic fields and sound pressure levels. However, the present validation is limited to a single aerodynamic configuration, and transfer to other domains may require adjustments to the surrogate architectures and uncertainty estimation methods. Demonstrating the framework on additional applications remains an important direction for future work. It should be noted, however, that the present study considers a four-dimensional input space and extending the approach to significantly higher dimensions may introduce practical challenges: GP surrogates require exponentially more data to maintain accuracy due to the curse of dimensionality, and global optimization of the acquisition function becomes increasingly expensive. The GNN field surrogate is less affected by the number of input parameters since it operates on the mesh graph, but it too will need more training data to learn the richer mapping. Investigating dimensionality-reduction techniques for the input space or scalable acquisition function optimizers could help mitigate these challenges in future work. Moreover, it would be interesting to investigate more rigorous uncertainty quantification methods for the employed graph neural networks since it has been observed that the drop-out based method employed herein sometimes resulted in rather erratic behavior throughout the infill process.

CRedit authorship contribution statement

Jigar Parekh: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Philipp Bekemeyer:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Philipp Bekemeyer reports financial support was provided by German Research Foundation. Jigar Parekh reports financial support was provided by German Research Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The funding of parts of these investigations from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy-EXC 2163/1-Sustainable and Energy Efficient Aviation-Project-ID 390881007 is gratefully acknowledged.

The authors gratefully acknowledge the scientific support and HPC resources provided by the German Aerospace Center (DLR). The HPC system CARA is partially funded by 'Saxon State Ministry for Economic Affairs, Labor and Transport' and 'Federal Ministry for Economic Affairs

and Climate Action'. The HPC system CARO is partially funded by "Ministry of Science and Culture of Lower Saxony" and "Federal Ministry for Economic Affairs and Climate Action".

Appendix A. Estimating epistemic uncertainty in GNN

Epistemic uncertainty in a model can be described as the uncertainty in the model's predictions due to the lack of knowledge about the underlying model parameters. In neural networks, this can be approximated by using Monte Carlo (MC) dropout during inference, where the dropout is kept active even during the prediction phase. The procedure for approximating epistemic uncertainty adopted in this study is as follows:

- The GNN is used in training mode to generate different outputs for the same input x .
- For each input, 10,000 MC dropout samples of the field output are generated, which are in turn integrated to obtain the desired scalars (aerodynamic coefficients). The mean $\mu_{GNN}(x)$ and variance $\sigma_{GNN}(x)$ of the scalar outputs are then computed.

The use of MC dropout for estimating epistemic uncertainty in the strategy discussed above can slow down computations during optimization in surrogate models. Specifically, using MC dropout during each infill step significantly increases the computational burden, as estimates of $\mu_{GNN}(x_i)$ and $\sigma_{GNN}(x_i)$ must be computed at several input points x_i . To mitigate this issue, we pre-compute two surrogate models before each infill step for the mean and standard deviation using 1000 input points, and use them during optimization. This reduces computational time by eliminating the need of performing MC dropout at thousands of probe locations during optimization.

It should be noted that MC-dropout carries several known limitations as an estimator of epistemic uncertainty. First, it approximates Bayesian inference by treating dropout masks as samples from a Bernoulli variational distribution over the network weights [33]. This variational family is relatively restrictive and may not capture the true posterior, leading to potential underestimation of predictive variance compared to richer approximations such as deep ensembles. Second, the dropout rate (here 5%) is typically chosen for regularization purposes rather than calibrated for uncertainty quantification, and different rates can yield substantially different variance estimates. Third, the resulting variance does not necessarily decrease monotonically as more training data is added, since retraining the network changes the weight distribution. This behavior was observed in the present study, where the GNN variance remained erratic during the infill process when used as the sole infill driver (cf. Fig. 6). These limitations motivated the coupled infill strategy, which uses the analytically grounded GP variance as the primary sampling driver and treats the GNN uncertainty as a supplementary consistency check.

References

- [1] M.C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, *J. R. Stat. Soc. Ser. B* 63 (3) (2001) 425–464. <https://doi.org/10.1111/1467-9868.00294>
- [2] T.J. Santner, B.J. Williams, W.I. Notz, *The Design and Analysis of Computer Experiments*, Springer, 2003. <https://doi.org/10.1007/978-1-4757-3799-8>
- [3] A.I.J. Forrester, A. Sobester, A.J. Keane, Engineering design via surrogate modelling: a practical guide, *Eng. Des. Surrogate Modell.* (2008). <https://doi.org/10.1002/9780470770801>
- [4] D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions, *J. Global Optim.* 13 (1998) 455–492. <https://doi.org/10.1023/A:1008306431147/METRICS>
- [5] J. Sacks, W.J. Welch, T.J. Mitchell, H.P. Wynn, Design and analysis of computer experiments, *Stat. Sci.* 4 (4) (1989) 409 – 423. <https://doi.org/10.1214/ss/1177012413>
- [6] J. Bect, D. Ginsbourger, L. Li, V. Picheny, E. Vazquez, Sequential design of computer experiments for the estimation of a probability of failure, *Stat. Comput.* 22 (3) (2012) 773–793. <https://doi.org/10.1007/s11222-011-9241-4>
- [7] R.B. Gramacy, H.K.H. Lee, Adaptive design and analysis of supercomputer experiments, *Technometrics* 51 (2) (2009) 130–145. <https://doi.org/10.1198/TECH.2009.0015>
- [8] E. Brochu, V.M. Cora, N. de Freitas, A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning, (2010) *arXiv preprint arXiv:1012.2599*.
- [9] B. Shahriari, K. Swersky, Z. Wang, R.P. Adams, N. de Freitas, Taking the human out of the loop: a review of Bayesian optimization, *Proc. IEEE* 104 (1) (2016) 148–175. <https://doi.org/10.1109/JPROC.2015.2494218>
- [10] B. Peherstorfer, K. Willcox, M. Gunzburger, Survey of multifidelity methods in uncertainty propagation, inference, and optimization, *SIAM Rev.* 60 (3) (2018) 550–591. <https://doi.org/10.1137/16M1082469>
- [11] D. Xiu, *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton University Press, 2010. <https://doi.org/10.1515/9781400835348>
- [12] M. Guénot, I. Lepot, C. Sainvitu, J. Goblet, R. Filomeno Coelho, Adaptive sampling strategies for non-intrusive POD-based surrogates, *Eng. Comput.* 30 (4) (2013) 521–547. <https://doi.org/10.1108/02644401311329352>
- [13] M.A. Álvarez, L. Rosasco, N.D. Lawrence, Kernels for vector-valued functions: a review, *Found. Trends Mach. Learn.* 4 (3) (2012) <https://doi.org/10.1561/22000000036>
- [14] V. Picheny, D. Ginsbourger, O. Roustant, L. Carraro, Gaussian process modeling with inequality constraints, in: *Advances in Intelligent Data Analysis XV*, 2016, pp. 1–14. <https://doi.org/10.5802/afst.1344>
- [15] T. Chugh, Scalarizing functions in Bayesian multiobjective optimization, (2019) *arXiv preprint arXiv:1904.05760*.
- [16] R. Graham, J. Cortés, Cooperative adaptive sampling of random fields with partially known covariance, *Int. J. Robust Nonlinear Contr.* 20 (9) (2010) 1010–1029. <https://doi.org/10.1002/rnc.1710>
- [17] N. Cressie, *Statistics for Spatial Data*, Wiley, 1993. <https://doi.org/10.1002/9781119115151>
- [18] M.L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*, Springer, 1999. <https://doi.org/10.1007/978-1-4612-1494-6>
- [19] L.-K. Song, F. Tao, X.-Q. Li, L.-C. Yang, Y.-P. Wei, M. Beer, Physics-embedding multi-response regressor for time-variant system reliability assessment, *Reliab. Eng. Syst. Saf.* 263 (2025) 111262. <https://doi.org/10.1016/j.res.2025.111262>
- [20] L.-K. Song, F. Tao, G.-Z. Peng, Mixed loss-guided modular regression for dependent system reliability, *Reliab. Eng. Syst. Saf.* 267 (2026) 111898. <https://doi.org/10.1016/j.res.2025.111898>
- [21] T.J. Sullivan, *Introduction to Uncertainty Quantification*, Springer, 2015. <https://doi.org/10.1007/978-3-319-23395-6>
- [22] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- [23] A. Bertram, Data-driven variable-fidelity reduced order modeling for efficient vehicle shape optimization (2018). <https://doi.org/10.24355/dbbs.084-201811231243-0>
- [24] S. Lophaven, H. Nielsen, J. Sondergaard, *DACE - A Matlab Kriging Toolbox*, Version 2.0, 2002.
- [25] M. Guo, J.S. Hesthaven, Data-driven reduced order modeling for time-dependent problems, *Comput. Methods Appl. Mech. Eng.* 345 (2019) 75–99. <https://doi.org/10.1016/j.cma.2018.10.029>
- [26] C. Eckart, G. Young, The approximation of one matrix by another of lower rank, *Psychometrika* 1 (3) (1936) 211–218. <https://doi.org/10.1007/BF02288367>
- [27] T. Pfaff, M. Fortunato, A. Sanchez-Gonzalez, P. Battaglia, Learning mesh-based simulation with graph networks, in: *International Conference on Learning Representations*, 2021. https://openreview.net/forum?id=roNqYLO_XP
- [28] D. Hines, P. Bekemeyer, Graph neural networks for the prediction of aircraft surface pressure distributions, *Aerosp. Sci. Technol.* 137 (2023) 108268. <https://doi.org/10.1016/j.ast.2023.108268>
- [29] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, P. Battaglia, Learning to simulate complex physics with graph networks, in: *Proceedings of the 37th International Conference on Machine Learning*, 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 8459–8468. <https://proceedings.mlr.press/v119/sanchez-gonzalez20a.html>
- [30] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, in: *Proceedings of the 34th International Conference on Machine Learning*, PMLR, 2017, pp. 1263–1272. <https://proceedings.mlr.press/v70/gilmer17a.html>
- [31] C. Sabater, P. Bekemeyer, S. Görtz, Efficient bilevel surrogate approach for optimization under uncertainty of shock control bumps, *AIAA J.* 58 (2020) 5228–5242. <https://doi.org/10.2514/1.J059480>
- [32] J. Parekh, P. Bekemeyer, S. Helm, D.G. François, C. Grabe, Surrogate based design space exploration and exploitation for an efficient airfoil optimization under uncertainties using transition models, *Aerosp. Sci. Technol.* 154 (2024) 109532. <https://doi.org/10.1016/j.ast.2024.109532>
- [33] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in: *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, PMLR, 2016, pp. 1050–1059. <https://proceedings.mlr.press/v48/gal16.html>
- [34] L. Breiman, Bagging predictors, *Mach. Learn.* 24 (2) (1996) 123–140. <https://doi.org/10.1007/BF00058655>
- [35] P. Gardner, T.J. Rogers, C.E. Lord, R.J. Barthorpe, Learning model discrepancy: a Gaussian process and sampling-based approach, *Mech. Syst. Signal Process.* (2021). <https://api.semanticscholar.org/CorpusID:230573913>
- [36] A. Garbo, J. Parekh, T. Rischmann, P. Bekemeyer, Multi-fidelity adaptive sampling for surrogate-based optimization and uncertainty quantification, *Aerospace* 11 (6) (2024). <https://www.mdpi.com/2226-4310/11/6/448>
- [37] J. Lin, Divergence measures based on the Shannon entropy, *IEEE Trans. Inf. Theory* 37 (1) (1991) 145–151. <https://doi.org/10.1109/18.61115>
- [38] S. Kullback, R.A. Leibler, On information and sufficiency, *Ann. Math. Stat.* 22 (1) (1951) 79–86. <https://doi.org/10.1214/aoms/1177729694>

- [39] L.-j. Chen, Y. Hong, S. Mangalathu, H.-y. Gou, Q.-h. Pu, Adaptive sampling approach based on Jensen-Shannon divergence for efficient reliability analysis, *J. Cent. South Univ.* 28 (9) (2021) 2407–2422. <https://doi.org/10.1007/s11771-021-4740-8>
- [40] M.N. Lynde, R.L. Campbell, Computational design and analysis of a transonic natural laminar flow wing for a wind tunnel model, in: *AIAA Aviation 2017 Forum*, 2017. AIAA 2017-3693, <https://ntrs.nasa.gov/citations/20170005688>.
- [41] M.N. Lynde, R.L. Campbell, M.B. Rivers, S.A. Viken, D.T. Chan, A.N. Watkins, S.L. Goodliff, Preliminary results from an experimental assessment of a natural laminar flow design method, *AIAA Scitech 2019 Forum* (2019). <https://doi.org/10.2514/6.2019-2298>
- [42] T. Gerhold, Overview of the hybrid RANS code TAU, *MEGAFLow Numer. Flow Simul. Aircr. Des.* (2005) 81–92. https://doi.org/10.1007/3-540-32382-1_5
- [43] F.R. Menter, M. Kuntz, R. Langtry, Ten years of industrial experience with the SST turbulence model, *Turbul. Heat Mass Transf.* 4 (1) (2003) 625–632. <https://api.semanticscholar.org/CorpusID:38105247>.
- [44] D.G. François, A. Krumbein, N. Krimmelbein, C. Grabe, Simplified stability-based transition transport modeling for unstructured computational fluid dynamics, *J. Aircr.* (2023) 1–12. <https://doi.org/10.2514/1.C037163>.
- [45] A.M. Krumbein, D.G. François, N. Krimmelbein, Transport-based transition prediction for the common research model natural laminar flow configuration, in: *AIAA Scitech 2022 Forum*, 2022. AIAA 2022-1541. <https://doi.org/10.2514/6.2022-1541>
- [46] J. Perraud, H. Deniau, G. Casalis, Overview of transition prediction tools in the elsA software, in: *ECCOMAS 2014*, 2014.
- [47] D.G. François, A. Krumbein, N. Krimmelbein, Crossflow extension of a simplified transition transport model for three-dimensional aerodynamic configurations, in: *AIAA AVIATION Forum*, 2022. <https://arc.aiaa.org/doi/abs/10.2514/6.2022-3678>.
- [48] S. Helm, D.G. François, C. Grabe, J. Parekh, P. Bekemeyer, CFD-basierte transitionsvorhersage für den entwurf von laminarflugzeugen, in: *Dt. Luft-und Raumfahrt Kongress DLRK 2023*, 2023.
- [49] J. van Ingen, A Suggested Semi-Empirical Method for the Calculation of the Boundary Layer Transition Region (1956). <https://api.semanticscholar.org/CorpusID:118788950>.
- [50] G. Schrauf, Industrial view on transition prediction, *Recent Results Laminar Turbul. Transition* (2004) 111–122. https://doi.org/10.1007/978-3-540-45060-3_9
- [51] T. Streit, S. Wedler, M. Kruse, DLR natural and hybrid transonic laminar wing design incorporating new methodologies, *Aeronaut. J.* 119 (2015) 1303–1326. <https://doi.org/10.1017/S0001924000011283>
- [52] M. Kruse, A. Küpper, R. Petzold, F. Munoz, Determination of the critical cross flow N-factor for the low-speed wind tunnel Braunschweig (DNW-NWB), *Notes Numer. Fluid Mech. Multi. Des.* 136 (2018) 251–261. https://doi.org/10.1007/978-3-319-64519-3_23
- [53] C. Sabater, P. Bekemeyer, S. Görtz, Robust design of transonic natural laminar flow wings under environmental and operational uncertainties, *AIAA J.* 60 (2022) 767–782. <https://doi.org/10.2514/1.J060676>
- [54] L.M. Mack, Transition prediction and linear stability theory, *ltt* (1977). <https://ui.adsabs.harvard.edu/abs/1977ltt..agarQ....M/abstract>.
- [55] R. Radespiel, W. Heinze, SFB 880: fundamentals of high lift for future commercial aircraft, *CEAS Aeronaut. J.* (2014). <https://doi.org/10.1007/s13272-014-0103-6>
- [56] M.N. Lynde, R.L. Campbell, S.A. Viken, Additional findings from the common research model natural laminar flow wind tunnel test, in: *AIAA Aviation 2019 Forum*, 2019. AIAA 2019-3292. <https://doi.org/10.2514/6.2019-3292>
- [57] P. Bekemeyer, A. Bertram, D.A.H. Chaves, M.D. Ribeiro, A. Garbo, A. Kiener, C.S. Campomanes, M. Stradtner, S. Wassing, M. Widhalm, S. Görtz, F. Jäckel, R. Hoppe, N. Hoffmann, Data-driven aerodynamic modeling using the DLR SMARTy toolbox, *American Institute of Aeronautics and Astronautics Inc, AIAA*, 2022. <https://doi.org/10.2514/6.2022-3899>