



Physics-Informed Deep Orthogonal Decomposition with Calibrated Uncertainty Quantification for Reduced-Order Modelling of Parametrized Partial Differential Equations

Umar Daraz¹, Sehrish Maryam^{2,*}, Khaleel Ahmad^{3,4}

¹ Department of Physics, University of Education Lahore, Pakistan

² Department of Mathematics, Bahauddin Zakariya University, Multan, Sub Campus Vehari, Pakistan

³ Department of Mathematics and Statistics, The University of Lahore, Pakistan

⁴ Center for Theoretical Physics, Khazar University, 41 Mehseti Str., Baku, AZ1096, Azerbaijan

ARTICLE INFO

Article history:

Received 18 April 2026

Received in revised form 25 July 2026

Accepted 1 September 2026

Available online 6 September 2026

Keywords:

Reduced-Order Modelling;
Adaptive Reduced Basis; Physics-Informed Learning; Uncertainty Quantification; Conformal Prediction; Kolmogorov Width

ABSTRACT

Solving parametrized partial differential equations thousands of times is the bottleneck of design, control, inversion, and digital-twin workflows. Reduced-order models answer it by replacing the expensive discretization with a fast surrogate. Adaptive-basis surrogates, such as the deep orthogonal decomposition, break the slow decay of the Kolmogorov width for problems whose parameters interact with space. This work develops physics-informed deep orthogonal decomposition, which supervises a continuously adaptive basis and the reduced coefficients with the discrete equation residual at unlabelled collocation parameters. The surrogate learns from points the high-fidelity solver never visits. A calibrated uncertainty layer couples a heteroscedastic aleatoric estimate with a deep-ensemble epistemic estimate. The linear decoder is proven to transport this uncertainty to the physical field exactly and without inflation. A solver-free trust indicator fuses the predictive variance, the equation residual, and a Grassmann distance to the training manifold, backed by an exact error decomposition, two-sided residual bounds, and a distribution-free coverage guarantee.

*Corresponding author.

E-mail address: sehrishmaryamvehari@gmail.com

<https://doi.org/10.66972/iscis202618>

© The Author(s) 2026 | [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

1. Introduction

The numerical solution of parametrized partial differential equations is a cornerstone of computational science and engineering, but its cost becomes prohibitive in the many-query and real-time settings that dominate design, control, inversion, and digital-twin applications, where the governing equations must be solved thousands or millions of times. Reduced-order modelling answers this challenge by replacing the expensive high-fidelity discretization, the high-fidelity model, with an inexpensive surrogate that preserves the essential input-to-output behaviour. The scope and the mathematical foundations of this field are surveyed by Benner *et al.* [1], who classify projection-based reduction methods for parametric systems. The reduced-basis methodology and its certification were developed in book form by Quarteroni *et al.* [2] as well as with a strong emphasis on a-posteriori error control by Hesthaven *et al.* [3]. The earlier survey of Rozza *et al.* [4] established the affine-decomposition and greedy machinery that made these methods practical for elliptic and parabolic problems. The practical importance of this acceleration is hard to overstate: an uncertainty-quantification study may require tens of thousands of forward solves to propagate input distributions, an inverse problem embeds the forward solve inside an outer optimization or sampling loop, and a digital twin must respond in real time to streaming data, so that a reduction of the per-solve cost by two or three orders of magnitude is often the difference between a tractable and an intractable analysis. It is precisely in these repeated-query regimes, rather than in a single high-fidelity simulation, that the reliability of the surrogate becomes as important as its speed, because errors are no longer inspected one at a time but are propagated blindly through a long computation.

Almost every reduced-order model begins with a linear compression of the solution set. The proper orthogonal decomposition, introduced in the turbulence context by Berkooz *et al.* [5], extracts an energy-optimal linear subspace from a family of solution snapshots. The method of snapshots of Sirovich [6] made its computation tractable for large discretizations. The convergence of the associated Galerkin reduced models was analyzed by Kunisch & Volkwein [7]. The accuracy attainable by any such linear method at a given latent dimension is governed by the Kolmogorov width of the solution manifold, whose role in the approximation of high-dimensional parametric problems is made precise by Cohen & DeVore [8]. For diffusion-dominated problems this width decays quickly and linear reduction is extremely effective, but for transport-dominated and geometry-dependent problems it decays slowly, a limitation analyzed for reduced-basis methods by Ohlberger & Rave [9] and quantified for wave-type problems by Greif & Urban [10]. This slow decay, often called the Kolmogorov barrier, motivated a family of nonlinear and adaptive alternatives, among them the transport-aware and registration-based constructions surveyed and advanced by Peherstorfer [11].

Several distinct responses to the barrier have been proposed. Deep convolutional autoencoders that learn a nonlinear latent manifold were introduced for model reduction by Lee & Carlberg [12]. The deep-learning reduced-order model that couples proper orthogonal decomposition with a deep decoder was developed by Fresca *et al.* [13] and refined for nonlinear parametric problems by Fresca & Manzoni [14]. A complementary line keeps the reduction linear but lets the subspace vary. For instance, interpolation of reduced-order bases on the Grassmann manifold was proposed by Amsallem & Farhat [15], while local bases assembled by clustering were introduced by Amsallem *et al.* [16]. A localized reduced-basis construction for problems with strong parameter dependence was studied by Pagani *et al.* [17]. A registration method that aligned coherent features across parameters was given by Taddei [18]. On the non-intrusive side, the proper-orthogonal-decomposition neural network of Hesthaven & Ubbiali [19] learns the map from parameters to reduced coefficients over a fixed global basis, and the Gaussian-process regression surrogate of Guo & Hesthaven [20] does the same for time-dependent problems. The deep orthogonal decomposition of Franco *et al.* [21] unifies the adaptive-basis and deep-learning viewpoints by learning a continuously parameter-dependent orthonormal

basis, and it is the direct starting point of the present work.

In parallel, scientific machine learning has produced tools for embedding physics directly into learned models. The physics-informed neural network of Raissi *et al.* [22] minimizes the equation residual at collocation points to solve forward and inverse problems. The deep operator network of Lu *et al.* [23] and the Fourier neural operator of Li *et al.* [24] learn solution operators between function spaces. The physics-informed operator learning of Wang *et al.* [25] combines the two ideas, and the general theory of neural operators is set out by Kovachki *et al.* [26]. The use of physics-informed learning specifically to improve reduced-order models was explored by Chen *et al.* [27], who regularize a global reduced representation with the governing equations. What has not been done, and what we do here, is to use the equation residual to supervise a continuously adaptive basis. The operator-learning models of Lu *et al.* [23], Li *et al.* [24], and Kovachki *et al.* [26] learn a global map between function spaces and are not organized around a reduced basis at all. Thus, they offer neither the interpretability of a small set of modes nor the certified projection that a linear decoder provides. On the other hand, the physics-informed reduced model of Chen *et al.* [27] keeps the basis global and therefore remains bounded by the width of that basis. Supervising a continuously adaptive basis with the residual combines the barrier-breaking capacity of the former line with the structure and certifiability of the latter.

The quantification of predictive uncertainty for such surrogates has its own mature toolbox. Deep ensembles were shown to give simple and scalable predictive uncertainty by Lakshminarayanan *et al.* [28]. Monte-Carlo dropout was interpreted as approximate Bayesian inference [29], and a lightweight Laplace approximation for deep networks was made practical by Daxberger *et al.* [30]. Distribution-free coverage is provided by conformal prediction, whose theory is developed in [31] and made accessible by the tutorial of Angelopoulos & Bates [32]. A recurring difficulty in heteroscedastic regression, namely that the variance head can absorb error from the mean, was analyzed by Seitzer *et al.* [33], whose remedy we adopt. Finally, because the object learned by an adaptive-basis model is a subspace, its natural geometry is that of the Grassmann manifold, whose numerical treatment was established by Edelman *et al.* [34] and placed in a general optimization framework by Absil *et al.* [35]. The multifidelity perspective that turns a calibrated surrogate into an estimator with a controlled budget is surveyed in [36], and the greedy a-posteriori sampling that our trust indicator makes possible dates back to [37].

Against this background three limitations of the current adaptive-basis surrogates stand out, and they are exactly the obstacles to using them inside the applications listed above. First, they are data hungry: the basis and coefficient networks are trained by a supervised reconstruction loss over high-fidelity snapshots, each of which is an expensive solve. Second, they are deterministic: they return a single field with no measure of confidence, whereas a surrogate placed inside uncertainty quantification or a digital twin must report how far it can be trusted at each query. Third, they fail silently outside the training distribution, degrading without any warning. The central observation of this paper is that a single ingredient, the discrete equation residual, addresses all three at once, because it is simultaneously a label-free training signal, a physical prior that improves extrapolation, and a computable and certified error surrogate that drives an uncertainty estimate. Building on this observation, our contributions are the following:

- (i) We introduce physics-informed deep orthogonal decomposition, which augments the training objective with the squared discrete residual evaluated at unlabelled collocation parameters, and thereby couples a continuously adaptive basis to the governing equations.
- (ii) We equip the surrogate with a calibrated uncertainty layer and prove that the linear reduced-order decoder propagates coefficient uncertainty to the physical field exactly and without inflation, in contrast with the biased and expansive propagation of nonlinear decoders.

- (iii) We construct a trust indicator that requires no high-fidelity solve, and we support it with an exact error decomposition, two-sided residual error bounds, a certified abstention rule, and a distribution-free coverage guarantee.
- (iv) We validate every claim on a parametrized advection diffusion benchmark that realizes the slow-global and fast-slice regime, and we release a reproducible implementation.

The framework as a whole is summarized in Figure 1. Labelled snapshots and unlabelled collocation parameters feed a single adaptive-basis model whose semi-supervised objective couples data to the equation residual; the trained surrogate yields three certificates, a calibrated uncertainty, a residual error bound, and a Grassmann novelty distance, that are fused into a trust indicator with an accept or abstain decision.

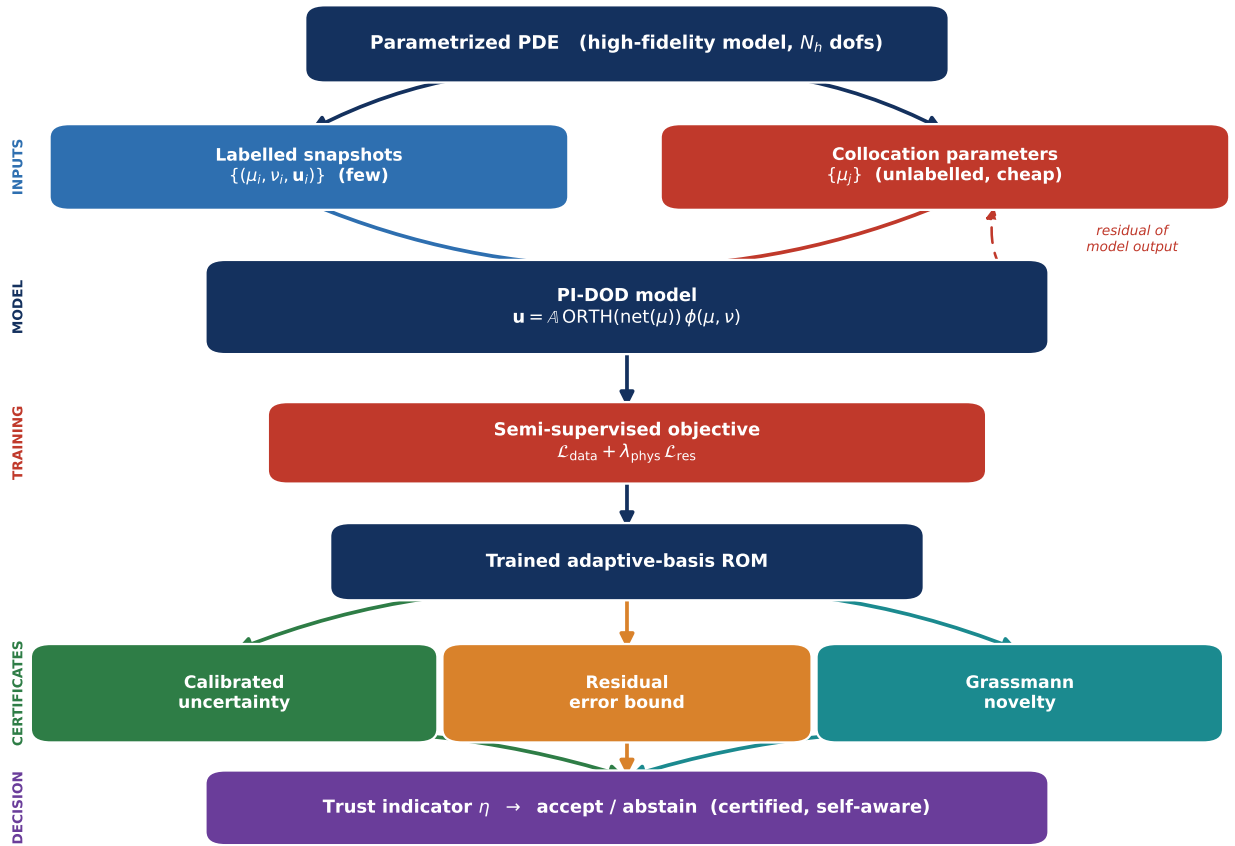


Fig. 1. Overview of the physics-informed deep orthogonal decomposition

The remainder of the paper is organized as follows. Section 2 fixes notation and states the definitions and assumptions we rely on. Section 3 develops the physics-informed training objective. Section 4 builds the uncertainty layer and the trust indicator. Section 5 collects the theoretical results and their proofs. Section 6 compares the method with existing approaches. Section 7 reports the numerical study, and Section 8 concludes.

2. Preliminaries

We fix notation and recall the notions on which the construction rests. A stable discretization of the parametrized problem on a bounded spatial domain $\Omega \subset \mathbb{R}^d$ produces a finite-dimensional state space V_h of dimension N_h with basis $\{\varphi_i\}_{i=1}^{N_h}$; a function $u \in V_h$ is identified with its vector of degrees

of freedom $\mathbf{u} \in \mathbb{R}^{N_h}$. The inner product of V_h yields the symmetric positive definite Gram matrix $\mathbb{G}_{ij} = \langle \varphi_i, \varphi_j \rangle$ and the discrete norm

$$\|\mathbf{u}\| = \sqrt{\mathbf{u}^\top \mathbb{G} \mathbf{u}}, \quad (1)$$

which coincides with the $L^2(\Omega)$ norm of u . The parameters are split into a group $\boldsymbol{\mu} \in \Theta \subset \mathbb{R}^p$ that interacts with space and a group $\boldsymbol{\nu} \in \Theta' \subset \mathbb{R}^{p'}$ whose effect is captured by a small linear subspace, both drawn from compact sets carrying probability distributions $\mathbb{P}$ and $\mathbb{Q}$. The parameter-to-solution map is $\mathcal{F} : (\boldsymbol{\mu}, \boldsymbol{\nu}) \mapsto \mathbf{u}_{\boldsymbol{\mu}, \boldsymbol{\nu}}$ and the solution manifold is $\mathcal{S} = \{\mathbf{u}_{\boldsymbol{\mu}, \boldsymbol{\nu}} : (\boldsymbol{\mu}, \boldsymbol{\nu}) \in \Theta \times \Theta'\}$. In the affine setting the discrete problem reads $\mathbb{A}(\boldsymbol{\mu})\mathbf{u}_{\boldsymbol{\mu}, \boldsymbol{\nu}} = \mathbf{f}(\boldsymbol{\mu}, \boldsymbol{\nu})$ with an invertible operator $\mathbb{A}(\boldsymbol{\mu}) \in \mathbb{R}^{N_h \times N_h}$, and the residual of an arbitrary field $\mathbf{u}$ is $\mathcal{R}(\boldsymbol{\mu}, \boldsymbol{\nu}, \mathbf{u}) = \mathbb{A}(\boldsymbol{\mu})\mathbf{u} - \mathbf{f}(\boldsymbol{\mu}, \boldsymbol{\nu})$. The recurring symbols are collected in Table 1.

Table 1
Principal notation used throughout the paper.

Symbol	Meaning
$\boldsymbol{\mu}, \boldsymbol{\nu}$	space-interacting and benign parameters
$\mathbf{u}_{\boldsymbol{\mu}, \boldsymbol{\nu}}$	high-fidelity solution in degrees of freedom
N_h, N_A, n	high-fidelity, ambient, and latent dimensions, $n \ll N_A \ll N_h$
$\mathbb{G}$	Gram matrix, with $\ \cdot\ $ the induced norm
$\mathbb{A}$	ambient basis, $\mathbb{G}$ -orthonormal, size $N_h \times N_A$
$\tilde{\mathbb{V}}_{\boldsymbol{\mu}}, \mathbb{V}_{\boldsymbol{\mu}}$	adaptive basis in ambient and full space, $\mathbb{V}_{\boldsymbol{\mu}} = \mathbb{A}\tilde{\mathbb{V}}_{\boldsymbol{\mu}}$
$\phi(\boldsymbol{\mu}, \boldsymbol{\nu})$	reduced coefficient network, segregated rank- m form
$\mathbf{c}_{\boldsymbol{\mu}, \boldsymbol{\nu}}$	exact coefficients $\mathbb{V}_{\boldsymbol{\mu}}^\top \mathbb{G} \mathbf{u}_{\boldsymbol{\mu}, \boldsymbol{\nu}}$
$d_n(\cdot)$	Kolmogorov width; $d(\cdot, \cdot)$ Grassmann projection metric
$\Sigma_{\mathbf{c}}, s^2$	coefficient and field predictive variances
ρ, v, δ, η	residual, uncertainty, novelty, and trust signals
λ_{phys}	physics-residual weight

The accuracy of any linear reduction is controlled by the Kolmogorov width, which we now recall.

Definition 1. The Kolmogorov n -width of a set $S \subset \mathbb{R}^{N_h}$ is

$$d_n(S) = \inf_{\mathbb{V} \in \mathbb{R}^{N_h \times n}} \sup_{\mathbf{u} \in S} \|\mathbf{u} - \mathbb{V}\mathbb{V}^\top \mathbb{G} \mathbf{u}\|, \quad (2)$$

the worst-case error of the best n -dimensional subspace, with $\mathbb{V}$ taken $\mathbb{G}$ -orthonormal [8].

Definition 2. Given snapshots $\{\mathbf{u}_i\}_{i=1}^N$, the proper orthogonal decomposition of dimension n is the $\mathbb{G}$ -orthonormal matrix $\mathbb{V} \in \mathbb{R}^{N_h \times n}$ whose columns are the leading eigenvectors of the snapshot correlation operator, equivalently the minimizer of $\sum_i \|\mathbf{u}_i - \mathbb{V}\mathbb{V}^\top \mathbb{G} \mathbf{u}_i\|^2$ over $\mathbb{G}$ -orthonormal subspaces [5].

Definition 3. The Grassmann manifold $\mathcal{G}_n(\mathbb{R}^{N_A})$ is the set of n -dimensional linear subspaces of $\mathbb{R}^{N_A}$; a point is represented, non-uniquely, by any orthonormal matrix whose columns span it [34].

Definition 4. For subspaces $\mathcal{V}, \mathcal{W} \in \mathcal{G}_n(\mathbb{R}^{N_A})$ with orthonormal representatives $\tilde{\mathbb{V}}, \tilde{\mathbb{W}}$, the projection metric is

$$d(\mathcal{V}, \mathcal{W}) = (1 - \sigma_{\min}^2(\tilde{\mathbb{V}}^\top \tilde{\mathbb{W}}))^{1/2} \in [0, 1], \quad (3)$$

where $\sigma_{\min}$ is the smallest singular value; it is independent of the chosen representatives [35].

Definition 5. A reduced-order model with basis $\mathbb{V}$ and coefficient map ϕ approximates the solution by $\mathbf{u}_{\mu,\nu}^{\text{ROM}} = \mathbb{V} \phi(\mu, \nu)$; when ϕ is obtained by projecting the governing equations onto $\text{span}(\mathbb{V})$ the model is intrusive, and when ϕ is learned from data it is non-intrusive [2].

Definition 6. Given a calibration sample of nonconformity scores $z_1, \dots, z_N$ and a level $1 - \alpha$, split-conformal prediction sets the radius $q_{1-\alpha}$ equal to the $\lceil (N+1)(1-\alpha) \rceil$ -th smallest score and forms the interval $[\hat{y} - q_{1-\alpha}, \hat{y} + q_{1-\alpha}]$ around a point prediction $\hat{y}$ [31].

Definition 7. A deep ensemble is a collection of M independently trained networks whose predictive distribution is the mixture of the members; its mean is the average prediction and its variance decomposes, by the law of total variance, into an average within-member variance and the between-member variance [28].

We work in the regime that makes an adaptive basis worthwhile, formalized by two assumptions on the decay of the width.

Assumption 1. There are constants $C > 0$ and $\alpha \in (0, 1]$ such that $d_n(\mathcal{S}) \leq C n^{-\alpha}$, that is, the global manifold has a slowly decaying width.

Assumption 2. For every μ the slice $\mathcal{S}_\mu = \{\mathbf{u}_{\mu,\nu} : \nu \in \Theta'\}$ satisfies $\sup_\mu d_n(\mathcal{S}_\mu) \leq C' n^{-\beta}$ with β much larger than α ; in the affine case where $\mathbf{f}(\mu, \nu)$ depends affinely on ν through q terms, the slice is contained in a q -dimensional subspace and its width vanishes for $n \geq q$.

A concrete instance clarifies the two assumptions. If the source is a superposition $\mathbf{f}(\mu, \nu) = \nu_1 \mathbf{g}_1(\mu) + \nu_2 \mathbf{g}_2(\mu)$ of two fields whose shapes depend on μ , then by linearity the solution is $\nu_1 \mathbf{u}_1(\mu) + \nu_2 \mathbf{u}_2(\mu)$, so for fixed μ the slice is spanned by two fields and Assumption 2 holds with vanishing width beyond $n = 2$; yet as μ varies the pair $\mathbf{u}_1(\mu), \mathbf{u}_2(\mu)$ rotates and translates through the state space, so no small fixed subspace captures the union over μ , and Assumption 1 holds with a slow rate. This is exactly the situation realized by the benchmark of section 7, and it is representative of transport and geometry-dominated problems in which a coherent feature is displaced by the parameter. Under Assumptions 1 and 2 a single global basis is inefficient on $\mathcal{S}$ but a μ -dependent basis can be efficient on every slice, which is the principle behind the deep orthogonal decomposition of Franco *et al.* [21] that we now extend.

3. Physics-Informed Deep Orthogonal Decomposition

The deep orthogonal decomposition builds a continuously adaptive basis inside an ambient space. A preliminary proper orthogonal decomposition of the training snapshots gives a $\mathbb{G}$ -orthonormal ambient matrix $\mathbb{A} \in \mathbb{R}^{N_h \times N_A}$ with $N_A \ll N_h$ that retains the manifold to a small tolerance. The adaptive basis is

$$\mathbb{V}_\mu = \mathbb{A} \tilde{\mathbb{V}}_\mu, \quad \tilde{\mathbb{V}}_\mu = \text{ORTH}(\text{net}(\mu)) \in \mathbb{R}^{N_A \times n}, \quad (4)$$

where net is a seed network followed by root networks that output the n raw columns and ORTH is a differentiable orthonormalization by reduced QR , which returns the factor with Euclidean-orthonormal columns, $\tilde{\mathbb{V}}_\mu^\top \tilde{\mathbb{V}}_\mu = \mathbb{I}_n$. The composed basis is then $\mathbb{G}$ -orthonormal, because $\mathbb{A}^\top \mathbb{G} \mathbb{A} = \mathbb{I}_{N_A}$ gives

$$\mathbb{V}_\mu^\top \mathbb{G} \mathbb{V}_\mu = \tilde{\mathbb{V}}_\mu^\top \mathbb{A}^\top \mathbb{G} \mathbb{A} \tilde{\mathbb{V}}_\mu = \tilde{\mathbb{V}}_\mu^\top (\mathbb{A}^\top \mathbb{G} \mathbb{A}) \tilde{\mathbb{V}}_\mu = \tilde{\mathbb{V}}_\mu^\top \tilde{\mathbb{V}}_\mu = \mathbb{I}_n, \quad (5)$$

so $\text{span}(\mathbb{V}_\mu)$ is a point of the Grassmann manifold and the reduced coefficients are measured in an orthonormal frame. The ambient space is a practical device rather than an approximation: it lowers

the output dimension from N_h to a moderate N_A so that the basis network is trainable, while retaining essentially all of the manifold, and it plays the same intermediate-compression role as in the deep-learning models of Fresca & Manzoni [14] and Lee & Carlberg [12]. The reduced coefficients are produced by a segregated network that separates the two parameter groups:

$$\phi(\boldsymbol{\mu}, \boldsymbol{\nu}) = \text{diag} [\phi_1(\boldsymbol{\mu})^\top \phi_2(\boldsymbol{\nu})] \in \mathbb{R}^n, \quad \phi_1 : \mathbb{R}^p \rightarrow \mathbb{R}^{m \times n}, \quad \phi_2 : \mathbb{R}^{p'} \rightarrow \mathbb{R}^{m \times n}, \quad (6)$$

with rank $m \geq q$, and the reduced-order field is $\mathbf{u}_{\boldsymbol{\mu}, \boldsymbol{\nu}}^{\text{ROM}} = \mathbb{A} \tilde{\mathbf{V}}_{\boldsymbol{\mu}} \phi(\boldsymbol{\mu}, \boldsymbol{\nu})$. The architecture reflects the two-scale structure of the problem. The seed network compresses the space-interacting parameter into a shared latent code, the root networks expand that code into the n raw basis vectors, and the orthonormalization projects the result onto the Grassmann manifold so that the output is always a valid subspace, which stabilizes training and makes the coefficients well conditioned. The segregated coefficient network keeps the two parameter groups apart, so that the benign parameters, which enter the solution through a small linear superposition, are represented by the low-rank factor ϕ_2 while the geometric parameters are carried by ϕ_1 ; this mirrors the separation of variables discussed below and is what allows the same trained model to be evaluated in batches over many benign parameters at once. The reason a rank of at least q is needed is made precise in Proposition 3.

The defining idea of the present work is to supervise this model with the governing equations where labels are unavailable. Let $\{(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i, \mathbf{u}_i)\}_{i=1}^{N_{\text{lab}}}$ be labelled snapshots and let $\{\boldsymbol{\mu}_j^c\}_{j=1}^{N_{\text{col}}}$ be collocation parameters that carry no high-fidelity solution. We train the composed model by minimizing:

$$\mathcal{L} = \mathcal{L}_{\text{data}} + \lambda_{\text{phys}} \mathcal{L}_{\text{res}}, \quad (7)$$

with a labelled reconstruction term and an unlabelled residual term:

$$\mathcal{L}_{\text{data}} = \frac{1}{N_{\text{lab}}} \sum_{i=1}^{N_{\text{lab}}} \frac{\|\mathbf{u}_i - \mathbb{A} \tilde{\mathbf{V}}_{\boldsymbol{\mu}_i} \phi(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)\|^2}{\|\mathbf{u}_i\|^2}, \quad (8)$$

$$\mathcal{L}_{\text{res}} = \frac{1}{N_{\text{col}}} \sum_{j=1}^{N_{\text{col}}} \mathbb{E}_{\boldsymbol{\nu} \sim \mathbf{Q}} \frac{\|\mathcal{R}(\boldsymbol{\mu}_j^c, \boldsymbol{\nu}, \mathbb{A} \tilde{\mathbf{V}}_{\boldsymbol{\mu}_j^c} \phi(\boldsymbol{\mu}_j^c, \boldsymbol{\nu}))\|^2}{\|\mathbf{f}(\boldsymbol{\mu}_j^c, \boldsymbol{\nu})\|^2}. \quad (9)$$

The residual term is evaluated through forward passes of the networks and a single application of the sparse discrete operator, so it costs a negligible fraction of a high-fidelity solve, and the collocation set may be taken as large and dense as desired. Two forms of the residual are available: a strong form obtained by automatic differentiation of the reconstructed field on the ambient grid, suited to operators that act cleanly on V_h , and a weak Galerkin form $\mathbb{A}(\boldsymbol{\mu}) \mathbf{u}^{\text{ROM}} - \mathbf{f}(\boldsymbol{\mu}, \boldsymbol{\nu})$ assembled with the finite-element machinery of the high-fidelity model, which is robust on complex geometries and exact in the affine case; our experiments use the weak form. The two forms differ in what they require: the strong form needs the differential operator to act on the ambient fields, which is immediate on a structured grid but delicate on unstructured meshes, whereas the weak form needs only the algebraic operator that the high-fidelity solver already assembles, so it inherits the stability of the underlying discretization and extends to any problem for which a residual can be formed. In both forms the cost is one operator application per collocation sample, which is why the collocation set can be made far larger than the labelled set at negligible expense, and it is this asymmetry between the cost of a residual and the cost of a solve that the method exploits.

To fit the predictive variance of Section 4 without corrupting the mean, we add a heteroscedastic term evaluated on detached residuals, following the analysis in [33]. Writing $\hat{\mathbf{u}}^{\text{ROM}} = \tilde{\mathbf{V}}_{\boldsymbol{\mu}} \phi$ for the ambient reconstruction and $\hat{\sigma}^2 = (\tilde{\mathbf{V}}_{\boldsymbol{\mu}}^{\odot 2}) \exp \ell$ for the propagated ambient variance produced by a

small network ℓ , this term is

$$\mathcal{L}_{\text{nl}} = \frac{1}{N_{\text{lab}}} \sum_i \frac{1}{2} \left[\frac{(\text{sg}[\hat{\mathbf{u}}_i^{\text{ROM}}] - \hat{\mathbf{u}}_i)^2}{\hat{\sigma}_i^2} + \log \hat{\sigma}_i^2 \right], \quad (10)$$

with sg the stop-gradient operator and $\hat{\mathbf{u}}_i = \mathbb{A}^\top \mathbb{G} \mathbf{u}_i$; it enters Equation 7 with a small weight. The training procedure is Algorithm 1, in which the physics weight is warmed up over the first iterations and, in the data-efficiency study, scaled inversely with the number of labelled points, so that the physical prior is strongest exactly when data are scarce.

Algorithm 1 Training of physics-informed deep orthogonal decomposition

- 1: **input:** labelled set, collocation set, ambient $\mathbb{A}$, latent dimension n , weight λ_{phys} .
 - 2: pre-compute ambient coordinates $\hat{\mathbf{u}}_i = \mathbb{A}^\top \mathbb{G} \mathbf{u}_i$ and cache the sparse operators and sources at the collocation parameters.
 - 3: **for** each optimization step **do**
 - 4: draw a mini-batch of labelled indices and, for each, a set of ν values;
 - 5: evaluate the adaptive basis (Equation (4)) and the coefficients (Equation (6));
 - 6: form the reconstruction term (Equation (8)) and the heteroscedastic term (Equation (10));
 - 7: draw a mini-batch of collocation indices and ν values, form the residual term (Equation (9));
 - 8: update all parameters on $\mathcal{L}_{\text{data}} + \lambda_{\text{phys}} \mathcal{L}_{\text{res}}$.
 - 9: **output:** trained adaptive basis and coefficient networks.
-

4. Uncertainty Quantification and the Trust Indicator

A surrogate that is to be used inside inference must report its own reliability. We model two sources of uncertainty and, crucially, exploit the linearity of the reduced-order decoder to transport them to the physical field without approximation. The aleatoric part, arising from measurement or discretization noise, is the heteroscedastic coefficient variance $\Sigma_c(\boldsymbol{\mu}, \boldsymbol{\nu}) = \text{diag}(\exp \ell)$ fitted by Equation (10). The epistemic part, which dominates away from the data, is captured by a deep ensemble in the sense of Definition 7 with members $\{(\tilde{\mathbf{V}}^{(k)}, \phi^{(k)})\}_{k=1}^M$ producing field means $\mathbf{u}^{(k)}$ and aleatoric field variances $a^{(k)}$. The predictive mean and variance are:

$$\bar{\mathbf{u}} = \frac{1}{M} \sum_k \mathbf{u}^{(k)}, \quad s^2 = \frac{1}{M} \sum_k a^{(k)} + \frac{1}{M} \sum_k (\mathbf{u}^{(k)} - \bar{\mathbf{u}})^2, \quad (11)$$

the two terms being the aleatoric and epistemic contributions. Because the map from coefficients to field, $\mathbf{c} \mapsto \mathbb{A} \tilde{\mathbf{V}}_\mu \mathbf{c}$, is linear and isometric, the field covariance is available in closed form:

$$\text{Cov}[\mathbf{u}^{\text{ROM}} \mid \boldsymbol{\mu}, \boldsymbol{\nu}] = (\mathbb{A} \tilde{\mathbf{V}}_\mu) \Sigma_c (\mathbb{A} \tilde{\mathbf{V}}_\mu)^\top, \quad (12)$$

and it is transported without inflation, a property established in Theorem 2 and contrasted there with the Lipschitz expansion incurred by a nonlinear decoder.

Both formulas follow from elementary conditioning, which we now carry out coordinate by coordinate. Let K be the uniform index of an ensemble member, and given $K = k$ let the coefficient vector have mean $\phi^{(k)}$ and diagonal covariance $\Sigma_c^{(k)} = \text{diag}(\exp \ell^{(k)})$, so that the member field mean is $\mathbf{u}^{(k)} = \mathbb{A} \tilde{\mathbf{V}}^{(k)} \phi^{(k)}$ and the member field variance at node i is $a_i^{(k)}$. The law of total variance applied to the i -th node gives

$$\text{Var}[u_i] = \underbrace{\mathbb{E}_K[\text{Var}(u_i \mid K)]}_{\text{aleatoric}} + \underbrace{\text{Var}_K[\mathbb{E}(u_i \mid K)]}_{\text{epistemic}} = \frac{1}{M} \sum_{k=1}^M a_i^{(k)} + \frac{1}{M} \sum_{k=1}^M (u_i^{(k)} - \bar{u}_i)^2, \quad (13)$$

because K is uniform on $\{1, \dots, M\}$, so the conditional expectation averages the member variances and the conditional variance is the empirical variance of the member means; stacking the nodes reproduces Equation (11). For a single member the aleatoric field covariance is obtained by pushing the diagonal coefficient covariance through the linear decoder with $\mathbf{u}^{\text{ROM}} = \mathbb{A} \tilde{\mathbb{V}}_{\mu} \mathbf{c}$, mean $\hat{\mathbf{c}}$, and $\text{Cov}[\mathbf{c}] = \Sigma_{\mathbf{c}}$:

$$\text{Cov}[\mathbf{u}^{\text{ROM}}] = \mathbb{A} \tilde{\mathbb{V}}_{\mu} \mathbb{E}[(\mathbf{c} - \hat{\mathbf{c}})(\mathbf{c} - \hat{\mathbf{c}})^{\top}] \tilde{\mathbb{V}}_{\mu}^{\top} \mathbb{A}^{\top} = (\mathbb{A} \tilde{\mathbb{V}}_{\mu}) \Sigma_{\mathbf{c}} (\mathbb{A} \tilde{\mathbb{V}}_{\mu})^{\top}, \quad (14)$$

which is Equation (12). Its i -th diagonal entry is the node-wise variance $a_i = \sum_{k=1}^n (\mathbb{A} \tilde{\mathbb{V}}_{\mu})_{ik}^2 \exp(\ell_k)$, a sum over modes of squared basis entries weighted by the coefficient variances; in ambient coordinates the same computation reads $\hat{\sigma}^2 = (\tilde{\mathbb{V}}_{\mu}^{\odot 2}) \exp \ell$, the propagated variance used in Equation (10). The scale of s is then corrected by a single factor through the conformal step, which is all that the isometry of Theorem 2 requires. To correct the absolute scale of the ensemble variance we apply the split-conformal procedure of Definition 6. On a held-out calibration set we form the standardized scores $z = |\mathbf{u} - \bar{\mathbf{u}}|/s$, take the empirical quantile $q_{1-\alpha}$ at level $1 - \alpha$, and report the interval:

$$\mathcal{C}_{1-\alpha}(\mu, \nu) = [\bar{\mathbf{u}} - q_{1-\alpha} s, \bar{\mathbf{u}} + q_{1-\alpha} s], \quad (15)$$

whose finite-sample coverage is guaranteed in Theorem 4. The calibration is summarized in Algorithm 2.

Algorithm 2 Split-conformal recalibration of the predictive intervals

- 1: **input:** ensemble mean $\bar{\mathbf{u}}$ and standard deviation s ; calibration set $\{(\mu_i, \nu_i, \mathbf{u}_i)\}_{i=1}^N$; level $1 - \alpha$.
 - 2: form the standardized scores $z_i = |\mathbf{u}_i - \bar{\mathbf{u}}|/s_i$ pooled over nodes;
 - 3: set $q_{1-\alpha}$ equal to the $\lceil (N+1)(1-\alpha) \rceil$ -th smallest score;
 - 4: **output:** at a new query return the interval $[\bar{\mathbf{u}} - q_{1-\alpha} s, \bar{\mathbf{u}} + q_{1-\alpha} s]$, with coverage at least $1 - \alpha$.
-

The trust indicator combines three quantities, each computable without a high-fidelity solve. The first is the relative residual of the reconstructed field, $\rho(\mu, \nu) = \|\mathbb{A}(\mu) \mathbf{u}^{\text{ROM}} - \mathbf{f}\| / \|\mathbf{f}\|$, which by Theorem 3 bounds the true error from both sides up to the conditioning of the operator. The second is the normalized predictive uncertainty $v(\mu, \nu) = \bar{s} / \|\bar{\mathbf{u}}\|$ obtained from Equation (11). The third is the geometric novelty of the queried subspace, the distance to the training bases in the metric of Definition 4:

$$\delta(\mu) = \min_i d(\mathbb{V}_{\mu}, \mathbb{V}_{\mu_i}), \quad (16)$$

which measures how far the model must extrapolate its basis. Standardizing each signal by its in-distribution statistics and summing gives the scalar trust indicator $\eta(\mu, \nu)$, which we use both to detect out-of-distribution queries and, by rejecting the queries with the largest η and returning them to the high-fidelity model, to trade accuracy against coverage with a certified fallback in the sense of Corollary 1.

Remark 1. The abstention rule of corollary 1 and the risk-coverage curve reported later are two readings of the same quantity. At deployment the rule decides which queries to trust, while during training the same indicator identifies the parameters at which a new high-fidelity solve is most informative, which turns the surrogate into an oracle for greedy data acquisition in the sense of [37]. This dual role makes the combined residual-and-uncertainty signal more than a diagnostic; it is an actionable control on the accuracy-cost trade-off of the whole pipeline.

5. Theoretical Analysis

Throughout this section $\|\cdot\|$ is the $\mathbb{G}$ -norm and all projections are $\mathbb{G}$ -orthogonal. We write $P_{\mathbb{A}} = \mathbb{A}\mathbb{A}^\top \mathbb{G}$ for the projection onto the ambient space and $P_{\mu} = \mathbb{V}_{\mu} \mathbb{V}_{\mu}^\top \mathbb{G}$ for the projection onto the adaptive subspace, and we use the identities $\mathbb{A}^\top \mathbb{G} \mathbb{A} = \mathbb{I}_{N_A}$, $\mathbb{V}_{\mu}^\top \mathbb{G} \mathbb{V}_{\mu} = \mathbb{I}_n$, and $\mathbb{V}_{\mu} = \mathbb{A} \tilde{\mathbb{V}}_{\mu}$. Define the ambient, adaptive-projection, coefficient, and total errors by $\mathcal{E}_{\text{amb}}^2 = \mathbb{E} \|\mathbf{u}_{\mu, \nu} - P_{\mathbb{A}} \mathbf{u}_{\mu, \nu}\|^2$, $\mathcal{E}_{\text{dod}}^2 = \mathbb{E} \|P_{\mathbb{A}} \mathbf{u}_{\mu, \nu} - P_{\mu} \mathbf{u}_{\mu, \nu}\|^2$, $\mathcal{E}_{\text{coef}}^2 = \mathbb{E} \|P_{\mu} \mathbf{u}_{\mu, \nu} - \mathbb{V}_{\mu} \phi(\boldsymbol{\mu}, \boldsymbol{\nu})\|^2$, and $\mathcal{E}_A^2 = \mathbb{E} \|\mathbf{u}_{\mu, \nu} - \mathbb{V}_{\mu} \phi(\boldsymbol{\mu}, \boldsymbol{\nu})\|^2$, with the expectation taken over $(\boldsymbol{\mu}, \boldsymbol{\nu})$.

Theorem 1. For every ambient space, every adaptive basis, and every coefficient map:

$$\mathcal{E}_A^2 = \mathcal{E}_{\text{amb}}^2 + \mathcal{E}_{\text{dod}}^2 + \mathcal{E}_{\text{coef}}^2. \quad (17)$$

Moreover $\mathcal{E}_{\text{coef}}^2 = \mathbb{E} \|\mathbf{c}_{\mu, \nu} - \phi(\boldsymbol{\mu}, \boldsymbol{\nu})\|_2^2$, where $\mathbf{c}_{\mu, \nu} = \mathbb{V}_{\mu}^\top \mathbb{G} \mathbf{u}_{\mu, \nu}$, and $\mathcal{E}_{\text{dod}}^2 \geq \mathbb{E} d_n(\mathcal{S}_{\mu})^2$.

Proof. Fix $(\boldsymbol{\mu}, \boldsymbol{\nu})$ and abbreviate $\mathbf{u} = \mathbf{u}_{\mu, \nu}$. We first record that $P_{\mathbb{A}}$ and P_{μ} are $\mathbb{G}$ -orthogonal projections. From $\mathbb{A}^\top \mathbb{G} \mathbb{A} = \mathbb{I}_{N_A}$:

$$P_{\mathbb{A}}^2 = \mathbb{A}(\mathbb{A}^\top \mathbb{G} \mathbb{A})\mathbb{A}^\top \mathbb{G} = \mathbb{A}\mathbb{A}^\top \mathbb{G} = P_{\mathbb{A}}, \quad \mathbb{G}P_{\mathbb{A}} = \mathbb{G}\mathbb{A}\mathbb{A}^\top \mathbb{G} = (\mathbb{G}\mathbb{A}\mathbb{A}^\top \mathbb{G})^\top = (\mathbb{G}P_{\mathbb{A}})^\top, \quad (18)$$

so $P_{\mathbb{A}}$ is idempotent and $\mathbb{G}$ -self-adjoint, meaning $\mathbb{G}P_{\mathbb{A}} = P_{\mathbb{A}}^\top \mathbb{G}$; the same two computations with $\mathbb{V}_{\mu}^\top \mathbb{G} \mathbb{V}_{\mu} = \mathbb{I}_n$ give $P_{\mu}^2 = P_{\mu}$ and $\mathbb{G}P_{\mu} = P_{\mu}^\top \mathbb{G}$. Because $\mathbb{V}_{\mu} = \mathbb{A} \tilde{\mathbb{V}}_{\mu}$ lies in the range of $\mathbb{A}$:

$$\mathbb{V}_{\mu}^\top \mathbb{G} P_{\mathbb{A}} = \tilde{\mathbb{V}}_{\mu}^\top \mathbb{A}^\top \mathbb{G} \mathbb{A} \mathbb{A}^\top \mathbb{G} = \tilde{\mathbb{V}}_{\mu}^\top (\mathbb{A}^\top \mathbb{G} \mathbb{A}) \mathbb{A}^\top \mathbb{G} = \tilde{\mathbb{V}}_{\mu}^\top \mathbb{A}^\top \mathbb{G} = \mathbb{V}_{\mu}^\top \mathbb{G}, \quad (19)$$

and left-multiplying by $\mathbb{V}_{\mu}$ yields the absorption identity $P_{\mu} P_{\mathbb{A}} = \mathbb{V}_{\mu} \mathbb{V}_{\mu}^\top \mathbb{G} = P_{\mu}$. Now decompose:

$$\mathbf{u} - \mathbb{V}_{\mu} \phi = e_1 + e_2 + e_3, \quad e_1 = \mathbf{u} - P_{\mathbb{A}} \mathbf{u}, \quad e_2 = P_{\mathbb{A}} \mathbf{u} - P_{\mu} \mathbf{u}, \quad e_3 = P_{\mu} \mathbf{u} - \mathbb{V}_{\mu} \phi, \quad (20)$$

and verify that the three cross terms in the energy inner product vanish. For e_1 against e_2 , use that $e_2 \in \text{range}(\mathbb{A})$ gives $e_2 = P_{\mathbb{A}} e_2$, and that $\mathbb{G}P_{\mathbb{A}} = P_{\mathbb{A}}^\top \mathbb{G}$, so:

$$\langle e_1, e_2 \rangle_{\mathbb{G}} = e_2^\top \mathbb{G} e_1 = (P_{\mathbb{A}} e_2)^\top \mathbb{G} e_1 = e_2^\top P_{\mathbb{A}}^\top \mathbb{G} e_1 = e_2^\top \mathbb{G} P_{\mathbb{A}} e_1 = e_2^\top \mathbb{G} (P_{\mathbb{A}} - P_{\mathbb{A}}^2) \mathbf{u} = 0, \quad (21)$$

and the identical computation with $e_3 = P_{\mathbb{A}} e_3$ gives $\langle e_1, e_3 \rangle_{\mathbb{G}} = 0$. For e_2 against e_3 , note $e_3 = \mathbb{V}_{\mu}(\mathbf{c} - \phi) \in \text{span}(\mathbb{V}_{\mu})$ with $\mathbf{c} := \mathbb{V}_{\mu}^\top \mathbb{G} \mathbf{u}$, while:

$$P_{\mu} e_2 = P_{\mu} (P_{\mathbb{A}} - P_{\mu}) \mathbf{u} = (P_{\mu} P_{\mathbb{A}} - P_{\mu}^2) \mathbf{u} = (P_{\mu} - P_{\mu}) \mathbf{u} = 0, \quad (22)$$

so e_2 is $\mathbb{G}$ -orthogonal to $\text{span}(\mathbb{V}_{\mu})$, which contains e_3 , and hence $\langle e_2, e_3 \rangle_{\mathbb{G}} = 0$. The Pythagorean identity then gives $\|\mathbf{u} - \mathbb{V}_{\mu} \phi\|^2 = \|e_1\|^2 + \|e_2\|^2 + \|e_3\|^2$; taking the expectation over $(\boldsymbol{\mu}, \boldsymbol{\nu})$ produces Equation (17). For the coefficient term, $P_{\mu} \mathbf{u} = \mathbb{V}_{\mu} \mathbb{V}_{\mu}^\top \mathbb{G} \mathbf{u} = \mathbb{V}_{\mu} \mathbf{c}_{\mu, \nu}$, whence $e_3 = \mathbb{V}_{\mu}(\mathbf{c}_{\mu, \nu} - \phi)$ and, by $\mathbb{G}$ -orthonormality:

$$\|e_3\|^2 = (\mathbf{c}_{\mu, \nu} - \phi)^\top \mathbb{V}_{\mu}^\top \mathbb{G} \mathbb{V}_{\mu} (\mathbf{c}_{\mu, \nu} - \phi) = \|\mathbf{c}_{\mu, \nu} - \phi\|_2^2, \quad (23)$$

which is the stated identity. Finally $e_2 = (P_{\mathbb{A}} - P_{\mu}) \mathbf{u}$ is the $\mathbb{G}$ -orthogonal residual of $P_{\mathbb{A}} \mathbf{u}$ onto the n -dimensional subspace $\text{span}(\mathbb{V}_{\mu})$; as $\boldsymbol{\nu}$ varies $P_{\mathbb{A}} \mathbf{u}$ traces the ambient projection of the slice $\mathcal{S}_{\mu}$, and replacing $\text{span}(\mathbb{V}_{\mu})$ by the best n -dimensional subspace can only lower the error, so $\mathbb{E} \|e_2\|^2 \geq \mathbb{E} d_n(\mathcal{S}_{\mu})^2$ by Definition 1. $\square$

Remark 2. The identity (17) is exact, not an inequality, because the three error contributions are mutually orthogonal in the energy inner product, and it is invariant under any orthogonal change of representative of the adaptive basis, since the projection P_μ depends only on the subspace and not on its matrix. In particular the decomposition holds for any differentiable orthonormalization used in Equation (4), and it separates the two quantities that a practitioner can influence, the ambient tolerance and the capacity of the two networks, from the one that is fixed by the problem, the slice width.

The decomposition (17) isolates the role of each component: the ambient error is set by the tolerance of the preliminary decomposition, the adaptive-projection error is bounded by the slice width and is what the basis network must learn, and the coefficient error is the statistical error of the coefficient network. The physics residual acts on the product $\mathbb{V}_\mu \phi$ and therefore constrains the last two terms at unlabelled parameters, which is the mechanism behind the data efficiency observed in Section 7.

Theorem 2. Fix (μ, ν) and let $\mathbf{c}$ be a random vector with mean $\hat{\mathbf{c}}$ and covariance $\Sigma_{\mathbf{c}}$, and set $\mathbf{u} = \mathbb{V}_\mu \mathbf{c}$. Then $\mathbf{u}$ has mean $\mathbb{V}_\mu \hat{\mathbf{c}}$ and covariance $\mathbb{V}_\mu \Sigma_{\mathbf{c}} \mathbb{V}_\mu^\top$; the total predictive variance measured in the $\mathbb{G}$ -norm equals the latent one, $\mathbb{E} \|\mathbf{u} - \mathbb{E}\mathbf{u}\|^2 = \text{tr}(\Sigma_{\mathbf{c}})$; and the decoder $\mathbf{c} \mapsto \mathbb{V}_\mu \mathbf{c}$ is a $\mathbb{G}$ -isometry, so its Lipschitz constant is one and no direction of the covariance is inflated. A nonlinear decoder Ψ with Lipschitz constant L_Ψ satisfies only $\mathbb{E} \|\Psi(\mathbf{c}) - \Psi(\hat{\mathbf{c}})\|^2 \leq L_\Psi^2 \text{tr}(\Sigma_{\mathbf{c}})$, which can exceed $\text{tr}(\Sigma_{\mathbf{c}})$ when $L_\Psi > 1$, and in general $\mathbb{E} \Psi(\mathbf{c}) \neq \Psi(\hat{\mathbf{c}})$.

Proof. Linearity of $\mathbf{c} \mapsto \mathbb{V}_\mu \mathbf{c}$ gives $\mathbb{E}\mathbf{u} = \mathbb{V}_\mu \mathbb{E}\mathbf{c} = \mathbb{V}_\mu \hat{\mathbf{c}}$ and, writing $\Delta = \mathbf{c} - \hat{\mathbf{c}}$:

$$\text{Cov}[\mathbf{u}] = \mathbb{E}[(\mathbf{u} - \mathbb{E}\mathbf{u})(\mathbf{u} - \mathbb{E}\mathbf{u})^\top] = \mathbb{V}_\mu \mathbb{E}[\Delta \Delta^\top] \mathbb{V}_\mu^\top = \mathbb{V}_\mu \Sigma_{\mathbf{c}} \mathbb{V}_\mu^\top. \quad (24)$$

For the total variance in the $\mathbb{G}$ -norm, expand the quadratic form and use the cyclic property of the trace together with $\mathbb{V}_\mu^\top \mathbb{G} \mathbb{V}_\mu = \mathbb{I}_n$:

$$\mathbb{E} \|\mathbf{u} - \mathbb{E}\mathbf{u}\|^2 = \mathbb{E}[\Delta^\top \mathbb{V}_\mu^\top \mathbb{G} \mathbb{V}_\mu \Delta] = \mathbb{E}[\Delta^\top \Delta] = \text{tr}(\mathbb{E}[\Delta \Delta^\top]) = \text{tr}(\Sigma_{\mathbf{c}}). \quad (25)$$

That the decoder is a $\mathbb{G}$ -isometry is the pointwise version of (25): for every $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbb{V}_\mu \mathbf{x}\|^2 = \mathbf{x}^\top \mathbb{V}_\mu^\top \mathbb{G} \mathbb{V}_\mu \mathbf{x} = \|\mathbf{x}\|_2^2$, so norms are preserved and the Lipschitz constant is exactly one. No principal variance is magnified either: if $\Sigma_{\mathbf{c}} \mathbf{w} = \sigma \mathbf{w}$ then $(\mathbb{V}_\mu \Sigma_{\mathbf{c}} \mathbb{V}_\mu^\top) \mathbb{G} (\mathbb{V}_\mu \mathbf{w}) = \mathbb{V}_\mu \Sigma_{\mathbf{c}} (\mathbb{V}_\mu^\top \mathbb{G} \mathbb{V}_\mu) \mathbf{w} = \sigma \mathbb{V}_\mu \mathbf{w}$, so the field covariance and $\Sigma_{\mathbf{c}}$ share the same nonzero spectrum and every latent principal variance is carried to the field unchanged. For a nonlinear decoder Ψ with Lipschitz constant L_Ψ in the $\mathbb{G}$ -norm, the pointwise bound $\|\Psi(\mathbf{c}) - \Psi(\hat{\mathbf{c}})\| \leq L_\Psi \|\mathbf{c} - \hat{\mathbf{c}}\|_2$ squared and averaged gives $\mathbb{E} \|\Psi(\mathbf{c}) - \Psi(\hat{\mathbf{c}})\|^2 \leq L_\Psi^2 \mathbb{E} \|\Delta\|_2^2 = L_\Psi^2 \text{tr}(\Sigma_{\mathbf{c}})$, which strictly exceeds $\text{tr}(\Sigma_{\mathbf{c}})$ once $L_\Psi > 1$; equality is attained by the linear map $\Psi(\mathbf{c}) = L_\Psi O \mathbf{c}$ with O orthogonal, showing the constant cannot be improved. The bias is a consequence of Jensen's inequality: if ψ is a strictly convex component of Ψ then $\mathbb{E} \psi(\mathbf{c}) > \psi(\mathbb{E}\mathbf{c})$ whenever $\text{Var} \mathbf{c} > 0$, so $\mathbb{E} \Psi(\mathbf{c}) \neq \Psi(\hat{\mathbf{c}})$ in general, whereas the linear decoder satisfies $\mathbb{E} \mathbb{V}_\mu \mathbf{c} = \mathbb{V}_\mu \hat{\mathbf{c}}$ with no bias. $\square$

Remark 3. Theorem 2 explains why a calibrated predictive distribution is natural for an adaptive linear decoder and only approximate for a nonlinear one. The isometric map transports the entire latent law to the field with neither bias nor variance inflation, so a single scalar recalibration through Equation (15) suffices, whereas a nonlinear decoder distorts the shape of the distribution and a global rescaling cannot repair it. This is the precise sense in which keeping the decoder linear, while making the basis adaptive, is advantageous for uncertainty quantification.

Theorem 3. Let $\mathbf{u}_{\mu,\nu}$ solve $\mathbb{A}(\boldsymbol{\mu})\mathbf{u} = \mathbf{f}(\boldsymbol{\mu}, \boldsymbol{\nu})$ with $\mathbb{A}(\boldsymbol{\mu})$ invertible, and let $\mathbf{u}^{\text{ROM}}$ be any approximation with residual $\mathbf{r} = \mathbb{A}(\boldsymbol{\mu})\mathbf{u}^{\text{ROM}} - \mathbf{f}(\boldsymbol{\mu}, \boldsymbol{\nu})$. Then:

$$\frac{\|\mathbf{r}\|}{\|\mathbb{A}(\boldsymbol{\mu})\|} \leq \|\mathbf{u}_{\mu,\nu} - \mathbf{u}^{\text{ROM}}\| \leq \|\mathbb{A}(\boldsymbol{\mu})^{-1}\| \|\mathbf{r}\|, \quad (26)$$

and with the condition number $\kappa(\boldsymbol{\mu}) = \|\mathbb{A}(\boldsymbol{\mu})\| \|\mathbb{A}(\boldsymbol{\mu})^{-1}\|$ the relative error and the relative residual $\rho = \|\mathbf{r}\| / \|\mathbf{f}\|$ satisfy $\kappa(\boldsymbol{\mu})^{-1} \rho \leq \|\mathbf{u} - \mathbf{u}^{\text{ROM}}\| / \|\mathbf{u}\| \leq \kappa(\boldsymbol{\mu}) \rho$.

Proof. Write $A = \mathbb{A}(\boldsymbol{\mu})$ for brevity. Subtracting the exact equation $A\mathbf{u} = \mathbf{f}$ from $A\mathbf{u}^{\text{ROM}} = \mathbf{f} + \mathbf{r}$ gives $A(\mathbf{u}^{\text{ROM}} - \mathbf{u}) = \mathbf{r}$, that is $\mathbf{u} - \mathbf{u}^{\text{ROM}} = -A^{-1}\mathbf{r}$. Submultiplicativity of the induced norm gives the upper bound:

$$\|\mathbf{u} - \mathbf{u}^{\text{ROM}}\| = \|A^{-1}\mathbf{r}\| \leq \|A^{-1}\| \|\mathbf{r}\|, \quad (27)$$

and applying A to the same identity gives $\|\mathbf{r}\| = \|A(\mathbf{u} - \mathbf{u}^{\text{ROM}})\| \leq \|A\| \|\mathbf{u} - \mathbf{u}^{\text{ROM}}\|$, that is the lower bound $\|\mathbf{u} - \mathbf{u}^{\text{ROM}}\| \geq \|\mathbf{r}\| / \|A\|$; together these are (26). For the relative form, divide (26) by $\|\mathbf{u}\|$ and bound $\|\mathbf{f}\| = \|A\mathbf{u}\|$ on both sides. On one side $\|A\mathbf{u}\| \leq \|A\| \|\mathbf{u}\|$, and on the other $\|\mathbf{u}\| = \|A^{-1}A\mathbf{u}\| \leq \|A^{-1}\| \|A\mathbf{u}\|$, so:

$$\frac{\|\mathbf{u}\|}{\|A^{-1}\|} \leq \|\mathbf{f}\| \leq \|A\| \|\mathbf{u}\|. \quad (28)$$

Using the upper bound of Equation (26) with the right inequality of Equation (28):

$$\frac{\|\mathbf{u} - \mathbf{u}^{\text{ROM}}\|}{\|\mathbf{u}\|} \leq \frac{\|A^{-1}\| \|\mathbf{r}\|}{\|\mathbf{u}\|} \leq \|A\| \|A^{-1}\| \frac{\|\mathbf{r}\|}{\|\mathbf{f}\|} = \kappa(\boldsymbol{\mu}) \rho,$$

and using the lower bound of Equation (26) with the left inequality of Equation (28):

$$\frac{\|\mathbf{u} - \mathbf{u}^{\text{ROM}}\|}{\|\mathbf{u}\|} \geq \frac{\|\mathbf{r}\|}{\|A\| \|\mathbf{u}\|} \geq \frac{1}{\|A\| \|A^{-1}\|} \frac{\|\mathbf{r}\|}{\|\mathbf{f}\|} = \kappa(\boldsymbol{\mu})^{-1} \rho,$$

which is the relative two-sided statement. $\square$

Theorem 3 has two consequences. Minimizing the physics residual drives down the true error even where no label is present, up to the conditioning of the operator, which is the theoretical basis of the data-efficiency gains, and the computable relative residual is a certified error surrogate, which is what licenses its use in the trust indicator.

Corollary 1. Under the hypotheses of Theorem 3, for every query $\|\mathbf{u} - \mathbf{u}^{\text{ROM}}\| / \|\mathbf{u}\| \leq \kappa(\boldsymbol{\mu}) \rho(\boldsymbol{\mu}, \boldsymbol{\nu})$; hence, for a target tolerance τ , the acceptance rule $\rho(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq \tau / \bar{\kappa}$ with $\bar{\kappa} = \sup_{\boldsymbol{\mu}} \kappa(\boldsymbol{\mu})$ guarantees a relative error of at most τ on every accepted query, the deferred queries being returned to the high-fidelity model.

Proof. The pointwise bound is the right inequality of Theorem 3. If a query satisfies $\rho(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq \tau / \bar{\kappa}$, then, since $\kappa(\boldsymbol{\mu}) \leq \bar{\kappa}$:

$$\frac{\|\mathbf{u} - \mathbf{u}^{\text{ROM}}\|}{\|\mathbf{u}\|} \leq \kappa(\boldsymbol{\mu}) \rho(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq \bar{\kappa} \cdot \frac{\tau}{\bar{\kappa}} = \tau,$$

so every accepted query meets the tolerance; queries with $\rho(\boldsymbol{\mu}, \boldsymbol{\nu}) > \tau / \bar{\kappa}$ are deferred to the high-fidelity model by construction. $\square$

Proposition 1. Order the test queries by increasing relative residual, $\rho_{(1)} \leq \dots \leq \rho_{(N)}$, and for a retained fraction $c \in (0, 1]$ accept the $\lceil cN \rceil$ queries of smallest residual. Then the certified worst-case relative error over the accepted set, $R(c) = \bar{\kappa} \rho_{(\lceil cN \rceil)}$, is nondecreasing in c ; hence abstaining on the queries of largest residual can only decrease the certified error of the retained set, which is the monotone risk-coverage trade-off.

Proof. By corollary 1 every accepted query obeys $\|\mathbf{u} - \mathbf{u}^{\text{ROM}}\| / \|\mathbf{u}\| \leq \bar{\kappa} \rho_{(\lceil cN \rceil)} = R(c)$, so $R(c)$ certifies the accepted set. Since the order statistics are nondecreasing and $c \mapsto \lceil cN \rceil$ is nondecreasing, $R(c)$ is nondecreasing in c ; letting c decrease removes the largest-residual queries first and lowers the certified error, which is exactly the behaviour of the risk-coverage curve reported in Section 7. $\square$

Theorem 4. Let $z_1, \dots, z_N$ be the calibration scores and z_{N+1} the score of a new query, and assume the $N + 1$ scores are exchangeable. With $q = z_{(\lceil (N+1)(1-\alpha) \rceil)}$ the interval (15) satisfies $1 - \alpha \leq \mathbb{P}(z_{N+1} \leq q) \leq 1 - \alpha + 1/(N + 1)$, the upper bound holding when the scores are almost surely distinct.

Proof. Let $R \in \{1, \dots, N + 1\}$ be the rank of z_{N+1} among the pooled scores $z_1, \dots, z_{N+1}$, that is $R = \sum_{i=1}^{N+1} \mathbf{1}\{z_i \leq z_{N+1}\}$. Exchangeability of the $N + 1$ scores makes every ordering equally likely, so R is uniform on $\{1, \dots, N + 1\}$ and $\mathbb{P}(R \leq k) = k/(N + 1)$ for each integer k . Set $k = \lceil (N + 1)(1 - \alpha) \rceil$, so that $q = z_{(k)}$ is the k -th smallest calibration score. If $R \leq k$ then z_{N+1} does not exceed the k -th smallest of all $N + 1$ scores, hence $z_{N+1} \leq z_{(k)} = q$; therefore $\{R \leq k\} \subseteq \{z_{N+1} \leq q\}$ and:

$$\mathbb{P}(z_{N+1} \leq q) \geq \mathbb{P}(R \leq k) = \frac{k}{N + 1} \geq \frac{(N + 1)(1 - \alpha)}{N + 1} = 1 - \alpha, \quad (29)$$

where the last inequality uses $k \geq (N + 1)(1 - \alpha)$ from the ceiling. When the scores are almost surely distinct there are no ties, the two events coincide, and $\mathbb{P}(z_{N+1} \leq q) = k/(N + 1)$. The ceiling also gives the strict bound $k < (N + 1)(1 - \alpha) + 1$, so:

$$\mathbb{P}(z_{N+1} \leq q) = \frac{k}{N + 1} < \frac{(N + 1)(1 - \alpha) + 1}{N + 1} = 1 - \alpha + \frac{1}{N + 1}, \quad (30)$$

which is the stated upper bound. $\square$

Proposition 2. Let $X_1, \dots, X_{N_r} \in [0, 1]$ be independent copies of a score X with mean θ , and let $\hat{\theta}$ be their average. Then $\mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) \leq 2e^{-2N_r\varepsilon^2}$, so $N_r \geq \frac{1}{2\varepsilon^2} \log \frac{2}{\delta}$ samples give $|\hat{\theta} - \theta| \leq \varepsilon$ with probability at least $1 - \delta$, and on that event $|\sqrt{\hat{\theta}} - \sqrt{\theta}| \leq \sqrt{\varepsilon}$.

Proof. Write $Y_i = X_i - \theta$, so $\mathbb{E}Y_i = 0$ and $Y_i \in [-\theta, 1 - \theta] \subset [-1, 1]$ with range one. By Hoeffding's lemma each centred, bounded variable has a sub-Gaussian moment generating function, $\mathbb{E}e^{sY_i} \leq e^{s^2/8}$ for all $s \in \mathbb{R}$, since the interval has length one. For the upper tail, Markov's inequality applied to $e^{s \sum_i Y_i}$ with $s > 0$ gives, using independence:

$$\mathbb{P}(\hat{\theta} - \theta > \varepsilon) = \mathbb{P}\left(\sum_i Y_i > N_r \varepsilon\right) \leq e^{-sN_r\varepsilon} \prod_{i=1}^{N_r} \mathbb{E}e^{sY_i} \leq e^{-sN_r\varepsilon + N_r s^2/8}. \quad (31)$$

Minimizing the exponent over s at $s = 4\varepsilon$ gives $\mathbb{P}(\hat{\theta} - \theta > \varepsilon) \leq e^{-2N_r\varepsilon^2}$; the lower tail is identical with $Y_i \rightarrow -Y_i$, and a union bound yields $\mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) \leq 2e^{-2N_r\varepsilon^2}$. Requiring the right side to be at most δ gives $2e^{-2N_r\varepsilon^2} \leq \delta$, that is $N_r \geq \frac{1}{2\varepsilon^2} \log \frac{2}{\delta}$, and on the complementary event $|\hat{\theta} - \theta| \leq \varepsilon$ of probability at least $1 - \delta$ the concavity of $t \mapsto \sqrt{t}$ makes it Hölder with constant one on $[0, 1]$, $|\sqrt{a} - \sqrt{b}| \leq \sqrt{|a - b|}$, so $|\sqrt{\hat{\theta}} - \sqrt{\theta}| \leq \sqrt{|\hat{\theta} - \theta|} \leq \sqrt{\varepsilon}$. $\square$

Proposition 3. Suppose the exact coefficients are affine in ν with two active terms, $c_i(\mu, \nu) = \nu_1 a_i(\mu) + \nu_2 b_i(\mu)$, where (a_i, b_i) is not collinear as a function of μ . Then no rank-one map $c_i = \phi_{1,i}(\mu)\phi_{2,i}(\nu)$ represents c_i on a set of positive measure, whereas a segregated map (6) of rank $m \geq 2$ does.

Proof. Suppose a rank-one representation $\nu_1 a_i(\mu) + \nu_2 b_i(\mu) = \phi_{1,i}(\mu) \phi_{2,i}(\nu)$ held on an open set of ν and a positive-measure set of μ . Differentiating in ν_1 and in ν_2 at fixed μ gives:

$$a_i(\mu) = \phi_{1,i}(\mu) \partial_{\nu_1} \phi_{2,i}(\nu), \quad b_i(\mu) = \phi_{1,i}(\mu) \partial_{\nu_2} \phi_{2,i}(\nu), \quad (32)$$

since the left side is affine in ν . On the set where $\phi_{1,i}(\mu) \neq 0$ and $\partial_{\nu_2} \phi_{2,i} \neq 0$ the ratio eliminates $\phi_{1,i}(\mu)$,

$$\frac{a_i(\mu)}{b_i(\mu)} = \frac{\partial_{\nu_1} \phi_{2,i}(\nu)}{\partial_{\nu_2} \phi_{2,i}(\nu)}, \quad (33)$$

whose right side is independent of μ ; hence $(a_i(\mu), b_i(\mu))$ points along a single fixed direction for almost every μ , contradicting the assumed non-collinearity of (a_i, b_i) (the residual sets where $\phi_{1,i}$ or $\partial_{\nu_2} \phi_{2,i}$ vanishes are lower-dimensional and do not affect the positive-measure conclusion). Conversely, the rank-two segregated choice $\phi_{1,i}(\mu) = (a_i(\mu), b_i(\mu))$ and $\phi_{2,i}(\nu) = (\nu_1, \nu_2)$ reproduces the coefficients exactly:

$$\sum_{k=1}^2 \phi_{1,ki}(\mu) \phi_{2,ki}(\nu) = \nu_1 a_i(\mu) + \nu_2 b_i(\mu) = c_i(\mu, \nu), \quad (34)$$

so a rank of $m \geq 2$ suffices and $m \geq q$ is the general requirement. $\square$

6. Comparison with Existing Approaches

It is useful to locate the proposed method among established reduced-order models along the axes that matter for the applications of Section 1; i.e., whether the basis is global or adaptive, whether the decoder is linear, whether the model is intrusive, whether it is data efficient, whether it returns calibrated uncertainty, and whether it carries a certified error estimate. The reduced-basis and proper-orthogonal Galerkin models of Quarteroni *et al.* [2] and Kunisch & Volkwein [7] use a global linear basis and are intrusive and certified, but they inherit the slow width decay of Assumption 1 and need hyper-reduction for nonlinear terms. The non-intrusive proper-orthogonal neural network of Hesthaven & Ubbiali [19] and the deep-learning model of Fresca & Manzoni [14] keep the basis global while learning the coefficients or a nonlinear decoder. The former is bounded below by the projection error of the global basis. The latter escapes that bound at the price of a nonlinear decoder whose uncertainty propagation is biased and expansive by Theorem 2. The basis-interpolation method of Amsallem & Farhat [15] and the local bases of Amsallem *et al.* [16] make the subspace vary but change it discretely and require a selection rule, whereas the deep orthogonal decomposition of Franco *et al.* [21] makes it vary continuously. None of these methods supplies a calibrated predictive distribution or a solver-free error estimate. The present method keeps the decoder linear and isometric. So, projection is certified by Theorem 1 and uncertainty propagates exactly by Theorem 2, while making the basis continuously adaptive. Thus, the width barrier is broken. It adds the equation residual as a soft penalty rather than an intrusive projection. So, no online assembly is required. Table 2 summarizes the comparison.

The comparison also clarifies a design trade-off. The intrusive reduced-basis models from [37] and Rozza *et al.* [4] deliver rigorous output bounds but require access to the discrete operators and a hyper-reduction of the nonlinear terms. This ties the surrogate to the solver and complicates deployment. The non-intrusive models of Hesthaven & Ubbiali [19] and Guo & Hesthaven [20] avoid this by learning the coefficient map. However, they discard the equation entirely and therefore cannot certify a single

prediction. The present method occupies a middle ground: it is non-intrusive at inference, since a query is a forward pass, yet it uses the operator during training and at the certification stage through the residual, which recovers a certified error estimate without paying the online cost of an intrusive projection.

Among the tools for uncertainty, the deep ensemble of Lakshminarayanan *et al.* [28] that we adopt is the most robust but also the most expensive, since it multiplies the training cost by the number of members. The Monte-Carlo dropout from [29] and the Laplace approximation of Daxberger *et al.* [30] are cheaper single-model alternatives that our framework accommodates without change, because the exact propagation of Theorem 2 holds for any coefficient-space covariance regardless of how it is obtained. The conformal calibration of Vovk *et al.* [31] sits on top of any of these and supplies the finite-sample guarantee of Theorem 4. Thus, the choice of epistemic estimator trades cost against sharpness while leaving the coverage guarantee intact.

Table 2

Comparison of reduced-order models along the axes relevant to many-query and inference applications. A dash denotes a property that the method does not provide

Method	Basis	Linear dec.	Data eff.	Calib. UQ	Cert. error	Reference
POD-Galerkin	global	yes	no	no	yes	[2]
POD neural network	global	yes	no	no	no	[19]
POD-DL-ROM	global	no	no	no	no	[14]
Autoencoder ROM	nonlinear	no	no	no	no	[12]
Basis interpolation	local	yes	no	no	no	[15]
Local reduced bases	local	yes	no	no	no	[16]
Deep orthogonal dec.	adaptive	yes	no	no	no	[21]
Present method	adaptive	yes	yes	yes	yes	this work

7. Numerical Experiments

We validate the method on a parametrized advection diffusion problem engineered to realize the regime of Assumptions 1 and 2. The equation $-\varepsilon \Delta u + \mathbf{b}(\boldsymbol{\mu}) \cdot \nabla u = g(\cdot; \boldsymbol{\mu}, \boldsymbol{\nu})$ is posed on the unit square with homogeneous Dirichlet data and discretized by a five-point Laplacian and first-order upwinding on a 48×48 interior grid, giving $N_h = 2304$ degrees of freedom. The space-interacting parameters $\boldsymbol{\mu} = (\theta, x_c, y_c)$ rotate the advection field and move a Gaussian source, while the benign parameters $\boldsymbol{\nu} = (\alpha_1, \alpha_2)$ are two source amplitudes that enter linearly. So, each slice is at most two-dimensional while the global manifold decays slowly. We draw 900 training, 300 test, and 1500 collocation parameters and report the mean relative L^2 error on the test set. The high-fidelity model, the data generation, and the network architectures are detailed in Appendix 8 and Appendix 8. The benchmark is chosen deliberately. Advection diffusion with a rotating field and a moving source produces a sharp, coherent feature whose position and orientation are controlled by the space-interacting parameters, which is the canonical mechanism that makes the global width decay slowly, while the linear dependence on the amplitudes keeps each slice low-dimensional, so both assumptions of section 2 hold by construction and can be verified rather than assumed. The two-dimensional affine setting further allows every quantity in the theory, the projections of Section 1, the residual of Theorem 3, and the coverage of Theorem 4, to be computed exactly. So, that the numerical study tests the method and

not the accuracy of an auxiliary approximation. Before turning to the quantitative study we collect in Figure 2 a set of three-dimensional views of the benchmark and of the objects the surrogate learns, computed from the same sparse solver used throughout. The surfaces make visible the coherent front that the space-interacting parameters displace, the way the adaptive modes move with the parameter, and the spatial structure of the reconstruction error and of the propagated uncertainty, and they are referenced again at the points where the corresponding quantity is measured.

The target structure is confirmed in Figure 3 and, in three dimensions, in panels 10 and 15 of Figure 2. In Figure 3, panels one and two show high-fidelity solution fields at two space-interacting parameters, in which a coherent front rotates and translates; panel three shows the two-Gaussian source. Panels four and five show the first two adaptive-basis fields at one parameter and panel six the leading basis field at a second parameter, which is displaced with the parameter. Panels seven, eight, and nine show the rank-four reduced reconstruction, its pointwise absolute error, and the predicted standard deviation obtained in closed form by Equation (12); the last two concentrate on the moving front. Panel 10 shows the slice projection error against latent size and orientation, with the slowly decaying global width and the rapidly decaying slice width overlaid. Panel 11 shows the mean relative error over the labelled-set size and the physics weight. Panel 12 and Panel 13 show the trust indicator and the Grassmann novelty distance over the parameter plane, both rising beyond the training band marked by the dashed edge. Panel 14 shows the relative residual against the true error clustering about the identity plane, separated into in-distribution and extrapolation queries. Panel 15 shows the proper-orthogonal energy spectrum together with its cumulative share. The global manifold has an algebraically slow width with fitted exponent close to one, so a global linear basis needs of the order of a hundred modes for a few-percent error, whereas each slice collapses to the level of round-off at $n = 2$, which is the affine dimension of the source. This is exactly the situation in which a continuously adaptive basis is expected to help.

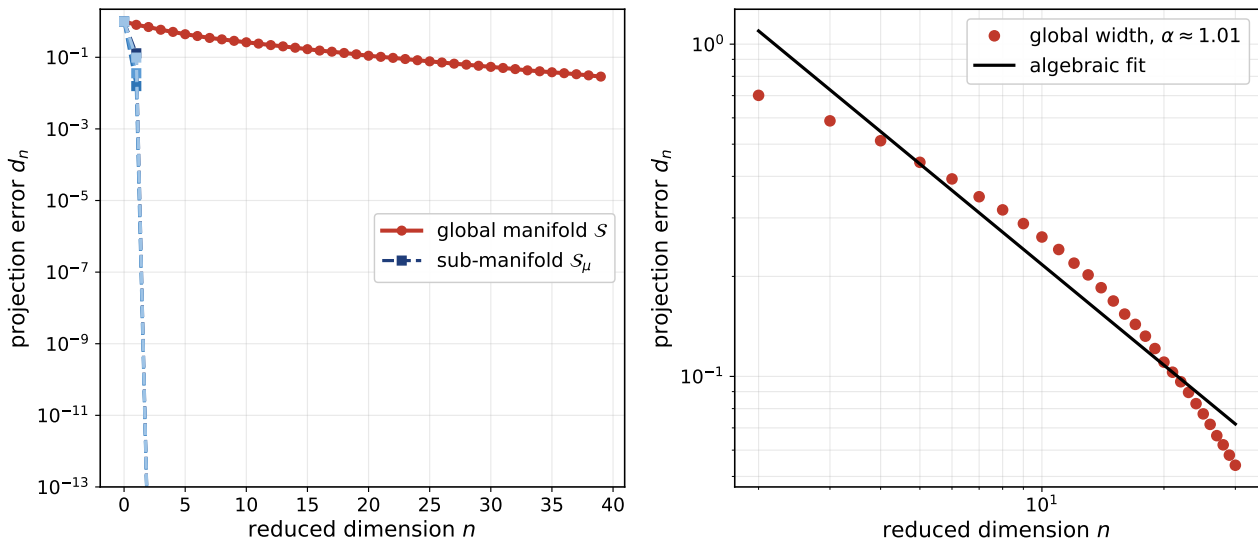


Fig. 3. Kolmogorov analysis of the benchmark; left, the projection error of the global manifold decays slowly while four representative slices collapse at $n = 2$; right, the global width follows an algebraic law with exponent close to one

The projection error of the adaptive basis is compared with that of the global proper orthogonal decomposition in Figure 4, with the values listed in Table 3. At every latent dimension the adaptive basis is far more accurate: it reaches 7.8% at $n = 2$ against 70.8% for the linear basis, and 5.3% at $n = 4$ against 52.1%, so a global basis needs about forty modes to match what the adaptive basis

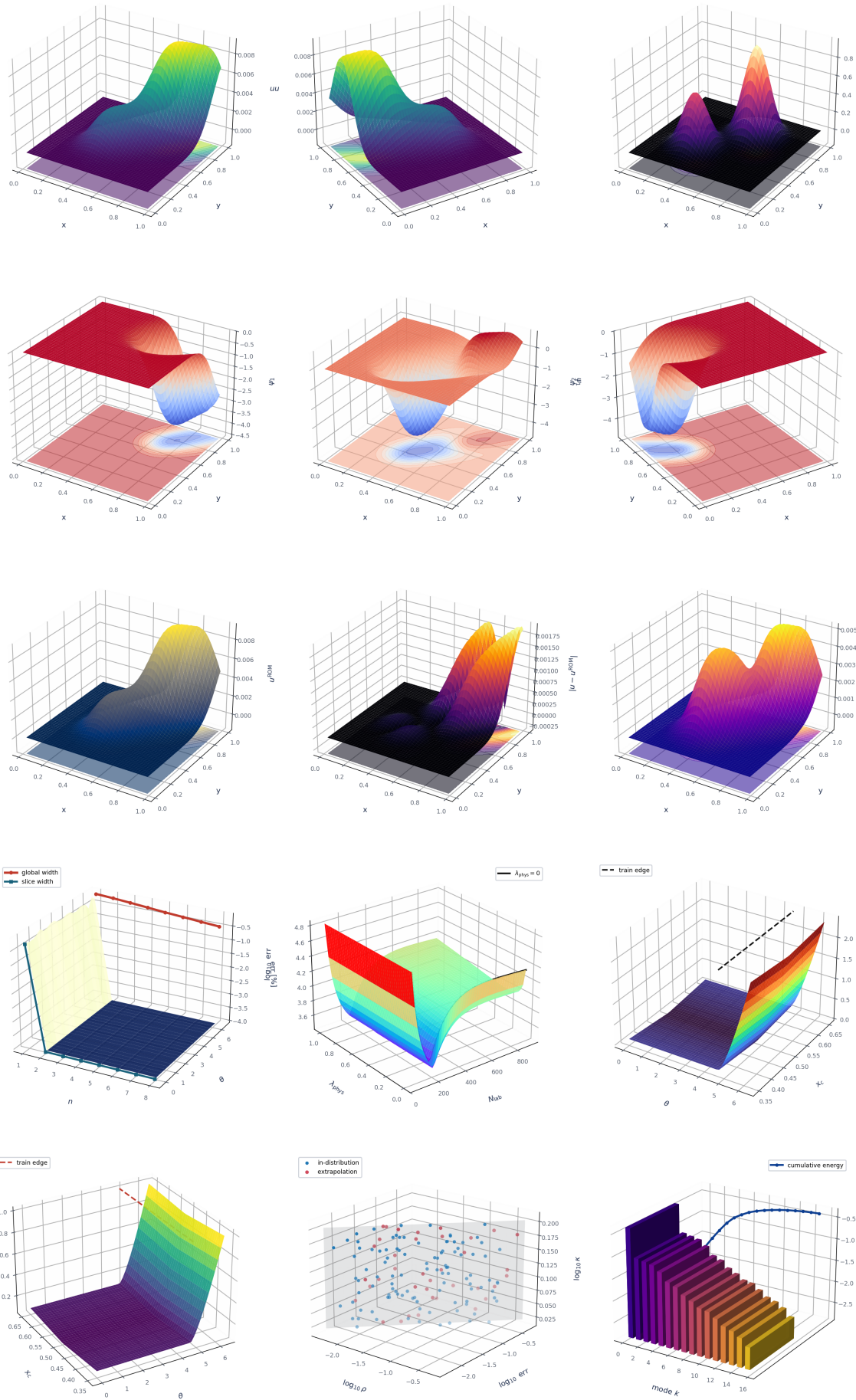


Fig. 2. Three-dimensional views of the benchmark and of the learned objects, in reading order (rows top to bottom, left to right)

achieves with four. The adaptivity score, the normalized variance of the basis over the parameter space in the metric of Definition 4, stays near 0.94, which confirms that the learned basis tracks the moving slice rather than collapsing to a constant one.

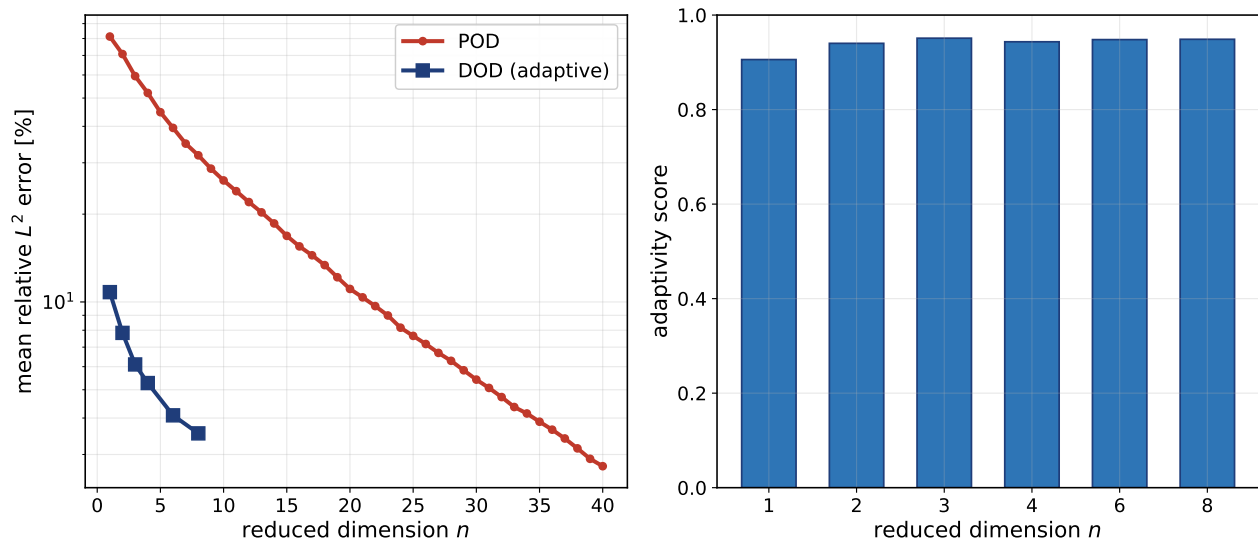


Fig. 4. Projection error and adaptivity score; left, mean relative projection error of the adaptive basis against the global proper orthogonal decomposition across latent dimensions; right, the adaptivity score of the learned basis

Table 3
Projection error and adaptivity score against latent dimension n

n	1	2	3	4	6	8
POD error [%]	81.3	70.8	59.5	52.1	39.5	31.8
DOD error [%]	10.8	7.83	6.11	5.27	4.09	3.54
Adaptivity	0.91	0.94	0.95	0.94	0.95	0.95

The adaptive basis is visualized in Figure 5, where for three values of the space-interacting parameter the first three basis fields rotate and translate to follow the advected plume, which is the mechanism by which four latent coordinates suffice. Panels four to six of Figure 2 give three-dimensional views of the same fields and of their displacement with the parameter.

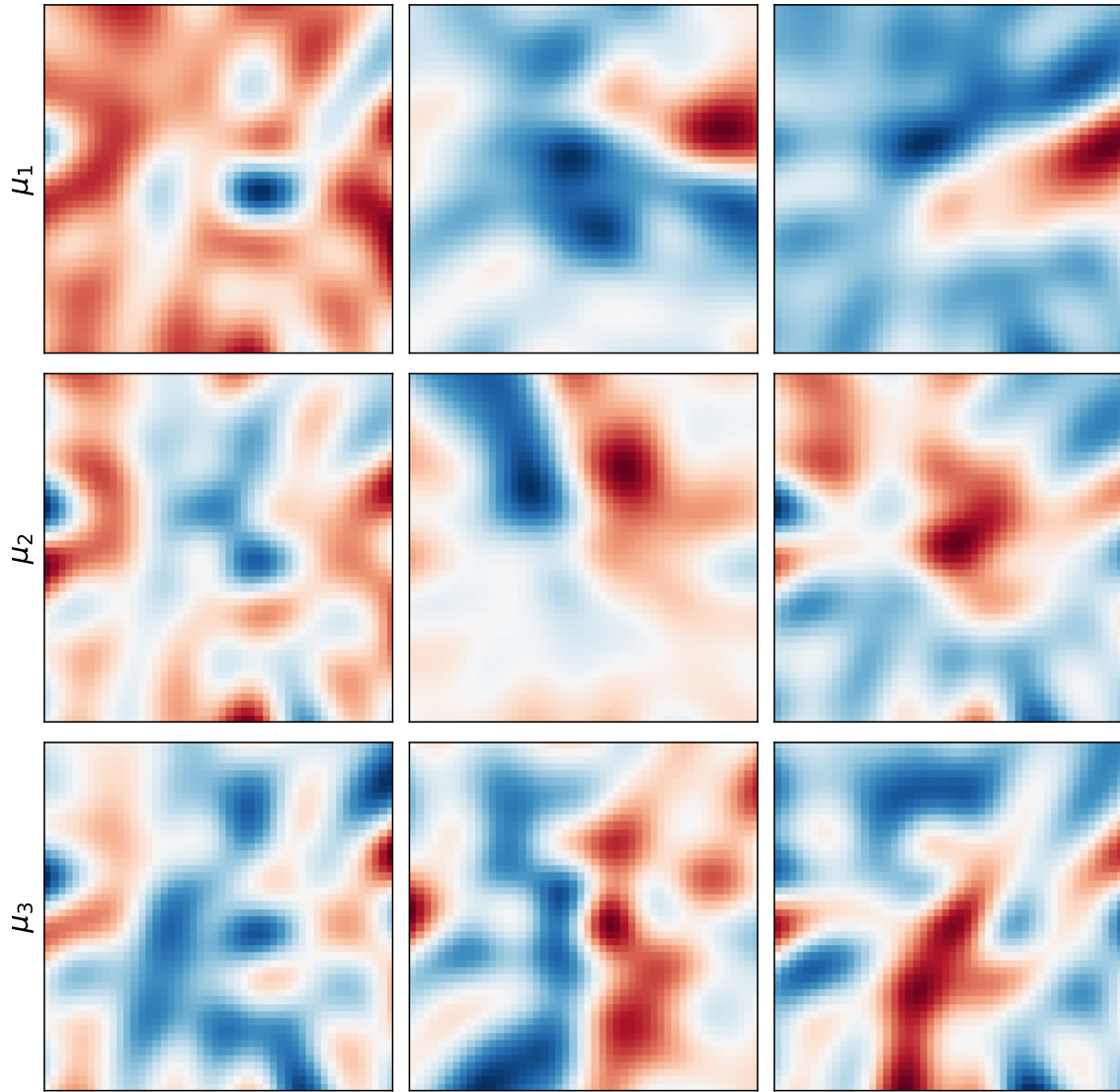


Fig. 5. First three adaptive-basis fields (columns) for three values of the space-interacting parameter (rows)

The central benefit of the physics term is data efficiency, shown in Figure 6 and summarized in Table 5. With a physics weight scaled inversely with the number of labelled snapshots, the physics-informed model is substantially more accurate than the purely data-driven one in the scarce-data regime that dominates the offline cost: with only 60 labelled solves it lowers the error from 34.0% to 20.9%, a relative improvement of about 39%, and as labelled data become abundant the two models converge, so the physical prior helps most when it is most needed and does no harm otherwise. The dashed line marks the lower bound set by the global basis at $n = 4$, which both adaptive models pass by a wide margin.

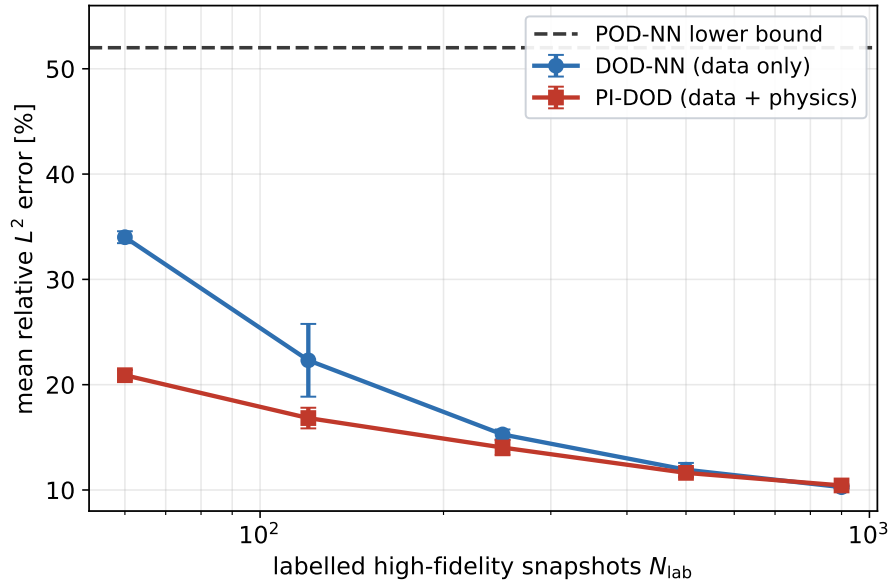


Fig. 6. Mean relative error against the number of labelled high-fidelity snapshots, averaged over two seeds with one standard deviation shown

Generalization beyond the training distribution is probed in Figure 7 by training on the angular band $\theta \in [0, 3\pi/2]$ and testing on the held-out band $\theta \in [3\pi/2, 2\pi]$, with the collocation set covering the full range. The residual, available on the unlabelled extrapolation region, curbs the error growth of the physics-informed model, which attains 26.3% over the extrapolation band against 58.4% for the data-driven model, a reduction by more than a factor of two, so the physical prior supplies exactly the inductive bias that a purely data-driven model lacks outside its training range.

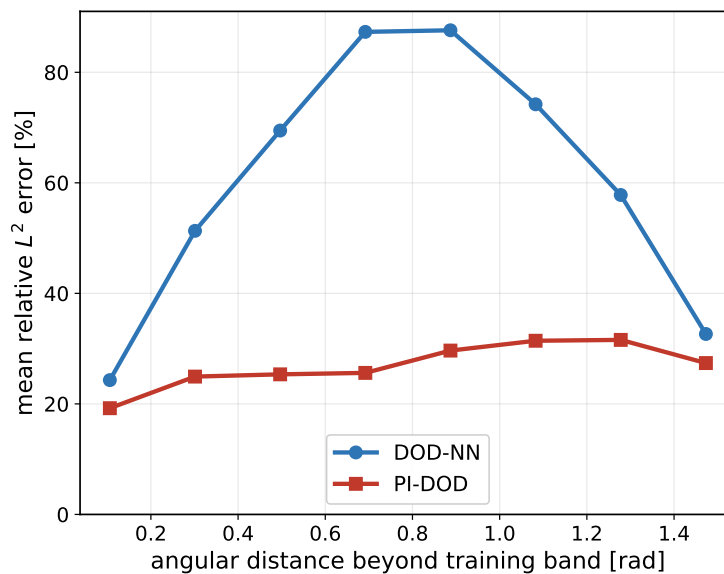


Fig. 7. Test error against angular distance beyond the training band

The sensitivity to the physics weight at a fixed labelled-set size is shown in Figure 8. A broad range of weights improves on the data-only model, and only very large weights degrade the fit.

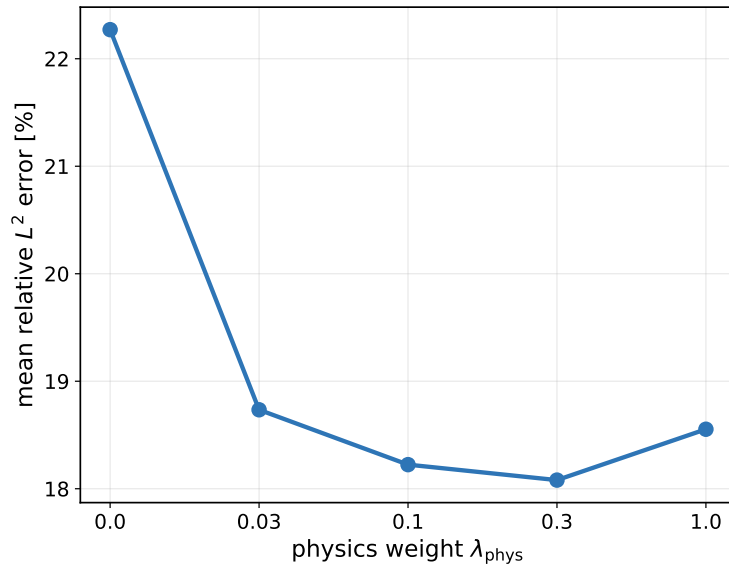


Fig. 8. Sensitivity to the physics weight at a fixed labelled-set size of 120

The uncertainty layer is assessed on a deep ensemble of five physics-informed models. The reliability diagram in Figure 9 shows that the conformally calibrated intervals track the diagonal closely, attaining an empirical coverage of 0.91 at the nominal level, whereas the raw ensemble is over-confident, and the scatter of predicted uncertainty against true error has a Pearson correlation of 0.75.

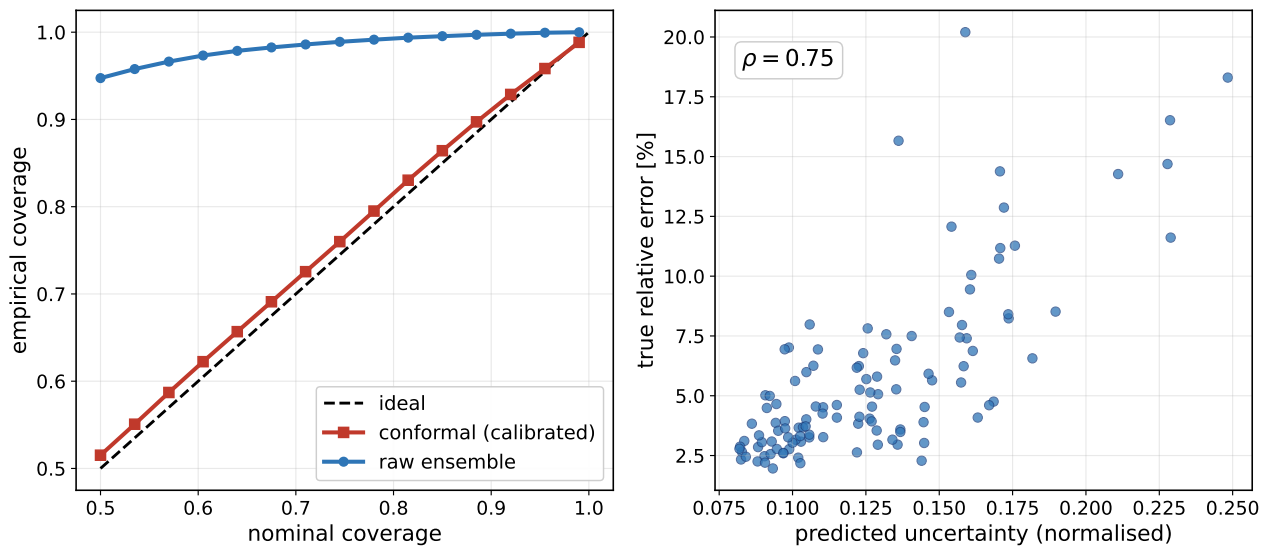


Fig. 9. Left, reliability diagram - the conformal intervals match nominal coverage while the raw ensemble does not; right, predicted uncertainty against true error with the Pearson correlation reported in the panel

The self-reported confidence is meaningful. The mechanism is worth stating: a raw ensemble of a handful of members underestimates the spread of the predictive distribution and is therefore over-confident, as the reliability curve above the diagonal shows. The conformal step of Algorithm 2 rescales the intervals by a single data-driven factor so that the empirical coverage matches the nominal level, which by Theorem 4 it is guaranteed to do up to a term of order one over the calibration size.

Because the decoder is isometric, this scalar recalibration is sufficient at the field level, a simplification that a nonlinear decoder would not permit.

The decomposition of the predictive variance into its aleatoric and epistemic parts by Equation (11) is shown in Figure 10. The predictions are displayed qualitatively in Figure 11 for two unseen parameters, with the ground truth, the posterior mean, the absolute error, and the predicted standard deviation; the predicted uncertainty is spatially structured and concentrates where the error is largest, near the moving front. So, the error bars are locally informative and not merely globally calibrated. This spatial structure is a direct consequence of the propagation formula (12): the field variance at a node is the sum over modes of the squared basis entries weighted by the coefficient variances, so it is largest where the adaptive modes have the most energy, which is exactly the neighbourhood of the advected front where the reconstruction is hardest; panels seven to nine of Figure 2 show the reconstruction, its error, and this predicted standard deviation as surfaces, with the last two sharing the same crest. The uncertainty map therefore does not merely report a global scalar but localizes the regions in which the surrogate is least sure, information that a downstream assimilation or control task can act on.

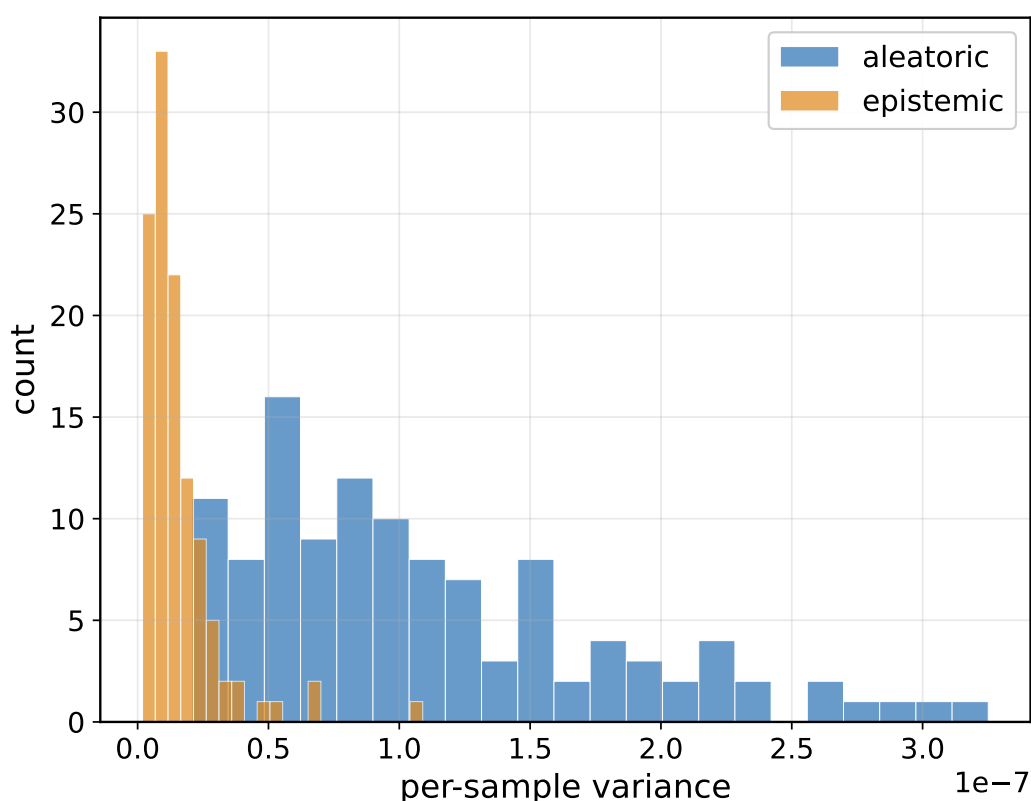


Fig. 10. Decomposition of the predictive variance into aleatoric and epistemic components by the law of total variance

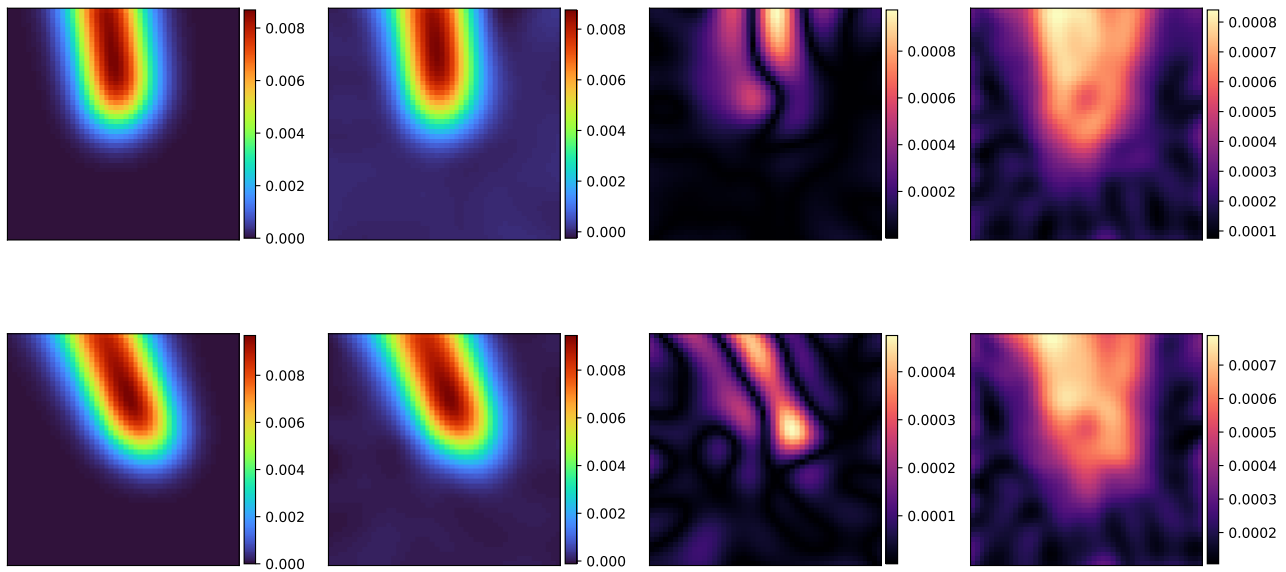


Fig. 11. Qualitative predictions for two unseen parameters (rows): ground truth, posterior mean, absolute error, and predicted standard deviation (columns)

Finally the trust indicator is evaluated. Figure 12 shows that both constituent signals, the predictive uncertainty and the Grassmann novelty distance, shift markedly between in-distribution and extrapolation queries, a shift also visible as the ridge that both signals develop beyond the training band in panels 12 and 13 of Figure 2.

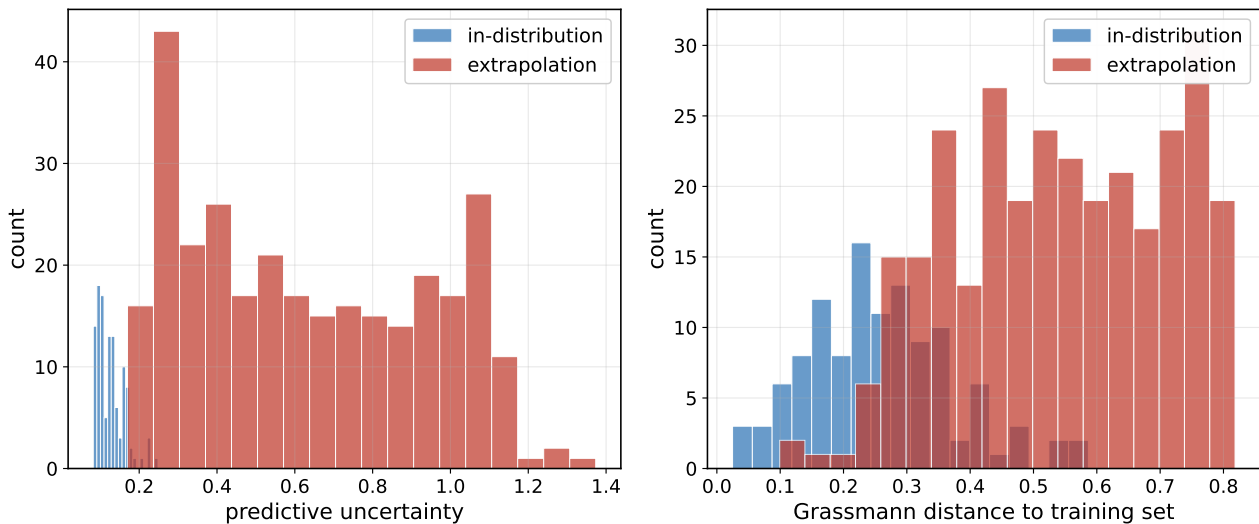


Fig. 12. Distributions of the two solver-free signals for in-distribution and extrapolation queries: predictive uncertainty (left) and Grassmann novelty distance (right). Both shift under distribution change.

Figure 13 reports the resulting out-of-distribution detection and the abstention behaviour. The indicator separates in-distribution from out-of-distribution queries with an area under the receiver

operating characteristic curve of 0.99, and the risk-coverage curve obtained by rejecting the queries with the largest indicator and returning them to the high-fidelity model shows that rejecting a modest fraction sharply reduces the error of those retained, realizing the certified and monotone trade-off of corollary 1 and proposition 1.

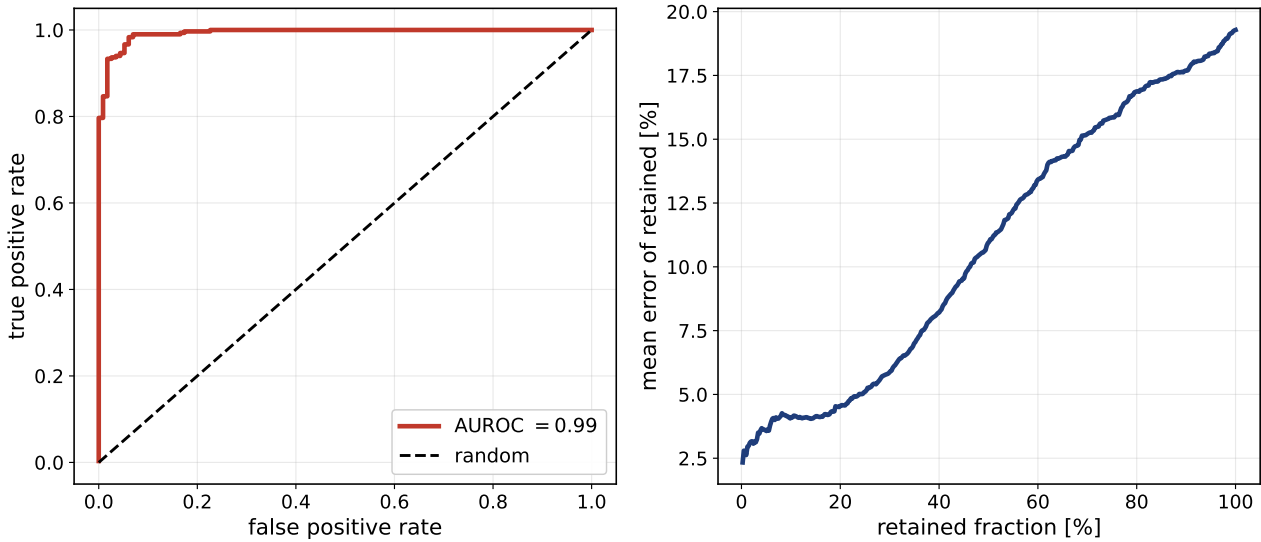


Fig. 13. Left, receiver operating characteristic for out-of-distribution detection by the trust indicator. Right, risk-coverage curve obtained by abstaining on the queries with the largest indicator.

Figure 14 verifies the residual bound of Theorem 3; i.e., across test queries the relative residual and the true error are tightly correlated and cluster about the diagonal. The same clustering about the identity plane seen in panel 14 of Figure 2, confirming that the residual is a faithful certified surrogate for the error.

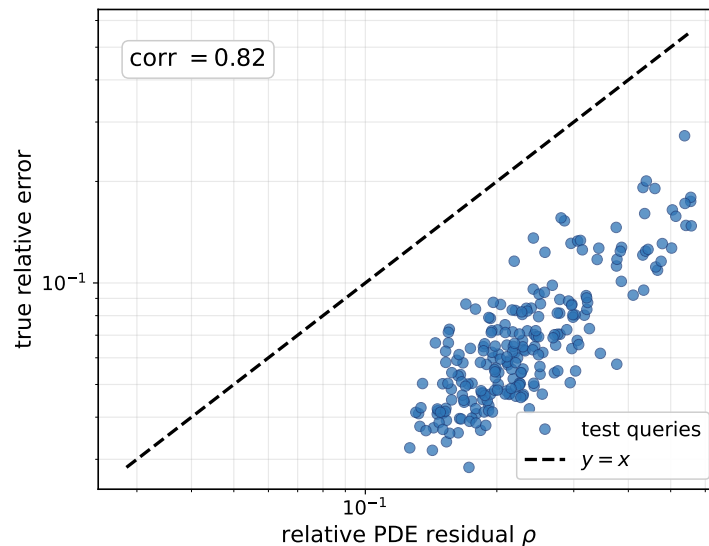


Fig. 14. Verification of the residual bound: true relative error against relative residual over test queries, clustering about the diagonal.

The error decomposition of Theorem 1 is verified directly in Table 4. For the trained model of latent

dimension four the ambient, adaptive-projection, and coefficient errors combine, in the root-sum-of-squares sense, to reproduce the total error to three decimal places, which is an exact instance of the identity (17). The breakdown shows that the dominant term is the adaptive-projection error, so further gains should come from a richer basis.

Table 4

Verification of the error decomposition of Theorem 1 for the model of latent dimension four

	Ambient	Adaptive-projection	Coefficient	Root-sum-of-squares vs total
relative error [%]	0.524	6.503	3.791	$7.545 = 7.545$

A summary of all models is given in Table 5. The wall-clock costs are listed in Table 6. A high-fidelity solve takes 22.7 milliseconds while an online query of the surrogate takes 0.067 milliseconds, a speed-up of 338 times, and the physics term adds only one sparse matrix-vector product per training step. The offline cost is dominated by the high-fidelity snapshots and by training. So, the data-efficiency gains of Figure 6 translate directly into offline savings: reaching a target accuracy with sixty rather than five hundred labelled solves removes hundreds of high-fidelity evaluations from the offline budget at no online cost, which is decisive when each high-fidelity evaluation is itself a coupled multiphysics simulation.

Table 5

Summary of the reduced-order models on the benchmark at latent dimension four

Model	Error in-dist.	Error extrap.	Error at $N_{\text{lab}}=60$	Calibrated UQ	Trust indicator
POD neural net	$\geq 52.0\%$	—	—	no	no
DOD-NN	10.3%	58.4%	34.0%	no	no
Present method	10.4%	26.3%	20.9%	yes ($\rho=0.75$)	yes (AUROC 0.99)

Taken together the experiments trace a single thread. The adaptive basis breaks the width barrier, the physics residual makes that basis learnable from few solves and robust beyond the training range, the linear decoder turns the learned coefficients into calibrated field-level uncertainty, and the residual and the Grassmann distance turn that uncertainty into a decision about whether to trust a prediction. Each link in this chain is supported by a theorem in Section 5 and confirmed by a figure in the present section. So, the framework is neither a collection of heuristics nor a single trick but a coherent construction in which the same physical quantity, the equation residual, serves as a training signal, an error certificate, and a novelty detector.

Table 6

Wall-clock costs on a single processor core

Operation	Cost
High-fidelity solve, per query	22.7 ms
Surrogate evaluation, per query	0.067 ms
Speed-up	338×
Physics term, per training step	One sparse matrix-vector product
Training, per ensemble member	About 190 s

8. Conclusion

Reduced-order modelling has reached a point where accuracy alone is no longer the only criterion; for a surrogate to be usable inside inference and control it must also be economical to build, honest about its confidence, and aware of its own domain of validity. We have developed a physics-informed and uncertainty-aware extension of the deep orthogonal decomposition that resolves the three shortcomings which have kept adaptive-basis surrogates out of demanding inference pipelines. By supervising the continuously adaptive basis and the reduced coefficients with the discrete equation residual at unlabelled collocation parameters, the method learns from points that the high-fidelity solver never visits, which lowers the number of expensive solves and improves extrapolation, and by exploiting the exact and inflation-free propagation guaranteed by the linear reduced-order decoder it reports calibrated predictive uncertainty at a negligible cost. The fusion of that uncertainty with a certified residual estimator and a Grassmann novelty distance yields a solver-free trust indicator that recognizes and abstains on out-of-distribution queries with a distribution-free coverage guarantee, and the exact error decomposition ties these pieces together and explains the observed data efficiency. On a benchmark built to expose the slow-global and fast-slice regime the method reduces the projection error of the linear basis by an order of magnitude at a fixed latent size, cuts the number of high-fidelity solves needed for a target accuracy, more than halves the extrapolation error, produces well-calibrated intervals, and detects out-of-distribution queries almost perfectly.

Several limitations delimit the scope of the present study. The residual bound of theorem 3 is sharp only up to the condition number of the operator, so for strongly advection-dominated or stiff regimes a preconditioned or residual-weighted norm would be needed to keep the certificate tight. The physics weight is a hyperparameter whose best value depends on the amount of labelled data, which is why we scale it inversely with the labelled-set size; an adaptive weighting driven by the residual estimator would remove this choice. The conformal guarantee of Theorem 4 certifies marginal coverage, and a localized variant would be required for coverage conditional on the parameter. Finally, the benchmark is affine and two-dimensional by design, chosen to expose the slow-global and fast-slice regime in a setting where every quantity can be computed exactly. Beyond these technical points, the broader message is methodological: the same equation residual that physics-informed learning uses to solve a single problem can, when attached to an adaptive reduced basis, serve simultaneously as a data-efficient training signal, a certified error estimate, and a novelty detector, and organizing a surrogate around this triple role is what yields a model that is at once fast, economical, calibrated, and self-aware. Natural continuations are the extension of the weak-residual route to strongly nonlinear operators through hyper-reduction, the treatment of time-dependent problems through a latent dynamical model, the use of the trust indicator as an oracle for greedy data acquisition in the spirit of [37] and of multifidelity estimation in the sense of [36], and the treatment of varying geometries through mesh-informed or operator-learning encoders in the sense of [26].

Funding

The authors received no external funding for this research.

Conflicts of Interest

The authors declare no conflicts of interest.

Appendix A: High-Fidelity Model and Data Generation

The benchmark solves the steady advection diffusion equation on the unit square with homogeneous Dirichlet data. With mesh size $h = 1/(N + 1)$ and interior values $u_{i,j}$, the diffusion term is the five-point Laplacian $\varepsilon h^{-2}(4u_{i,j} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1})$ and the advection term uses first-order upwinding consistent with the sign of $\mathbf{b} = (b_x, b_y)$, which produces a monotone M -matrix $\mathbb{A}(\boldsymbol{\mu})$ and a well-posed discrete problem for all $\boldsymbol{\mu}$. The advection field is $\mathbf{b}(\boldsymbol{\mu}) = \beta_0(\cos \theta, \sin \theta)$ with $\beta_0 = 30$, the diffusivity is $\varepsilon = 10^{-2}$, and the source is a superposition of two Gaussians of width 0.09 whose centres depend on (θ, x_c, y_c) ; the amplitudes enter linearly, so each slice is at most two-dimensional. Because the lumped-mass Gram matrix is $\mathbb{G} = h^2 \mathbb{I}$, the discrete norm coincides with the L^2 norm and relative errors equal relative Euclidean errors. A single sparse factorization per $\boldsymbol{\mu}$ yields both source responses, so thousands of solves are affordable; a gallery of solutions is shown in Figure 15, and we verified that the assembled residual of every computed field is at the level of round-off, so the physics residual used in training is exact.

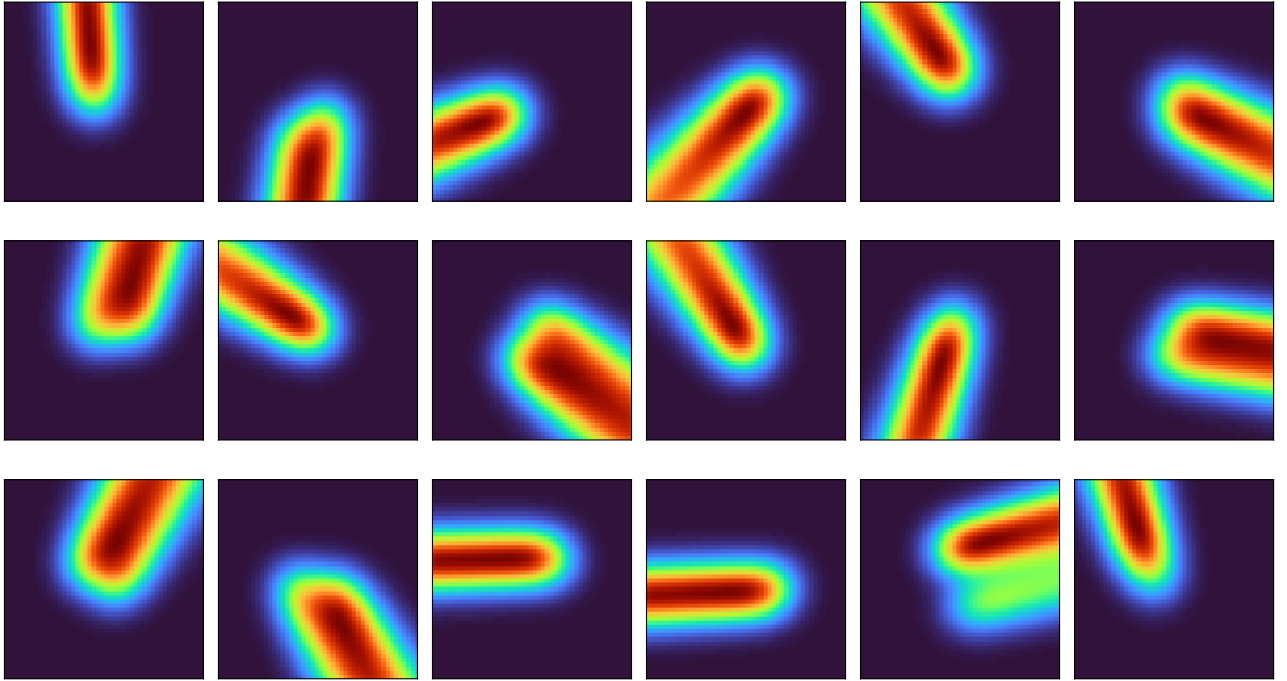


Fig. 15. Representative high-fidelity solutions sampled from the parameter space. The advected front rotates and translates with the space-interacting parameters

Appendix B: Network Architectures and Training

The seed network maps the periodic feature vector $(\cos \theta, \sin \theta, \cos 2\theta, \sin 2\theta, x_c, y_c)$ through three hidden layers of width 96 with leaky rectified-linear activations, and the root block outputs the raw basis of size $N_A \times n$, orthonormalized by a differentiable reduced QR with a sign convention. The coefficient map uses the segregated form (6) with rank eight, and a separate two-layer head predicts the log-variance. The ambient dimension is $N_A = 68$, fixed by a projection tolerance of one half of one percent. Training uses the Adam optimizer with an initial learning rate of 2×10^{-3} and cosine annealing, mini-batches of 256 parameters with four amplitude samples each, and 48 collocation points per step; the heteroscedastic term is weighted by 0.05 and the physics term is warmed up over the

first thirty percent of iterations and, in the data-efficiency study, scaled inversely with the number of labelled points. The hyperparameters are collected in Table 7. The ensemble uses five members with independent initializations.

Table 7
Hyperparameters shared by all experiments

Component	Setting
Seed network	$6 \rightarrow 96 \rightarrow 96 \rightarrow 96$, leaky rectified-linear
Root block	$96 \rightarrow 96 \rightarrow 96 \rightarrow N_A n$, reduced QR
Coefficient map	rank eight, segregated form (6)
Ambient dimension	$N_A = 68$, tolerance 0.5%
Optimizer	Adam, learning rate 2×10^{-3} , cosine annealing
Mini-batch	256 parameters, four amplitudes each
Collocation points	48 per step
Physics weight	warm-up, scaled as $1/N_{\text{lab}}$
Observation noise	1.5% relative
Ensemble	five members
High-fidelity grid	48×48 , $N_h = 2304$, $\varepsilon = 10^{-2}$, $\beta_0 = 30$

Appendix C: Additional Tables

We record the per-seed data behind the data-efficiency and physics-weight studies. Table 8 lists the two-seed errors at each labelled-set size together with the data-adaptive physics weight. Table 9 lists the full physics-weight sweep.

Table 8

Data efficiency: mean relative error in percent for two seeds at each labelled-set size, with the physics weight used by the physics-informed model

N_{lab}	λ_{phys}	DOD so	DOD s1	DOD mean	PI so	PI s1	PI mean
60	0.300	33.44	34.57	34.0	21.39	20.41	20.9
120	0.150	25.77	18.86	22.3	17.81	15.85	16.8
250	0.072	15.75	14.80	15.3	14.76	13.30	14.0
500	0.036	12.57	11.28	11.9	12.06	11.19	11.6
900	0.020	10.40	10.14	10.3	10.53	10.33	10.4

Table 9

Sensitivity to the physics weight at a labelled-set size of 120

λ_{phys}	0.00	0.03	0.10	0.30	1.00
Error [%]	22.27	18.73	18.22	18.08	18.55

References

- [1] Benner, P., Gugercin, S., & Willcox, K. (2015). A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Review*, 57(4), 483–531. <https://doi.org/10.1137/130932715>
- [2] Quarteroni, A., Manzoni, A., & Negri, F. (2016). *Reduced basis methods for partial differential equations: An introduction*. Springer.
- [3] Hesthaven, J. S., Rozza, G., & Stamm, B. (2016). *Certified reduced basis methods for parametrized partial differential equations*. Springer.
- [4] Rozza, G., Huynh, D. B. P., & Patera, A. T. (2008). Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Archives of Computational Methods in Engineering*, 15(3), 229–275. <https://doi.org/10.1007/BF03024948>
- [5] Berkooz, G., Holmes, P., & Lumley, J. L. (1993). The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 25(1), 539–575.
- [6] Sirovich, L. (1987). Turbulence and the dynamics of coherent structures. part i: Coherent structures. *Quarterly of Applied Mathematics*, 45(3), 561–571.
- [7] Kunisch, K., & Volkwein, S. (2002). Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. *SIAM Journal on Numerical Analysis*, 40(2), 492–515. <https://doi.org/10.1137/S0036142900382612>
- [8] Cohen, A., & DeVore, R. (2015). Approximation of high-dimensional parametric pdes. *Acta Numerica*, 24, 1–159. <https://doi.org/10.1017/S0962492915000033>
- [9] Ohlberger, M., & Rave, S. (2016). Reduced basis methods: Success, limitations and future challenges. *Proceedings of the Conference Algorithm*, 1–12. <https://doi.org/10.48550/arXiv.1511.02021>
- [10] Greif, C., & Urban, K. (2019). Decay of the Kolmogorov N-width for wave problems. *Applied Mathematics Letters*, 96, 216–222. <https://doi.org/10.1016/j.aml.2019.05.013>
- [11] Peherstorfer, B. (2022). Breaking the Kolmogorov barrier with nonlinear model reduction. *Notices of the American Mathematical Society*, 69(5), 725–733.
- [12] Lee, K., & Carlberg, K. T. (2020). Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders. *Journal of Computational Physics*, 404, 108973. <https://doi.org/10.1016/j.jcp.2019.108973>
- [13] Fresca, S., Dede', L., & Manzoni, A. (2021). A comprehensive deep learning-based approach to reduced order modeling of nonlinear time-dependent parametrized pdes. *Journal of Scientific Computing*, 87(2), 61. <https://doi.org/10.1007/s10915-021-01462-7>
- [14] Fresca, S., & Manzoni, A. (2022). POD-DL-ROM: Enhancing deep learning-based reduced order models for nonlinear parametrized PDEs by proper orthogonal decomposition. *Computer Methods in Applied Mechanics and Engineering*, 388, 114181. <https://doi.org/10.1016/j.cma.2021.114181>
- [15] Amsallem, D., & Farhat, C. (2008). Interpolation method for adapting reduced-order models and application to aeroelasticity. *AIAA Journal*, 46(7), 1803–1813. <https://doi.org/10.2514/1.35374>
- [16] Amsallem, D., Zahr, M. J., & Farhat, C. (2012). Nonlinear model order reduction based on local reduced-order bases. *International Journal for Numerical Methods in Engineering*, 92(10), 891–916. <https://doi.org/10.1002/nme.4371>
- [17] Pagani, S., Manzoni, A., & Quarteroni, A. (2018). Numerical approximation of parametrized problems in cardiac electrophysiology by a local reduced basis method. *Computer Methods in*

- Applied Mechanics and Engineering*, 340, 530–558. <https://doi.org/10.1016/j.cma.2018.06.003>
- [18] Taddei, T. (2020). A registration method for model order reduction: Data compression and geometry reduction. *SIAM Journal on Scientific Computing*, 42(2), A997–A1027. <https://doi.org/10.1137/19M1271270>
 - [19] Hesthaven, J. S., & Ubbiali, S. (2018). Non-intrusive reduced order modeling of nonlinear problems using neural networks. *Journal of Computational Physics*, 363, 55–78. <https://doi.org/10.1016/j.jcp.2018.02.037>
 - [20] Guo, M., & Hesthaven, J. S. (2019). Data-driven reduced order modeling for time-dependent problems. *Computer Methods in Applied Mechanics and Engineering*, 345, 75–99. <https://doi.org/10.1016/j.cma.2018.10.029>
 - [21] Franco, N. R., Manzoni, A., Zunino, P., & Hesthaven, J. S. (2026). Deep orthogonal decomposition: A continuously adaptive neural network approach to model order reduction of parametrized partial differential equations. *Advances in Computational Mathematics*, 52, 31. <https://doi.org/10.1007/s10444-026-10295-7>
 - [22] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686–707. <https://doi.org/10.1016/j.jcp.2018.10.045>
 - [23] Lu, L., Jin, P., Pang, G., Zhang, Z., & Karniadakis, G. E. (2021). Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3), 218–229. <https://doi.org/10.1038/s42256-021-00302-5>
 - [24] Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2021). Fourier neural operator for parametric partial differential equations. *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2010.08895>
 - [25] Wang, S., Wang, H., & Perdikaris, P. (2021). Learning the solution operator of parametric partial differential equations with physics-informed DeepONets. *Science Advances*, 7(40), eabi8605. <https://doi.org/10.1126/sciadv.abi8605>
 - [26] Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2023). Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89), 1–97.
 - [27] Chen, W., Wang, Q., Hesthaven, J. S., & Zhang, C. (2021). Physics-informed machine learning for reduced-order modeling of nonlinear problems. *Journal of Computational Physics*, 446, 110666. <https://doi.org/10.1016/j.jcp.2021.110666>
 - [28] Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30. <https://doi.org/10.48550/arXiv.1612.01474>
 - [29] Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, 1050–1059. <https://doi.org/10.48550/arXiv.1506.02142>
 - [30] Daxberger, E., Kristiadi, A., Immer, A., Eschenhagen, R., Bauer, M., & Hennig, P. (2021). Laplace Redux: Effortless Bayesian deep learning. *Advances in Neural Information Processing Systems*, 34. <https://doi.org/10.48550/arXiv.2106.14806>
 - [31] Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer.
 - [32] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494–591. <https://doi.org/10.1561/22000000101>

- [33] Seitzer, M., Tavakoli, A., Antic, D., & Martius, G. (2022). On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2203.09168>
- [34] Edelman, A., Arias, T. A., & Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2), 303–353. <https://doi.org/10.1137/S0895479895290954>
- [35] Absil, P.-A., Mahony, R., & Sepulchre, R. (2008). *Optimization algorithms on matrix manifolds*. Princeton University Press.
- [36] Peherstorfer, B., Willcox, K., & Gunzburger, M. (2018). Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3), 550–591. <https://doi.org/10.1137/16M108246>
- [37] Prud'homme, C., Rovas, D. V., Veroy, K., Machiels, L., Maday, Y., Patera, A. T., & Turinici, G. (2002). Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods. *Journal of Fluids Engineering*, 124(1), 70–80. <https://doi.org/10.1115/1.1448332>