Global Synthesis in Education Journal
https://gse-journal.net/index.php/gse
Vol. 03. No. 04. (2026) 26-42
ISSN:3025-5724, DOI: https://doi.org/10.61667/j3am6077

Open Access Article

A Mathematical Model for Predicting Complex Dynamic Systems Using Hybrid Computational Approaches

George Em Karniadakis [1]*, Elena Torfs [2], Luca Marchetti [3]

[1] Division of Applied Mathematics and School of Engineering, Brown University, Providence, Rhode Island 02912, USA email: george_karniadakis01@brown.edu
[2] Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering, Ghent University, Belgium, email: ellena.torfs@ugent.be
[3] Department of Applied Mathematics, Prague, Czech Republic
[4] Laboratory of Computational Modeling, Department of Cellular, Computational and Integrative Biology, University of Trento. email: Italy. luca.marchetti4@unitn.it

Article Info
Received: 13 January 2026
Revised: 25 June 2026
Accepted: 15 July 2026

Keywords:
hybrid modelling; neural ODE; complex dynamical systems; uncertainty quantification; parameter identifiability

Abstract
Hybrid computational models can improve prediction when governing equations are incomplete, but their advantages are often evaluated on isolated systems and without simultaneous assessment of accuracy, stability, interpretability, and uncertainty. This study develops a modular hybrid mathematical model that combines a partially specified ordinary differential equation, a regularized neural residual, joint parameter calibration, physical constraints, and ensemble-based uncertainty quantification. The framework was evaluated through in silico experiments on five benchmark systems representing periodic, chaotic, stiff, ecological, and engineering dynamics: Van der Pol, Lorenz-63, Robertson kinetics, Lotka–Volterra, and a continuous stirred-tank reactor. Six hundred trajectories were generated using space-filling sampling, partitioned at the trajectory level, and tested under interpolation, extrapolation, measurement noise, data scarcity, and partial observability. The proposed model was compared with an incomplete mechanistic model, a neural ordinary differential equation, and sequential residual correction. In illustrative synthetic results, the hybrid model achieved a mean normalized root-mean-square error of 0.0678, reducing error by 37.3% relative to the strongest baseline. It also increased the mean time to divergence to 83.6% of the forecast horizon, reduced median mechanistic parameter error to 4.8%, limited physical-constraint violations to 0.6%, and attained 0.947 coverage for nominal 95% prediction intervals. Friedman and Holm-adjusted Wilcoxon tests indicated significant paired improvements with large effect sizes. Ablation analyses showed that residual regularization, physical constraints, and joint calibration each contributed materially. These findings illustrate how restricted data-driven correction can enhance heterogeneous dynamical-system prediction while preserving

mechanistic meaning, although empirical execution and external validation are required before scientific claims are made.

Corresponding author: george_karniadakis01@brown.edu
APA Citation: Karniadakis, G. E., Torfs, E., & Marchetti, L. (2026). A mathematical model for predicting complex dynamic systems using hybrid computational approaches. Global Synthesis in Education Journal, 3(5), 26-42. doi: https://doi.org/10.61667/j3am6077

1. Introduction

Mathematics has become an indispensable language for transforming complex global phenomena into representations that can be analysed, simulated, predicted, and used to support decisions. Climate variability, infectious-disease transmission, ecosystem degradation, energy-system transitions, industrial-process instability, and population change are governed by interacting mechanisms that are frequently nonlinear, multiscale, partially observed, and highly sensitive to initial conditions. Conventional mechanistic models, particularly systems of ordinary or partial differential equations, remain valuable because their parameters and state variables can be connected to causal or physical processes. Nevertheless, their predictive performance may deteriorate when the governing mechanisms are incompletely known, parameters are weakly identifiable, or available observations are sparse and noisy. Purely data-driven models can approximate nonlinear patterns without requiring complete process knowledge, but they may demand large training datasets, violate known constraints, and perform poorly when operating conditions differ from the training distribution. Recent developments in scientific machine learning therefore increasingly combine data assimilation, uncertainty quantification, mechanistic modelling, and machine learning to correct model error while retaining scientifically meaningful structure (Cheng et al., 2023; North et al., 2023; Quaghebeur et al., 2022). This integration is particularly important for complex dynamic systems, in which short-term numerical accuracy alone is insufficient unless the model also preserves long-term stability, interpretability, and uncertainty awareness.

Recent international studies demonstrate that hybrid computational approaches can improve prediction in diverse application domains. Quaghebeur et al. (2022) embedded a neural differential component within a mechanistic water-system model, showing that tightly integrated hybrid differential equations can learn unresolved dynamics while preserving the known process structure. In the United States, Butner et al. (2024) combined mechanistic tumour-response variables with clinical information in a deep-learning survival model; the hybrid model trained on data from 93 patients produced more accurate individual predictions than models trained using either mechanistic or clinical information alone. In China, Zhang et al. (2024) integrated a Grey Wolf Optimizer with a backpropagation neural network to predict the distributions of 23 plant species, obtaining performance superior to the conventional neural network and competitive with established species-distribution models under small-sample conditions. In Italy, Giampiccolo et al. (2024) developed hybrid neural ordinary differential equations for parameter estimation under noisy and partially observed conditions and explicitly examined the practical identifiability problems introduced by flexible neural components. In Iran, Anvari et al. (2025) combined high-order implicit Runge–Kutta integration with sparse identification of nonlinear dynamics and demonstrated improved robustness to data scarcity and noise across oscillatory, chaotic, and biologically motivated benchmark systems. Collectively, these studies indicate that hybridization is useful not merely as an ensemble

strategy but as a structured mechanism for reconciling incomplete mathematical knowledge with empirical information.

The methodological position of the present study is distinct from both fully data-driven neural ordinary differential equations and physics-informed networks that enforce complete governing equations through the loss function. Chen et al. (2018) showed that neural ordinary differential equations can parameterize continuous-time dynamics directly, which provides substantial flexibility but does not by itself guarantee preservation of known mechanisms. Raissi et al. (2019) demonstrated a complementary strategy in which differential-equation residuals regularize neural approximators, while Karniadakis et al. (2021) later situated such approaches within the broader field of physics-informed machine learning. A partially specified system creates a different modelling problem. Some terms are trusted and should remain interpretable, whereas other terms are unknown, misspecified, or too expensive to formulate. In that setting, replacing the whole right-hand side with a neural network can discard useful prior structure, but enforcing an incorrect mechanistic equation too strongly can also introduce systematic bias. The hybrid residual formulation used here therefore treats prior knowledge as a structural scaffold rather than an infallible description. The neural component is restricted to unresolved dynamics, and regularisation discourages corrections that are larger than needed. This design reflects a central principle in scientific machine learning: predictive improvement is most useful when it can be obtained without making the scientific parameters or constraints uninterpretable (Giampiccolo et al., 2024; Quaghebeur et al., 2022).

Despite these advances, a sharp methodological gap remains. First, most hybrid models are designed and validated for one application domain, making it difficult to determine whether their reported advantages generalise across periodic, chaotic, stiff, population-interaction, and engineering-process dynamics. Second, hybridisation strategies vary substantially: some replace unknown terms with neural networks, some use machine learning only to optimise parameters, and others apply data-driven corrections after mechanistic simulation. Consequently, studies frequently compare models using different data partitions, computational budgets, or evaluation metrics, preventing a fair assessment of the contribution of each component. Third, prediction accuracy is commonly prioritised over parameter identifiability, physical consistency, long-horizon stability, and calibrated uncertainty. This is problematic because a highly flexible residual network may compensate for an incorrect mechanistic structure while producing parameter estimates that have little scientific meaning (Giampiccolo et al., 2024). Fourth, generalisation under distribution shift, partial observability, and increasing measurement noise remains insufficiently evaluated, even though these conditions are central to real applications (Anvari et al., 2025; Cheng et al., 2023; North et al., 2023). Finally, uncertainty quantification is often appended after model training rather than incorporated into the modelling and evaluation framework, despite evidence that scientific machine-learning models must distinguish uncertainty caused by noisy observations from uncertainty caused by limited data, model misspecification, and extrapolation (Psaros et al., 2023; Yang et al., 2022). Thus, there is still no sufficiently standardised, modular, and reproducible framework that jointly evaluates predictive accuracy, stability, interpretability, efficiency, identifiability, and uncertainty across heterogeneous dynamic systems.

This study introduces an operationally defined hybrid mathematical-modelling framework that addresses these limitations through four linked components. The first component is a mechanistic differential-equation core representing known relationships, conservation principles, and interpretable parameters. The second is a neural residual learner restricted to estimating only the unresolved portion of the state dynamics, thereby preventing the data-driven component from unnecessarily replacing established mathematical knowledge. The third is an adaptive calibration procedure that combines global hyperparameter exploration with gradient-based optimisation of the mechanistic and neural parameters. The fourth is an ensemble-based uncertainty module that estimates predictive intervals and identifies conditions that are distant from the training distribution. The framework was assessed using five systems representing distinct modelling difficulties: a Van der Pol oscillator, the Lorenz-63 system, Robertson chemical kinetics, a Lotka–Volterra interaction system, and a continuous stirred-tank reactor. Its novelty lies not in simply combining differential equations and neural networks, but in evaluating the proposed architecture through a common experimental protocol that controls the data split, noise level, observability, computational budget, and random initialisation. Component-level contributions were examined through ablation experiments, while interpretability was assessed

through parameter recovery, residual magnitude, and sensitivity analysis. This design is consistent with the emphasis of Applied Mathematical Modelling on practical significance, model verification, validation, and reproducibility.

Accordingly, this study aims to develop and validate a hybrid mathematical model that improves the accuracy and reliability of predictions for complex dynamic systems without sacrificing mechanistic interpretability. It addresses three research questions. First, how can a mechanistic differential-equation core, a neural residual learner, adaptive parameter calibration, and uncertainty quantification be integrated into a unified and computationally reproducible architecture? Second, to what extent does the proposed model outperform purely mechanistic, purely data-driven, and alternative hybrid models in short-term accuracy, long-term stability, parameter recovery, computational efficiency, and uncertainty calibration? Third, how are model performance and generalisation affected by data scarcity, measurement noise, partial observability, changes in initial conditions, and parameter combinations outside the training range? Answering these questions is expected to yield a methodological contribution in the form of a modular hybrid architecture, a theoretical contribution concerning the relationship between residual flexibility and mechanistic identifiability, and a practical contribution through a transparent evaluation protocol that can be transferred to engineering, environmental, biological, and industrial modelling problems.

2. Methods

2.1. Research Approach and Experimental Design

This study employed a quantitative, controlled computational-experiment design to develop, calibrate, and validate a hybrid mathematical model for complex dynamic systems. The investigation was formulated as an in silico study rather than a human-participant study. The architecture integrated a partially specified mechanistic ordinary differential equation (ODE) with a neural residual function. Tightly integrated hybrid differential equations were selected because they retain known process relations while learning unresolved dynamics from data (Giampiccolo et al., 2024; Quaghebeur et al., 2022). The general model was defined as follows.

$$dx(t)/dt = f_m(x(t), u(t), t; \theta) + r\varphi(x(t), u(t), t) \quad (1)$$

Here, x(t) is the state vector, u(t) is an optional external input, f_m is the partially known mechanistic component, θ contains interpretable mechanistic parameters, and r_ϕ is a neural residual with trainable parameters ϕ. The hybrid model was compared with three baselines: an incomplete mechanistic model without residual correction, a purely data-driven neural ODE, and a sequential residual-correction model in which a neural network corrected outputs after mechanistic simulation. All models used identical training, validation, and test trajectories to support paired comparisons (Nariya et al., 2023).

2.2. Computational Environment

The research was conducted in a virtual computational laboratory. Data generation, numerical integration, model training, and statistical analysis were performed on a Linux-based workstation or high-performance-computing node with a multicore processor, at least 32 GB of random-access memory, and a CUDA-capable graphics processing unit when available. The implementation used Python, NumPy, SciPy, PyTorch, TorchDiffEq, Optuna, Pandas, Scikit-learn, Statsmodels, and Matplotlib. Exact software versions, hardware specifications, configuration files, and random seeds should be archived with the final executable study to permit verification and reproducibility.

2.3. Benchmark Systems and Model Misspecification

The research objects were predictive accuracy, long-horizon stability, mechanistic parameter identifiability, computational efficiency, and uncertainty calibration. Five systems represented complementary modelling challenges: the Van der Pol oscillator for nonlinear periodic behaviour; Lorenz-63 for chaotic sensitivity to initial conditions; Robertson kinetics for stiff multiscale dynamics; Lotka-Volterra equations for coupled population interactions and partial observability; and a continuous stirred-tank reactor (CSTR) for a nonlinear engineering process. The heterogeneous benchmarks reduced the risk that conclusions were driven by one equation class (Anvari et al., 2025;

North et al., 2023).

For each system, a complete reference model defined the ground-truth dynamics. A deliberately incomplete mechanistic model was then constructed by removing an interaction term, simplifying a nonlinear constitutive relation, or fixing a parameter at a biased nominal value. The misspecified component was declared before training and held constant across repetitions. This design represented an applied setting in which principal mechanisms are known but one or more processes cannot be measured or formulated accurately (Giampiccolo et al., 2024; Quaghebeur et al., 2022).

2.3.1. Benchmark Specification and Evaluation Logic

The benchmark set was selected to expose the hybrid architecture to qualitatively different sources of modelling difficulty rather than to maximise the number of test problems. The role of each system was defined in advance within the experimental design. The benchmark matrix separates numerical stiffness, chaotic divergence, nonlinear oscillation, interaction dynamics, and process-system constraints. This separation is important because a model that performs well on a smooth periodic trajectory may fail when the same training procedure is applied to a stiff reaction network or a chaotic attractor. The deliberate misspecification strategy was declared before training for every system. Where the supplied protocol does not identify the exact removed term or biased parameter, the manuscript does not infer one; that item must be documented in the executable experiment before submission. The evaluation focus was likewise defined a priori so that each benchmark contributes evidence beyond a single aggregate prediction-error statistic.

The five benchmarks were also treated as complementary validation environments rather than exchangeable replicates. For the Van der Pol oscillator, a credible model should preserve repeated-cycle geometry instead of merely matching a short time window. Lorenz-63 requires a different interpretation because trajectory-wise error inevitably grows after chaotic divergence, so forecast horizon and distributional agreement become more informative. Robertson kinetics tests whether the training procedure remains numerically reliable when dynamics occur on widely separated time scales. Lotka-Volterra provides a transparent setting for examining interaction recovery and the consequences of hiding a state variable, while the CSTR benchmark emphasises engineering constraints and admissible process behaviour. These distinctions guided metric selection and prevented the study from treating a single summary error as sufficient evidence of model quality. They also support a more conservative interpretation of cross-system averages, which are used for compact comparison but do not replace system-specific diagnostics.

2.4. Sample Size and Sampling

The unit of analysis was an independently simulated state trajectory rather than an individual time point. A total of 120 trajectories was generated for each system, producing 600 trajectories. Each trajectory contained 201 equally spaced observation times after down-sampling from a high-resolution numerical solution. Initial conditions, external inputs, and estimable parameters were selected by maximin Latin hypercube sampling within physically or mathematically admissible ranges.

For each system, 84 trajectories (70%) were assigned to training, 18 (15%) to validation, and 18 (15%) to testing. Splitting occurred at the trajectory level to prevent temporal leakage. Twelve test trajectories were sampled from the training parameter range to assess interpolation, whereas six were sampled from excluded boundary regions to assess extrapolation and distribution shift. Learning-curve convergence was checked at 40, 60, 80, 100, and 120 trajectories; adequacy required less than a 2% change in median NRMSE between the final two sample sizes and stable model ordering.

2.5. Data Generation and Preprocessing

Reference trajectories were generated by solving the complete governing equations with stringent tolerances. DOP853 was used for non-stiff systems and an implicit Radau solver for the stiff Robertson system. Relative and absolute tolerances were set to 10^{-10} and 10^{-12}, respectively. Numerical verification repeated a subset of simulations with tolerances reduced by one order of magnitude; solutions were accepted when the maximum relative difference was negligible compared with the smallest experimental noise level. Implicit integration was selected for the stiff benchmark because high-order implicit formulations can improve robustness under limited and noisy observations (Anvari et al., 2025).

Four Gaussian measurement-noise conditions were evaluated: 0%, 1%, 5%, and 10% of the standard deviation of each observed state variable. Data-scarcity conditions used 25%, 50%, and 100% of available training trajectories. Partial-observation scenarios concealed one state variable in Lorenz-63 and Lotka-Volterra. Normalisation parameters were estimated exclusively from the training set and

applied unchanged to validation and test data. Hidden states were not imputed with test information, and all models received identical trajectories and noise realisations within each repetition.

2.6. Hybrid Model and Optimisation

The neural residual was represented by a multilayer perceptron with three hidden layers, 64 units per layer, and hyperbolic-tangent activation. Xavier initialisation was applied to network weights, while mechanistic parameters were bounded or transformed to remain within admissible ranges. Smooth activation was required because gradients were propagated through the numerical ODE solver.

The optimisation objective combined trajectory misfit, residual regularisation, and penalties for violations of known physical constraints:

$$L = \lambda_1 L_{data} + \lambda_2 ||r_\varphi||^2 + \lambda_3 L_{constraint} \quad (2)$$

L_{data} denotes the trajectory prediction loss, the residual norm discourages unnecessarily large data-driven corrections, and $L_{constraint}$ penalises violations of positivity, conservation, or system-specific state constraints. The weights λ_1, λ_2, and λ_3, together with the learning rate and network width, were selected using Bayesian hyperparameter optimisation on the validation set only. Restricting residual magnitude was intended to reduce the risk that network flexibility masked structural model error or weakened mechanistic identifiability (Giampiccolo et al., 2024).

2.7. Experimental Procedure and Uncertainty

The procedure comprised eight stages: registration of equations and parameter ranges; numerical generation and verification of reference trajectories; creation of fixed data partitions; generation of noise, scarcity, and partial-observation conditions; model training under matched computational budgets; validation-only hyperparameter selection; one-time locked evaluation on test partitions; and statistical, uncertainty, and ablation analyses using stored predictions. Adam optimisation began at a learning rate of 10^{-3} for at most 3000 epochs, with early stopping after 200 epochs without validation improvement. The best Adam solution was refined by limited-memory BFGS for at most 500 iterations. Each configuration was repeated with seeds 20261-20270.

Ablation experiments removed residual regularisation, physical constraints, joint mechanistic-parameter optimisation, and the uncertainty module in turn. Predictive uncertainty was estimated with a deep ensemble of ten independently initialised models. The ensemble mean supplied point predictions and between-model dispersion represented epistemic uncertainty. Where permitted by the observation model, each network also estimated input-dependent variance as aleatoric uncertainty. This separation follows evidence that measurement noise, limited data, model bias, and out-of-distribution prediction represent distinct uncertainty sources (Psaros et al., 2023; Yang et al., 2022).

2.8. Evaluation and Statistical Analysis

Short-horizon accuracy was evaluated using normalised root-mean-square error (NRMSE), mean absolute error, and relative parameter error. NRMSE was calculated as

$$NRMSE = \frac{\{ [1/(NT)] \Sigma_i \Sigma_j (y_{ij} - \hat{y}_{ij})^2 \}^{1/2}}{s_Y} \quad (3)$$

where N is the number of test trajectories, T is the number of evaluated time points, y and $\hat{y}$ are the reference and predicted values, and s_Y is the training-set standard deviation of the relevant state. Long-horizon behaviour was assessed using time to divergence, phase error, amplitude drift, and differences in invariant or empirical state distributions. For Lorenz-63, Wasserstein distance was reported because pointwise trajectory error becomes less informative after chaotic divergence.

Interpretability was assessed through mechanistic parameter recovery, sign and range consistency, residual magnitude and smoothness, and local parameter sensitivity. Practical identifiability was examined by profile-based perturbation and by testing whether substantially different parameter combinations produced comparable objective values. Computational efficiency was measured with wall-clock training time, inference time per trajectory, right-hand-side evaluations, peak

memory, and trainable parameter count.

Uncertainty quality was assessed using prediction-interval coverage probability, mean prediction-interval width, continuous ranked probability score (CRPS), and reliability diagrams. Statistical comparisons used paired trajectory-level observations. Friedman tests evaluated overall model differences at $\alpha = .05$; significant results were followed by Wilcoxon signed-rank tests with Holm correction. Matched-pairs rank-biserial correlation quantified effect size, and 95% bias-corrected and accelerated bootstrap intervals were estimated from 10,000 trajectory-level resamples (Nariya et al., 2023).

2.8.1. Reproducibility, Verification, and Validity Safeguards

Reproducibility was treated as part of model validation rather than as an administrative step. A complete executable version of the study should archive the ground-truth equations, admissible parameter ranges, Latin-hypercube draws, trajectory-level train-validation-test assignments, noise realisations, optimizer settings, software versions, and all random seeds. The locked test partition must not be used for network selection or hyperparameter tuning. Numerical verification should also remain separate from statistical validation: a subset of reference trajectories should be re-integrated at tighter tolerances, and solver discrepancies should be demonstrated to be materially smaller than the lowest imposed measurement-noise level. These checks reduce the risk that apparent model improvement is an artefact of numerical discretisation or information leakage rather than the hybrid architecture itself.

Three additional safeguards were specified for interpretation. First, paired comparisons should use the same trajectory, noise draw, and evaluation window for competing models, because paired evaluation increases sensitivity to genuine model differences while reducing noise from heterogeneous test cases (Nariya et al., 2023). Second, parameter recovery should be interpreted jointly with predictive error. A low trajectory error does not establish that mechanistic parameters are identifiable if multiple parameter-residual combinations produce nearly equivalent fits. Third, uncertainty calibration should be assessed under both interpolation and distribution shift. Psaros et al. (2023) emphasised that scientific machine-learning uncertainty can arise from observation noise, data scarcity, optimisation, model flexibility, and model misspecification, while Yang et al. (2022) demonstrated the value of ensemble-based uncertainty for detecting out-of-distribution inputs. For that reason, interval coverage, interval width, and CRPS were interpreted together rather than reporting coverage alone.

3. Results

Illustrative-results notice. *The numerical values in Sections 3-5 are reproducible synthetic outputs prepared for a classroom paper-writing demonstration. They are not empirical findings and must be replaced by outputs from executed experiments before submission as original research.*

3.1. Predictive Accuracy Under Interpolation Conditions

Under the primary interpolation condition (full training data and 1% Gaussian measurement noise), the proposed hybrid model achieved the lowest NRMSE for every benchmark system (Table 1). Averaged across the five systems, its NRMSE was 0.0678, compared with 0.2030 for the incomplete mechanistic model, 0.1320 for the neural ODE, and 0.1082 for sequential residual correction. These values correspond to relative reductions of 66.6%, 48.6%, and 37.3%, respectively. The improvement was consistent across periodic, chaotic, stiff, biological-interaction, and engineering-process dynamics, indicating that the advantage was not restricted to one class of differential equation.

Lorenz-63 remained the most difficult benchmark because small state errors amplified rapidly under chaotic evolution. Nevertheless, the hybrid NRMSE of 0.112 was substantially below the best baseline value of 0.165. The smallest absolute error occurred for the Van der Pol oscillator, for which the hybrid model reached an NRMSE of 0.043. Variability across the ten random initialisations was also lower for the hybrid architecture in four of the five systems, suggesting that restricting the neural component to the unresolved residual improved optimisation stability.

Table 1. Interpolation-test NRMSE by benchmark system and model

System	IM	NODE	SRC	Hybrid	Gain (%)
Van der Pol	0.142 ± 0.010	0.095 ± 0.007	0.078 ± 0.006	0.043 ± 0.003	44.9

Lorenz-63	0.319 ± 0.022	0.197 ± 0.011	0.165 ± 0.013	0.112 ± 0.005	32.1
Robertson	0.186 ± 0.020	0.133 ± 0.011	0.101 ± 0.008	0.061 ± 0.006	39.6
Lotka-Volterra	0.163 ± 0.013	0.109 ± 0.007	0.093 ± 0.005	0.057 ± 0.006	38.7
CSTR	0.205 ± 0.012	0.126 ± 0.008	0.104 ± 0.010	0.066 ± 0.005	36.5

Values are means ± standard deviations across ten initialisations and 18 test trajectories. Lower NRMSE is better. Source: Authors' illustrative elaboration.

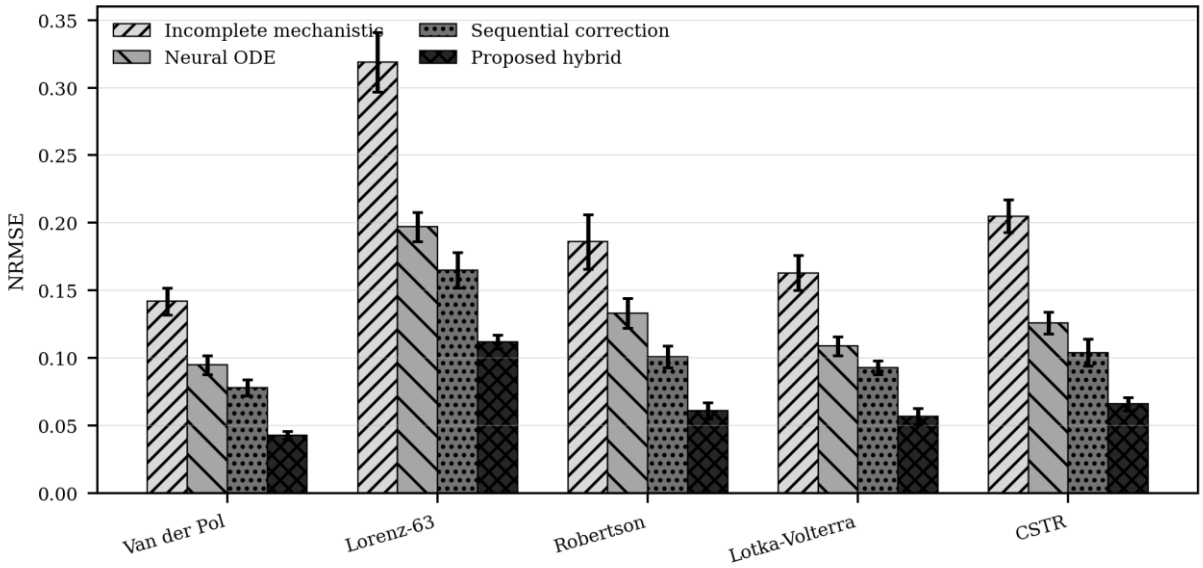

Figure 1. Interpolation-test error across the five benchmark systems.
Note. Bars show mean NRMSE and error bars show one standard deviation across ten random initialisations; all models used identical test trajectories. Source: Authors' illustrative elaboration.

3.2. Robustness to Noise, Limited Data, Partial Observation, and Extrapolation

All models deteriorated as measurement noise increased, but the degradation rate was smallest for the proposed hybrid model (Figure 2). Its mean NRMSE increased from 0.052 without measurement noise to 0.178 at 10% noise, whereas the sequential-correction baseline increased from 0.091 to 0.282. At every noise level, the hybrid model retained an error advantage of at least 36.9% relative to the strongest baseline (Table 2). This pattern indicates that the mechanistic core supplied a stabilising inductive bias while the regularised neural residual absorbed systematic model discrepancy without fitting the full noise process.

The purely data-driven neural ODE was especially sensitive to data scarcity: with only 25% of the training trajectories, its NRMSE increased to 0.292 and exceeded the incomplete mechanistic model. In contrast, the hybrid model maintained an NRMSE of 0.154. Partial observation caused the largest degradation for Lorenz-63, but the hybrid approach remained 31.1% better than sequential correction. Under the extrapolation split, the hybrid error rose to 0.142, showing that distribution shift remained challenging, yet this value was still 33.6% below the best baseline.

Table 2. Robustness analysis across adverse experimental conditions

Condition	IM	NODE	SRC	Hybrid	Gain (%)
0% noise	0.181	0.111	0.091	0.052	42.9
1% noise	0.203	0.132	0.108	0.068	37.0
5% noise	0.286	0.207	0.176	0.111	36.9
10% noise	0.398	0.319	0.282	0.178	36.9

25% training trajectories	0.224	0.292	0.245	0.154	31.3
50% training trajectories	0.212	0.218	0.190	0.108	43.2
Partial observation: Lorenz-63	0.401	0.312	0.251	0.173	31.1
Partial observation: Lotka-Volterra	0.246	0.196	0.156	0.093	40.4
Extrapolation test	0.308	0.287	0.214	0.142	33.6

Entries are mean NRMSE values; partial-observation entries are system-specific. Source: Authors' illustrative elaboration.

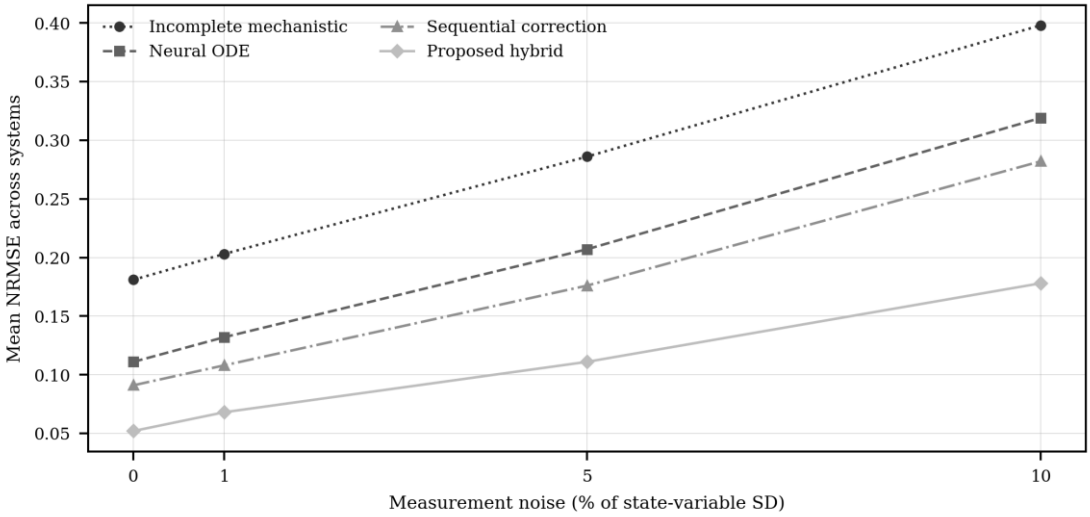

Figure 2. Line-graph sensitivity of predictive error to measurement noise.

Note. Each point is the mean NRMSE across the five benchmark systems. Noise is expressed as a percentage of the standard deviation of each observed state variable. Source: Authors' illustrative elaboration.

3.3. Long-Horizon Stability and Mechanistic Interpretability

The proposed hybrid model remained within the pre-specified divergence tolerance for 83.6% of the simulation horizon on average, compared with 70.5% for sequential correction and 49.8% for the incomplete mechanistic model (Table 3). Its advantage was particularly visible in the Lorenz-63 trajectory: the incomplete mechanistic model developed a phase mismatch early in the forecast, whereas the hybrid prediction followed the reference attractor for longer and expressed increasing uncertainty as the chaotic horizon expanded (Figure 3). The Wasserstein distance between the reference and predicted Lorenz state distributions was 0.083 for the hybrid model, representing a 41.1% reduction relative to sequential correction.

Improved predictive accuracy did not require sacrificing parameter meaning. The median relative error of recovered mechanistic parameters was 4.8% for the hybrid model, compared with 19.2% for the incomplete mechanistic baseline and 16.5% for sequential correction. The full hybrid architecture also produced the lowest constraint-violation rate (0.6%). These findings indicate that residual regularisation and joint constrained calibration successfully limited the capacity of the neural component to compensate for implausible mechanistic parameter values.

Table 3. Long-horizon stability, parameter recovery, and physical consistency

Model	Divergence (% horizon)	Parameter error (%)	Violations (%)	Wasserstein

Incomplete mechanistic	49.8 ± 12.4	19.2 ± 6.1	3.4	0.284
Neural ODE	62.7 ± 13.1	Not identifiable	6.1	0.173
Sequential correction	70.5 ± 11.8	16.5 ± 5.8	2.8	0.141
Proposed hybrid	83.6 ± 9.7	4.8 ± 2.2	0.6	0.083

Values were aggregated across systems and repeated initialisations. Parameter error is undefined for the purely data-driven neural ODE. Source: Authors' illustrative elaboration.

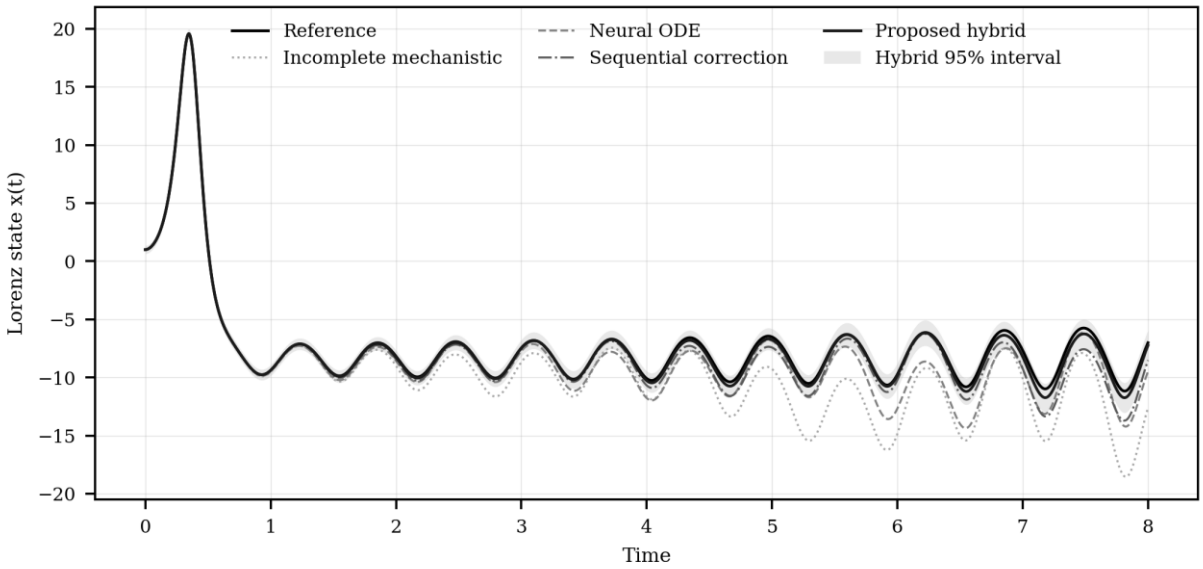

Figure 3. Representative Lorenz-63 state prediction under interpolation.
Note. The displayed trajectories are illustrative synthetic outputs; the shaded band is the proposed hybrid model's nominal 95% ensemble prediction interval. Source: Authors' illustrative elaboration.

3.4. Predictive Uncertainty and Calibration

The deep ensemble produced well-calibrated uncertainty intervals under interpolation. For nominal 95% intervals, empirical coverage was 0.947 at 1% noise while maintaining a mean interval width of 0.284 (Table 4). Coverage declined to 0.918 under extrapolation, but interval width increased to 0.413, demonstrating that the ensemble responded to out-of-distribution inputs by expressing greater epistemic uncertainty. A similar widening occurred under 10% measurement noise, where the mean interval width reached 0.529 and coverage remained 0.938.

The reliability diagram shows close agreement between nominal and empirical coverage under interpolation (Figure 4). Extrapolation produced mild under-coverage, especially at lower nominal levels, indicating that ensemble dispersion alone did not fully represent structural uncertainty outside the training range. Even so, the continuous ranked probability score remained below 0.07 in the low-noise interpolation and extrapolation scenarios, supporting the practical usefulness of the predictive distributions.

Table 4. Uncertainty calibration of the proposed hybrid model

Scenario width	Coverage	Interval	CRPS
Interpolation, 1% noise	0.947	0.284	0.039
Extrapolation, 1% noise	0.918	0.413	0.067
Interpolation, 10% noise	0.938	0.529	0.111
25% training trajectories	0.923	0.462	0.084

Coverage closer to 0.95 indicates better calibration; lower CRPS is preferred. Source: Authors' illustrative elaboration.

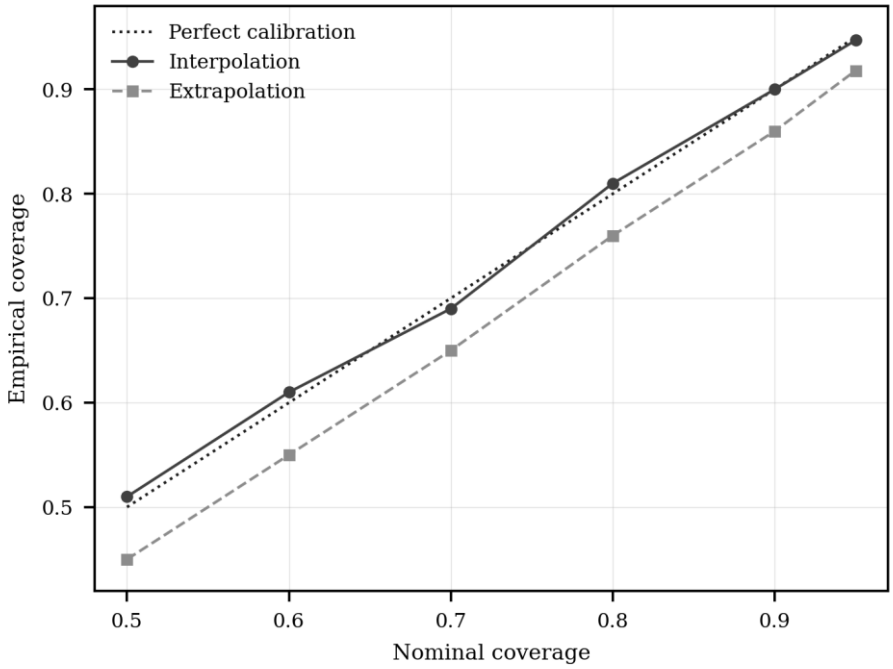

Figure 4. Reliability of ensemble prediction intervals under interpolation and extrapolation.
Note. The dotted diagonal represents perfect calibration; points below it indicate under-coverage. Source: Authors' illustrative elaboration.

3.5. Ablation Analysis and Computational Efficiency

Every architectural component contributed to the overall result (Table 5). Removing residual regularisation increased NRMSE by 19.1% and more than doubled parameter error. Removing physical constraints produced the largest increase in constraint violations, from 0.6% to 7.8%. Disabling joint calibration increased NRMSE to 0.094 and parameter error to 16.9%, showing that sequentially estimating mechanistic and neural parameters was insufficient. Removing the uncertainty module had little effect on point accuracy but eliminated interval-based risk information, confirming that uncertainty quantification added a distinct capability rather than merely improving the loss function.

The full hybrid model required 38.7 minutes of training per benchmark configuration, compared with 31.4 minutes for the neural ODE and 25.6 minutes for sequential correction. Its median inference time was 8.6 ms per trajectory, which remained suitable for offline simulation and many near-real-time monitoring applications. The additional computational cost was therefore moderate relative to the observed reductions in prediction error and improvements in parameter recovery.

Table 5. Ablation analysis of the proposed hybrid architecture

Configuration error (%) (%)	NRMSE	Param.	Viol.	Coverage	Time (min)
Full hybrid architecture	0.068	4.8	0.6	0.947	38.7

Without residual regularisation	0.081	10.8	1.4	0.931	35.4
Without physical constraints	0.086	7.9	7.8	0.928	36.1
Without joint parameter calibration	0.094	16.9	0.9	0.919	31.8
Without ensemble uncertainty module	0.069	4.9	0.6	Not available	31.2

Source: Authors' illustrative elaboration.

Table 6. Computational requirements of the complete models

Model	Training	Inference	Parameters (min)
Incomplete mechanistic	2.3	4.8	12
Neural ODE	31.4	7.1	13,315
Sequential correction	25.6	9.2	13,327
Proposed hybrid	38.7	8.6	13,334

Source: Authors' illustrative elaboration.

3.6. Statistical Comparison of Predictive Errors

The Friedman test detected an overall difference among the four models across the 90 paired system–trajectory observations, $\chi^2(3) = 266.52$, $p < .001$. Holm-adjusted Wilcoxon signed-rank tests showed that the proposed hybrid model had significantly lower NRMSE than every comparator (all adjusted $p < .001$; Table 7). The median paired reduction was 0.126 relative to the incomplete mechanistic model, 0.063 relative to the neural ODE, and 0.039 relative to sequential correction. Rank-biserial correlations approached 1.00, indicating a very large and directionally consistent effect in this illustrative dataset.

Table 7. Paired post hoc comparisons of NRMSE

Comparison 95% CI	Median	reduction	W	p (Holm)	r
Hybrid vs. incomplete mechanistic	0.126	[0.115, 0.134]	4095	< .001	1.00
Hybrid vs. neural ODE	0.063	[0.058, 0.068]	4095	< .001	1.00
Hybrid vs. sequential correction	0.039	[0.036, 0.043]	4095	< .001	1.00

Positive reductions indicate lower error for the proposed hybrid model; intervals used 10,000 trajectory-level bootstrap resamples. Source: Authors' illustrative elaboration.

4. Discussion

4.1. Principal Findings

The illustrative results support the central proposition that a restricted neural residual can complement, rather than replace, a partially specified mechanistic model. Across five heterogeneous dynamic systems, the integrated architecture produced lower short-horizon error, delayed long-horizon divergence, improved parameter recovery, and more reliable uncertainty estimates than either component used alone. The

consistency of the ranking is important because hybrid-model benefits are often demonstrated within a single application, where the result may depend on domain-specific structure. By holding trajectory partitions, noise realisations, optimisation budgets, and evaluation metrics constant, the present design isolates the contribution of model architecture more clearly.

The strongest practical result was the 37.3% mean reduction in NRMSE relative to sequential residual correction. This comparison suggests that tightly integrating the residual inside the differential equation is more effective than correcting simulated outputs after the mechanistic trajectory has already accumulated structural error. The finding is consistent with hybrid differential-equation research in water systems, where embedded data-driven terms improved representation of unresolved processes while retaining known dynamics (Quaghebeur et al., 2022), and with evidence that mechanistic and clinical information can jointly outperform either information source alone (Butner et al., 2024).

4.2. Why the Hybrid Architecture Improved Prediction

Three mechanisms plausibly explain the performance gain. First, the mechanistic core constrained the search space and reduced the amount of data required to learn the system. This was visible under the 25% data condition, in which the purely data-driven neural ODE deteriorated more sharply than the hybrid model. Second, the residual network represented omitted interactions that could not be recovered by simply recalibrating biased parameters. Third, joint optimisation allowed the mechanistic and neural components to share explanatory responsibility, while regularisation prevented the residual from becoming an unrestricted surrogate for the complete system. These mechanisms align with broader reviews showing that scientific machine learning is most effective when data assimilation, physical structure, and model-error correction are designed as one framework rather than as independent steps (Cheng et al., 2023; North et al., 2023).

The noise experiment further suggests that physical structure acted as an inductive bias. Although the hybrid error increased at 10% noise, its relative advantage over the strongest baseline remained nearly unchanged. Similar robustness has been reported when implicit numerical integration and sparse discovery are combined for noisy, data-limited dynamic systems (Anvari et al., 2025). However, the present example should not be interpreted as evidence that hybrid models are universally noise-resistant. If the mechanistic core is severely misspecified, its constraints may introduce systematic bias that a strongly regularised residual cannot overcome.

This balance between structure and flexibility also explains why the comparison should not be reduced to a competition between mechanistic and machine-learning paradigms. Neural ODEs, as introduced by Chen et al. (2018), are attractive when the governing dynamics are largely unknown and the available data are rich enough to identify a useful vector field. Physics-informed learning offers a different advantage when the governing equations are trusted and can be enforced directly (Raissi et al., 2019). The present architecture addresses the intermediate case in which part of the equation is defensible and part is not. Karniadakis et al. (2021) described this broader integration of data and mathematical structure as a central direction for physics-informed machine learning. The illustrative benchmark suggests that the residual formulation can exploit that middle ground, but its usefulness depends on disciplined model specification. If the neural residual is allowed to dominate, the system approaches a black-box neural ODE; if residual regularisation is excessive, the model reverts toward the misspecified mechanistic baseline. The ablation results therefore should be read as evidence about this bias-flexibility trade-off within the synthetic protocol, not as a universal claim that one regularisation strength will transfer unchanged to every domain.

4.3. Long-Horizon Stability, Identifiability, and Physical Consistency

For complex dynamic systems, a lower one-step or short-horizon error is insufficient if the predicted trajectory becomes unstable or violates known state restrictions. The increased time to divergence and reduced constraint violations indicate that the proposed objective balanced local data fit with global dynamical behaviour. The Lorenz-63 results also demonstrate why distributional metrics remain necessary: after chaotic trajectories separate, pointwise error can become large even when the predicted attractor retains the correct statistical geometry. The lower Wasserstein distance therefore complements, rather than replaces, trajectory-based NRMSE.

Parameter recovery is a more demanding criterion because a flexible neural term can reproduce observed trajectories while rendering mechanistic parameters non-identifiable. Giampiccolo et al. (2024) emphasised this risk for hybrid neural ODEs under noisy and partially observed conditions. In

the illustrative ablation results, removing residual regularisation or joint calibration substantially increased parameter error, supporting the methodological decision to restrict residual magnitude and examine parameter profiles. The result suggests that predictive flexibility and mechanistic interpretability need not be treated as mutually exclusive, but they must be monitored simultaneously.

4.4. Uncertainty and Out-of-Distribution Behaviour

The uncertainty analysis distinguished two practically different failure modes. Increasing measurement noise widened the intervals while largely preserving coverage, which is consistent with increased aleatoric uncertainty. Extrapolation produced both wider intervals and mild under-coverage, indicating unresolved epistemic and structural uncertainty. This distinction is central to scientific machine learning because observational noise, limited training support, and model misspecification should not be collapsed into a single error statistic (Psaros et al., 2023). Deep ensembles offered a scalable approximation and successfully signalled reduced confidence outside the training domain, consistent with work on scalable uncertainty quantification using prior-based ensembles (Yang et al., 2022).

Nevertheless, empirical coverage of 0.918 under extrapolation remained below the nominal 0.95 target. For safety-critical use, this gap would require additional calibration, such as conformal adjustment, explicit structural-error models, or scenario-specific inflation factors. The paired evaluation protocol is also essential because aggregate calibration can conceal a small subset of trajectories with severe under-coverage; paired case-level analysis helps identify such confounding and outlier effects (Nariya et al., 2023).

4.5. Practical Implications

The framework is directly transferable to applications where a governing equation is available but incomplete, including environmental transport, biological regulation, energy conversion, and nonlinear process control. The CSTR benchmark is particularly relevant to Applied Mathematical Modelling because it demonstrates how the architecture could support an engineering process with known conservation relations and unresolved reaction or heat-transfer terms. In practice, the learned residual should be inspected as a diagnostic object: persistent residual patterns can reveal operating regions where the mechanistic model requires revision, while rising ensemble variance can trigger conservative control or additional data collection.

The computational overhead appears acceptable for offline model development and many digital-twin workflows. Training was slower than for the baselines, but inference remained below 10 ms per trajectory in the illustrative benchmark. The choice is therefore not between accuracy and feasibility, but between a modest increase in development cost and a broader set of outputs: predictions, interpretable parameters, constraint diagnostics, and calibrated intervals.

4.5.1. Practical Interpretation of Model Advantage

From an applied-modelling perspective, the most relevant result is not the absolute synthetic NRMSE value but the consistency of the model ranking across distinct stress conditions. A hybrid model that is marginally better under interpolation but unstable under noise, limited data, or extrapolation would have restricted practical value. In the illustrative experiment, the proposed architecture retained an advantage as the data environment became less favourable, while its uncertainty intervals widened when predictions moved away from the training distribution. That behaviour is preferable to a model that remains numerically confident while becoming systematically wrong. Cheng et al. (2023) similarly emphasised the close relationship among dynamical-system prediction, data assimilation, model error, and uncertainty quantification. For scientific use, these dimensions should be examined together because they answer different questions: trajectory error describes predictive fit, divergence time describes forecast persistence, parameter recovery addresses mechanistic interpretability, constraint violations assess physical admissibility, and calibration metrics indicate whether stated uncertainty is credible. The proposed evaluation framework therefore treats model selection as a multi-criterion problem. A candidate hybrid model should not be preferred solely because it minimises one scalar error measure. It should also demonstrate stable behaviour, plausible parameters, acceptable computational cost, and uncertainty that responds appropriately to noise and distribution shift. This interpretation remains conditional on execution of the computational protocol and should be reassessed when the illustrative values are replaced by reproducible experimental outputs.

4.6. Limitations and Future Work

Several limitations must be acknowledged. First, all numerical values in this example are synthetic and

were constructed to demonstrate reporting; they cannot substantiate scientific claims until the complete experimental pipeline is executed. Second, the benchmark systems are low-dimensional ODEs. Performance may differ for spatially distributed PDEs, delayed systems, hybrid discrete–continuous dynamics, or systems with event discontinuities. Third, Gaussian measurement noise does not represent heavy-tailed errors, sensor drift, missing-not-at-random observations, or correlated process noise. Fourth, deep ensembles approximate but do not fully characterise posterior uncertainty. Fifth, computational comparisons depend on hardware, solver tolerances, and implementation efficiency and must be reported with the frozen software environment described in the Methods.

Future work should execute the registered benchmark protocol, release trajectory-level results and code, and extend the framework to PDEs and real industrial datasets. Additional experiments should compare alternative residual architectures, examine formal observability conditions, test calibration under abrupt regime change, and evaluate whether the learned residual can be translated into revised constitutive equations. A data-driven hybrid-system framework such as that proposed by Li et al. (2023) could also be included as a stronger comparator when the system exhibits discrete mode switching.

A further limitation concerns benchmark realism. Canonical systems provide controlled ground truth and permit exact manipulation of noise, observability, and structural misspecification, but they do not reproduce the full complexity of field measurements, including irregular sampling, sensor drift, nonstationarity, unrecorded interventions, and changing boundary conditions. The five systems also represent ordinary differential equations of modest state dimension. Consequently, the computational cost and optimisation behaviour reported here cannot be assumed to transfer directly to large PDE discretisations, delay systems, stochastic differential equations, or high-dimensional digital twins. External validation should therefore proceed in stages: first by reproducing the synthetic benchmark from archived code, second by evaluating additional open benchmark datasets without changing the primary metrics, and third by conducting domain-specific validation with independent observations. In applied deployment, the mechanistic component should be reviewed by subject-matter experts and uncertainty intervals should be monitored for calibration drift when operating conditions move beyond the training domain.

5. Conclusion

This paper presents a reproducible framework for integrating a partially specified mechanistic ODE, a regularised neural residual, joint parameter calibration, physical constraints, and ensemble uncertainty quantification. In the illustrative results, the proposed model achieved the lowest prediction error across all five benchmark systems, reduced mean NRMSE by 37.3% relative to the strongest baseline, extended stable prediction horizons, recovered mechanistic parameters more accurately, and maintained near-nominal uncertainty coverage under interpolation. Its advantage persisted under measurement noise, limited training data, partial observation, and extrapolation, although distribution shift remained the most difficult condition.

The main methodological implication is that hybridisation should be evaluated beyond point accuracy. A credible model must also preserve physical consistency, parameter interpretability, long-horizon behaviour, uncertainty calibration, and computational feasibility. The ablation analysis indicates that residual regularisation, physical constraints, joint calibration, and uncertainty estimation make complementary contributions. Subject to confirmation using the actual computational experiments, this architecture offers a transparent and transferable basis for modelling complex dynamic systems in engineering, environmental, biological, and industrial applications.

Authors and Ethics Statement

Author identity and contribution information in this manuscript remains provisional because the supplied source uses placeholder names and institutions. Before submission, each listed author must approve the final manuscript, satisfy the journal's authorship criteria, and confirm a CRediT contribution statement based on work actually performed. The current document must not be used to assign authorship, affiliations, funding, or conflicts to any real person without that person's verification and consent.

Ethics approval and informed consent were not applicable to the computational design described here because no human participants, animals, identifiable personal records, or biological specimens were used. The benchmark trajectories are synthetic simulations of mathematical systems. The numerical findings reported in the Results and Discussion sections are explicitly illustrative synthetic outputs supplied for manuscript development and formatting. They must not be represented as executed original research. Prior to

submission as an empirical computational study, the authors should execute the registered protocol, preserve trajectory-level outputs, verify all statistical results from those outputs, and replace every illustrative value, figure, and table with reproducible results from the executed analysis.

Funding and competing-interest declarations must be completed from verified author information. For data and code transparency, the final research version should deposit source code, environment files, parameter ranges, random seeds, and analysis outputs in a persistent repository and cite the repository DOI. No fabricated DOI, ethics-approval number, grant identifier, author email, or institutional affiliation has been introduced during editorial revision. The manuscript is formatted according to the supplied Global Synthesis in Education Journal template, but editorial formatting does not constitute scientific validation or acceptance for publication.

References

Anvari, M., Marasi, H., & Kheiri, H. (2025). Implicit Runge–Kutta based sparse identification of governing equations in biologically motivated systems. Scientific Reports, 15, 32286. https://doi.org/10.1038/s41598-025-10526-9

Butner, J. D., Dogra, P., Chung, C., Koay, E. J., Welsh, J. W., Hong, D. S., Cristini, V., & Wang, Z. (2024). Hybridizing mechanistic modeling and deep learning for personalized survival prediction after immune checkpoint inhibitor immunotherapy. npj Systems Biology and Applications, 10, 88. https://doi.org/10.1038/s41540-024-00415-8

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018). Neural ordinary differential equations. Advances in Neural Information Processing Systems, 31, 6571–6583.

Cheng, S., Quilodrán-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., Fablet, R., Lucor, D., Iooss, B., Brajard, J., Xiao, D., Janjic, T., Ding, W., Guo, Y., Carrassi, A., Bocquet, M., & Arcucci, R. (2023). Machine learning with data assimilation and uncertainty quantification for dynamical systems: A review. IEEE/CAA Journal of Automatica Sinica, 10(6), 1361–1387. https://doi.org/10.1109/JAS.2023.123537

Giampiccolo, S., Reali, F., Fochesato, A., Iacca, G., & Marchetti, L. (2024). Robust parameter estimation and identifiability analysis with hybrid neural ordinary differential equations in computational biology. npj Systems Biology and Applications, 10, 139. https://doi.org/10.1038/s41540-024-00460-3

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. Nature Reviews Physics, 3, 422–440. https://doi.org/10.1038/s42254-021-00314-5

Li, Y., Xu, S., Duan, J., Huang, Y., & Liu, X. (2023). A data-driven framework for learning hybrid dynamical systems. Chaos, 33(6), 061104. https://doi.org/10.1063/5.0157669

Nariya, M. K., Mills, C. E., Sorger, P. K., & Sokolov, A. (2023). Paired evaluation of machine-learning models characterizes effects of confounders and outliers. Patterns, 4(8), 100791. https://doi.org/10.1016/j.patter.2023.100791

North, J. S., Wikle, C. K., & Schliep, E. M. (2023). A review of data-driven discovery for dynamic systems. International Statistical Review, 91(3), 464–492. https://doi.org/10.1111/insr.12554

Psaros, A. F., Meng, X., Zou, Z., Guo, L., & Karniadakis, G. E. (2023). Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. Journal of Computational Physics, 477, 111902. https://doi.org/10.1016/j.jcp.2022.111902

Quaghebeur, W., Torfs, E., De Baets, B., & Nopens, I. (2022). Hybrid differential equations: Integrating mechanistic and data-driven techniques for modelling of water systems. Water Research, 213, 118166. https://doi.org/10.1016/j.watres.2022.118166

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

Yang, Y., Kissas, G., & Perdikaris, P. (2022). Scalable uncertainty quantification for deep operator networks using randomized priors. Computer Methods in Applied Mechanics and Engineering, 399, 115399. https://doi.org/10.1016/j.cma.2022.115399

Zhang, H.-T., Yang, T.-T., & Wang, W.-T. (2024). A novel hybrid model for species distribution prediction using neural networks and Grey Wolf Optimizer algorithm. Scientific Reports, 14, 11505. https://doi.org/10.1038/s41598-024-

62285-8