
APIC: Amortized Physics-Informed Calibration using Neural Processes

Aishwarya Venkataramanan^{*1}

Sai Karthikeya Vemuri^{*1}

Joachim Denzler¹

¹Computer Vision Group, Friedrich Schiller University Jena, Germany

Abstract

Physics models are inherently imperfect due to misspecified or missing mechanisms, resulting in systematic discrepancies between model predictions and real-world observations. The Kennedy–O’Hagan (KOH) framework addresses this issue through explicit discrepancy modeling. However, its non-amortized, per-instance formulation limits scalability across families of related systems. We introduce Amortized Physics-Informed Calibration (APIC), a population-level extension of KOH that leverages Neural Processes to perform scalable Bayesian inference across realizations. Our framework employs a two-branch latent architecture to disentangle instance-specific physical parameters from shared, state-dependent structural discrepancies. By integrating differentiable physics into an amortized inference backbone, APIC enables rapid calibration of unseen realizations from sparse observations while quantifying uncertainty. Experiments on the damped spring oscillator, the Lotka–Volterra system, and the advection–diffusion PDE with misspecified physics demonstrate improved parameter recovery and consistent identification of the systemic discrepancy structure compared to other calibration approaches.

1 INTRODUCTION

Physical simulators are ubiquitous in scientific modeling, enabling the prediction of complex phenomena ranging from climate dynamics [Lai et al., 2025] to aerostructural mechanics [Catalani et al., 2024]. Yet even high-fidelity computational models inevitably exhibit systematic discrepancies with real-world observations due to unresolved physics, modeling assumptions, or numerical approxima-

tions [Kennedy and O’Hagan, 2001]. In practice, observed system behavior reflects two distinct sources of uncertainty. The first is parametric uncertainty, corresponding to the “known unknowns” of the model: the parameters explicitly represented in the simulator’s equations but whose numerical values are unknown. The second is structural discrepancy, representing the “unknown unknowns”: systematic deviations arising from model inadequacy, where relevant physical processes or dependencies are absent from the governing equations altogether. Disentangling these two sources of variation is generally non-identifiable without additional structural assumptions. During model calibration, i.e., adjusting parameters to fit observed data, structural errors can mimic or compensate for parameter misspecification, and conversely, parameter adjustments can absorb unmodeled discrepancy. In such cases, the optimizer may find parameters that minimize the residual error but lead to non-identifiability and compromise physical interpretability [Brynjarsdottir and O’Hagan, 2014].

The Kennedy–O’Hagan (KOH) [Kennedy and O’Hagan, 2001] framework is a seminal approach for disentangling parametric uncertainty from structural discrepancy in physical models. It explicitly decomposes observed data y into three components: a physical simulator $\mathcal{F}$, a systematic discrepancy term, and observational noise, formalized as

$$y = \mathcal{F}(\theta) + \delta + \epsilon, \quad (1)$$

where θ represents the optimal (but unknown) physical parameters, δ captures systematic model discrepancy, and ϵ represents observational noise. Here, δ is intended to capture only variability that cannot be explained by the simulator. If the discrepancy is sufficiently flexible to reproduce the simulator’s dynamics, it can compensate for misspecified physical parameters. This results in non-identifiability between δ and θ and thus compromises physical interpretability.

KOH employs Gaussian Processes [Rasmussen and Williams, 2006] to simultaneously estimate unknown parameters and learn the discrepancy. Despite its theoretical el-

^{*} Equal contribution.

egance, the classical framework is non-amortized. For each new realization (e.g., a new experiment or system instance), inference over θ and δ must be performed from scratch. This paradigm does not exploit the fact that although parameters may vary across instances, the underlying physical mechanisms that govern the discrepancy are often shared across a population.

In parallel, the development of hybrid modeling has aimed to bridge the gap between purely physical simulators and data-driven approaches by integrating physical laws with deep learning [Reichstein et al., 2019, Benson et al., 2025]. Such approaches include Physics-Informed Neural Networks (PINNs) [Raissi et al., 2017], residual learning frameworks [Wang et al., 2022, Takeishi and Kalousis, 2021], and neural operator-based corrections [Li et al., 2020], which correct imperfect simulators using observed data. While these methods often improve predictive accuracy, they primarily optimize pointwise reconstruction objectives and do not typically adopt a fully probabilistic formulation. As a result, uncertainty quantification is often limited or heuristic, and identifiability between physical parameters and learned discrepancy remains insufficiently addressed.

We propose *Amortized Physics-Informed Calibration (APIC)*: a framework that unites the probabilistic rigor of KOH with the amortized efficiency of deep generative models [Jha et al., 2023]. APIC models a population of related system realizations as arising from a shared physical simulator and a learned, state-dependent correction for structured discrepancy over the input domain. The model consists of a dual-latent encoder architecture. One latent branch captures instance-specific physical parameters, which are mapped to the simulator’s input. A second latent branch encodes residual discrepancies. Given sparse observations from a new realization, APIC infers both latent variables, combines the simulator output with the discrepancy network, and produces predictive distributions with principled uncertainty estimates. To encourage meaningful separation between the physics and discrepancy components, the training proceeds in two stages. We first learn the mapping between observations and physical parameters using simulator-generated data with known ground-truth parameters. We then train on real observations while regularizing the discrepancy component to capture systematic errors without undermining physical interpretability. The model is presented in Figure 1.

We evaluate APIC on three representative dynamical systems: a damped spring oscillator, a predator-prey ordinary differential equation (ODE) model, and an advection-diffusion system exhibiting systematic discrepancies. Across the systems, we assess reconstruction accuracy, predictive uncertainty, and the accuracy of parameter and discrepancy recovery. Our results show that APIC achieves improved predictive performance while maintaining accurate and physically interpretable parameter recovery.

In summary, our contributions are threefold: (1) We introduce APIC, a probabilistic framework that integrates physics-based simulation with amortized latent-variable inference for population-level calibration. (2) We propose a dual-latent architecture that disentangles instance-specific physical parameters from structured model discrepancies, enabling physically interpretable inference under sparse observations. (3) We demonstrate that APIC achieves improved predictive performance and reliable uncertainty quantification across multiple dynamical systems. The code is available in: <https://github.com/cvjena/APIC.git>.

2 RELATED WORKS

The foundational framework for computer model calibration was established by Kennedy and O’Hagan [Kennedy and O’Hagan, 2001], who formulated calibration as a decomposition of observations into a parametric simulator output and a systematic discrepancy term representing missing physics. While the KOH framework provides a basis for uncertainty quantification, it suffers from two major limitations: identifiability, where the discrepancy term confounds inference of physically meaningful parameters [Brynjarsdottir and O’Hagan, 2014], and computational intractability. Classical implementations rely on Gaussian Processes (GPs) and Markov Chain Monte Carlo (MCMC), which scale cubically with data size and require full re-inference for each new experimental realization [Higdon et al., 2004]. Subsequent work introduced orthogonality constraints to mitigate identifiability issues [Bayarri et al., 2007, Tuo and Wu, 2015, Plumlee, 2017], but per-instance inference remains a computational bottleneck for large-scale or real-time applications.

Recent advances in deep learning, particularly Physics-Informed Neural Networks (PINNs) [Raissi et al., 2017, Vemuri et al., 2024], have shifted physics modeling toward gradient-based optimization by embedding physical residuals directly into the loss function. Extensions have incorporated uncertainty modeling and structural error correction. For example, Bayesian PINNs (B-PINNs) [Yang et al., 2021] quantify uncertainty in inverse problems, while [Zou et al., 2024, Wang et al., 2022, Spitieris and Steinsland, 2023] introduce secondary networks or Deep Gaussian Processes to account for misspecified differential equations. Although effective, these methods remain optimization-based: they solve a new inverse problem for each dataset and do not learn an amortized mapping from observations to parameters. This limits their scalability in settings requiring rapid calibration across many realizations.

Neural Processes (NPs) [Garnelo et al., 2018a,b] provide a probabilistic, amortized framework for learning distributions over functions from sparse context observations. By encoding context–target relationships into latent variables, NPs enable fast adaptation to new tasks without per-

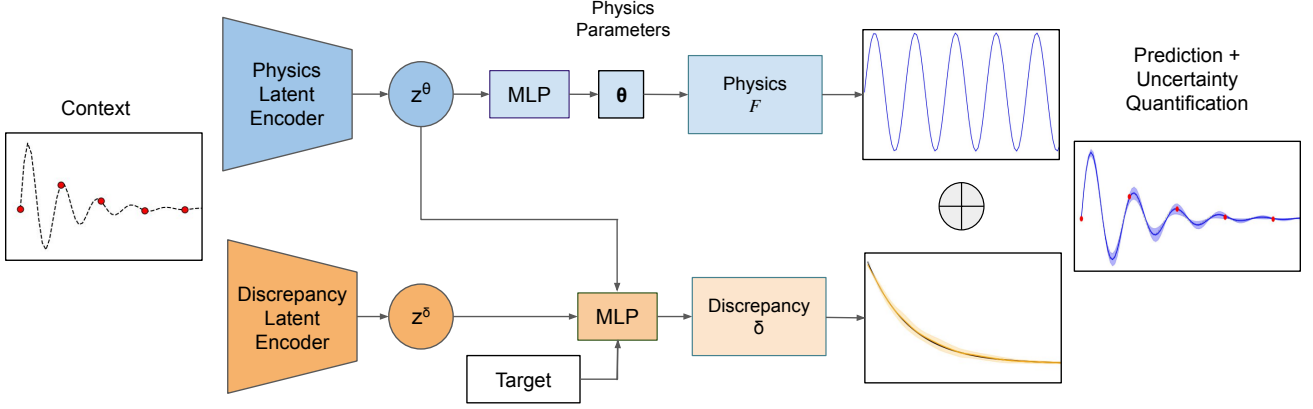


Figure 1: **Overview of Amortized Physics-Informed Calibration (APIC):** The framework employs a dual-branch latent-variable Neural Process architecture to disentangle observed system behavior into parametric physics and a structural discrepancy. The *Physics Branch* (top) infers instance-specific parameters θ to drive a differentiable simulator F , while the *Discrepancy Branch* (bottom) learns a shared, state-dependent correction model δ . By combining these components, APIC achieves calibration and uncertainty quantification for unseen realizations of imperfect physical systems.

task optimization. Subsequent extensions [Kim et al., 2019, Venkataramanan and Denzler, 2026] improve expressiveness and uncertainty calibration. However, standard NPs are purely data-driven: they do not incorporate known physical simulators or account for structured model discrepancies.

Amortized approaches [Cranmer et al., 2020, Gonçalves et al., 2020] learn direct mappings from data to posterior distributions, shifting computational cost to an offline training phase and enabling fast inference. Neural operator methods, such as DeepONet [Lu et al., 2021] and Fourier Neural Operators [Li et al., 2020], further enable resolution-independent mappings between function spaces. While promising, these methods generally treat the underlying physics either as a black box (ignoring known physical constraints) or as a perfect white box (ignoring structural discrepancies).

To the best of our knowledge, there is currently no unified framework that integrates deep learning-based amortized inference with the KOH calibration principle. We address this gap by introducing an amortized, physics-informed Neural Process that explicitly disentangles physical parameter inference from structured discrepancy modeling across a family of related system realizations. This formulation preserves physical interpretability while enabling efficient inference and principled uncertainty quantification.

3 METHOD

We present Amortized Physics-Informed Calibration (APIC), a population-level extension of the KOH framework that generalizes single-instance Bayesian calibration to families of related tasks. The method employs a dual-encoder Neural Process architecture to model realization-specific physical parameters and structural discrepancy, together with a two-stage curriculum training strategy that stabilizes

learning and promotes their disentanglement.

3.1 PROBLEM FORMULATION: META-LEVEL CALIBRATION

We extend the classical KOH framework to a family of related realizations $\mathcal{Y} = \{\mathbf{y}_i\}_{i=1}^N$. Observations for each realization are modeled as

$$\mathbf{y}_i = \mathcal{F}(\mathbf{x}_i; \boldsymbol{\theta}_i) \circledast \delta(\mathbf{x}_i; \boldsymbol{\theta}_i) + \epsilon_i, \quad (2)$$

where $\circledast$ denotes an operator that specifies how structural model error modifies the nominal solver output. Classical KOH corresponds to the additive case, while multiplicative corrections are also accommodated in our formulation. $\mathcal{F}$ is a differentiable physics solver, $\mathbf{x}_i$ denotes the system state (e.g., time or observable variables), $\boldsymbol{\theta}_i$ are realization-specific physical parameters, and ϵ_i represents observation noise. Our objective is to jointly infer the latent physical parameters $\boldsymbol{\theta}_i$ and the corresponding discrepancy functions $\delta(\mathbf{x}_i; \boldsymbol{\theta}_i)$ across realizations.

To this end, we employ a Neural Process framework to amortize inference over the family of realizations. This enables mapping sparse observations to posterior distributions over latent variables governing both physical parameters and structured discrepancy. The model is illustrated in Figure 1.

3.2 APIC MODEL ARCHITECTURE

For each realization i , we observe a set of T_i state-observation pairs $\{(\mathbf{x}_{ij}, \mathbf{y}_{ij})\}_{j=1}^{T_i}$. Following the NP framework, we partition these observations into a context set $\mathcal{C}_i = \{(\mathbf{x}_{ij}^{(c)}, \mathbf{y}_{ij}^{(c)})\}_{j=1}^{C_i}$ and a target set $\mathcal{T}_i = \{(\mathbf{x}_{ij}^{(t)}, \mathbf{y}_{ij}^{(t)})\}_{j=1}^{T_i}$, such that $\mathcal{C}_i \subseteq \mathcal{T}_i$. The context set provides sparse observa-

tions used for inference, while the target set specifies the states at which predictions are required.

The encoder consists of 2 latent inference networks that operate on the context set $\mathcal{C}_i$. The physics encoder infers a latent representation $\mathbf{z}_i^{(\theta)}$ that captures instance-specific physical properties. This is further mapped to the solver parameters θ_i . In parallel, a discrepancy encoder infers a latent variable $\mathbf{z}_i^{(\delta)}$ that parameterizes a state-dependent discrepancy function. The decoder combines the inferred physical parameters, discrepancy latent, and target states to produce predictive distributions at the target inputs.

Physics Latent Encoder. The physics latent encoder infers a latent representation of the instance-specific physical parameters from the context set $\mathcal{C}_i$. Each context pair $(\mathbf{x}_{ij}^{(c)}, \mathbf{y}_{ij}^{(c)})$ is mapped through a shared encoder network $h_\theta(\cdot)$ to obtain a representation. The individual context representations are aggregated into a single latent embedding that summarizes information from the context set. Depending on the specific NP variant, this aggregation can be a simple mean or sum over embeddings, or a learned attention mechanism. This aggregated representation is used to produce a Gaussian posterior over the physics latent variable:

$$q(\mathbf{z}_i^{(\theta)} | \mathcal{C}_i) = \mathcal{N}(\boldsymbol{\mu}_\theta(\mathcal{C}_i), \boldsymbol{\Sigma}_\theta(\mathcal{C}_i)), \quad (3)$$

where $\boldsymbol{\mu}_\theta(\cdot)$ and $\boldsymbol{\Sigma}_\theta(\cdot)$ are neural networks producing the mean and diagonal covariance, respectively.

A sample from this posterior is then mapped through a neural network

$$\hat{\theta}_i = g_\phi(\mathbf{z}_i^{(\theta)}), \quad (4)$$

yielding the instance-specific solver parameters for the physics model $\mathcal{F}$.

Discrepancy Latent Encoder. Similar to the physics latent encoder, the discrepancy latent encoder maps the context set $\mathcal{C}_i$ to the parameters of a variational posterior over the discrepancy latent variable $\mathbf{z}_i^{(\delta)}$:

$$q(\mathbf{z}_i^{(\delta)} | \mathcal{C}_i) = \mathcal{N}(\boldsymbol{\mu}_\delta(\mathcal{C}_i), \boldsymbol{\Sigma}_\delta(\mathcal{C}_i)), \quad (5)$$

where $\boldsymbol{\mu}_\delta(\cdot)$ and $\boldsymbol{\Sigma}_\delta(\cdot)$ are neural networks shared across instances. The latent variable $\mathbf{z}_i^{(\delta)}$ captures instance-specific variations in the structured discrepancy, representing systematic deviations not explained by the physics solver.

A sample from $\mathbf{z}_i^{(\delta)}$, together with the physics latent $\mathbf{z}_i^{(\theta)}$ and a target state $\mathbf{x}_i^{(t)}$, is provided to the discrepancy function

$$\delta_i(\mathbf{x}_i^{(t)}; \mathbf{z}_i^{(\theta)}, \mathbf{z}_i^{(\delta)}) = f_\psi(\mathbf{x}_i^{(t)}, \mathbf{z}_i^{(\theta)}, \mathbf{z}_i^{(\delta)}). \quad (6)$$

f_ψ is modeled as a neural network and represents systematic discrepancies shared across realizations, while allowing instance-specific corrections that depend on the system state and physical parameters.

Generative Model. Given the inferred latent variables $\mathbf{z}_i^{(\theta)}$ and $\mathbf{z}_i^{(\delta)}$, the decoder defines the conditional distribution of an observation $\mathbf{y}_i^{(t)}$ at a target input $\mathbf{x}_i^{(t)}$ as

$$p(\mathbf{y}_i^{(t)} | \mathbf{x}_i^{(t)}, \mathbf{z}_i^{(\theta)}, \mathbf{z}_i^{(\delta)}) = \mathcal{N}(\mathcal{F}(\mathbf{x}_i^{(t)}; \hat{\theta}_i) + \delta(\mathbf{x}_i^{(t)}; \mathbf{z}_i^{(\theta)}, \mathbf{z}_i^{(\delta)}), \Sigma), \quad (7)$$

where Σ is the observation noise.

The posterior predictive distribution conditioned on the context set is

$$p(\mathbf{y}_i^{(t)} | \mathbf{x}_i^{(t)}, \mathcal{C}_i) = \int p(\mathbf{y}_i^{(t)} | \mathbf{x}_i^{(t)}, \mathbf{z}_i^{(\theta)}, \mathbf{z}_i^{(\delta)}) q(\mathbf{z}_i^{(\theta)}, \mathbf{z}_i^{(\delta)} | \mathcal{C}_i) d\mathbf{z}_i^{(\theta)} d\mathbf{z}_i^{(\delta)}, \quad (8)$$

where we adopt a mean-field variational approximation,

$$q(\mathbf{z}_i^{(\theta)}, \mathbf{z}_i^{(\delta)} | \mathcal{C}_i) = q(\mathbf{z}_i^{(\theta)} | \mathcal{C}_i) q(\mathbf{z}_i^{(\delta)} | \mathcal{C}_i). \quad (9)$$

Note that although the approximate posterior factorizes, interaction between the two latent variables is reintroduced through the discrepancy function network (Equation 6), which models their joint influence on the observations.

3.3 TRAINING AND INFERENCE

We train the model using a two-stage procedure to encourage a meaningful decomposition between physics and discrepancy components. The training data consists of two sources: (i) a nominal (or reference) dataset and (ii) a real observation dataset. The nominal dataset contains realizations generated using the physics solver with known parameters and without unmodeled discrepancies. This provides ground-truth supervision for the physics parameters θ . In contrast, the real observation dataset consists of measurements from the true system, where discrepancies between the physics solver and reality may be present. The nominal data is used to anchor the physics latent to physically meaningful parameters during pretraining, while the real observations are used to train the full probabilistic model with discrepancy correction.

Stage 1: Physics Pretraining on Nominal Data. In the first stage, we train only the physics pathway using a nominal dataset, for which ground-truth solver parameters θ_i are available. The discrepancy network and decoder are disabled during this stage.

We minimize the supervised parameter loss

$$\mathcal{L}_\theta = \sum_{i=1}^{N_{\text{nom}}} \|\theta_i - \hat{\theta}_i\|_2^2, \quad (10)$$

where N_{nom} denotes the number of realizations in the nominal dataset and $\hat{\theta}_i$ is the predicted solver parameter. This stage encourages $\mathbf{z}^{(\theta)}$ to encode physically meaningful structure before introducing a discrepancy model.

Stage 2: Joint Variational Training. In the second stage, we activate the discrepancy network and decoder, and train the full model via variational inference. This stage leverages both the real observation dataset and the nominal dataset.

We place standard normal priors on both latent variables:

$$p(\mathbf{z}^{(\theta)}) = \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad p(\mathbf{z}^{(\delta)}) = \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

For each realization i in the real observation dataset, we maximize the evidence lower bound (ELBO):

$$\begin{aligned} \mathcal{L}_i^{\text{ELBO}} = & \mathbb{E}_{q(\mathbf{z}_i^{(\theta)} | \mathcal{C}_i) q(\mathbf{z}_i^{(\delta)} | \mathcal{C}_i)} [\log p(\mathbf{y}_i^{(t)} | \mathbf{x}_i^{(t)}, \mathbf{z}_i^{(\theta)}, \mathbf{z}_i^{(\delta)})] \\ & - \text{KL}(q(\mathbf{z}_i^{(\theta)} | \mathcal{C}_i) \| p(\mathbf{z}_i^{(\theta)})) - \text{KL}(q(\mathbf{z}_i^{(\delta)} | \mathcal{C}_i) \| p(\mathbf{z}_i^{(\delta)})), \end{aligned} \quad (11)$$

To encourage disentanglement between the physics and discrepancy components, we retain the physics parameter supervision from Stage 1 on the nominal dataset to anchor the physics encoder and mapping network. Simultaneously, we regularize the discrepancy function via

$$\mathcal{L}_\delta = \sum_{i=1}^N \mathcal{R}(\delta_i(\cdot)). \quad (12)$$

$\mathcal{R}$ is a functional penalty that encodes prior assumptions about the discrepancy structure. This can include standard L_1/L_2 penalties or more structured regularizers such as orthogonality or smoothness constraints [Gulrajani et al., 2017, Yoshida and Miyato, 2017]. The regularization discourages the discrepancy from absorbing variation that can be explained by the physics encoder, thereby encouraging disentanglement between physics and residual corrections. Based on empirical evidence from ablation studies, we found that L_2 regularization provides the best trade-off between predictive performance and disentanglement, and is therefore used in all experiments. The final objective is

$$\mathcal{L} = \sum_{i=1}^N -\mathcal{L}_i^{\text{ELBO}} + \lambda_\delta \mathcal{L}_\delta + \lambda_\theta \mathcal{L}_\theta, \quad (13)$$

where $\lambda_\theta, \lambda_\delta$ are hyperparameters controlling the relative weight of each term.

Inference. At test time, given a new realization with sparse context set $\mathcal{C}_i$, the encoders infer approximate posterior distributions over $\mathbf{z}_i^{(\theta)}$ and $\mathbf{z}_i^{(\delta)}$. Samples from these distributions are mapped to solver parameters $\hat{\theta}_i$ and propagated through $\mathcal{F}$, while the discrepancy network produces state-dependent corrections. Monte Carlo samples of the combined output yield predictive means and variances, providing uncertainty estimates over the target observations.

4 EXPERIMENTS

Experimental Setup. We consider three benchmark systems: a spring oscillator, a Lotka-Volterra ODE, and an advection-diffusion PDE. For each benchmark, the learner has access to a nominal physics simulator, while observations are generated from a true process that includes a structured, state-dependent discrepancy. The discrepancy is systematic rather than random noise, and it depends on task-specific quantities such as physical parameters or initial conditions. This setup reflects realistic scientific scenarios in which simulators capture known physics but omit structured, unmodeled effects. All models are implemented in PyTorch.

Baselines. We compare against three physics model calibration baselines. The first baseline is the classical KOH framework Kennedy and O’Hagan [2001], where the discrepancy is modeled as a Gaussian Process Rasmussen and Williams [2006]. We use a zero-mean RBF kernel and estimate the physical parameters (θ) together with the GP hyperparameters, via marginal likelihood maximization for each realization. Predictive uncertainty is obtained from the GP posterior. Bayesian Calibration with Physics-Informed Priors (BCPI; Spitieris and Steinland [2023]) instantiates a KOH-style model but incorporates physics-informed priors within the GP to improve identifiability and extrapolation. Correction-PINN Zou et al. [2024] is a PINN-based method in which a secondary correction network augments the governing equations or residual terms, and predictive uncertainty is approximated using deep ensembles of independently trained models.

To evaluate the influence of the NP backbone, we instantiate APIC with three architectures: CNP, LNP, and ANP. In all variants, the physics-informed structure, dual-latent formulation, and discrepancy modeling remain identical, and only the context aggregation mechanism differs. **APIC-CNP** uses a global mean aggregation and a deterministic encoding [Garnelo et al., 2018a], yielding a fixed task representation. **APIC-LNP** and **APIC-ANP** both employ stochastic task-level embeddings as described in section 3. The difference lies in the aggregation mechanism: APIC-LNP uses global mean aggregation in the encoder [Garnelo et al., 2018b], whereas APIC-ANP employs attention-based aggregation to obtain target-dependent context weighting [Kim et al., 2019]. Models are trained using the Adam optimizer with learning rate 5×10^{-4} . The values of λ_θ and λ_δ are set to 1. Details on the model architecture and training hyperparameters are provided in Appendix A.

Evaluation Metrics. We evaluate performance along three axes: predictive reconstruction, discrepancy identification, and parameter recovery. For predictive reconstruction, we report the Mean Absolute Error (MAE) to measure point-wise reconstruction error and the Predictive Log-Likelihood (LL) to measure the probabilistic accuracy of the predic-

Table 1: Quantitative comparison for damped spring system. $\dagger$ methods are corrected per realization and not amortized as our method. Best results are shown in bold, and the second-best results are underlined.

Method	Reconstruction			Discrepancy (δ) prediction			Parameter (θ) prediction
	MAE ($\downarrow$)	LL ($\uparrow$)	ECE ($\downarrow$)	MAE ($\downarrow$)	LL ($\uparrow$)	ECE ($\downarrow$)	MAE ($\downarrow$)
KOH-GP †	0.017 ± 0.009	0.738 ± 0.086	0.413 ± 0.010	0.760 ± 0.118	0.338 ± 0.049	0.411 ± 0.016	2.957 ± 1.487
BCPI †	0.033 ± 0.016	1.079 ± 0.267	0.206 ± 0.035	0.621 ± 0.272	0.916 ± 0.122	0.475 ± 0.051	1.395 ± 0.512
Correction-PINN †	0.094 ± 0.068	-0.718 ± 0.203	0.471 ± 0.021	1.546 ± 0.675	-0.771 ± 0.278	0.420 ± 0.043	0.654 ± 0.375
APIC-CNP	0.013 ± 0.003	1.133 ± 0.004	0.247 ± 0.001	0.009 ± 0.006	<u>1.130 ± 0.006</u>	0.298 ± 0.032	<u>0.097 ± 0.020</u>
APIC-LNP	<u>0.015 ± 0.006</u>	<u>1.121 ± 0.017</u>	0.128 ± 0.027	<u>0.011 ± 0.006</u>	1.136 ± 0.009	0.168 ± 0.008	0.081 ± 0.028
APIC-ANP	0.023 ± 0.002	0.932 ± 0.027	<u>0.147 ± 0.021</u>	0.037 ± 0.004	0.663 ± 0.051	<u>0.241 ± 0.063</u>	0.104 ± 0.060

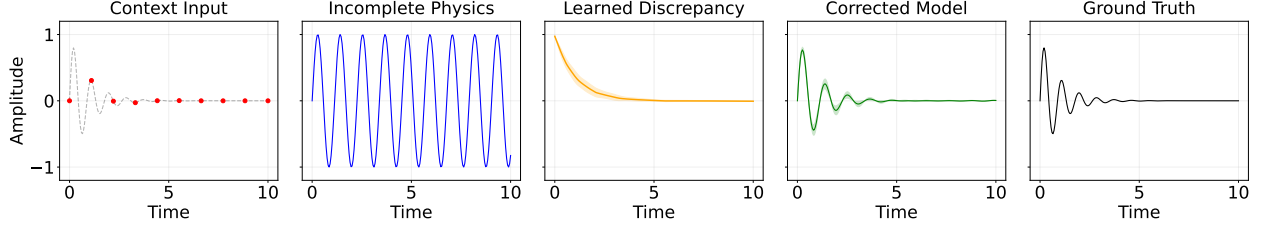


Figure 2: Qualitative correction pipeline for the damped spring system. From left to right: sparse context observations; prediction from the misspecified nominal simulator; learned discrepancy; corrected prediction obtained by combining simulator and discrepancy; and the ground-truth solution. APIC identifies and compensates for the structured model bias.

tive distributions. Uncertainty calibration is assessed using the Expected Calibration Error (ECE), computed as the weighted discrepancy between nominal and empirical coverage across multiple confidence levels Kuleshov et al. [2018]. For discrepancy identification, we compare the inferred discrepancy function to the ground-truth discrepancy and report MAE and LL evaluated over target inputs. Calibration of discrepancy uncertainty is likewise quantified using ECE. Finally, parameter recovery is evaluated using MAE between the inferred parameters and the ground-truth parameters.

4.1 DAMPED SPRING SYSTEM

Dataset. The nominal simulator $\mathcal{F}$ corresponds to an undamped harmonic oscillator, $u''(t) + (\omega)^2 u(t) = 0$. The initial conditions are chosen such that the analytical solution is $\mathcal{F}(t; \omega) = \sin(\omega t)$, where the frequency ω is sampled independently per task from $\mathcal{U}(3, 8)$. The true data-generating process introduces parameter-dependent exponential damping, $\mathcal{G}(t; \omega) = \sin(\omega t) \exp(-\lambda \omega t)$, $\lambda = 0.15$. This introduces a structured discrepancy $\delta(t) = \mathcal{G}(t; \omega) - \mathcal{F}(t; \omega) = \sin(\omega t)(\exp(-\lambda \omega t) - 1)$. Trajectories are evaluated on $t \in [0, 4]$ using $T = 100$ uniformly spaced points. For each task, observations are partitioned into a context set of randomly sampled pairs and a target set comprising the remaining time points. During training, the context size is uniformly sampled from 15 to 100, while at test time, we randomly sample 30 context points. Results are averaged over 200 independent function realizations.

Results. Quantitative results are reported in Table 1. All APIC variants outperform classical calibration methods across reconstruction, discrepancy prediction, and parameter recovery. We observe that APIC achieves a significantly lower discrepancy and parameter MAE while maintaining a higher predictive log-likelihood. Moreover, classical calibration methods are non-amortized; predictions are computed per sample, whereas APIC performs amortized calibration. This indicates that disentangling physical parameters from structured discrepancy mitigates bias under model misspecification. The stochastic variants (APIC-LNP and APIC-ANP) also exhibit improved uncertainty calibration, as demonstrated by lower ECE scores. Figure 2 provides a qualitative analysis of a sample test trajectory using APIC-LNP. The first panel shows the sparse spatio-temporal context observations. The second panel presents the prediction obtained from the misspecified physics model, which fails to capture the systematic damping term. The third panel displays the learned discrepancy field, which accurately identifies the structured bias between the nominal simulator and the true dynamics. When this correction is combined with the simulator output (fourth panel), the resulting prediction closely matches the ground-truth field shown in the final panel. These results demonstrate that APIC successfully isolates and learns structured model error, enabling accurate recovery of the full spatio-temporal dynamics despite misspecified physics and sparse observations.

In terms of computational time, during inference, APIC-LNP requires only 0.42 ms per realization at test time, whereas BCPI requires 12.6 s, KOH-GP requires 33.6 s, and

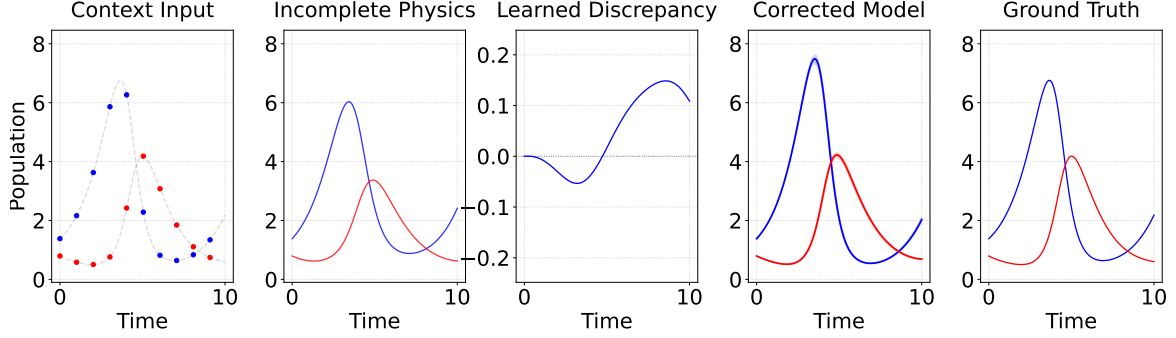


Figure 3: Qualitative correction pipeline for the Lotka-Volterra ODE System. From left to right: sparse context observations; prediction from the misspecified nominal simulator; learned discrepancy; corrected prediction obtained by combining simulator and discrepancy; and the ground-truth solution. The blue plot refers to the **prey** and the red plot is the **predator**.

Table 2: Quantitative results for Lotka-Volterra ODE System. $\dagger$ methods are corrected per realization and not amortized as our method. Best results are shown in bold, and the second-best results are underlined.

Method	Reconstruction			Discrepancy (δ) prediction			Parameter (θ) prediction
	MAE ($\downarrow$)	LL ($\uparrow$)	ECE ($\downarrow$)	MAE ($\downarrow$)	LL ($\uparrow$)	ECE ($\downarrow$)	MAE ($\downarrow$)
KOH-GP †	0.197 ± 0.066	0.163 ± 0.026	0.439 ± 0.008	0.197 ± 0.115	0.668 ± 0.438	0.347 ± 0.045	0.163 ± 0.120
BCPI †	0.484 ± 0.028	-4.235 ± 0.917	0.436 ± 0.056	0.284 ± 0.101	-4.762 ± 0.265	0.492 ± 0.008	0.283 ± 0.118
Correction-PINN †	0.088 ± 0.045	-4.834 ± 1.305	0.290 ± 0.076	0.301 ± 0.076	-3.030 ± 0.294	0.488 ± 0.020	0.147 ± 0.037
APIC-CNP	0.544 ± 0.086	0.680 ± 0.329	0.143 ± 0.005	0.048 ± 0.007	0.085 ± 0.016	0.206 ± 0.007	0.049 ± 0.006
APIC-LNP	0.036 ± 0.003	1.454 ± 0.177	0.251 ± 0.009	0.030 ± 0.007	<u>1.268 ± 0.153</u>	0.279 ± 0.042	<u>0.059 ± 0.005</u>
APIC-ANP	<u>0.061 ± 0.005</u>	<u>1.369 ± 0.164</u>	<u>0.248 ± 0.003</u>	<u>0.048 ± 0.007</u>	1.738 ± 0.071	<u>0.273 ± 0.016</u>	0.102 ± 0.016

Correction-PINN requires 150 s (with 20 ensembles) per realization, which represents a reduction of approximately 4-5 orders of magnitude.

Ablation analyses on the impact of context point size, observation noise, regularization, λ_θ and λ_δ are provided in Appendix B.2.

4.2 LOTKA-VOLTERRA ODE SYSTEM

Dataset. The nominal simulator $\mathcal{F}$ corresponds to the Lotka-Volterra system with state $u(t) = (x(t), y(t))$. $x(t)$ denotes the prey population and $y(t)$ denotes the predator population.

$$\dot{x} = \alpha x - \beta xy, \quad \dot{y} = \eta xy - \gamma y.$$

The task-specific parameters $\theta = (\alpha, \beta, \eta, \gamma)$ are independently sampled from $\mathcal{U}(0.5, 1.5)$ along with the initial condition $u_0 \sim \mathcal{U}(0.7, 1.8)$. The true dynamics $\mathcal{G}$ introduce a structured additive forcing in the prey channel,

$$\dot{x} = \alpha x - \beta xy + \delta_x(t), \quad \dot{y} = \eta xy - \gamma y,$$

where $\delta_x(t) = \Phi(t)^\top w$ is parameterized by an RBF time basis $\Phi(t)$ with $\Phi(0) = 0$. The coefficients w are generated via a fixed systemic map from task features $[\log(\theta), u_0]$, yielding structured, task-dependent discrepancy with $\delta_x(0) = 0$.

Both systems are integrated using RK4 on $t \in [0, 10]$ with $T = 100$ time steps. At training, the context size is uniformly sampled from 20 to 100 and fixed to 30 at test time. Results are averaged over 200 independently sampled tasks.

Results. Table 2 provides quantitative results. APIC-LNP achieves the best overall performance, with the lowest reconstruction MAE and highest log-likelihood. At the same time, it accurately captures the discrepancy and recovers parameters. APIC-ANP performs competitively with APIC-LNP, whereas APIC-CNP exhibits a lower performance in terms of reconstruction and discrepancy MAE and LL, possibly due to its deterministic aggregation. Classical per-instance methods show reasonable reconstruction but higher discrepancy and parameter errors. Qualitative reconstructions for the correction pipeline for predator and prey are provided in Figure 3 using APIC-LNP. APIC produces predictions closely aligned with the ground truth. Additional qualitative results for uncertainty and discrepancy estimation are provided in Appendix B.1.

4.3 ADVECTION-DIFFUSION PDE

Dataset. We consider a periodic one-dimensional field $u(t, x)$ on $x \in [0, 1)$ governed by the nominal advec-

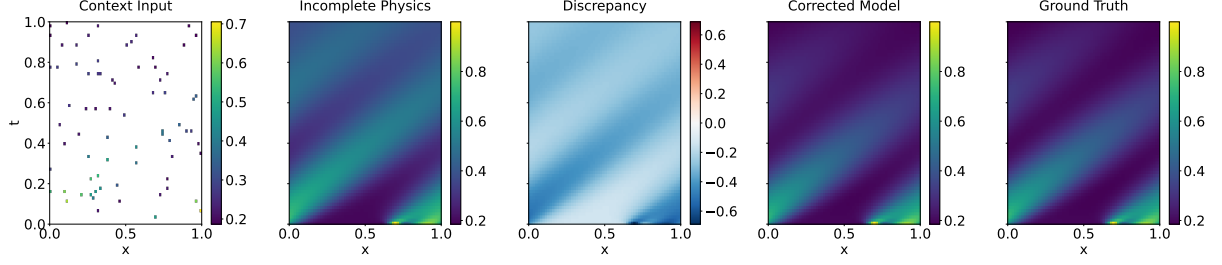


Figure 4: Qualitative correction pipeline for the advection–diffusion system. From left to right: sparse context observations; prediction from the misspecified nominal simulator; learned discrepancy field; corrected prediction obtained by combining simulator and discrepancy; and the ground-truth solution. APIC identifies and compensates for the structured model bias, recovering the true spatio-temporal dynamics.

Table 3: Quantitative results for advection-diffusion system. $\dagger$ methods are corrected per realization and not amortized as our method. Best results are shown in bold, and the second-best results are underlined.

Method	Reconstruction			Discrepancy (δ) prediction			Parameter (θ) prediction MAE	
	MAE ($\downarrow$)	LL ($\uparrow$)	ECE ($\downarrow$)	MAE ($\downarrow$)	LL ($\uparrow$)	ECE ($\downarrow$)	κ ($\downarrow$)	v ($\downarrow$)
KOH-GP †	0.128 ± 0.047	0.276 ± 0.179	0.420 ± 0.017	0.214 ± 0.062	0.628 ± 0.083	0.382 ± 0.022	0.865 ± 0.187	0.759 ± 0.241
BCPI †	0.031 ± 0.019	0.933 ± 0.212	0.896 ± 0.357	0.362 ± 0.176	-1.336 ± 1.059	0.934 ± 0.395	0.020 ± 0.019	0.390 ± 0.294
Correction-PINN †	0.021 ± 0.003	-1.552 ± 0.523	0.136 ± 0.072	0.070 ± 0.018	-3.259 ± 0.655	0.268 ± 0.100	0.021 ± 0.020	0.121 ± 0.086
APIC-CNP	0.054 ± 0.011	1.177 ± 0.186	0.041 ± 0.008	0.025 ± 0.014	1.475 ± 0.178	0.170 ± 0.060	<u>0.019 ± 0.001</u>	0.371 ± 0.084
APIC-LNP	0.010 ± 0.002	2.713 ± 0.224	0.064 ± 0.004	0.007 ± 0.006	2.925 ± 0.412	0.209 ± 0.045	0.004 ± 0.000	0.059 ± 0.008
APIC-ANP	0.053 ± 0.012	<u>1.183 ± 0.191</u>	<u>0.043 ± 0.010</u>	0.026 ± 0.014	<u>1.502 ± 0.167</u>	<u>0.173 ± 0.049</u>	<u>0.019 ± 0.001</u>	0.361 ± 0.091

tion–diffusion equation

$$u_t + v u_x = \kappa u_{xx}, \quad \theta = (v, \kappa),$$

where $v \sim \mathcal{U}(0.3, 2.0)$ and $\kappa \sim \mathcal{U}(0.002, 0.08)$ are sampled independently per task, and the initial condition $u(0, x) = u_0(x)$ is a smooth random field.

The true data-generating process includes an additional parameter-dependent reaction (sink) term,

$$u_t + v u_x = \kappa u_{xx} - c(\theta)u,$$

where the reaction rate follows a systematic law $c(\theta) = \text{softplus}(a^\top \log(\theta) + b)$, clipped to a maximum value. This induces a structured discrepancy $\delta(t, x) = \mathcal{G}(t, x; \theta) - \mathcal{F}(t, x; \theta)$.

Both systems are solved using an exact Fourier spectral method under periodic boundary conditions. Solutions are evaluated on a uniform grid with $N_t = 64$ time steps and $N_x = 64$ spatial points. For each task, context observations are randomly subsampled from the full spatio-temporal grid, and performance is evaluated on the remaining target points.

Results. Table 3 reports quantitative results. APIC-LNP achieves the best overall performance. Parameter recovery further highlights the advantage of amortized inference. APIC-LNP significantly outperforms classical per-instance approaches, particularly in estimating the advection coefficient v , where errors are reduced by an order of magnitude

relative to KOH-GP. APIC-CNP and APIC-ANP yield comparable performance. In contrast, per-instance calibration methods exhibit larger parameter and calibration errors, especially in discrepancy prediction. Figure 4 illustrates the full correction pipeline for the advection–diffusion system. We see that the resulting prediction after accounting for the model discrepancy closely matches the ground-truth field shown in the final panel. Additional qualitative analyses are provided in Appendix B.1.

5 CONCLUSION

We introduced APIC, a population-level extension of the calibration principle in Kennedy and O’Hagan [2001] that replaces per-realization Bayesian calibration with amortized inference across related systems. Instead of solving a new inverse problem for each realization, APIC leverages a dual-latent architecture Neural Process Garnelo et al. [2018b] backbone to map sparse context observations to predictive distributions while separating instance-specific physics parameters from structural discrepancy. Experimental results show that APIC improves reconstruction and discrepancy identification while preserving accurate parameter recovery and calibrated uncertainty in low-data regimes. These findings support amortized calibration as an effective organizing principle for hybrid modeling, reducing the need for repeated per-instance optimization while enabling adaptation of imperfect models to observed data. APIC assumes differentiable simulator models, sufficient simulation data

for meta-training, and appropriate choices of discrepancy parameterization, which may limit applicability in black-box, simulation-scarce, or structurally misspecified settings. Addressing these constraints provides a natural direction for future work. We discuss this further in Appendix C.

References

- M. J. Bayarri, J. O. Berger, J. Cafeo, G. Garcia-Donato, F. Liu, J. Palomo, R. J. Parthasarathy, R. Paulo, J. Sacks, and D. Walsh. Computer model validation with functional output. *The Annals of Statistics*, 35(5):1874–1906, 2007. ISSN 00905364. URL <http://www.jstor.org/stable/25464566>.
- Vitus Benson, Ana Bastos, Christian Reimers, Alexander J Winkler, Fanny Yang, and Markus Reichstein. Atmospheric transport modeling of co2 with neural networks. *Journal of Advances in Modeling Earth Systems*, 17(2):e2024MS004655, 2025.
- Jenny Brynjarsdottir and Anthony O’Hagan. Learning about physical parameters: the importance of model discrepancy. *Inverse Problems*, 30, 10 2014. doi: 10.1088/0266-5611/30/11/114007.
- Giovanni Catalani, Siddhant Agarwal, Xavier Bertrand, Frederic Tost, Michael Bauerheim, and Joseph Morlier. Neural fields for rapid aircraft aerodynamics simulations. *Scientific Reports*, 14(1):25496, 2024.
- Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. doi: 10.1073/pnas.1912789117. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1912789117>.
- Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International conference on machine learning*, pages 1704–1713. PMLR, 2018a.
- Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.
- Pedro J Gonçalves, Jan-Matthis Lueckmann, Michael Deistler, Marcel Nonnenmacher, Kaan Öcal, Giacomo Bassetto, Chaitanya Chintaluri, William F Podlaski, Sara A Haddad, Tim P Vogels, David S Greenberg, and Jakob H Macke. Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLife*, 9:e56261, sep 2020. ISSN 2050-084X. doi: 10.7554/eLife.56261. URL <https://doi.org/10.7554/eLife.56261>.
- Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.
- Dave Higdon, Marc Kennedy, James C. Cavendish, John A. Cafeo, and Robert D. Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466, 2004. doi: 10.1137/S1064827503426693. URL <https://doi.org/10.1137/S1064827503426693>.
- Saurav Jha, Dong Gong, Xuesong Wang, Richard E. Turner, and Lina Yao. The neural process family: Survey, applications and perspectives, 2023. URL <https://arxiv.org/abs/2209.00517>.
- Marc C. Kennedy and Anthony O’Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001. doi: <https://doi.org/10.1111/1467-9868.00294>. URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00294>.
- Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.
- Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR, 2018.
- Ching-Yao Lai, Pedram Hassanzadeh, Aditi Sheshadri, Maike Sonnewald, Raffaele Ferrari, and Venkatramani Balaji. Machine learning for climate physics and simulations. *Annual Review of Condensed Matter Physics*, 16(1):343–365, 2025.
- Zong-Yi Li, Nikola B. Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *ArXiv*, abs/2010.08895, 2020. URL <https://api.semanticscholar.org/CorpusID:224705257>.
- Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deepnet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, Mar 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00302-5. URL <https://doi.org/10.1038/s42256-021-00302-5>.
- Matthew Plumlee. Bayesian calibration of inexact computer models. *Journal of the American Statistical Association*, 112(519):1274–1285, 2017. doi: 10.1080/

- 01621459.2016.1211016. URL <https://doi.org/10.1080/01621459.2016.1211016>.
- Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part ii): Data-driven discovery of nonlinear partial differential equations, 11 2017.
- Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.
- Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and F Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743): 195–204, 2019.
- Michail Spitieris and Ingelin Steinsland. Bayesian calibration of imperfect computer models using physics-informed priors. *J. Mach. Learn. Res.*, 24(1), January 2023. ISSN 1532-4435.
- Naoya Takeishi and Alexandros Kalousis. Physics-integrated variational autoencoders for robust and interpretable generative modeling. *Advances in Neural Information Processing Systems*, 34:14809–14821, 2021.
- Rui Tuo and C. F. Jeff Wu. Efficient calibration for imperfect computer models. *The Annals of Statistics*, 43 (6):2331–2352, 2015. ISSN 00905364. URL <http://www.jstor.org/stable/43818853>.
- Sai Karthikeya Vemuri, Tim Büchner, Julia Niebling, and Joachim Denzler. Functional tensor decompositions for physics-informed neural networks. In *Pattern Recognition: 27th International Conference, ICPR 2024, Kolkata, India, December 1–5, 2024, Proceedings, Part XXV*, page 32–46, Berlin, Heidelberg, 2024. Springer-Verlag. doi: 10.1007/978-3-031-78389-0_3.
- Aishwarya Venkataramanan and Joachim Denzler. Distance-informed neural processes. *Advances in Neural Information Processing Systems*, 38:12456–12485, 2026.
- Zheng Wang, Wei Xing, Robert Kirby, and Shandian Zhe. Physics informed deep kernel learning. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 1206–1218. PMLR, 28–30 Mar 2022. URL <https://proceedings.mlr.press/v151/wang22a.html>.
- Liu Yang, Xuhui Meng, and George Em Karniadakis. B-pinns: Bayesian physics-informed neural networks for forward and inverse pde problems with noisy data. *Journal of Computational Physics*, 425:109913, 2021. ISSN 0021-9991. doi: <https://doi.org/10.1016/j.jcp.2020.109913>. URL <https://www.sciencedirect.com/science/article/pii/S0021999120306872>.
- Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.
- Zongren Zou, Xuhui Meng, and George Em Karniadakis. Correcting model misspecification in physics-informed neural networks (pinns). *Journal of Computational Physics*, 505:112918, 2024. ISSN 0021-9991. doi: <https://doi.org/10.1016/j.jcp.2024.112918>. URL <https://www.sciencedirect.com/science/article/pii/S0021999124001670>.

APIC: Amortized Physics-Informed Calibration using Neural Processes (Supplementary Material)

Aishwarya Venkataramanan^{*1}

Sai Karthikeya Vemuri^{*1}

Joachim Denzler¹

¹Computer Vision Group, Friedrich Schiller University Jena, Germany

A IMPLEMENTATION DETAILS

Damped Spring System. For the damped spring system, we employ two independent multilayer perceptron (MLP) latent encoders for physics and discrepancy inference. Each encoder first flattens the context trajectory and processes it through a two-layer MLP with hidden dimensions $128 \rightarrow 64$, producing a hidden embedding of dimension 64. Linear heads output the mean and log-variance of 1-dimensional Gaussian latent variables. The physical frequency ω is inferred from the physics latent via a two-layer MLP with hidden dimension 128. The discrepancy function is modeled multiplicatively. The predictive distribution is Gaussian with learned mean and variance. Observation noise in the decoder is modeled with a Gaussian distribution with diagonal covariance Σ . Training proceeds in two stages. In the first stage (3,000 steps), the model is trained to recover physical parameters from trajectories without discrepancy correction to anchor parameter identification. In the second stage (4,000 steps), full variational training is performed, jointly optimizing parameter inference and discrepancy modeling. Both stages use a batch size of 64.

Lotka–Volterra ODE System. We employ a stochastic MLP encoder that maps observed predator–prey trajectories to a Gaussian latent variable of dimension 64. The encoder first flattens the input trajectory and processes it through a three-layer MLP with hidden dimensions $512 \rightarrow 256 \rightarrow 256$, producing a hidden embedding of dimension 256. Linear heads produce the mean and log-variance of the latent distribution. Physical parameters $\theta = (\alpha, \beta, \delta, \gamma)$ are predicted from the latent via a linear decoder in log-space with bounded sigmoid scaling. The discrepancy head consists of a two-layer MLP (hidden size 128) with Softplus activation. Corrected dynamics are integrated using a differentiable RK4 solver. Observation noise in the decoder is modeled with a Gaussian distribution with diagonal covariance Σ . The first stage is trained for 3000 steps, and the second stage for 4000 steps. The batch size is 64 for both stages.

Advection-Diffusion PDE. For the advection-diffusion benchmark, the latent encoders employ a three-layer MLP with hidden dimensions $512 \rightarrow 256 \rightarrow 256$, producing a hidden embedding of dimension $d_h = 256$. A linear layer maps this embedding to produce the mean and log-variance of a Gaussian latent distribution with dimension 64. Physical parameters (v, κ) are inferred from the latent variable via a linear head, constrained to valid ranges through a sigmoid transformation in log-space. The discrepancy is parameterized through a three-layer MLP with hidden dimension 256 and outputs the mean and variance. Observation noise in the decoder is modeled with a Gaussian distribution with diagonal covariance Σ . The first stage is trained for 4000 steps, and the second stage for 10000 steps. The batch size is 64 for both stages.

B ADDITIONAL RESULTS

B.1 QUALITATIVE RESULTS

B.1.1 Lotka–Volterra ODE System

Figure 5 shows the reconstructed prey and predator populations using APIC-LNP, along with the predicted uncertainty and the ground truth. We observe that the model accurately captures the oscillatory dynamics of both populations, and the uncertainty covers the observed data points. We further examine the learned discrepancy functions for the Lotka–Volterra system. This is illustrated in Figure 6. We observe that APIC can capture the structured deviation between the nominal simulator and the true dynamics. The learned discrepancy closely follows the ground-truth correction across time. Uncertainty bands appropriately widen in regions with limited context observations, while remaining sharp where data is informative.

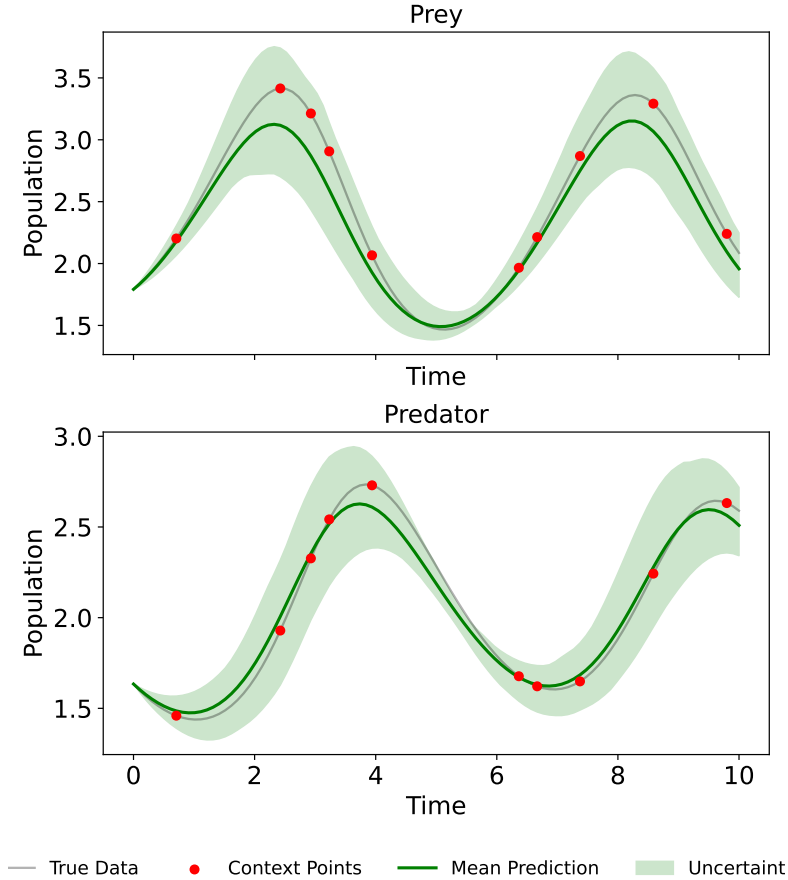


Figure 5: Qualitative reconstruction results for the Lotka–Volterra system. Ground-truth predator and prey populations are shown alongside APIC’s posterior predictive mean and ± 2 standard deviations. The model accurately captures the oscillatory dynamics, and the uncertainty bands reliably cover the true trajectories under sparse observations.

B.1.2 Advection-Diffusion PDE

Figure 7 shows qualitative spatial reconstructions at four representative time snapshots using APIC-LNP. It accurately captures the transported and diffused profiles across time, closely matching the ground truth while exhibiting well-calibrated uncertainty.

Figure 8 further analyzes the learned discrepancy field. We compare the predicted discrepancy with the ground-truth correction and visualize the absolute error between them. The model accurately recovers the dominant spatio-temporal structure of the discrepancy. Importantly, regions exhibiting higher absolute error coincide with increased predictive

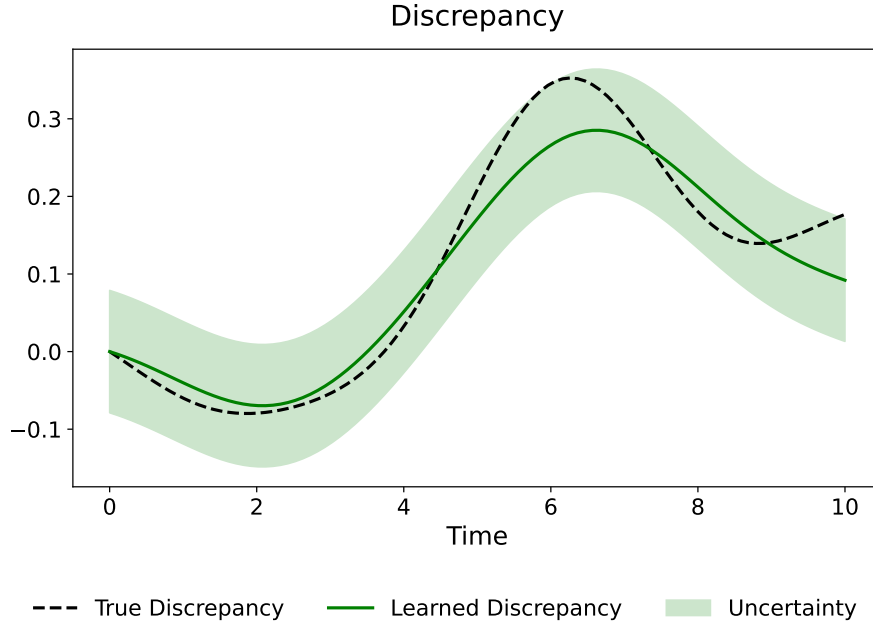


Figure 6: Qualitative discrepancy prediction for the Lotka–Volterra system. The learned discrepancy for predator and prey populations is shown alongside the ground-truth correction and predictive uncertainty (± 2 standard deviations).

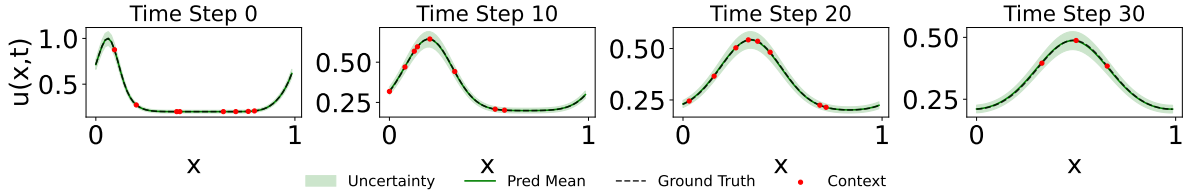


Figure 7: Qualitative reconstruction for the advection–diffusion system using APIC-LNP. Each panel shows the spatial field $u(x, t)$ at different time steps. The predictive mean closely follows the ground truth, while uncertainty bands (± 2 standard deviations) cover the true reconstructions.

uncertainty, indicating that the model appropriately expresses uncertainty where the discrepancy is more difficult to infer. This alignment between error and uncertainty suggests that the learned posterior provides meaningful and well-calibrated confidence estimates.

B.2 ABLATION STUDY

In this study, we assess how the different components of APIC impact its performance. All experiments are performed on the damped-spring system using APIC-LNP.

Effect of Context Set Size. We first investigate how the number of context observations influences predictive performance on the damped spring system. The number of context points is varied from 10 to 100, and we report MAE for reconstruction, discrepancy prediction, and physics parameter estimation.

As shown in Figure 9, increasing the number of context points significantly improves performance in the low-data regime. In particular, all three error metrics exhibit a sharp decrease as the context size increases up to approximately 30 points. Beyond this threshold, improvements gradually saturate, and the MAE stabilizes. This suggests that APIC efficiently extracts task-level information from relatively few observations, and that around 30 context points are sufficient to reliably identify both the underlying dynamics and the discrepancy structure for this system.

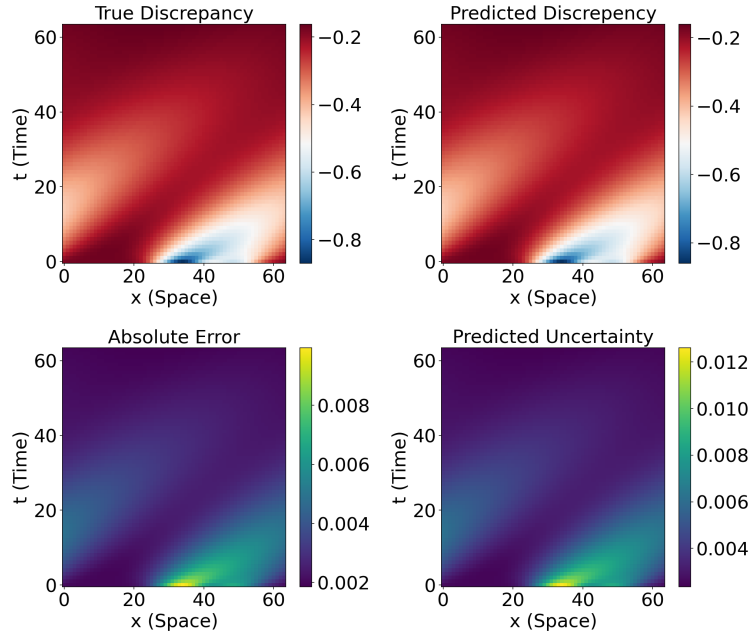


Figure 8: Discrepancy analysis for the advection–diffusion system. The plot shows the predicted discrepancy, the ground-truth discrepancy, the absolute error, and the predictive uncertainty. Regions with higher prediction error correspond to increased uncertainty, indicating calibrated confidence in the learned correction.

Effect of Observation Noise. We further evaluate robustness to observation noise by varying the noise level added to the damped spring observations. As shown in Figure 10, reconstruction, discrepancy, and parameter errors increase progressively with noise magnitude.

While performance degrades as expected, the increase is gradual rather than abrupt, indicating that APIC maintains stable inference under moderate noise levels. In particular, parameter estimation remains relatively robust in the low-to-medium noise regime, suggesting that the latent-variable formulation effectively separates structural dynamics from observational perturbations.

Effect of λ_θ and λ_δ . We investigate the impact of the regularization parameters λ_θ (green curves) and λ_δ (yellow curves) in Figure 11 on three key metrics: reconstruction MAE, discrepancy MAE, and physics parameter θ MAE. Each metric is visualized with two curves: one showing the effect of varying λ_θ while keeping $\lambda_\delta = 1$, and the other showing the effect of varying λ_δ with $\lambda_\theta = 1$.

Reconstruction MAE Figure 11 (a): We observe that reconstruction error remains relatively stable across the range of λ_θ and λ_δ , with a slight decrease as either parameter increases. Notably, the reduction in reconstruction MAE is more pronounced when increasing λ_θ , suggesting that emphasizing the physics parameter regularization slightly improves the reconstruction quality.

Discrepancy MAE Figure 11 (b): The discrepancy MAE exhibits a stronger dependence on the regularization parameters. When $\lambda_\theta = 0$, the error is initially high, but increases in λ_θ result in a sharp drop in MAE, particularly at $\lambda_\theta = 0.2$, after which the error gradually saturates with a modest decrease for larger values. In contrast, increasing λ_δ leads to a smoother, more gradual reduction in discrepancy MAE, indicating a consistent improvement in aligning predictions with observations as the discrepancy regularization is strengthened.

Physics Parameter θ MAE Figure 11 (c): The trend for parameter MAE closely mirrors that of discrepancy MAE. Increasing λ_θ produces a rapid initial reduction in error followed by saturation, while higher λ_δ values yield a steady, gradual decline in MAE. This suggests that both regularization terms contribute to better estimation of the underlying physics parameters, with λ_θ having a more immediate impact and λ_δ providing stable long-term improvements.

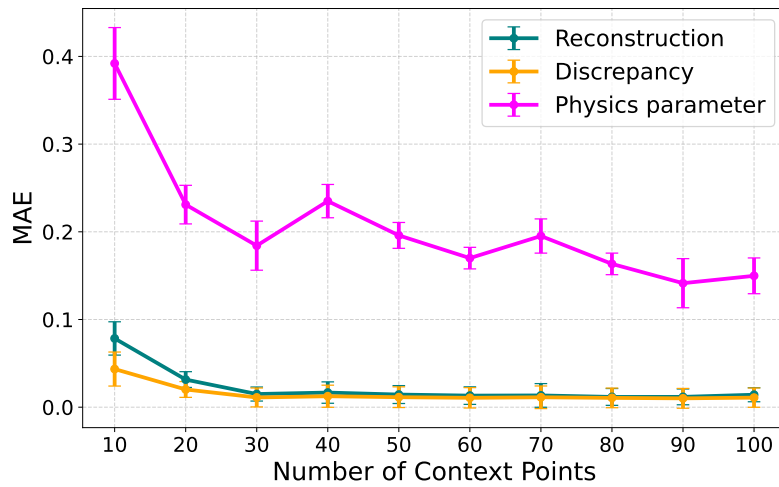


Figure 9: Effect of context set size on predictive performance for the damped spring system. Reconstruction, discrepancy, and parameter MAE are shown as the number of context points increases from 10 to 100. Errors decrease sharply in the low-data regime and stabilize beyond approximately 30 context points.

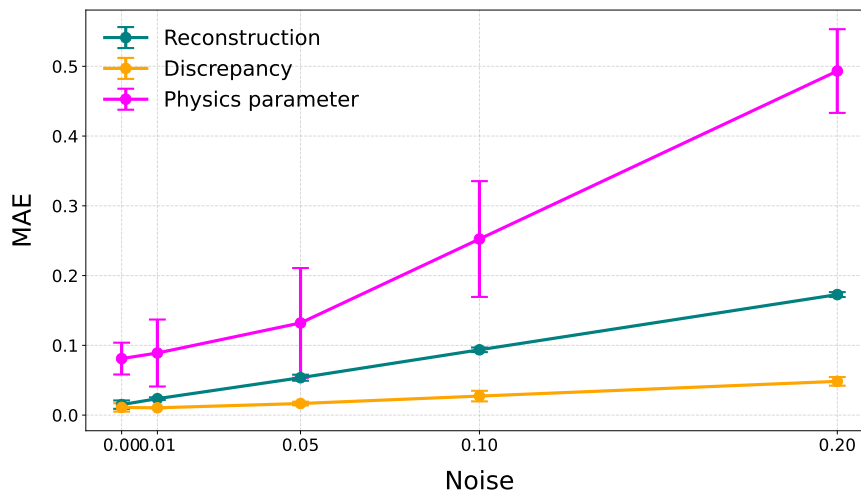


Figure 10: Effect of observation noise on predictive performance for the damped spring system. Reconstruction, discrepancy, and parameter MAE are reported as noise magnitude increases. Errors grow gradually with noise, indicating stable and robust inference under moderate perturbations.

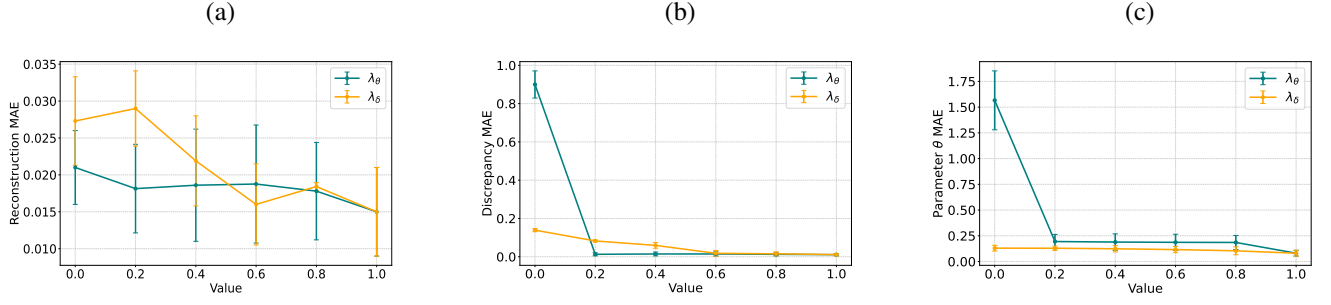


Figure 11: Ablation analysis on the influence of λ_θ (green) and λ_δ (yellow) on (a) reconstruction MAE, (b) Discrepancy MAE and (c) Physics parameter θ MAE. With increasing values of λ_θ and λ_δ , the MAE values decrease.

Table 4: Ablation study on the different regularization methods for the discrepancy network (δ).

Regularization	Reconstruction MAE	δ MAE	θ MAE
L1	0.045 $\pm$ 0.017	0.077 $\pm$ 0.034	0.204 $\pm$ 0.106
L2	0.015$\pm$0.006	0.011$\pm$0.006	0.081$\pm$0.028
Orthogonal reg.	0.029 $\pm$ 0.009	0.023 $\pm$ 0.019	0.198 $\pm$ 0.086
Spectral weight reg.	0.021 $\pm$ 0.008	0.015 $\pm$ 0.010	0.121 $\pm$ 0.099
Gradient penalty	0.037 $\pm$ 0.002	0.084 $\pm$ 0.015	0.244 $\pm$ 0.147

Regularization methods. In this study we evaluate the different regularization strategies $\mathcal{R}$ to obtain $\mathcal{L}_\delta$. L_1 and L_2 regularization penalize large discrepancies, encouraging sparsity (L_1) or smooth, small-magnitude discrepancies (L_2). Orthogonal regularization encourages the discrepancy network to be orthogonal to the physics latent, promoting disentanglement. This is given as $\|\mathbf{z}^{(\delta)T} \cdot \mathbf{z}^{(\theta)}\|^2$. It encourages the discrepancy network to represent residual variation that is independent of the physics latent, thereby promoting disentanglement. Spectral weight regularization [Venkataramanan and Denzler, 2026] bounds the largest and smallest singular values of the weight matrices to prevent the discrepancy network from dominating the physics contribution. Gradient penalty [Gulrajani et al., 2017] discourages large gradients of the discrepancy with respect to the inputs, promoting smooth outputs. As shown in Table 4, L2 and spectral weight regularization achieve the best balance, producing low reconstruction and discrepancy errors while preserving accurate recovery of the physics parameters. Orthogonal regularization can enforce disentanglement but may reduce the accuracy of physics parameters, whereas gradient penalties overly constrain the network, resulting in higher errors. These results suggest that moderate capacity constraints are more effective than strict disentanglement or smoothness penalties for encouraging separation of physics and discrepancy roles.

Effect of number of trajectory sample. This experiment studies how APIC performance scales with available data for the damped spring system. The results are shown in Table 5. Performance improves with more training realizations but begins to plateau beyond 64,000, suggesting that the gains are not simply due to data volume but reflect the amortized inference framework learning shared structure across realizations.

Table 5: Ablation study on the number of trajectory samples.

N	Recon MAE	Param MAE	LL
8000	0.033 $\pm$ 0.015	1.263 $\pm$ 0.027	0.424 $\pm$ 0.019
16,000	0.019 $\pm$ 0.010	0.625 $\pm$ 0.032	0.778 $\pm$ 0.017
32,000	0.019 $\pm$ 0.008	0.241 $\pm$ 0.024	0.751 $\pm$ 0.018
64,000	0.017 $\pm$ 0.008	0.118 $\pm$ 0.028	0.883 $\pm$ 0.020
96,000	0.016 $\pm$ 0.006	0.102 $\pm$ 0.022	0.966 $\pm$ 0.019
128,000	0.015 $\pm$ 0.006	0.081 $\pm$ 0.028	1.121 $\pm$ 0.017

C LIMITATIONS AND FUTURE WORK

Our current formulation assumes that the nominal simulator is a white-box, differentiable end-to-end model, allowing gradients to flow through the physics solver during training and calibration. This assumption is increasingly reasonable in modern hybrid modeling, where differentiable programming and neural surrogates are commonly used. Still, many high-fidelity industrial simulators remain black-box (e.g., legacy codes, proprietary solvers, or pipelines with non-differentiable components), which limits the direct applicability of our approach. A natural direction is to extend the framework to black-box simulators, for example, via gradient-free calibration, implicit differentiation with adjoints when available, surrogate-assisted optimization, or likelihood-free/simulation-based inference.

We assume that phase-1 training can access many simulator samples, implicitly treating simulation as cheap. This is a limitation for settings with expensive forward solves, large parameter spaces, or restricted access to simulation. Future work should address efficient learning under limited simulation budgets, including active task selection, multifidelity training, and the use of pretrained foundation surrogates. In particularly constrained regimes, amortized “zero-shot” calibration may effectively become a few-shot adaptation problem, where a pretrained calibrator is fine-tuned on a small number of observations per task. A principled treatment of this continuum from zero-shot to few-shot to per-task optimization would clarify when amortization is beneficial and how to allocate compute between pretraining and adaptation.

Like classic Kennedy–O’Hagan-style approaches, we treat the discrepancy as global, i.e., potentially present everywhere in the domain. In many scientific systems, misspecification is localized: confined to a subregion, boundary segment, time window, or regime (e.g., turbulence regions, contact events, switching dynamics). A promising extension is to incorporate localized discrepancy priors or gating mechanisms, allowing expert knowledge to specify candidate subdomains where discrepancy can act, or to automatically learn sparse/compact supports. This could improve identifiability and yield more interpretable corrections.

We allow different structural forms for the correction, additive or multiplicative, depending on the domain. This choice is consequential: it shapes both the learned correction and its interpretation. For example, in the spring system, a multiplicative envelope is natural, whereas in other settings, an additive forcing term is more appropriate. However, when the true misspecification is not well matched to the chosen form, the learned discrepancy can become harder to interpret. Future work could study model selection over discrepancy parameterizations, mixtures of additive/multiplicative forms, or physics-guided parameterizations that encode conservation laws and invariances.