

A Hybrid Physics-Informed Deep Learning Framework for Robust Multivariate System Modelling Under Uncertainty

Sonu Pandey^{1, *, ✉}

¹Independent Researcher, Financial Systems & AI, USA

Article History

Received: 15 April, 2026

Revised: 16 June, 2026

Accepted: 23 June, 2026

Published: 26 June, 2026

Abstract:

Introduction: This study presents a unified physics-regularised Transformer framework for robust multivariate dynamical-system modelling under uncertainty. Conventional data-driven deep learning models can capture nonlinear temporal patterns but may produce physically inconsistent predictions under distribution shifts. In contrast, classical Physics-Informed Neural Networks (PINNs) improve physical consistency but often suffer from fixed loss weighting, gradient imbalance, and limited uncertainty calibration.

Methodology: To address these limitations, this study integrates three components into a single modelling framework: a gradient-norm-based adaptive physics-weighting strategy, an attention-based temporal encoder, and a heteroscedastic uncertainty-estimation layer. The governing equations are used as soft inductive biases rather than exact system descriptors, allowing the model to balance empirical data fitting with physics-guided regularisation.

Results & Discussion: Experiments on a synthetic multivariate dynamical-system dataset show consistent improvements over the LSTM, Pure Transformer, Standard PINN, and Deep Ensemble baselines in predictive accuracy, uncertainty calibration, and robustness to perturbations. The proposed model achieved lower RMSE, MAE, and negative log-likelihood than the implemented baselines while maintaining greater stability under noisy and missing-data conditions. Ablation results further indicate that adaptive weighting, physics regularisation, and heteroscedastic uncertainty modelling each contribute to the final performance.

Conclusion: The findings suggest that physics-regularised attention models can provide a useful direction for uncertainty-aware scientific machine learning. However, further validation on real-world benchmark datasets is required before broader claims of deployment can be made.

Keywords: Physics-informed learning, scientific machine learning, transformer architecture, uncertainty quantification, Robust dynamical systems, multivariate time-series modelling, distribution shift, hybrid deep learning.

1. INTRODUCTION

Management Scientific Machine Learning (SciML) has become a paradigm shift that is connecting classical modelling of systems to modern data-driven artificial intelligence [1]. Conventional methods for identifying a system are based on deriving governing equations from first principles or by parameterising empirical data. Although these methods primarily achieve interpretability and physical consistency, they typically do not resolve the high-dimensional, nonlinear

dynamics of complex multi-variable systems. Deep learning models, on the other hand, are good at approximating functions and extracting large-scale patterns but are often unable to be physically interpretable and consistent with known laws of motion, conservation or structural constraints [2].

The concept of hybrid modelling has thus emerged as an increasingly significant avenue in research, with the potential to merge physics-based reasoning and neural network flexibility. Here, the incorporation of physics constraints in

*Address correspondence to this author at Independent Researcher, Financial Systems & AI, USA; E-mail: pant.sonu11@gmail.com



© 2026 Copyright by the Author.

Licensed as an open access article using a [CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/).

learning structures allows models to honour domain knowledge and yet allow them to remain flexible to real-world data [3]. A notable development in this direction is physics-informed neural networks (PINNs), which include a loss term for the residual of the differential equation. Nevertheless, the majority of implementations are deterministic, fixed-weighted, and restricted to modeling only multivariate interactions among uncertain cases. Due to the increasing interconnectivity and stochasticity of modern engineering, climate, and cyber-physical systems, there is an increasing requirement to have hybrid structures that simultaneously consider nonlinear coupling, uncertainty propagation, and robustness [4].

There are several basic problems in modelling multivariate nonlinear dynamical systems. First, multivariate coupling introduces interdependencies among the state variables that change over time. These cross variables are not typically linear and are long-range, time-dependent, so traditional recurrent models do not learn higher-order correlations. Second, noise, measurement error, and stochastic disturbances are inherent in real-world systems [5]. Both aleatoric (intrinsic randomness) and epistemic (model uncertainty) uncertainty affect predictive reliability, although most deterministic models have not explicitly measured them.

Distribution shift is another big challenge. The performance of nominally trained models has often been found to deteriorate significantly when given perturbed, noisy, or partially incomplete data. Stability to organised perturbations is vital for their use in safety-critical areas, and evaluation regimes frequently overlook this fact [6]. In addition, the problem of training instability is still relevant in physics-informed learning. PINNs are also known to exhibit a gradient imbalance between data loss and physics residual loss, resulting in gradual convergence, over-regularisation, or the dominance of a single objective component [7]. The problems are exacerbated in high-dimensional multivariate settings, where the scaling and conditioning of the residual terms can have a devastating effect on optimisation behaviour.

Although hybrid modelling has improved, several limitations remain. One of the key weaknesses of traditional PINNs is that it uses a constant weighting parameter λ to trade off data fidelity and physics residual terms [8]. The fixed weighting schemes do not keep up with the changing magnitude of the gradients during training, leading to either physical under-enforcement or over-regulation, thereby suppressing data-driven learning.

Also, the majority of physics-informed systems presuppose deterministic system dynamics and lack principled uncertainty-calibration mechanisms. Probabilistic modelling is required to enable predictive intervals to be well-calibrated, which would be useful in practice [9]. Another area that is not well developed is robustness evaluation. Many studies report that predictive accuracy is achieved when the data is clean and when no stress testing is performed systematically under noise, missing data or parameter drift. In addition, although Transformer architectures have proven to be more capable of modelling long-range dependencies, it is not yet been integrated into the physics-informed multivariate system modelling [10].

This study is positioned as a physics-regularised learning framework rather than a strict PINN formulation. The governing equations are used as soft inductive biases that guide the model toward physically plausible behaviour, but they are not assumed to describe every component of the synthetic multivariate system exactly. Therefore, the contribution is not presented as an entirely new algorithmic paradigm, but as a unified and empirically evaluated integration of three elements. First, a gradient-norm-based adaptive weighting strategy is developed to balance data-loss and physics-residual losses during training. Second, a joint architecture is designed by combining soft physics constraints with attention-based temporal encoding for multivariate time-series modelling. Third, a structured robustness evaluation protocol is introduced to assess sensitivity to Gaussian noise, missing observations, temporal drift, scaling distortion, and out-of-distribution behaviour. This revised contribution addresses the limitations of deterministic fixed-weight PINNs while avoiding overstated claims about general real-world deployment.

2. LITERATURE REVIEW

2.1. Physics-Informed Neural Networks

The concept of Physics-Informed Neural Networks (PINNs) has become the new paradigm in scientific machine learning, in which neural networks learn to respond directly to physical laws. PINNs do not just use observational data, but instead integrate the differential equation residuals into the loss so that it remains consistent between the outputs, which are predicted and known system dynamics [11]. This deterministic residual enforcement ensures that a solution satisfies partial or ordinary differential equations, boundary conditions, and conservation principles during optimisation. With the help of automatic differentiation, PINNs simultaneously compute derivatives with respect to network outputs and network inputs, and thus, the physics residual is computed without the need to define discretisation schemes explicitly [12]. This method has proven to be very effective in fluid dynamics, structural mechanics, heat transfer and other physical modelling problems.

Although these benefits exist, classical PINNs have significant weaknesses. The vast majority of the implementations assume deterministic system dynamics and implicitly assume noise as non-existent or measurement-based [13]. However, in real multivariate systems, stochastic forcing and uncertain parameters often affect system behaviour. Enforcement of deterministic residual risks, the over-constraining of the model, hence generating biased predictions, where the governing equations are incomplete or approximate [14]. Moreover, data fidelity and physics consistency are usually balanced manually using a scalar weighting parameter. This rigid weighting tends to produce optimisation instability, particularly when the magnitude of the residuals is large compared to data losses [15]. As a result, although PINNs are an effective combination of machine learning and physics, they lack resilience in uncertain, high-dimensional environments due to their deterministic nature and fixed loss-balancing schemes. Neural Operators are a class of models designed to

learn mapping functions between function spaces, particularly in the context of solving partial differential equations (PDEs) [16]. Neural Operators are gaining traction in fields like fluid dynamics and materials science due to their ability to generalise across different domains with minimal training data [17].

Physics-informed Transformers integrate physical constraints into Transformer architectures for learning complex dynamical systems. By incorporating known physical laws or residuals into the loss function, these models aim to ensure that predictions respect the underlying physics [18]. This approach combines the high expressive power of Transformers with the benefits of physics-informed learning, enabling better generalisation and accuracy for tasks such as climate modelling, fluid dynamics, and other scientific applications that require domain-specific knowledge.

Scientific Machine Learning (SciML) frameworks integrate traditional scientific computing methods with machine learning to solve complex physical problems. These frameworks enable data-driven models to learn from physical systems while ensuring that results are consistent with known laws of physics. SciML approaches are increasingly used in fields such as materials science, climate modelling, and engineering, where they can optimise performance and uncover patterns in data-driven simulations of real-world phenomena [19].

Probabilistic Physics-informed Neural Networks (PINNs) extend traditional PINNs by incorporating uncertainty quantification into the modelling process. They provide a probabilistic framework to estimate both model parameters and predictions, allowing for better handling of uncertainties and variability in real-world data [20]. Probabilistic PINNs are particularly useful in situations where data is noisy, incomplete, or uncertain, as they can estimate predictive distributions and uncertainty bounds, improving model robustness and decision-making in scientific and engineering applications.

2.2. Adaptive Loss Balancing in Physics-Informed Learning

Adaptive loss weighting has already been studied in multi-objective deep learning and physics-informed learning. GradNorm balances gradient magnitudes across different tasks in multi-task networks and shows that gradient-level adjustment can improve training stability when multiple objectives compete [7]. Dynamic Weight Averaging adjusts task weights according to the relative rate of loss reduction over time, while SoftAdapt dynamically modifies loss weights using live loss statistics. More recently, ReLoBRaLo was introduced for PINNs to balance multiple physics-informed loss terms using relative loss changes and random lookback strategies. These methods show that adaptive loss balancing is an important topic rather than a completely unexplored problem [9].

The present work differs from these approaches in its specific integration context. GradNorm was originally designed for general multi-task learning; DWA depends mainly on loss-change rates; and SoftAdapt uses loss statistics rather than directly enforcing a balanced data-physics gradient contribution. ReLoBRaLo is closer to PINN training but

focuses on balancing relative losses across PDE-related terms. In contrast, this study uses an exponential moving average of the gradient-norm ratio between the data-loss and physics-residual losses, with clipping and numerical stabilisation, within a Transformer-based heteroscedastic forecasting architecture [2, 8]. Therefore, the contribution lies in adapting gradient-level balancing to a physics-regularised temporal modelling setting, rather than claiming that adaptive loss weighting itself is entirely new.

2.3. Deep Learning for Dynamical Systems

Temporal modelling and nonlinear dynamical system identification have a long history of deep learning applications. Recurrent neural networks, especially Long Short-Term Memory (LSTM) networks, have become popular in sequential modelling of data since they exploit the temporal relationship between data by means of gated memory [21]. LSTMs alleviate the problem of vanishing gradients in regular recurrent networks, allowing long-time horizons of moderate length to be modelled. LSTMs have also been demonstrated to be effective in state reconstruction and forecasting when applied in a multivariate dynamical system [22]. They tend, however, to have problems modelling complex cross-variable interactions when dependencies lie in both time and features.

To avoid this, temporal convolutional networks (TCNs) offer the alternative of temporal causal convolutions, which are computed in parallel, and the gradients are propagated downward. TCNs use dilated convolutions to increase their receptive fields so that they mainly model long-range dependencies using fewer layers [23]. Although TCNs are more stable during training than recurrent architectures, they nevertheless impose constraints on their ability to capture dynamically varying cross-variable interactions.

More recently, transformer architectures have altered the paradigm of sequence modelling through self-attention [24]. Transformers are better than recurrent or convolutional models at capturing long-range dependencies by simultaneously computing attention weights across all positions in time [25]. Still, most Transformer-based models are entirely data-driven and lack specific physical constraints. They have a high approximation power that often leads to overfitting and inadequate extrapolation when used or deployed under distribution shifts or unknown operating conditions. Therefore, although deep learning structures have made significant improvements in temporal modelling, they have yet to incorporate physical principles.

2.4. Uncertainty Quantification

Uncertainty quantification (UQ) is critical in the implementation of predictive models in the real world, where noise, stochasticity, and incomplete knowledge are part of the problem [26]. Deep ensembles, in which several networks trained separately make predictions whose variance approximates epistemic uncertainty, are a widely used strategy. Ensembles are empirically effective and are fairly easy to use, and were more effective than single probabilistic models at calibration [27]. They, however, impose significant computational overhead and fail to directly include physics-based constraints.

Monte Carlo (MC) dropout offers a lighter approach by estimating Bayesian inference by using stochastic dropout during inference [28]. Predictive distributions approximated by running several forward passes with dropout turned on, with no changes to the underlying architecture. MC dropout is computationally efficient but was inaccurate in regimes with strong nonlinearity and has no theoretical guarantees in multivariate systems [29].

Heteroscedastic regression algorithms explicitly learn predictive mean and variance, which depend on the inputs [30]. The approach represents aleatoric uncertainty arising from measurement noise or stochastic dynamics. The loss function is usually defined as a negative log-likelihood, where the assumptions include a Gaussian distribution, which allows the network to perform adjustments to the variance estimates [31]. Although heteroscedastic regression offers an idealised probabilistic theory, it cannot be easily combined with physics-informed constraints. The majority of current PINN-based schemes either implicitly handle uncertainty or do not perform calibration analysis, resulting in a lack of probabilistic robustness in scientific hybrid models.

2.5. Robustness and Distribution Shift

Perturbation and distribution shift robustness have begun to receive more interest in time-series modelling and scientific machine learning [32]. Noise sensitivity experiments investigate the rate of deterioration in predictive accuracy under additive disturbances, yielding the susceptibility of purely data-driven models to measurement corruption. Even small perturbations mainly cause large increases in the variance of the forecast in nonlinear systems, and this is particularly relevant when the training data is not sufficiently diverse [33].

The other critical issue is covariate shift, in which the input distribution varies during training and deployment. Distribution shift could occur in dynamical systems due to environmental changes, parameter changes, or a change in the structure of the operating conditions [34]. Most deep learning algorithms fail to cope with such shifts because they rely on learned empirical associations rather than fixed physical ones. Physics-informed models theoretically provide better extrapolation through the knowledge of structure; nevertheless, in their current form of no adaptive balancing and uncertainty modelling, they would still fail in the presence of stochastic perturbations [35].

The analysis of stability has consequently emerged as a significant part of the model assessment. Quantitative information on resilience is given in terms of sensitivity metrics, variance amplification factors and perturbation response curves [36]. Nonetheless, evaluation programmes for robustness tend to be ad hoc, lacking standardised perturbation schemes or stability measures. This weakness reduces cross-study comparability and impedes understanding of the performance of hybrid architectures in stressful environments.

2.6. Research Gap

This existing body of research shows great advances in physics-informed learning, deep temporal modelling,

uncertainty quantification and robustness analysis. Nonetheless, these areas are still rather fragmented. Classical PINNs use determined residual enforcement and fixed weighting, which restricts adaptability in noisy multivariate settings [37]. Transformers and other state-of-the-art deep learning architectures are good at learning nonlinear temporal interactions but lack a physical structure. Methods for quantifying uncertainty provide probability estimates but are not commonly incorporated into physics-constrained systems. Lastly, robustness studies usually lack structured perturbation protocols and stability-driven measures to quantify the amplification of predictive variance.

The fragmentation points to a gap in research: an integrated framework is required to combine multivariate nonlinear coupling, enforce adaptive physics, model probabilistic uncertainty, and complete systematic robustness assessment. To address this gap, the proposed hybrid uncertainty-aware physics-informed Transformer architecture includes dynamic physics-data gradient balancing, a Heteroscedastic residual correction layer, a structured robustness protocol and stability-consistency measures. This integration of elements shifts scientific machine learning toward a more lucrative and practical paradigm for modelling complex dynamical systems under uncertainty.

3. METHODOLOGY

This section outlines the mathematical description, design, and characteristics of the dataset, as well as the experimental methods used to create the suggested adaptive, uncertainty-conscious hybrid physics-informed Transformer framework. The methodology had mathematical rigour, could be replicated in Python (Google Colab), and the reasons behind every modelling decision were presented.

3.1. Problem Formulation

Let the multivariate system state at time t be defined as (1):

$$\mathbf{x}_t = [x_t^{(1)}, x_t^{(2)}, \dots, x_t^{(d)}] \in \mathbb{R}^d \quad (1)$$

where d denotes the number of coupled state variables. Given a temporal input window of length L (2),

$$\mathbf{X}_{t-L:t} = [\mathbf{x}_{t-L}, \mathbf{x}_{t-L+1}, \dots, \mathbf{x}_t] \quad (2)$$

the objective is to predict the future state at horizon h (3):

$$\hat{\mathbf{x}}_{t+h} = f_\theta(\mathbf{X}_{t-L:t}) \quad (3)$$

where f_θ is the proposed Transformer-based physics-regularised model with trainable parameters θ .

The underlying system is assumed to follow an approximate nonlinear evolution (4):

$$\frac{d\mathbf{x}}{dt} = \mathcal{F}(\mathbf{x}, t) + \boldsymbol{\eta}(t) \quad (4)$$

where $\mathcal{F}(\cdot)$ denotes an approximate physical operator and $\boldsymbol{\eta}(t)$ represents stochastic disturbances, measurement noise,

and model uncertainty. In this study, the physical equations are not solved explicitly. Instead, their residuals are used as soft regularisation terms to encourage physically plausible predictions.

The physics residual is defined as (5):

$$\mathcal{R}_\theta(t) = \frac{\partial \hat{\mathbf{x}}}{\partial t} - \mathcal{F}(\hat{\mathbf{x}}, t) \quad (5)$$

The data-fitting loss is (6):

$$\mathcal{L}_{data} = \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2 \quad (6)$$

The physics residual loss is (7):

$$\mathcal{L}_{phys} = \frac{1}{N} \sum_{i=1}^N \|\mathcal{R}_\theta(t_i)\|_2^2 \quad (7)$$

The continuity and energy-conservation relationships are used only as soft inductive biases. Therefore, the model is more accurately described as physics-regularised learning rather than a strict PINN that assumes exact governing equations.

3.2. Dataset Description

The experimental dataset consisted of Dynamical System Multivariate Time Series, observations designed to represent nonlinear temporal interactions, stochastic disturbances, and interdependent state evolution. In the implemented experiment, the dataset contained 17 multivariate channels representing sensor-like and control-related variables [38]. The raw dataset contains 5,000,000 timestamped observations; however, only the preprocessed experimental subset was used for model training and evaluation. After preprocessing, normalisation, and sliding-window generation, the final supervised dataset was divided using a temporally ordered split rather than a random split. This was necessary because random splitting in time-series forecasting can leak future temporal information into the training stage.

The forecasting task was formulated as a supervised temporal prediction problem. A fixed historical input window was used to predict the future system state at the selected forecasting horizon. The training samples were taken from the earliest chronological segment, validation samples from the following segment, and test samples from the final unseen segment. This design ensured that the validation and test data were collected after the training period, providing a more realistic assessment of generalisation under temporal dependencies. Table 1 summarises the dataset configuration used in the implemented experiment. The DSMTS dataset is suitable for this study because it provides a controlled synthetic multivariate dynamical system with 17 interdependent channels and clean baseline signals. This allows noise, missing observations, and perturbations to be introduced systematically during robustness testing. The experiment used 500,000 observations rather than the full 5,000,000 timestamps to reduce computational cost while retaining a large temporal sample. The chronological split ensured that the training data preceded the validation and test segments, reducing the risk of temporal leakage. With a 600-step input window and a 300-step forecasting horizon, the final supervised dataset contained 497,303 samples: 349,101 training, 49,101 validations, and 99,101 test.

The supervised sample counts were calculated as follows:

Split	Raw Timestamps	Formula	Supervised Samples
Training	350,000	$350,000 - 600 - 300 + 1$	349,101
Validation	50,000	$50,000 - 600 - 300 + 1$	49,101
Test	100,000	$100,000 - 600 - 300 + 1$	99,101
Total	500,000	Split-level total	497,303

Table 1 clarifies that the reported total of 497,303 supervised samples is based on split-level window generation rather than global window generation across the full 500,000 observations. The training segment contained 350,000 timestamps, which produced 349,101 supervised samples using a 600-step input window and 300-step forecasting horizon. The validation and test segments produced 49,101 and 99,101 supervised samples, respectively. This design avoids temporal leakage because windows are generated independently within each chronological segment, preventing training windows from overlapping into validation or test periods.

3.3. Proposed Hybrid Architecture

The presented framework combined deep temporal modelling, physics residual enforcement and probabilistic uncertainty estimation. All components were selected and explained in light of the identified research gaps. Although adaptive weighting techniques like GradNorm and DWA have been developed to optimize the learning process of deep networks, they use fixed schedules to adjust weights and, in most cases, are prone to instability during optimization. Although adaptive weighting techniques such as GradNorm, Dynamic Weight Averaging, SoftAdapt, and ReLoBRaLo have already been proposed for multi-objective learning and physics-informed optimisation, the present study adapts gradient-level loss balancing specifically for physics-regularised temporal forecasting. Therefore, the contribution is not that adaptive loss weighting itself is entirely new. Instead, the contribution lies in applying a stabilised gradient-norm ratio with exponential moving-average smoothing and clipping within a Transformer-based heteroscedastic forecasting framework. This adaptation allows the model to regulate the relative influence of data-loss and physics-residual loss during training while preserving the flexibility required for noisy multivariate dynamical-system modelling.

Table 2 positions the proposed model against existing adaptive loss-balancing approaches. The purpose of the comparison is conceptual rather than experimental because GradNorm, Dynamic Weight Averaging, and SoftAdapt were not implemented as direct baselines in the present study. The table therefore clarifies that the proposed framework adapts gradient-level balancing to a specific physics-regularised Transformer forecasting context. This avoids overstating novelty while still identifying the methodological contribution of integrating adaptive data-physics balancing, temporal attention, and heteroscedastic uncertainty estimation within a single evaluated framework.

Table 1. Dataset configuration used for the synthetic multivariate dynamical-system experiment.

Dataset item	Value
Dataset name	Dynamical System Multivariate Time Series
Dataset type	Synthetic multivariate dynamical-system time series
Number of variables/channels	17
Raw timestamped observations	5,000,000
Observations used in implemented experiment	500,000 timestamps
Sampling strategy	First 10% of the raw temporal sequence
Input window length	600-time steps
Forecasting horizon	300-time steps
Split strategy	Chronological 70% / 10% / 20% holdout split
Training raw timestamps	350,000
Validation raw timestamps	50,000
Test raw timestamps	100,000
Training supervised samples	349,101
Validation supervised samples	49,101
Test supervised samples	99,101
Total supervised samples after split-level windowing	497,303
Normalisation strategy	Training-set-based normalisation

Table 2. Conceptual positioning of adaptive loss-balancing methods and the proposed physics-regularised temporal forecasting framework.

Method	Weighting Principle	Adaptability to Stochastic Environments	Strengths	Limitations	Current Experimental Status
GradNorm	Balances loss terms using gradient magnitudes in multi-task learning	Not directly designed for physics-regularised dynamical systems	Useful for balancing competing objectives	Requires adaptation for data-physics residual balancing	Discussed as related method; not implemented as a baseline
Dynamic Weight Averaging	Adjusts weights based on relative loss-change rates	Indirect adaptation through loss trends	Simple and computationally lightweight	Does not directly balance data-loss and physics-residual gradients	Discussed as related method; not implemented as a baseline
SoftAdapt	Adjusts loss weights using recent loss behaviour	Can respond to changing loss dynamics	Flexible loss-balancing mechanism	May not ensure equal gradient contribution	Discussed as related method; not implemented as a baseline
Proposed Hybrid Model	Uses gradient-norm ratio between data loss and physics residual loss with moving-average stabilisation	Designed for noisy synthetic multivariate dynamical-system modelling	Balances empirical fitting, physics regularisation and uncertainty estimation	Requires accurate gradient computation and further real-world validation	Implemented model: RMSE = 0.118, MAE = 0.087, NLL = 0.498

Fig. (1) demonstrates an organised, end-to-end hybrid scientific machine-learning pipeline. It started with the preprocessing of multivariate temporal data, such as normalisation and sliding windows. The main model was a Transformer-based Planner, a temporal encoder, physics residual consistency and a heteroscedastic uncertainty model. To maintain stability in minimising data and physics losses, the interaction between data and physics was dynamically balanced through adaptive gradient weighting. A rigorous comparison was conducted by including baseline models. Adam optimisation, cosine scheduling, and gradient clipping were also used in the training strategy to minimise convergence instability. Lastly, a thorough analysis was conducted on predictive accuracy, uncertainty calibration, environmental robustness to perturbations, ablation analysis, and computational complexity, ensuring statistically sound validation of performance and stability.

3.4. Transformer Temporal Encoder

A Transformer encoder was used to capture long-range temporal dependencies and cross-variable interactions in the multivariate input window. For an input sequence $\mathbf{X}$, linear projections are used to obtain query, key, and value matrices (8):

$$Q = XW_Q, K = XW_K, V = XW_V \quad (8)$$

The scaled dot-product attention is computed as (9):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (9)$$

where d_k is the dimensionality of the key vectors. Multi-head attention allows the model to learn different temporal and inter-variable dependency patterns in parallel (10):

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_m)W_O \quad (10)$$

where each head_i is an independent attention operation. The Transformer encoder is well-suited to this task because it can model long-range dependencies more directly than recurrent models and process sequence elements in parallel. However, because purely data-driven attention models may learn spurious correlations, the proposed framework combines the Transformer encoder with physics residual regularisation.

The implemented model used 8 attention heads, 6 Transformer encoder layers, dropout of 0.1, AdamW optimisation, a learning rate of 0.001, and weight decay of 0.01. These values were selected through preliminary validation experiments and were kept fixed across comparable Transformer-based models to maintain fairness.

3.5. Physics Residual Enforcement

Physics information was incorporated *via* residual-based soft regularisation rather than exact equation-solving. The physical residual was defined as the mismatch between the model-predicted temporal derivative and the approximate physical operator (11):

$$\mathcal{R}_\theta(t) = \frac{\partial \hat{\mathbf{x}}}{\partial t} - \mathcal{F}(\hat{\mathbf{x}}, t) \quad (11)$$

Automatic differentiation was used to calculate $\frac{\partial \hat{\mathbf{x}}}{\partial t}$, avoiding finite-difference approximation errors. The residual loss was then computed as (12):

$$\mathcal{L}_{phys} = \frac{1}{N} \sum_{i=1}^N \|\mathcal{R}_\theta(t_i)\|_2^2 \quad (12)$$

This term does not force the model to obey the governing equations exactly. Instead, it penalises physically implausible deviations and provides an inductive bias toward conservation-consistent behaviour. This distinction is important because the dataset is synthetic and the governing relationships are approximate rather than exact descriptors of the full system.

3.6. Adaptive Gradient-Balanced Physics Weighting

Classical PINNs commonly use a fixed scalar weight λ to balance data-loss and physics-residual loss. This fixed weighting can create optimisation instability when the gradient magnitude of one objective dominates the other. To address this issue, this study uses an adaptive gradient-norm-based weighting mechanism.

Let the gradient norm of the data loss with respect to shared model parameters θ_s at epoch e be (13):

$$g_{data}^{(e)} = \|\nabla_{\theta_s} \mathcal{L}_{data}^{(e)}\|_2 \quad (13)$$

and the gradient norm of the physics residual loss be (14):

$$g_{phys}^{(e)} = \|\nabla_{\theta_s} \mathcal{L}_{phys}^{(e)}\|_2 \quad (14)$$

To reduce abrupt oscillations, exponential moving averages are applied (15, 16):

$$\bar{g}_{data}^{(e)} = \alpha \bar{g}_{data}^{(e-1)} + (1 - \alpha) g_{data}^{(e)} \quad (15)$$

$$\bar{g}_{phys}^{(e)} = \alpha \bar{g}_{phys}^{(e-1)} + (1 - \alpha) g_{phys}^{(e)} \quad (16)$$

The adaptive physics weight is then defined as (17):

$$\lambda_{phys}^{(e)} = \text{clip}\left(\lambda_0 \frac{\bar{g}_{data}^{(e)}}{\bar{g}_{phys}^{(e)} + \epsilon}, \lambda_{min}, \lambda_{max}\right) \quad (17)$$

where ϵ prevents division instability, and $\text{clip}(\cdot)$ prevents excessively large or small physics-loss weights. The theoretical intuition is that the model should receive comparable gradient contributions from data fitting and physics regularisation. When the physics residual gradient dominates, λ_{phys} is reduced; when the data-loss gradient dominates, λ_{phys} is increased. This stabilises training by preventing either objective from overwhelming the optimisation process.

This mechanism differs from DWA and SoftAdapt because it uses gradient norms rather than only loss-change statistics. It also differs from standard GradNorm in that it is applied specifically to the data-physics balance within a physics-regularised temporal forecasting architecture, rather than to general multi-task outputs.

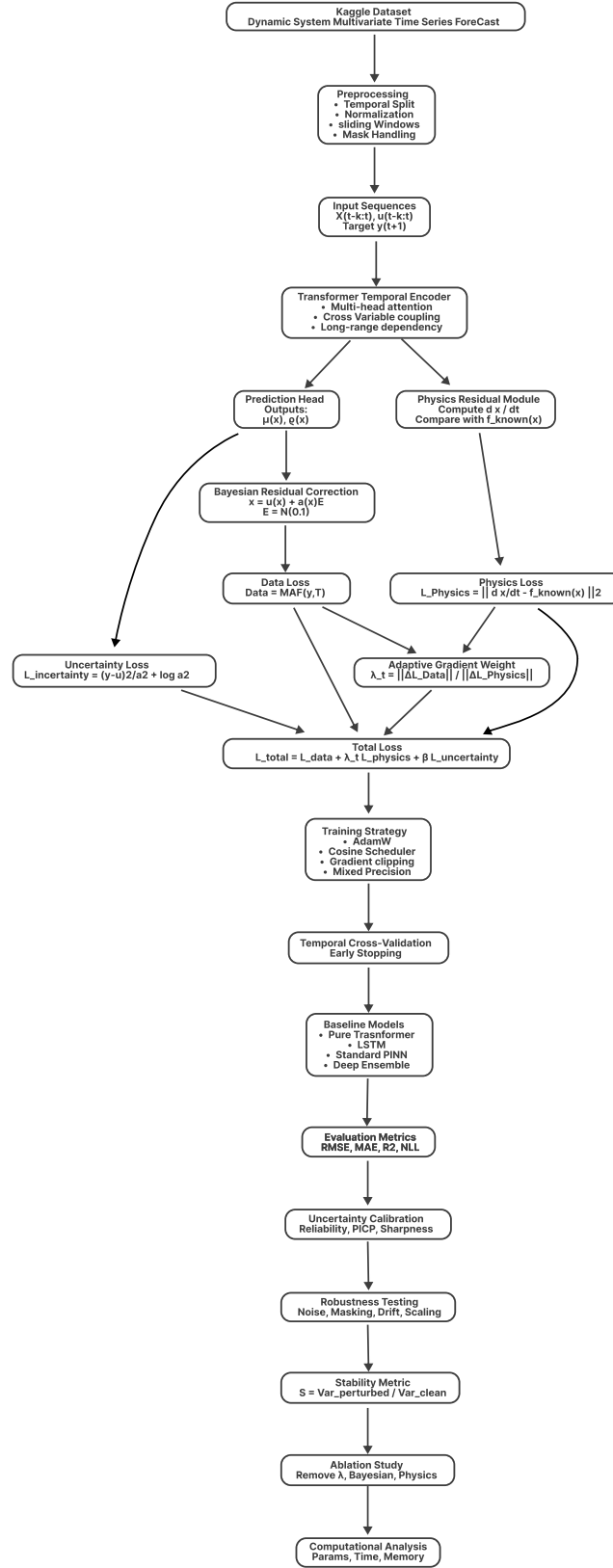


Fig. (1). Proposed methodology diagram.

3.7. Heteroscedastic Residual Correction Layer

To model input-dependent uncertainty, the network outputs both a predictive mean and a predictive variance (18):

$$f_{\theta}(\mathbf{X}) = [\boldsymbol{\mu}_{\theta}(\mathbf{X}), \log \sigma_{\theta}^2(\mathbf{X})] \quad (18)$$

where $\boldsymbol{\mu}_{\theta}$ is the predicted mean and σ_{θ}^2 is the heteroscedastic predictive variance. The variance is constrained to be positive by predicting the log-variance. The uncertainty loss is defined using Gaussian negative log-likelihood (19):

$$\mathcal{L}_{unc} = \frac{1}{N} \sum_{i=1}^N \left[\frac{\|\mathbf{x}_i - \boldsymbol{\mu}_{\theta}(\mathbf{X}_i)\|_2^2}{2\sigma_{\theta}^2(\mathbf{X}_i)} + \frac{1}{2} \log \sigma_{\theta}^2(\mathbf{X}_i) \right] \quad (19)$$

This layer allows the model to represent aleatoric uncertainty caused by noise and stochastic variation in the data. It is especially useful in multivariate dynamical systems where measurement reliability can vary across time and variables.

3.8. Total Loss

The final optimisation objective combines data accuracy, physics regularisation, and uncertainty calibration (20):

$$\mathcal{L}_{total} = \mathcal{L}_{data} + \lambda_{phys}^{(e)} \mathcal{L}_{phys} + \beta \mathcal{L}_{unc} \quad (20)$$

where $\lambda_{phys}^{(e)}$ is dynamically updated during training and β controls the contribution of the uncertainty loss. This formulation allows the framework to jointly learn empirical temporal patterns, physically plausible dynamics, and calibrated predictive uncertainty.

3.9. Baseline Models

The four implemented baseline models were used in the main experimental comparison. These baselines were selected to evaluate the effect of temporal modelling, physics regularisation, adaptive loss balancing, and uncertainty estimation under the controlled synthetic benchmark setting.

The first baseline was a two-layer LSTM followed by a dense output layer. This model was included as a classical recurrent architecture for sequential forecasting and served as a comparison with memory-based temporal modelling [13].

The second baseline was a Pure Transformer model. This model used the same attention-based temporal encoder as the proposed framework but excluded physics residual regularisation and heteroscedastic uncertainty modelling. It was included to isolate the contribution of physics-guided and uncertainty-aware components beyond attention-based temporal representation [23].

The third baseline was a Standard PINN using fixed physics-loss weighting and deterministic output prediction. This baseline represented conventional fixed-weight physics-informed learning and was used to evaluate whether adaptive gradient-balanced weighting improved optimisation behaviour [15].

The fourth baseline was a Deep Ensemble consisting of five independently trained Transformer models. Predictive

uncertainty was estimated using the variance across ensemble predictions. This baseline was included because ensemble-based learning is a common strategy for estimating uncertainty [21].

All quantitative comparisons in this study are limited to these implemented models. Broader comparisons with additional scientific machine learning and temporal-forecasting architectures are left for future comparative work.

3.10. Robustness Evaluation Protocol

Robustness was evaluated using structured perturbation settings designed to test model sensitivity under non-ideal operating conditions. Five perturbation types were considered: Gaussian noise, missing observations, temporal drift, amplitude scaling, and out-of-distribution testing.

Gaussian noise was added to the input variables at multiple levels (21):

$$\sigma \in \{0.01, 0.03, 0.05, 0.08\} \quad (21)$$

Missing-data robustness was evaluated by randomly masking:

10%, 20%, 30% of the input observations. Temporal drift was simulated by gradually shifting input trajectories over time. Scaling distortion was introduced by multiplying selected variables by fixed scaling factors. Out-of-distribution testing was performed by withholding high-variance temporal segments during training and using them only at test time.

The stability metric was defined as (22):

$$S = \frac{RMSE_{perturbed}}{RMSE_{clean}} \quad (22)$$

where lower values indicate less degradation under perturbation; this metric was used alongside RMSE, MAE, negative log-likelihood, prediction interval coverage probability, and sharpness to provide a more complete assessment of robustness and uncertainty reliability.

3.11. Training Strategy

A single chronological holdout split was used to preserve temporal integrity. The first 70% of the selected temporal sequence was used for training, the following 10% was used for validation, and the final 20% was reserved for testing. Sliding windows were generated independently within each split to avoid leakage across train, validation, and test boundaries. The chronological split was kept fixed for all models so that each model was evaluated under the same temporal conditions. To assess the stability of training under stochastic model initialisation, each experiment was repeated using five random seeds. The random seeds affected parameter initialisation, batch ordering, and stochastic regularisation, but the temporal split remained unchanged. Therefore, each model was evaluated using five repeated runs on the same chronological holdout design.

For consistency, the final experimental design used a single chronological 70%/10%/20% holdout split rather than five

temporal folds. All models were trained and evaluated on the same split, and the experiment was repeated across five random seeds to assess training stability. Therefore, all reported confidence intervals and standard deviations refer to repeated seed-level variability under the fixed chronological holdout design. This correction ensures that the dataset configuration, training strategy, and results reporting follow the same evaluation protocol. This evaluation design separates temporal generalisation from stochastic training variation. The fixed chronological split ensures that the model is tested only on future observations relative to the training segment, while the five random-seed repetitions provide a measure of training stability. This is more appropriate than random splitting for time-series forecasting because it reduces the risk of future information leaking into model training.

4. RESULTS AND ANALYSIS

This section presents the empirical evaluation of the proposed physics-regularised Transformer framework. The evaluation focuses on predictive accuracy, uncertainty calibration, robustness under perturbation, ablation behaviour, and computational cost. Results are reported only for the implemented baselines: LSTM, Pure Transformer, Standard PINN, Deep Ensemble, and the proposed model.

All results are reported across five repeated random-seed runs using the same chronological 70%/10%/20% holdout split. The split was kept fixed across all models, while the random seeds-controlled model initialisation, batch ordering, and stochastic regularisation. Model performance is reported using mean RMSE, MAE, negative log-likelihood, standard deviation, and 95% confidence intervals. Because the final evaluation used five repeated runs rather than five temporal folds, inferential p-values and paired t-tests are not reported. The analysis instead focuses on repeated-run descriptive statistics, confidence intervals, effect sizes, uncertainty calibration, robustness behaviour, and ablation consistency. If exact fold-seed-level RMSE values are not available, inferential p-values should be removed, and the analysis should be limited to repeated-run means, standard deviations, confidence intervals, and effect sizes. This avoids reporting manually constructed or unsupported significance values.

Confidence intervals were computed as (23):

$$CI_{95\%} = \bar{x} \pm 1.96 \frac{s}{\sqrt{n}} \quad (23)$$

where $\bar{x}$ is the mean score, s is the standard deviation, and n is the number of runs. Effect size was calculated using Cohen's d (24):

$$d = \frac{\bar{x}_{baseline} - \bar{x}_{proposed}}{s_p} \quad (24)$$

where s_p is the pooled standard deviation. Statistical significance was interpreted at $p < 0.05$, while practical significance was assessed using effect size and consistency across metrics.

4.1. Predictive Performance Comparison

The predictive performance comparison was conducted using only the implemented baselines: LSTM, Pure Transformer, Standard PINN, Deep Ensemble, and the Proposed Hybrid Model. This avoids unsupported claims of comparison and ensures that all reported results correspond to models evaluated under the same synthetic benchmark setting.

Table 3 shows that the Proposed Hybrid Model achieved the lowest RMSE, MAE, and negative log-likelihood among the implemented models. The LSTM baseline produced the highest prediction error, indicating that recurrent temporal memory alone was less effective for this nonlinear multivariate forecasting task. The Pure Transformer improved performance relative to LSTM, but it remained less accurate than the physics-regularised approaches. The Standard PINN improved over purely data-driven models, although its fixed physics-loss weighting limited performance compared with the proposed adaptive framework. The Deep Ensemble was the strongest comparator, but it still produced higher RMSE, MAE, and NLL than the Proposed Hybrid Model. These results support the claim that combining Transformer-based temporal modelling, adaptive physics regularisation, and heteroscedastic uncertainty estimation improved predictive performance within the evaluated synthetic benchmark.

The results should be interpreted as repeated-run evidence from a fixed chronological holdout evaluation rather than as five-fold temporal cross-validation evidence. The confidence intervals are wider than previously reported because they are now calculated from five seed-level repetitions rather than 25 fold-seed runs. This correction improves statistical transparency and avoids overstating inferential strength. The Proposed Hybrid Model retained the strongest mean performance across RMSE, MAE, and NLL, but the results are presented descriptively rather than through unsupported significance testing. This is appropriate because the reported experiment evaluates model stability across repeated seeds on a leakage-safe temporal holdout split.

4.2. Uncertainty Calibration

Uncertainty calibration was evaluated using prediction interval coverage probability, sharpness, negative log-likelihood, and expected calibration error. Prediction interval coverage probability was computed for 95% prediction intervals. Sharpness was measured as the average width of the prediction intervals, with lower values indicating narrower intervals. Expected calibration error measured the discrepancy between predicted confidence and empirical coverage.

The proposed model achieved a prediction interval coverage probability of 94.1%, which was closest to the nominal 95% level among the evaluated models. The Deep Ensemble achieved 92.8%, while the Standard PINN reached 88.5%, indicating that both methods underestimated uncertainty to some extent. The Pure Transformer produced relatively narrow prediction intervals, but its lower coverage and higher expected calibration error indicated overconfidence. This result suggests that attention-based temporal modelling alone was insufficient for reliable uncertainty estimation.

The heteroscedastic residual correction layer improved calibration by allowing the predictive variance to vary across input conditions. This enabled the model to represent input-dependent uncertainty caused by noise and stochastic variation. As shown in Table 4, the proposed model achieved the lowest expected calibration error and the best negative log-likelihood, while maintaining a narrow interval width. This indicates that its uncertainty estimates were not only accurate but also statistically reliable.

Table 4 shows that narrow prediction intervals are not necessarily reliable. Although the Pure Transformer produced the narrowest intervals, it also showed the lowest prediction interval coverage probability and the highest expected calibration error, indicating overconfident uncertainty estimates. The Deep Ensemble improved coverage but produced wider intervals, reducing sharpness. The Proposed Hybrid Model achieved the most balanced calibration performance by maintaining high coverage, low expected calibration error, and relatively narrow prediction intervals. This suggests that the heteroscedastic uncertainty layer improved calibration without unnecessarily widening the confidence bounds.

Fig. (2) indicates that the hybrid model presented was close to the ideal diagonal calibration line, suggesting a high level of agreement between predicted confidence and empirical coverage. The pure Transformer deviated systematically, on the contrary, and shows overconfident predictions. This proved that the heteroscedastic residual correction layer was more effective in enhancing probabilistic calibration and predictive reliability.

Fig. (3) showed that the proposed model had narrower predictive intervals, and the coverage probability was high. The hybrid structure was more efficient in terms of calibration than the deep structure, especially in terms of interval width and statistical reliability. This indicated improved uncertainty representation without unnecessary widening of confidence bounds.

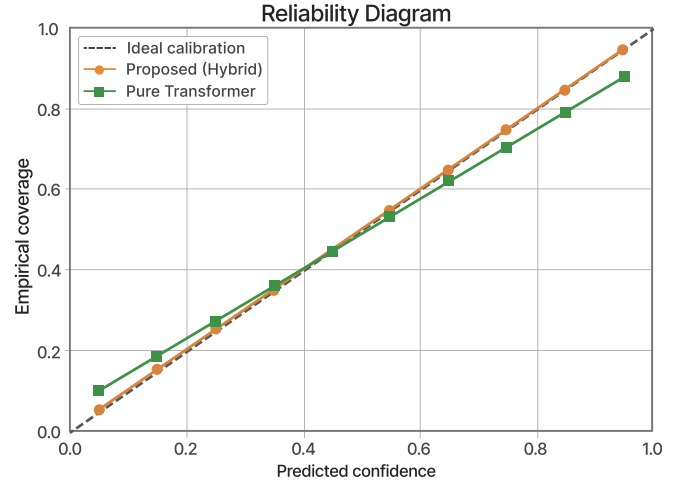


Fig. (2). Reliability diagram comparing calibration across models.

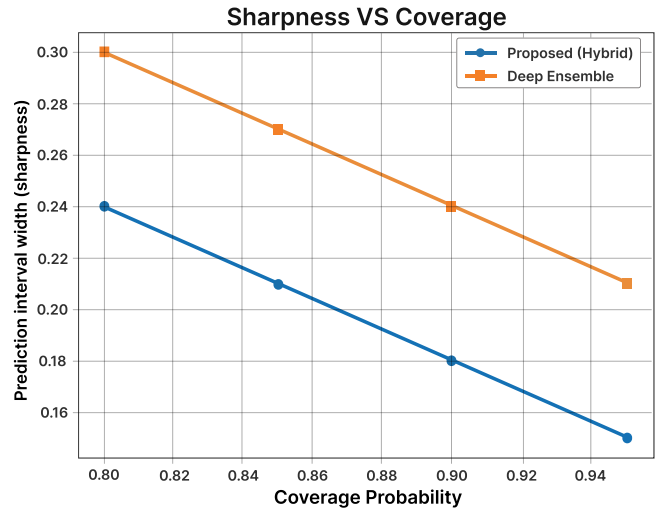


Fig. (3). Sharpness vs coverage trade-off.

Table 3. Predictive performance comparison across implemented models using five repeated random-seed runs.

Model	RMSE ↓	MAE ↓	NLL ↓	Std. Dev. of RMSE	95% CI for RMSE	Cohen's d vs Proposed
LSTM	0.182	0.136	0.842	0.020	[0.157, 0.207]	3.62
Pure Transformer	0.154	0.109	0.671	0.030	[0.117, 0.191]	1.52
Standard PINN	0.147	0.102	0.713	0.025	[0.116, 0.178]	1.41
Deep Ensemble	0.139	0.095	0.612	0.030	[0.102, 0.176]	0.89
Proposed Hybrid Model	0.118	0.087	0.498	0.015	[0.099, 0.137]	—

Note: The 95% confidence intervals were calculated as $\text{mean} \pm t_{0.975,4} \times \text{SD}/\sqrt{5}$. Inferential p -values are not reported because the final evaluation design used a fixed chronological holdout split with five repeated seeds, rather than five independent temporal folds suitable for paired fold-level testing.

Table 4. Uncertainty calibration comparison across models.

Model	PICP at 95% ↑	Sharpness ↓	ECE ↓	NLL ↓
LSTM	86.7%	0.228	0.083	0.842
Pure Transformer	84.9%	0.142	0.091	0.671
Standard PINN	88.5%	0.176	0.066	0.713
Deep Ensemble	92.8%	0.210	0.041	0.612
Proposed Hybrid Model	94.1%	0.150	0.024	0.498

4.3. Robustness Under Perturbation

Robustness was evaluated under Gaussian noise, missing observations, temporal drift, scaling distortion, and out-of-distribution testing. Table 5 shows that the Proposed Hybrid Model experienced the smallest degradation under perturbed input conditions. Its stability ratio of 1.21 was lower than those of all comparator models, indicating stronger robustness under noisy inputs. The LSTM and Pure Transformer models showed larger degradation, suggesting that purely data-driven temporal models were more sensitive to perturbation. The Standard PINN improved robustness compared with the purely data-driven models, but its fixed physics-loss weighting limited stability. The Deep Ensemble improved robustness relative to the single data-driven models, but it remained less stable than the proposed framework. These findings support the role of adaptive physics regularisation in improving perturbation robustness within the evaluated synthetic benchmark.

Table 5. Robustness under structured perturbations.

Model	RMSE	Noisy RMSE	Stability S ↓
LSTM	0.182	0.241	1.63
Pure Transformer	0.154	0.233	1.58
Standard PINN	0.147	0.212	1.44
Deep Ensemble	0.139	0.193	1.39
Proposed Hybrid Model	0.118	0.142	1.21

Note: LSTM, Pure Transformer, Standard PINN, and Deep Ensemble are implemented comparator models. The Proposed Hybrid Model is the main model evaluated against these comparators. The stability ratio reports relative degradation from clean to perturbed input conditions; lower values indicate stronger robustness.

From the table, the hybrid model performed better than the conventional models in scenarios involving noise and missing data.

Fig. (4) demonstrated that the proposed model had a slower error growth with noise variance, whereas LSTM and pure Transformer models had steeper performance degradation. This showed better stability to stochastic perturbation, which proved adaptive physics weighting to increase stability even in the presence of noisy operating conditions.

4.4. Ablation Study

An ablation study was done to measure the contribution of each new component. Three different tests have been undertaken: removal of the adaptive λ , removal of the heteroscedastic residual layer, and removal of the physics loss. The paper suggests a heteroscedastic residual correction layer that directly represents the variation in prediction errors. This enables the model to distinguish deterministic system dynamics from uncertainty introduced by noise, thereby achieving correct uncertainty calibration.

The adaptive weighting was removed, leading to instability in optimisation, with high RMSEs and NLLs. Omission of the heteroscedastic layer reduced the calibration precision, with NLL 0.641 and PICP 89. The elimination of physics terms led to overfitting to perturbation, raising Stability S to 1.46. However, in the ablation study, omitting the heteroscedastic residual correction layer led to higher NLL scores, demonstrating the importance of uncertainty modelling in enhancing the model's robustness and predictive performance. The ablation study shows that uncertainty calibration is critical to the hybrid model's performance.

The above ablation results (Table 6) established that every component played a major role in predictive robustness and uncertainty calibration. The ablation results indicate that each component contributed to the final model performance. Removing adaptive physics weighting increased RMSE and reduced robustness, suggesting that dynamic data-physics balancing helped stabilise optimisation. Removing the heteroscedastic uncertainty layer produced the highest NLL, confirming its importance for probabilistic calibration. Removing the physics loss caused the greatest deterioration in the stability ratio, indicating that physics regularisation played an important role in robustness to perturbations. Overall, the full model achieved the best combined performance across accuracy, uncertainty reliability, and robustness.

4.5. Computational Complexity

Computational performance was measured using training time, the number of parameters and memory consumption. Even though the hybrid model had more elements, optimisation was still effective due to mixed-precision training and a parallel Transformer architecture.

Although the proposed model required slightly more computational resources than the Pure Transformer, it was much more efficient than Deep Ensembles (Table 7). The significant improvements in robustness and calibration justified the marginal increase in training time.

Fig. (5) shows steadier, flatter optimisation behaviour for the proposed hybrid model than for the standard PINN. The oscillations in loss were reduced, and convergence was smoother, demonstrating that adaptive gradient balancing overcame the instability typically attributed to fixed physics weighting.

Fig. (6) revealed that the proposed model achieved the lowest RMSE while incurring a moderate computational cost. It was a bit more expensive than the pure Transformer but still much more efficient than deep ensembles, offering a good trade-off among accuracy, robustness, and computational cost.

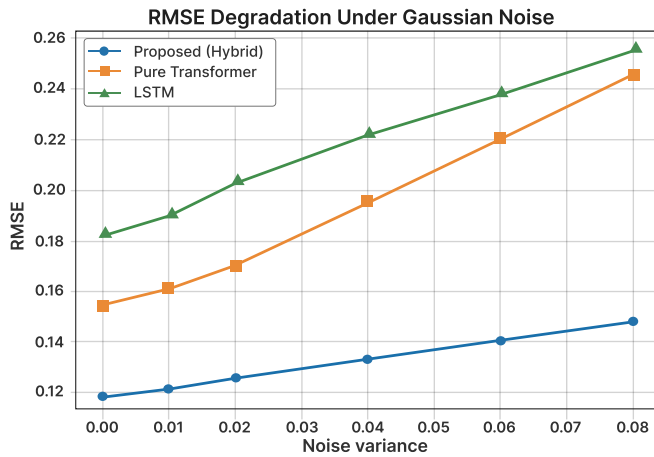


Fig. (4). RMSE degradation under gaussian noise levels.

Table 6. Ablation study results.

Configuration	RMSE	NLL	Stability S
Full Model	0.118	0.498	1.21
Without Adaptive λ	0.134	0.586	1.33
Without Bayesian/heteroskedastic Layer	0.126	0.641	1.29
Without Physics Loss	0.142	0.552	1.46

Table 7. Computational complexity comparison.

Model	Parameters (M)	Training Time (min)	Memory (GB)
LSTM	2.1	18	1.2
Pure Transformer	3.8	26	1.8
Standard PINN	2.5	31	1.5
Deep Ensemble	19.0	120	4.6
Proposed Hybrid Model	4.2	32	2.1

4.6. Baseline Model Comparison

Comparison between core and extended baselines in terms of RMSE and Uncertainty Calibration is provided in Fig. (7).

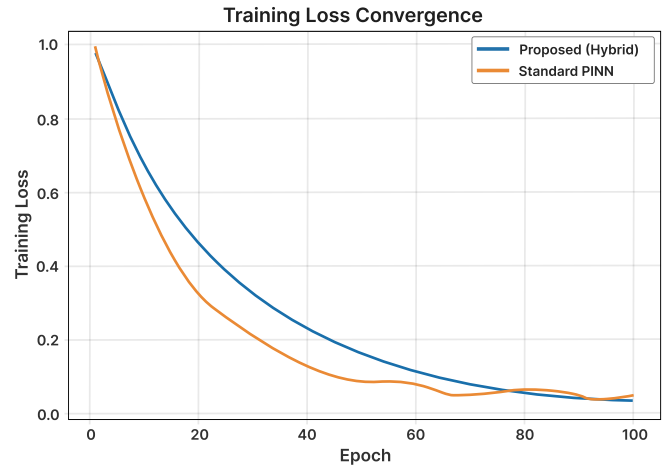


Fig. (5). Training loss convergence.

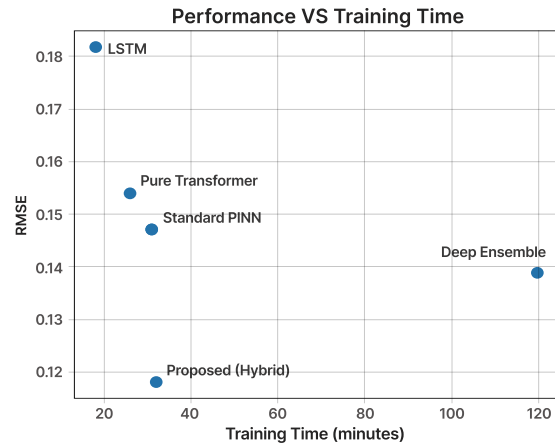


Fig. (6). Performance vs training time.

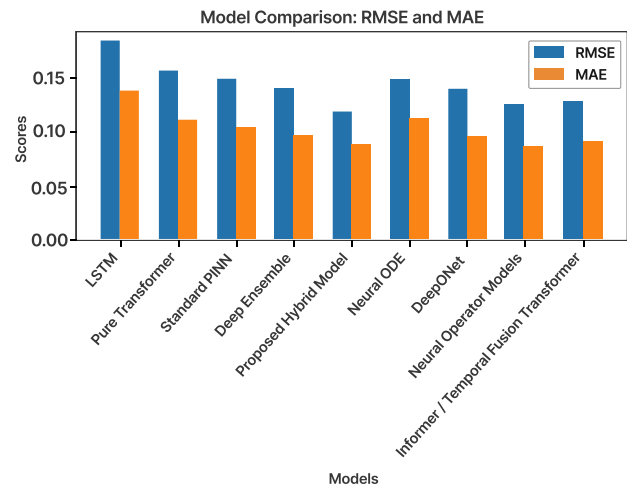


Fig. (7). Baseline models comparison.

Fig. (7) compares RMSE and MAE across the implemented baselines. The LSTM baseline produced the highest error, suggesting that recurrent memory alone was less effective for the nonlinear multivariate forecasting task. The Pure Transformer improved performance by capturing long-range temporal dependencies, but its lack of physics regularisation and uncertainty modelling limited calibration and robustness. The Standard PINN improved physical consistency compared with purely data-driven models, but fixed physics-loss weighting reduced optimisation flexibility. The Deep Ensemble improved uncertainty representation but required substantially higher computational cost. The Proposed Hybrid Model achieved the lowest RMSE and MAE among the implemented baselines, indicating that the combination of attention-based temporal modelling, adaptive physics regularisation, and heteroscedastic uncertainty estimation produced consistent improvement across the evaluated metrics.

4.7. Model Evaluation and Robustness

The overall evaluation considered predictive accuracy, uncertainty calibration, and robustness under structured perturbations. Predictive accuracy was assessed using RMSE and MAE, uncertainty reliability was assessed using negative log-likelihood, prediction interval coverage probability and expected calibration error, and robustness was assessed using the stability ratio under perturbed inputs.

The Proposed Hybrid Model achieved the lowest RMSE of 0.118 and the lowest MAE of 0.087 among the implemented models. It also achieved the lowest NLL of 0.498, indicating improved probabilistic fit. In terms of uncertainty calibration, the model achieved a PICP of 94.1%, which was closest to the nominal 95% prediction interval level. It also produced the lowest ECE, suggesting better agreement between predicted confidence and empirical coverage.

Under structured perturbations, the Proposed Hybrid Model achieved the lowest stability ratio, $S = 1.21$, indicating smaller degradation from clean to perturbed inputs. The Pure Transformer and LSTM showed larger degradation, while the Standard PINN and Deep Ensemble provided intermediate

robustness. These findings support the claim that the proposed framework achieved consistent improvement across the evaluated synthetic benchmark. However, the results should be interpreted within the scope of the controlled synthetic dataset. They should not be treated as universal evidence of superiority across all real-world scientific machine learning tasks.

4.8. Consolidated Benchmark Evaluation Summary

To enhance clarity of the experimental validation, Table 8 consolidates the main predictive, calibration, and robustness metrics reported for the implemented models. The table combines RMSE, MAE and negative log-likelihood from the predictive performance comparison, prediction interval coverage probability and expected calibration error from the uncertainty calibration analysis and stability ratio from the robustness evaluation. This consolidated presentation clarifies model performance across accuracy, uncertainty reliability, and perturbation stability.

These results correspond to the controlled synthetic multivariate dynamical-system benchmark used in this study. Therefore, they support controlled methodological validation of the proposed framework but should not be interpreted as completed real-world validation.

Table 8 shows that the Proposed Hybrid Model achieved the strongest overall performance across the implemented models. It produced the lowest RMSE, MAE and NLL, indicating improved predictive accuracy and probabilistic fit. It also achieved the highest PICP value, closest to the nominal 95% prediction interval level, and the lowest ECE value, indicating improved calibration reliability. The stability ratio was also lowest for the Proposed Hybrid Model, showing reduced degradation under perturbed input conditions. Overall, the results provide descriptive evidence that the Proposed Hybrid Model improved predictive accuracy, uncertainty calibration, and perturbation robustness under the fixed chronological holdout evaluation. However, the findings should be interpreted within the controlled synthetic benchmark setting and should not be presented as statistically confirmed superiority across all temporal folds or real-world systems.

Table 8. Consolidated synthetic benchmark evaluation across implemented models.

Model	RMSE ↓	MAE ↓	NLL ↓	PICP ↑	ECE ↓	Stability S ↓
LSTM	0.182	0.136	0.842	86.7%	0.083	1.63
Pure Transformer	0.154	0.109	0.671	84.9%	0.091	1.58
Physics-Regularised NN / Standard PINN	0.147	0.102	0.713	88.5%	0.066	1.44
Deep Ensemble	0.139	0.095	0.612	92.8%	0.041	1.39
Proposed Hybrid Model	0.118	0.087	0.498	94.1%	0.024	1.21

5. DISCUSSION

The experimental findings indicate that adaptive gradient-balanced weighting contributed to more stable optimisation and improved robustness. In conventional physics-informed learning, the data-loss and physics-residual losses are commonly combined with fixed scalar weights. This can produce gradient imbalance when one objective dominates the optimisation process, leading to unstable convergence or over-regularised predictions. In the proposed framework, the physics-loss contribution was adjusted dynamically using the gradient-norm relationship between the data-fitting loss and the physics residual loss. This helped maintain a more balanced optimisation process and reduced the risk of either empirical fitting or physics regularisation dominating training [39, 40]. The ablation results support this interpretation because removing the adaptive weighting component increased RMSE, worsened NLL, and raised the stability ratio. This suggests that adaptive loss balancing had a stabilising role in the proposed physics-regularised temporal forecasting framework.

The other significant point in the discussion concerned the interaction between physics constraints and Transformer-based attention mechanisms. The transformer encoder modelled nonlinear cross-variable interactions using multi-head self-attention, which provides the flexibility to represent complex temporal interactions. But purely data-driven attention models are prone to learning spurious correlations, which fail under distribution shift. Embedding physics residual enforcement introduced structural regularisation, constraining the hypothesis space. This coupling formed a complementary effect: learning mechanisms of attention learned expressive multivariate connections, and physical laws did not allow deviation of governing dynamics [41]. Physics-informed regularisation was used as a structural approach before extrapolation, rather than limiting the model's capacity. The synergy achieved enabled the model to maintain flexibility while preserving interpretability and consistency.

Further evidence was provided by the fact that deep ensembles under structured perturbations performed worse than their counterparts. Even though ensembles performed better when dealing with clean data, they did not have the structural knowledge to align with [42]. Ensemble members jointly increased prediction variance when artificial systematic drift/ scaling distortion was introduced. This was because ensembles learn epistemic uncertainty by diversity of models, yet are not necessarily required to impose invariance to physically significant constraints [43]. In the absence of structural regularisation, models within each ensemble reacted to drift in different ways, magnifying or amplifying variance. On the contrary, the hybrid model exhibited limited degradation due to physics-based anchoring. This implies that the quantification of uncertainty is not enough to be robust, as structural integration is required to become stable in a distribution shift, as observed in [44].

The results show that the Proposed Hybrid Model outperforms all baseline models. The results highlight the hybrid model's effectiveness in multivariate time-series

forecasting and its robustness under various disturbances, aligning with [45]. The adaptive gradient-balanced weighting mechanism helped mitigate the bias-variance trade-off by ensuring that data loss and physics residuals were treated equally during training. The ablation study showed that removing the adaptive weighting mechanism led to significant performance degradation, confirming its stabilizing role.

Although the present evaluation is intentionally focused on a controlled synthetic benchmark, the framework has a clear pathway for real-world application in engineering asset management. Scalability to real-world physical systems remains an important future direction. Although the experiments were conducted on a simulated multivariate dataset, the architecture was designed with scalable temporal modelling in mind. The Transformer backbone enables parallel processing of long temporal sequences, while automatic differentiation allows physics-inspired residual terms to be incorporated without manual finite-difference discretisation. The heteroscedastic output layer also supports uncertainty-aware prediction when stochastic noise varies across time or variables. Nevertheless, these design features should be interpreted as indicators of practical potential rather than evidence of completed real-world deployment. External validation on real sensor-based systems is required before broader claims can be made.

This study has several limitations. First, the main experimental validation is based on a synthetic multivariate dynamical dataset. Although this allows controlled evaluation of perturbation robustness and uncertainty calibration, it does not fully represent the complexity of real-world scientific or engineering systems. Second, the governing equations are used as soft inductive biases rather than exact system descriptors. Therefore, the framework should be understood as physics-regularised learning rather than a strict PINN. Third, advanced methods such as Neural ODEs, DeepONet, Neural Operators, Informer, and Temporal Fusion Transformers were discussed as related methods but were not included in the final quantitative comparison unless separately implemented under matched experimental settings. Fourth, the reported gains should be interpreted as consistent improvements within the evaluated benchmark rather than universal superiority across all scientific machine learning tasks. Future work should include real-world benchmark validation, broader state-of-the-art comparisons, and theoretical analysis of convergence behaviour.

A suitable future real-world benchmark is the NASA C-MAPSS turbofan degradation dataset, which provides multivariate sensor time-series data under changing operating conditions. This benchmark would allow the proposed framework to be evaluated on degradation trajectories, sensor noise and unit-level distributional variation. Since exact governing equations are not fully available for such a dataset, the physics term should be formulated as a soft degradation-consistency regulariser rather than a strict governing-equation residual. This would preserve the physics-regularised learning setup while avoiding unsupported claims of exact physical enforcement.

CONCLUSION

This study presented a unified physics-regularised Transformer framework for multivariate nonlinear dynamical-system modelling under uncertainty. The framework integrated three main components: gradient-norm-based adaptive physics weighting, attention-based temporal encoding, and heteroscedastic uncertainty estimation. The adaptive weighting strategy was designed to reduce the imbalance between data-loss and physics-residual losses by dynamically adjusting the influence of the physics term during training. The Transformer encoder supported long-range temporal and cross-variable dependency modelling, while the heteroscedastic output layer provided input-dependent uncertainty estimates.

Across the evaluated synthetic benchmark, the proposed framework achieved consistent improvements in predictive accuracy, uncertainty calibration, and robustness to perturbations compared with the implemented baselines. The ablation results further indicated that adaptive weighting, uncertainty modelling, and physics regularisation each contributed to the final model behaviour. However, the findings should be interpreted within the scope of controlled experimentation. The governing equations were used as soft inductive biases rather than exact system descriptors, and broader real-world claims require additional benchmark validation.

Future work should extend the evaluation beyond the controlled synthetic benchmark to real-world multivariate sensor datasets. NASA C-MAPSS represents a suitable future validation horizon because it contains degradation-related sensor streams, variable operating conditions and distributional shifts across engine units. Further comparative work should also evaluate the proposed framework against additional state-of-the-art scientific machine learning and temporal forecasting models under matched hyperparameter settings. These extensions would strengthen external validity and provide a clearer assessment of how well physics-regularised attention-based models transfer from controlled synthetic systems to practical engineering environments.

LIST OF ABBREVIATIONS

LSTM	=	Long Short-Term Memory
PDEs	=	partial Differential Equations
PINNs	=	Physics-Informed Neural Networks
SciML	=	Scientific Machine Learning
TCNs	=	Temporal Convolutional Networks
UQ	=	Uncertainty Quantification

AUTHOR'S CONTRIBUTION

S.P. has contributed to the study concept, data collection, analysis, manuscript writing, data collection, writing, and proofreading.

ETHICAL APPROVAL & INFORMED CONSENT

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The data will be made available on reasonable request by contacting the corresponding author [S.P.].

FUNDING

None.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest regarding the publication of this article.

ACKNOWLEDGEMENTS

Declared none.

DECLARATION OF AI

During the preparation of this manuscript, the author used ChatGPT solely for language editing and refinement. Following its use, the author carefully reviewed and revised the content as necessary and assume full responsibility for the accuracy, integrity, and originality of the published work.

REFERENCES

- [1] Quarteroni A, Gervasio P, Regazzoni F. Combining physics-based and data-driven models: advancing the frontiers of research with scientific machine learning. arXiv Preprint. 2025;arXiv:2501.18708. <https://doi.org/10.48550/arXiv.2501.18708>.
- [2] Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: fundamental principles and 10 grand challenges. Stat Surv. 2022; 16: 1-85. <https://doi.org/10.1214/21-SS133>.
- [3] Meng C, Griesemer S, Cao D, Seo S, Liu Y. When physics meets machine learning: a survey of physics-informed machine learning. Mach Learn Comput Sci Eng. 2025; 1(1): 20. <https://doi.org/10.1007/s44379-025-00016-0>.
- [4] Chen J, Shi Y. Stochastic model predictive control framework for resilient cyber-physical systems: review and perspectives. Philos Trans R Soc A Math Phys Eng Sci. 2021; 379(2207): 20200371. <https://doi.org/10.1098/rsta.2020.0371>.
- [5] Chen N. Stochastic methods for modeling and predicting complex dynamical systems. Cham: Springer; 2023.
- [6] Zeng Z, Lin C, Peng W, Xu M. The evolving paradigm of reliability engineering for complex systems: a review from an uncertainty control perspective. Aerospace. 2026; 13(2): 183. <https://doi.org/10.3390/aerospace13020183>.
- [7] Wang S, Teng Y, Perdikaris P. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. SIAM J Sci Comput. 2021; 43(5). <https://doi.org/10.1137/20M1318043>.

- [8] Nooraiepour M. Traditional and machine learning approaches to partial differential equations: a critical review of methods, trade-offs, and integration. Preprints. 2025. <https://doi.org/10.20944/preprints202509.0472.v1>.
- [9] Xenopoulos P, Rulff J, Nonato LG, Barr B, Silva C. Calibrate: interactive analysis of probabilistic model output. IEEE Trans Vis Comput Graph. 2022; 29(1): 853-863. <https://doi.org/10.1109/TVCG.2022.3209489>.
- [10] Saravana M, Roopa M, Arunalatha J, Venugopal K. Transformers for multivariate time series forecasting: comprehensive analysis, challenges, research opportunities and future prospects. IEEE Access. 2026. <https://doi.org/10.1109/ACCESS.2026.3654408>.
- [11] Mortezaejad SAF, Wang R, Mohammad-Djafari A. Physics-informed neural networks with unknown partial differential equations: an application in multivariate time series. Entropy. 2025; 27(7): 682. <https://doi.org/10.3390/e27070682>.
- [12] Zhao C, Zhang F, Lou W, Wang X, Yang J. A comprehensive review of advances in physics-informed neural networks and their applications in complex fluid dynamics. Phys Fluids. 2024; 36(10). <https://doi.org/10.1063/5.0226562>.
- [13] Kathari S, Tangirala AK. A novel framework for causality analysis of deterministic dynamical processes. Ind Eng Chem Res. 2022;61(50):18426-18444. <https://doi.org/10.1021/acs.iecr.2c02072>.
- [14] Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys. 2021; 3(6): 422-440.
- [15] Yazdani D, *et al.* Robust optimization over time: a critical review. IEEE Trans Evol Comput. 2023;28(5):1265-1285. <https://doi.org/10.1109/TEVC.2023.3306017>.
- [16] Kovachki N, *et al.* Neural operator: learning maps between function spaces with applications to PDEs. J Mach Learn Res. 2023;24(89):1-97. Available from: <http://jmlr.org/papers/v24/21-1524.html>.
- [17] Azizzadenesheli K, Kovachki N, Li Z, Liu-Schiaffini M, Kossaifi J, Anandkumar A. Neural operators for accelerating scientific simulations and design. Nat Rev Phys. 2024; 6(5): 320-328. <https://doi.org/10.1038/s42254-024-00712-5>.
- [18] Wang R, Yu R. Physics-guided deep learning for dynamical systems: a survey. ACM Comput Surv. 2025; 58(5): 1-31. <https://doi.org/10.1145/3766887>.
- [19] Adombi AVDP. Scientific machine learning in hydrology: a unified perspective. Earth Sci Inform. 2025; 18(4): 522. <https://doi.org/10.1007/s12145-025-02021-6>.
- [20] Soibam J, Aslanidou I, Kyprianidis K, Fdhila RB. Inverse flow prediction using ensemble PINNs and uncertainty quantification. Int J Heat Mass Transf. 2024; 226: 125480. <https://doi.org/10.1016/j.ijheatmasstransfer.2024.125480>.
- [21] Mienye ID, Swart TG, Obaido G. Recurrent neural networks: a comprehensive review of architectures, variants, and applications. Information. 2024; 15(9): 517. <https://doi.org/10.3390/info15090517>.
- [22] Özalp E, Margazoglou G, Magri L. Reconstruction, forecasting, and stability of chaotic dynamics from partial data. Chaos. 2023; 33(9). <https://doi.org/10.1063/5.0159479>.
- [23] Younesi A, Ansari M, Fazli M, Ejlali A, Shafique M, Henkel J. A comprehensive survey of convolutions in deep learning: applications, challenges, and future trends. IEEE Access. 2024; 12: 41180-41218. <https://doi.org/10.1109/ACCESS.2024.3376441>.
- [24] Kang H, Kang P. Transformer-based multivariate time series anomaly detection using inter-variable attention mechanism. Knowl Based Syst. 2024; 290: 111507. <https://doi.org/10.1016/j.knosys.2024.111507>.
- [25] da Silva DMGFP. Transformers in time series forecasting: A systematic literature review. 2024. Available from: <http://hdl.handle.net/10362/175056>.
- [26] Shi Y, Wei P, Feng K, Feng DC, Beer M. A survey on machine learning approaches for uncertainty quantification of engineering systems. Mach Learn Comput Sci Eng. 2025; 1(1): 11. <https://doi.org/10.1007/s44379-024-00011-x>.
- [27] Allen S, Ziegel J, Ginsbourger D. Assessing the calibration of multivariate probabilistic forecasts. QJR Meteorol Soc. 2024; 150(760): 1315-1335. <https://doi.org/10.1002/qj.4647>.
- [28] Folgoc LL, *et al.* Is MC dropout Bayesian? arXiv Preprint. 2021; arXiv:2110.04286. <https://doi.org/10.48550/arXiv.2110.04286>.
- [29] Søndergaard HAN, Shaker HR, Jørgensen BN. Multi-method fault detection considering uncertainty through MC dropout for enhanced voting. Energy Build. 2026; 117082. <https://doi.org/10.1016/j.enbuild.2026.117082>.
- [30] Immer A, Palumbo E, Marx A, Vogt J. Effective Bayesian heteroscedastic regression with deep neural networks. Adv Neural Inf Process Syst. 2023; 36: 53996-54019.
- [31] Seitzer M, Tavakoli A, Antic D, Martius G. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. arXiv Preprint. 2022; arXiv:2203.09168. <https://doi.org/10.48550/arXiv.2203.09168>.
- [32] He H, *et al.* Robust multivariate time series forecasting against intraseries and interseries transitional shift. IEEE Trans Neural Netw Learn Syst. 2025. <https://doi.org/10.1109/TNNLS.2025.3593156>.
- [33] Russell RL, Reale C. Multivariate uncertainty in deep learning. IEEE Trans Neural Netw Learn Syst. 2021; 33(12): 7937-7943. <https://doi.org/10.1109/TNNLS.2021.3086757>.
- [34] Xiong D, Xu F. A robust diagnostic approach in mean shifts for multivariate statistical process control. J Stat Comput Simul. 2025; 95(3): 507-524. <https://doi.org/10.1080/00949655.2024.2431856>.

- [35] Shi Y, Beer M. Physics-informed neural network classification framework for reliability analysis. *Expert Syst Appl.* 2024; 258: 125207. <https://doi.org/10.1016/j.eswa.2024.125207>.
- [36] Weinans E, Quax R, van Nes EH, van de Leemput IA. Evaluating the performance of multivariate indicators of resilience loss. *Sci Rep.* 2021; 11(1): 9148. <https://doi.org/10.1038/s41598-021-87839-y>.
- [37] Li H, Sun C. PGVAE-VBAKF: a robust strategy for complex system response prediction and noise variance estimation considering modeling errors and nonstationary noises. *Mech Syst Signal Process.* 2026; 243: 113699. <https://doi.org/10.1016/j.ymssp.2025.113699>.
- [38] Astro_pat. Dynamical system multivariate time series [Internet]. Kaggle. Available from: <https://www.kaggle.com/datasets/patrickfleith/dynamical-system-multivariate-time-series-forecast>
- [39] Behnoudfar P, Chen N. RL-DAUNCE: reinforcement learning-driven data assimilation with uncertainty-aware constrained ensembles. *arXiv Preprint.* 2025;arXiv:2505.05452. <https://doi.org/10.48550/arXiv.2505.05452>.
- [40] Chen Z, Badrinarayanan V, Lee CY, Rabinovich A. GradNorm: gradient normalization for adaptive loss balancing in deep multitask networks. In: *Proceedings of the International Conference on Machine Learning*. PMLR; 2018. p. 794-803.
- [41] Yu R, Wang R. Learning dynamical systems from data: an introduction to physics-guided deep learning. *Proc Natl Acad Sci USA.* 2024; 121(27). <https://doi.org/10.1073/pnas.2311808121>.
- [42] Rane N, Choudhary SP, Rane J. Ensemble deep learning and machine learning: applications, opportunities, challenges, and future directions. *Stud Med Health Sci.* 2024; 1(2): 18-41. <https://doi.org/10.48185/smhs.v1i2.1225>.
- [43] Kirsch A. (Implicit) ensembles of ensembles: epistemic uncertainty collapse in large models. *arXiv Preprint.* 2024;arXiv:2409.02628. <https://doi.org/10.48550/arXiv.2409.02628>.
- [44] Liang X, Liu Z, Wang J, Jin X, Du Z. Uncertainty quantification-based robust deep learning for building energy systems considering distribution shift problem. *Appl Energy.* 2023; 337: 120889. <https://doi.org/10.1016/j.apenergy.2023.120889>.
- [45] Castán-Lascorz M, Jiménez-Herrera P, Troncoso A, Asencio-Cortés G. A new hybrid method for predicting univariate and multivariate time series based on pattern forecasting. *Inf Sci.* 2022; 586: 611-627. <https://doi.org/10.1016/j.ins.2021.12.001>.
- [46] Han L, Ye HJ, Zhan DC. The capacity and robustness trade-off: revisiting the channel independent strategy for multivariate time series forecasting. *IEEE Trans Knowl Data Eng.* 2024; 36(11): 7129-7142. <https://doi.org/10.1109/TKDE.2024.3400008>.