

Uncertainty and decision reliability in physics-informed engineering predictions

Vladimir Shapovalov
The Polytechnic University of Turin

Abstract

Physics-informed predictions require uncertainty statements tied to the engineering quantity and decision being supported. Small average field error does not establish calibrated uncertainty in a peak temperature, integrated force, inferred parameter, or defect classification. This review distinguishes observation error, uncertainty within a model family, physical discrepancy, and numerical approximation. It compares Bayesian, ensemble, and conformal approaches according to their assumptions and develops analytical examples for correlated field integration, simultaneous coverage, exact physical constraints, and rare-event validation. A proposed assessment protocol separates calibration from testing and evaluates uncertainty under operating shifts, sensor faults, and downstream threshold decisions.

1. Specify the uncertain object

A prediction interval is meaningful only after identifying what it covers. An interval for a latent noise-free temperature differs from an interval for a future noisy sensor reading. An interval for a constitutive parameter differs from a predictive interval for stress under a future loading path. A spatially averaged error measure supplies none of these interpretations by itself. Physics-informed learning provides a broad framework for combining observations and equations [1, 2], but the uncertainty target remains an additional modeling choice.

Bayesian PINNs introduce probabilistic inference for forward and inverse problems with noisy data [3]. Deep ensembles use multiple fitted predictors as a practical uncertainty approach [4]. These methods can be compared only after specifying the same quantity, noise model, and test population. Dispersion among fitted networks is not automatically a complete account of physical uncertainty when every network shares a biased simulator or an incorrect boundary condition.

The frequency-compensated fluid study [5] provides an engineering setting in which velocity and pressure predictions can be assessed across time and design conditions. Its deterministic field comparisons should not be read as probabilistic calibration results. An uncertainty extension would need to identify a target such as a sensor reading, a flow field, or a derived force, then supply independent evidence for the coverage or probability claims made about that target.

2. A hierarchy of error sources

Use an observation model $y = H(u) + \delta + \varepsilon$, where H maps the physical field to the instrument response, δ represents systematic discrepancy, and ε represents random measurement error. Unknown model parameters and limited training data introduce additional uncertainty in u . Numerical discretization, optimization, and surrogate approximation add further errors. These sources can interact, so a decomposition should be understood as a model of uncertainty rather than an experimentally unique partition.

Bayesian calibration of computer models [6] gives an important reference for separating parameter fitting from discrepancy. A flexible discrepancy can compensate for many parameter values, weakening identifiability. Conversely, omitting discrepancy can force physical parameters to absorb systematic error. Both choices can produce apparently precise predictions within the calibration conditions while failing under a new excitation or operating regime.

Multifidelity learning [7] and broader multifidelity methodology [8] combine sources with different costs and error structures. Low-fidelity simulations are not independent measurements of the real system. A large volume of biased simulation data should not drive uncertainty toward zero merely because the statistical model counts each sample as equally reliable evidence. Thermal reconstruction [9] illustrates a setting where partial observations and unknown parameters coexist; its measurement assumptions must remain visible in any uncertainty extension.

The law of total variance helps organize a probabilistic model. If θ denotes uncertain parameters and D the available data, then the following identity is exact whenever the relevant variances exist.

$$\text{Var}(Y \mid D) = E[\text{Var}(Y \mid \theta, D) \mid D] + \text{Var}(E[Y \mid \theta, D] \mid D).$$

Interpreting the first term as observation uncertainty and the second as parameter uncertainty requires the conditional model to have that meaning. Unrepresented physical discrepancy appears in neither term. The identity therefore provides an accounting rule, not proof that a particular estimator captures every engineering error source.

3. Training variability is not a physical error certificate

PINN optimization can fail because gradients from different loss terms compete, because training emphasizes different modes, or because the optimization trajectory does not resolve the intended solution [10, 11, 12]. Repeated seeds help assess this numerical variability. A collection of similarly biased fits, however, can have small ensemble spread. Conversely, widely dispersed failed fits may produce a large interval that lacks a coherent physical interpretation.

Training guidance [13], adaptive loss balancing [14], and self-adaptive weighting [15] provide candidate ways to improve the deterministic foundation. They should be evaluated before attributing all variation to uncertainty about the physical system. Report residual, boundary, and measurement errors separately, together with solver termination criteria. Selecting only the best seed after inspecting test errors understates the variability of the actual training procedure.

Residual-adaptive sampling [16] and importance sampling [17] change the numerical information supplied to the optimizer. A low sampled residual is not a confidence interval, particularly if unsampled regions contain localized errors. Theoretical convergence or inverse-error estimates [18, 19] require their stated regularity, sampling, and stability assumptions. They cannot be transferred unchanged to an arbitrary noisy manufacturing problem with uncertain constitutive laws.

4. Correlation matters when fields become engineering quantities

Let a discretized field U have mean μ and covariance Σ . For a linear engineering quantity $Q = w^T U$, the mean is $w^T \mu$ and the variance is $w^T \Sigma w$. Applying independent pointwise error bars while discarding covariance changes that variance and can change the engineering conclusion. A force integral, average temperature, or accumulated energy should therefore be computed for joint field realizations, or through an explicitly justified covariance model.

An exact two-point example shows the effect. Suppose U_1 and U_2 each have variance one and correlation ρ , and $Q = (U_1 + U_2)/2$. Then $\text{Var}(Q) = (1 + \rho)/2$. Perfect positive correlation gives variance one, independence gives one-half, and correlation minus one gives zero. The same two pointwise standard deviations are consistent with all three outcomes. The algebraic example isolates how correlation changes uncertainty in the average.

Neural operator models [20, 21, 22] can generate fields efficiently across many inputs, but joint uncertainty does not follow automatically from a deterministic operator. Sampling uncertain inputs propagates the specified input distribution through that operator; it does not include omitted surrogate error. Mixed-variable flow models [23] similarly make pressure or stress-like quantities available as outputs without automatically calibrating their joint uncertainty.

For a nonlinear quantity such as maximum temperature, use the maximum of each joint field realization before forming a distribution. Taking the maximum of pointwise means gives a different object, and taking the maximum of pointwise upper bounds does not by itself establish a stated simultaneous coverage level. Differentiation creates another concern: uncertainty in heat flux depends on uncertainty in temperature gradients and conductivity, not only on temperature values.

5. Pointwise and simultaneous coverage

Suppose separate intervals each cover one of m field locations with probability 0.95. If their coverage events were independent, the probability that all locations are covered would be 0.95^m . For $m = 100$ this is approximately 0.0059. Physical fields generally have correlated errors, so this is an illustration of the distinction rather than a realistic independence assumption for every field.

A dependence-free union bound provides another calculation. If each location has failure probability at most α/m , then the probability that at least one location fails is at most α . This can be conservative and becomes demanding on a fine grid. It also guarantees coverage only on the assessed grid unless additional regularity or interpolation bounds control what happens between points. A continuous-domain maximum requires attention to spatial resolution.

Sparse sensing [24] and hidden-flow reconstruction [25] highlight why observed and unobserved locations should be evaluated separately. High coverage at sensor positions can coexist with poor coverage between them. Report interval behavior by distance from observations, near boundaries, and in regions with strong gradients. Pooling all coordinates into one number can hide precisely the regions where a decision quantity is most sensitive.

6. What exact physical constraints do to uncertainty

Assume an unconstrained Gaussian field vector has covariance Σ and is conditioned on an exact linear relation $CU = d$. If $C\Sigma C^T$ is invertible, the conditional covariance is the expression below. This is the standard Gaussian conditioning result, stated here for an exact known constraint.

$$\Sigma_c = \Sigma - \Sigma C^T (C\Sigma C^T)^{-1} C \Sigma.$$

The removed covariance component lies in directions constrained by the observation. This calculation explains how valid physical information can reduce uncertainty. It does not show that every penalty term in a neural loss has the same probabilistic meaning. A finite penalty leaves a different distribution, and an incorrect exact constraint can create unwarranted confidence by excluding the true state.

Hard boundary constructions [26, 27] are useful when the enforced values are actually known. If boundary temperature or loading is uncertain, that uncertainty should be represented rather than eliminated by an exact ansatz with one nominal value. Variational physics formulations [28] change how equations are enforced but likewise need a clear statistical interpretation before their residual weights are treated as noise precisions.

If CU is observed with Gaussian covariance R_c instead of known exactly, conditioning replaces the inverse above with $(C\Sigma C^T + R_c)^{-1}$. The amount of uncertainty reduction then depends on information quality. Treating thousands of collocation points as independent physical observations with arbitrarily small noise can substantially exaggerate information about a misspecified model. Equations constrain the chosen model; they are not repeated independent experiments on the actual part.

7. Calibration with independent physical cases

Conformal prediction supplies prediction sets with finite-sample marginal coverage under the required exchangeability conditions [29]. In a split construction, fit the predictor on training cases, compute nonconformity scores on separate calibration cases, and use a finite-sample adjusted score quantile. The relevant exchangeable unit in engineering may be an entire part or experiment, not each correlated sensor reading. A guarantee for the chosen target and sampling population does not automatically persist under operating shift.

Conformalized quantile regression combines learned lower and upper quantiles with held-out calibration [30]. It provides a way to adapt interval widths to varying noise or difficulty while retaining a marginal coverage statement under its assumptions. For engineering quantities, calibrate the actual quantity of interest or define a field-level score that matches the intended

coverage event. Pointwise calibration and simultaneous field calibration require different scores and should have different labels in a report.

Empirical assessment should pair coverage with interval width. An interval that covers nearly everything by being extremely broad may have little decision value. Evaluate coverage within predeclared physical groups, such as load ranges or geometry families, and report the number of independent cases in each group. Marginal coverage can conceal undercoverage in a small but important operating region. Repeatedly changing the method based on calibration results also consumes that data's role in model selection.

Classification calibration [31] concerns probabilities of discrete outcomes. The graph-based defect system described in WO2024205617A1 [32] instead supplies an architecture for reconstruction or classification; calibrated deployment probabilities require additional evidence. An anomaly score is a measure of departure from learned nominal behavior [33], and nominal-feature methods such as PatchCore [34] do not turn that score into a universal defect probability. Graph architectures [35] likewise do not remove dependence on labeling quality and defect prevalence.

8. Threshold decisions and the cost of uncertainty

Suppose an engineering quantity must remain below a limit q_{lim} . A point predictor accepts whenever its mean is below that limit. A probabilistic policy might accept only if an upper prediction bound is below the limit and otherwise request further assessment. The latter policy changes both the error rate and the number of referred cases. Its evaluation should include the cost and delay of referrals, rather than reporting accuracy only on the easy accepted subset.

For a binary inspection decision, let p be a calibrated defect probability, C_{FN} the cost of accepting a defective part, and C_{FP} the cost of rejecting a conforming one. Under this simplified two-action loss, rejection has expected cost $(1 - p)C_{\text{FP}}$ and acceptance has expected cost pC_{FN} . Reject when p exceeds $C_{\text{FP}}/(C_{\text{FP}} + C_{\text{FN}})$. This threshold follows from the stated loss model; real decisions may require additional actions, constraints, or consequence categories.

Uncertainty about p and about the costs can change the preferred action. Report sensitivity to those assumptions, especially when the operating point is close to the threshold. A calibrated probability in one population may also be inappropriate after defect prevalence changes. An engineering uncertainty study should therefore freeze its deployment population and decision policy before final evaluation, then test explicitly how plausible shifts affect the result.

9. Rare events and validation sample size

Observing no failures in a small validation set does not establish a zero failure probability. For n independent Bernoulli trials with zero observed failures, an exact one-sided 95 percent upper confidence limit for the failure probability is $1 - 0.05^{(1/n)}$. This follows by solving $(1 - p)^n = 0.05$. For $n = 100$, the limit is about 0.0295; for $n = 1,000$ it is about 0.00299. The familiar approximation is roughly three divided by n .

These limits require the trials and event definition to match the claim. One thousand correlated pixels from the same part are not one thousand independent parts. A test set that omits a relevant defect type cannot bound the failure probability for that missing mechanism merely by containing many nominal examples. State the sampled population, physical grouping, and confirmed ground-truth method alongside the count.

Simulation can expose failure scenarios efficiently, but its event frequencies are not automatically production frequencies. Multifidelity estimates need explicit treatment of surrogate bias and sampling design. Physical benchmarks such as PDEBench [36] and PINNacle [37] support controlled computational comparisons; they do not on their own establish the reliability of a factory acceptance policy. The distinction should remain clear when translating benchmark results into an engineering recommendation.

10. A proposed uncertainty assessment protocol

The proposed study compares a deterministic baseline, an ensemble, and a probabilistic or calibrated interval method under a fixed information budget. Use independent training, calibration, and final test groups. Declare the target quantity, unit of independence, observation covariance, discrepancy assumptions, and acceptance policy before running the final assessment. Compare predictive error, coverage, interval width, referral frequency, and complete computational cost.

Challenge	Keep fixed	Assess separately
Increased sensor noise	Physical case distribution	Latent-field and future-reading coverage
New operating range	Measurement procedure	Bias, uncertainty, and decision errors
Changed boundary condition	Declared model family	Consequences of model discrepancy
Missing sensor cluster	Defect or load distribution	Coverage near unobserved regions
New defect mechanism	Inspection threshold	Misses, false alarms, and referral rate

Maintain failed runs in the record and identify whether their cause is numerical, informational, or physical. Computational comparisons [38] motivate including training, tuning, and inference cost for the specific task rather than relying on a single prediction-time figure. Report where the uncertainty procedure is useful and where additional measurement or a more appropriate physical model is required.

11. Precision of the calibration assessment itself

The measured coverage fraction is an estimate with sampling uncertainty. If interval coverage is a Bernoulli event for independent physical cases, the estimated fraction has standard error approximately equal to the square root of $p(1 - p)/n$, with p replaced by the observed fraction for a simple large-sample approximation. At coverage near 0.95 and $n = 100$, this standard error is about 0.0218. An observed fraction of 0.95 therefore does not establish that population coverage is exactly 0.95, especially in small operating subgroups.

This calculation helps plan validation rather than serving as a substitute for an appropriate confidence interval. Cases near a boundary, rare geometries, and uncommon process conditions may each require separate sampling. A global test set of substantial size can still leave the uncertainty procedure almost untested in a small region that dominates engineering risk. Report the subgroup denominators and avoid selecting favorable groups only after observing their results.

When cases are grouped in batches, resampling for an empirical uncertainty assessment should preserve that grouping. Resampling individual coordinates from a field destroys their spatial dependence and can make performance estimates appear much more precise than they are. Repeated training seeds supply information about optimization variability, while repeated independent parts supply information about generalization across the production population. Neither repetition should be silently substituted for the other.

There is also uncertainty in the reference quantity. A validation sensor with non-negligible calibration error can make a correct latent-field interval appear to miss the measured reading. Conversely, a reference solver using the same constitutive assumptions as the surrogate can conceal shared discrepancy. Identify whether the reference is an analytical solution, a numerically converged simulation, or a calibrated physical measurement. Where feasible, compare more than one reference type and explain which claim each comparison supports.

The final assessment should therefore include both predictive uncertainty and uncertainty in its measured performance. This distinction keeps a finite validation exercise from being presented as a universal guarantee. Engineering reliability depends on a chain from observations through physical assumptions to a decision. The calculations in this review show why pointwise accuracy, agreement among models, and small residuals each address only part of that chain. A defensible uncertainty statement identifies its target, preserves relevant dependence, states its assumptions, and demonstrates its decision value on independent physical cases.

References

- [1] Maziar Raissi, Paris Perdikaris, George Em Karniadakis. 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 (2019). DOI.
- [2] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, et al.. Physics-informed machine learning. *Nature Reviews Physics* 3, 422-440 (2021). DOI.
- [3] Liu Yang, Xuhui Meng, George Em Karniadakis. B-PINNs: Bayesian Physics-Informed Neural Networks for Forward and Inverse PDE Problems with Noisy Data. arXiv:2003.06097 (2020). arXiv.

- [4] Balaji Lakshminarayanan, Alexander Pritzel, Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. arXiv:1612.01474 (2016). arXiv.
- [5] Tongtao Zhang, Biswadip Dey, Pratik Kakkar, Arindam Dasgupta, Amit Chakraborty. Frequency-compensated PINNs for Fluid-dynamic Design Problems. ML4Eng Workshop, NeurIPS 2020. arXiv:2011.01456 (2020). arXiv.
- [6] M. C. Kennedy, A. O'Hagan. Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2001). DOI.
- [7] Xuhui Meng, George Em Karniadakis. A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems. arXiv:1903.00104 (2019). arXiv.
- [8] Benjamin Peherstorfer, Karen Willcox, Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. arXiv:1806.10761 (2018). arXiv.
- [9] Shuheng Liao, Tianju Xue, Jihoon Jeong, et al.. Hybrid thermal modeling of additive manufacturing processes using physics-informed neural networks for temperature prediction and parameter identification. arXiv:2206.07756 (2022). arXiv.
- [10] Sifan Wang, Yujun Teng, Paris Perdikaris. Understanding and mitigating gradient pathologies in physics-informed neural networks. arXiv:2001.04536 (2020). arXiv.
- [11] Sifan Wang, Xinling Yu, Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. arXiv:2007.14527 (2020). arXiv.
- [12] Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, et al.. Characterizing possible failure modes in physics-informed neural networks. arXiv:2109.01050 (2021). arXiv.
- [13] Sifan Wang, Shyam Sankaran, Hanwen Wang, Paris Perdikaris. An Expert's Guide to Training Physics-informed Neural Networks. arXiv:2308.08468 (2023). arXiv.
- [14] Rafael Bischof, Michael Kraus. Multi-Objective Loss Balancing for Physics-Informed Deep Learning. arXiv:2110.09813 (2021). arXiv.
- [15] Levi McClenny, Ulisses Braga-Neto. Self-Adaptive Physics-Informed Neural Networks using a Soft Attention Mechanism. arXiv:2009.04544 (2020). arXiv.
- [16] Chenxi Wu, Min Zhu, Qinyang Tan, et al.. A comprehensive study of non-adaptive and residual-based adaptive sampling for physics-informed neural networks. arXiv:2207.10289 (2022). arXiv.
- [17] Mohammad Amin Nabian, Rini Jasmine Gladstone, Hadi Meidani. Efficient training of physics-informed neural networks via importance sampling. arXiv:2104.12325 (2021). arXiv.
- [18] Yeonjong Shin, Jerome Darbon, George Em Karniadakis. On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type PDEs. arXiv:2004.01806 (2020). arXiv.
- [19] Siddhartha Mishra, Roberto Molinaro. Estimates on the generalization error of Physics Informed Neural Networks (PINNs) for approximating a class of inverse problems for PDEs. arXiv:2007.01138 (2020). arXiv.
- [20] Lu Lu, Pengzhan Jin, George Em Karniadakis. DeepONet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. arXiv:1910.03193 (2019). arXiv.
- [21] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, et al.. Fourier Neural Operator for Parametric Partial Differential Equations. arXiv:2010.08895 (2020). arXiv.
- [22] Zongyi Li, Hongkai Zheng, Nikola Kovachki, et al.. Physics-Informed Neural Operator for Learning Partial Differential Equations. arXiv:2111.03794 (2021). arXiv.
- [23] Chengping Rao, Hao Sun, Yang Liu. Physics-informed deep learning for incompressible laminar flows. arXiv:2002.10558 (2020). arXiv.
- [24] Krithika Manohar, Bingni W. Brunton, J. Nathan Kutz, Steven L. Brunton. Data-Driven Sparse Sensor Placement for Reconstruction. arXiv:1701.07569 (2017). arXiv.
- [25] Maziar Raissi, Alireza Yazdani, George Em Karniadakis. Hidden Fluid Mechanics: A Navier-Stokes Informed Deep Learning Framework for Assimilating Flow Visualization Data. arXiv:1808.04327 (2018). arXiv.

- [26] N. Sukumar, Ankit Srivastava. Exact imposition of boundary conditions with distance functions in physics-informed deep neural networks. arXiv:2104.08426 (2021). arXiv.
- [27] Lu Lu, Raphael Pestourie, Wenjie Yao, et al.. Physics-informed neural networks with hard constraints for inverse design. arXiv:2102.04626 (2021). arXiv.
- [28] E. Kharazmi, Z. Zhang, G. E. Karniadakis. Variational Physics-Informed Neural Networks For Solving Partial Differential Equations. arXiv:1912.00873 (2019). arXiv.
- [29] Anastasios N. Angelopoulos, Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. arXiv:2107.07511 (2021). arXiv.
- [30] Yaniv Romano, Evan Patterson, Emmanuel J. Candès. Conformalized Quantile Regression. arXiv:1905.03222 (2019). arXiv.
- [31] Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Q. Weinberger. On Calibration of Modern Neural Networks. arXiv:1706.04599 (2017). arXiv.
- [32] Pratik Kakkar, Bruno Paes Leao, Julian Busch, Shinjan Ghosh. Applying physics to a neural network model for detection of manufacturing defects. WO2024205617A1. PCT/US2023/030943. Published 3 October 2024 (2024). Source.
- [33] Guansong Pang, Chunhua Shen, Longbing Cao, Anton van den Hengel. Deep Learning for Anomaly Detection: A Review. arXiv:2007.02500 (2020). arXiv.
- [34] Karsten Roth, Latha Pemula, Joaquin Zepeda, et al.. Towards Total Recall in Industrial Anomaly Detection. arXiv:2106.08265 (2021). arXiv.
- [35] Thomas N. Kipf, Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. arXiv:1609.02907 (2016). arXiv.
- [36] Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, et al.. PDEBENCH: An Extensive Benchmark for Scientific Machine Learning. arXiv:2210.07182 (2022). arXiv.
- [37] Zhongkai Hao, Jiachen Yao, Chang Su, et al.. PINNacle: A Comprehensive Benchmark of Physics-Informed Neural Networks for Solving PDEs. arXiv:2306.08827 (2023). arXiv.
- [38] Tamara G. Grossmann, Urszula Julia Komorowska, Jonas Latz, Carola-Bibiane Schönlieb. Can Physics-Informed Neural Networks beat the Finite Element Method?. arXiv:2302.04107 (2023). arXiv.
- [39] Rajat Arora, Pratik Kakkar, Amit Chakraborty, Biswadip Dey. Physics-informed neural networks for modeling rate- and temperature-dependent plasticity. Machine Learning and the Physical Sciences Workshop, NeurIPS 2022. arXiv:2201.08363v3 (2022). arXiv.