forecasting

MDPI

Article

Simulator-Grounded Benchmarking and a Corpus-Distilled Physics-Informed Forecaster for Fire Hazard-State and Damage Forecasting

Dohun Kim, Seonghee Lee and In-Hwan Lee *

Electronics and Telecommunications Research Institute (ETRI), 22, Daewangpangyo-ro 712beon-gil, Bundang-gu, Seongnam-si 13488, Gyeonggi-do, Republic of Korea; dohun@etri.re.kr (D.K.); slee0003@etri.re.kr (S.L.)
* Correspondence: ihlee@etri.re.kr

Highlights

What are the main findings?

- A corpus-distilled physics-informed forecaster (CD-PINN) estimates a data-optimal, reduced-order physics residual from a simulated corpus through physics-guided regression and uses it as a physics prior, reaching the accuracy of corpus-trained forecasters while remaining interpretable and corpus-light.
- Across 26 independent laboratory compartment-fire experiments, the large in-distribution performance gaps between models shrink substantially, indicating that in-distribution leaderboards alone do not predict laboratory-fire performance.

What are the implications of the main findings?

- Forecasting complexity should be matched to the data regime: parsimonious models approach the best accuracy with a small corpus, whereas physics-informed models are most valuable when data are scarce or the forecasting problem is underdetermined.
- Forecasters should be validated on independent, out-of-distribution laboratory-fire data before deployment because in-distribution benchmarks can overstate model differences. For downstream damage assessment, the label definition should also be selected carefully.

Abstract

Forecasting research repeatedly finds that simple methods can match or beat complex ones out of sample. We test this in a safety-critical domain, near-real-time prediction of fire hazard state and structural damage, using a simulator-grounded benchmark: high-fidelity computational fluid dynamics (Fire Dynamics Simulator, FDS) provides reference data, an FDS-calibrated zone model (CFAST) generates a large corpus cheaply, and the temperature trajectories drive a finite-element model (OpenSees) and a HAZUS/Eurocode-informed damage rule. Under one protocol we compare simple, deep (PatchTST, TimesNet), and physics-informed (PINN, PIKAN) forecasters. Complex models do not dominate: a small corpus lets parsimonious models approach best accuracy, and physics helps mainly when data are scarce (crossover near twenty scenarios). We propose CD-PINN, which identifies a data-optimal reduced-order physics residual from the corpus by physics-guided regression over a candidate library and uses it as the physics constraint. This lifts a per-event physics model to the accuracy of corpus-trained forecasters while staying interpretable. On 26 laboratory-fire experiments, however, in-distribution rankings do not transfer: the large accuracy spread collapses to near-parity, so a leaderboard poorly predicts laboratory-fire accuracy. For

Check for updates

Academic Editors: Foteini Kyriazi, Dimitrios D. Thomakos and Stelios Arvanitis

Received: 25 June 2026
Revised: 6 August 2026
Accepted: 8 August 2026
Published: 11 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

downstream damage, we further show that the label definition, not the model class, sets the achievable ceiling.

Keywords: benchmark forecasting; model selection; predictive complexity; physics-informed neural networks; time series forecasting; surrogate modeling; fire hazard; structural damage; FDS; CFAST; OpenSees

1. Introduction

Forecasting practice has long been shaped by a simple but powerful insight: uncomplicated methods often perform best out of sample. Naive forecasts, exponential smoothing, and low-order autoregressive models are repeatedly found to be robust, transparent, and scalable. They adapt to structural change, avoid overfitting, and remain usable in practice [1]. The enduring role of simple and combined statistical benchmarks is documented in recent methodological reviews [2]. Whether this lesson carries over to modern deep forecasters in safety-critical engineering settings is far less clear, and is rarely tested on data whose reference is physically validated.

We study this question for the near-real-time prediction of fire hazard state and the structural damage it causes. When a fire develops, running computational fluid dynamics (CFD) or finite-element structural solvers in real time is infeasible. A lightweight surrogate must instead forecast how hot it will get and whether the structure is at risk. The natural design question is which surrogate: a simple benchmark, a complex deep forecaster, or a physics-informed model that embeds conservation laws? Unusually for forecasting, here the reference can be generated and validated by simulators, enabling a controlled benchmark across data regimes.

The contributions of this paper are: (i) a simulator-grounded benchmark for fire hazard-state and damage forecasting that compares two data sources, a single high-fidelity FDS model and a lightweight FDS-calibrated CFAST combination, coupled to OpenSees and HAZUS-style damage labels, with fixed splits and seeds, (ii) a comparative study along two axes: corpus-free per-event models (classical and physics) vs. corpus-trained models, and simple vs. complex models, establishing when complexity pays off and showing that parsimonious models suffice in most regimes, (iii) a proposed model, a corpus-distilled physics-informed forecaster (CD-PINN) that identifies a data-optimal reduced-order physics residual from the corpus by physics-guided regression over a candidate library and uses it as the physics constraint, lifting a per-event physics model to corpus-model accuracy (0.153 $\rightarrow$ 0.091 nRMSE), (iv) an external validation on independent laboratory-fire experiments showing that in-distribution rankings do not transfer, the large synthetic accuracy spread collapsing to near-parity on the laboratory fires, and (v) an analysis of why these outcomes arise, including an exposure-integrated member-temperature analysis showing the peak-based damage label is a conservative approximation, and the finding that the label definition, not the model class, sets the ceiling for downstream damage prediction. Figure 1 gives an overview of the framework.

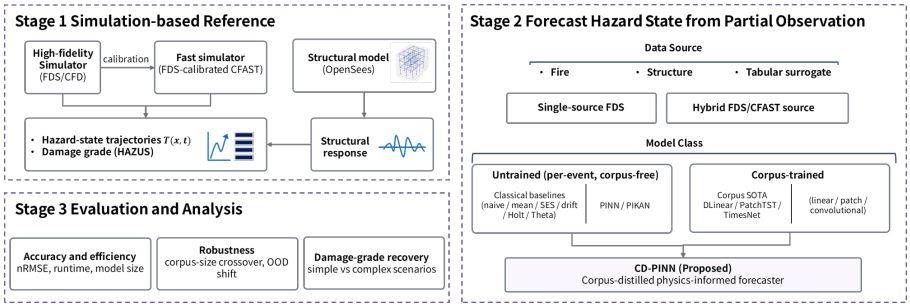


Figure 1. Overview of the simulator-grounded benchmark. Stage 1 generates the reference data, Stage 2 forecasts the hazard state along two axes (data source × model class, with the proposed CD-PINN bridging the model classes), and Stage 3 evaluates accuracy, robustness, and damage-grade recovery.

2. Related Work

2.1. Forecasting Models: Simple, Deep, and Physics-Informed

A long tradition establishes that simple and combined statistical benchmarks remain difficult to beat out of sample. Across successive M-competitions, elaborate stand-alone models rarely improved on exponential smoothing, low-order autoregressive baselines, and naive forecasts on truly unseen series. The most robust performers were statistical methods and, especially, combinations and hybrids of them [1,3]. The advantage is usually attributed to better adaptation to structural change, lower estimation variance, and reduced overfitting, properties that matter most precisely when the deployment distribution differs from the training one. This "simplicity-wins" lesson, however, is established almost entirely on economic and demographic series, and it is far less clear whether it survives in safety-critical engineering settings, where the dynamics are governed by physical conservation laws and the forecast horizon must be extrapolated from a short observed window rather than interpolated within a stationary regime.

Against this backdrop, transformer and convolutional architectures have driven much of the recent progress in long-sequence forecasting. PatchTST tokenizes a series into sub-sequence patches and treats channels independently [4], while TimesNet reshapes a 1D series into 2D tensors to capture multiple periodicities with convolutional inception blocks [5]. Both build on a line of efficient transformers that reduced the quadratic attention cost (Informer [6]) or injected series-decomposition and auto-correlation priors (Autoformer [7]), and on deep basis-expansion models such as N-BEATS [8]. Strikingly, this race toward complexity has been repeatedly questioned: a one-layer linear model, DLinear, matches or exceeds these transformers on standard long-horizon benchmarks [9], suggesting that much of the reported gain may stem partly from preprocessing, normalization, and benchmark-protocol choices rather than from architectural depth alone. Most recently, this line has extended to pretrained time-series foundation models that forecast unseen series zero-shot, such as the decoder-only TimesFM [10] and the tokenization-based Chronos [11]. These, however, presuppose a large and diverse pretraining corpus and offer no physical interpretability, in contrast to the data-scarce, physics-grounded regime we target. The same tension recurs in time-series classification, where a random-convolution baseline (ROCKET [12]) is strong and computationally efficient, whereas a deeper inception-based ensemble (InceptionTime [13]) reaches strong accuracy at markedly higher cost. These debates, however, are typically settled on large corpora. How the ranking changes when only a handful of trajectories are available, or when the test distribution shifts, is rarely quantified, and is one of the questions we address.

A third family embeds physics directly. Physics-informed neural networks (PINNs) add a differential-equation residual to the training loss so that predictions obey the governing conservation laws, which is especially valuable when data are scarce or the inverse

problem is under-determined [14,15]. The coordinate-network backbone is a design choice in its own right, ranging from a plain multilayer perceptron to Fourier-feature mappings that combat spectral bias [16], sinusoidal representations (SIREN [17]), and the recent Kolmogorov–Arnold network with learnable edge activations (KAN [18]), whose physics-informed variant (PIKAN) has been implemented and compared directly against MLP backbones for differential equations [19,20]. A related but distinct approach, operator learning such as DeepONet, instead amortizes a family of solutions into a single input-to-solution map [21]. A complementary and data-driven route does not assume the equation at all but discovers it from data by sparse regression over a library of candidate terms (SINDy [22]). Crucially, in all of these methods the governing law is either hand-specified or learned in isolation. Our proposed model instead distills the residual from the forecasting corpus itself, coupling the data-driven and physics-informed families that are usually studied separately. We include representatives of each family (together with the modern PINN backbones above) and evaluate them against simple alternatives under one protocol.

2.2. Simulation-Grounded Surrogates for Fire and Structures

Realistic reference for fire hazard comes from physics simulators that span a wide fidelity–cost range. The Fire Dynamics Simulator (FDS) is the de facto high-fidelity reference for compartment and tunnel fires, solving the low-Mach Navier–Stokes equations with large-eddy simulation, combustion, and thermal radiation [23]. It is accurate but expensive, ranging in our implementation from about two minutes per scenario at an engineering-grade mesh to hours at fine, validation-grade resolution (the reference guide describes the model rather than these problem-specific run times). At the other extreme, CFAST is a two-zone control-volume model that divides each room into a hot upper and a cool lower layer and integrates mass and energy balances [24], trading spatial detail for the sub-second run times we observe. For tunnels, a semi-empirical correlation for the maximum ceiling-jet temperature provides an even cheaper reduced-order description [25]. The resulting temperature histories drive the structural side, where nonlinear finite-element solvers (OpenSees [26], with the thermal-analysis extensions of Jiang and Usmani [27]) combine mechanical and thermal strains with temperature-dependent constitutive laws, using the Eurocode strength-reduction factors for concrete and steel under fire [28,29]. The resulting demands are mapped to discrete damage states through HAZUS-style fragility and interstory-drift thresholds [30]. The thermal-property factors and heat-balance equations follow the 2004/2005 editions of the Eurocodes used in the implementation.

Replacing such solvers at run time is the province of surrogate modeling [31], and when a single fidelity is too costly, multi-fidelity schemes combine a few high-fidelity runs with many cheap low-fidelity ones, calibrating the low-fidelity model against the reference [32]. This conceptually motivates the FDS-calibrated-CFAST strategy we adopt to build a large corpus affordably. Despite this rich toolbox, most fire loss-estimation work is one-shot rather than time-resolved, evaluates a single model class instead of benchmarking the forecasting family against simple baselines, and stops at the hazard state without closing the loop to structural damage. To our knowledge, no prior study provides a simulator-grounded, multi-fidelity benchmark that compares simple, complex, and physics-informed forecasters for hazard-state prediction, couples them to structural damage, and distills a deployment rule for when complexity is worthwhile. This is the gap this paper addresses.

3. Materials and Methods

3.1. Dataset: Simulator-Grounded Reference

3.1.1. Task

Given the first K steps of the compartment temperature $T(t)$ and static building/scenario features, we forecast the remaining horizon and recover the structural damage grade. We use a resampled length $L = 40$ with $K = 8$ (20% observed) unless stated. The forecast error is the normalized root-mean-square error (nRMSE) over the unobserved horizon, defined per event as

$$\mathrm{nRMSE} = \frac{\sqrt{\dfrac{1}{L-K}\displaystyle\sum_{i=K+1}^{L}\left(\hat{T}_i - T_i\right)^2}}{\max_i T_i - \min_i T_i}, \tag{1}$$

that is, the RMSE of the prediction $\hat{T}$ against the reference T normalized by that event's own temperature range (peak minus minimum), then averaged over events. Because each event is normalized by its own range, this metric is scale-invariant across events. The higher absolute peaks of the laboratory fires (Section 4) therefore do not by themselves inflate the reported laboratory-fire nRMSE. The identical per-event definition is applied to the simulated and the laboratory-fire datasets.

3.1.2. High-Fidelity Reference and Fast Calibrated Corpus

FDS solves, per cell and step, the reacting low-Mach Navier–Stokes system (detailed below). It is accurate but heavy. CFAST approximates each room by hot/cool layers governed by mass/energy balance ordinary differential equations. It is orders of magnitude faster but systematically under-predicts the fire-room peak. Over 34 paired FDS↔CFAST cases (111 compartments) the raw fire-room mean absolute error (MAE) is 676.6 °C. A single multiplicative calibration $T_{\mathrm{cal}} = T_{\mathrm{amb}} + k\,(T_{\mathrm{CFAST}} - T_{\mathrm{amb}})$ with $k = 2.39$ reduces it to 189.0 °C, a 72.1% reduction, with overall correlation $r = 0.86$ (Figure 2, Table 1). Calibrated CFAST therefore substantially improves fire-room agreement with FDS at zone-model cost and generates the bulk corpus (2000 building-fire scenarios). FDS supplies the high-fidelity subset (34 compartment + 40 tunnel cases, Figure 3). Individual forecasting experiments subsample this corpus (typically 10–120 scenarios for training and 30–50 held out, over five seeds) so that the corpus size is an experimental variable and per-seed variation remains visible. The full 2000-case corpus underlies the damage-label analyses, whereas the FDS–CFAST calibration itself is derived on the 34 FDS-paired compartment fire-room cases above and applied unchanged to the corpus.

To confirm that this calibration generalizes rather than merely fitting the same 34 cases, we performed leave-one-case-out cross-validation: for each case the coefficient is estimated from the other 33 cases and evaluated on the held-out case. The coefficient is highly stable across the 34 folds (standard deviation ≈0.03), and the out-of-sample fire-room MAE reduction (71.8%, 676.6 → 191.1 °C) essentially matches the in-sample value (72.1%), as does the out-of-sample fire-room damage-grade agreement (79%, vs. 0% for raw CFAST). The out-of-sample calibration residual, defined as the calibrated CFAST temperature minus the FDS reference and thus a signed error rather than a temperature, is nearly unbiased (mean −5.8 °C), with a median absolute residual of 121 °C and a 90% interval of $[-142, +262]$ °C. These figures are large in absolute terms but modest relative to the several-hundred-degree fire-room temperatures they correct, and the per-case calibration ratio has interquartile range $[2.09, 2.45]$ with a few high-ratio outliers. The single-coefficient calibration is therefore an independent, generalizing correction rather than an in-sample fitting artifact, though not an exact one. We treat the calibrated corpus as a substantially improved but imperfect

reference and note that, because independent per-scenario FDS is not available for all 2000 corpus scenarios, this does not by itself establish exact transfer to every scenario.

The calibration matters downstream, not only at the temperature level. Mapping each model's fire-room temperature to the (temperature-based) damage grade and comparing against the FDS reference grade (Table 2), raw CFAST agrees with FDS on 0% of the 34 fire rooms (it under-predicts so severely that every case is graded one level too low), whereas the calibrated combination agrees on 79% (grade mean-absolute-error 1.82 $\rightarrow$ 0.21). Figure 4 places the three data sources in accuracy–cost space: single FDS is accurate but costs about 120 s per scenario at the engineering-grade mesh (0.25 m cells) used here, versus hours for fine, validation-grade meshes. Raw CFAST runs in under 1 s but is unusable for damage (0% grade agreement), and the FDS-calibrated CFAST combination keeps most of the FDS-level damage-grade agreement (79%) at the same sub-second cost, roughly a 130$\times$ speed-up over FDS. The lightweight combination is therefore a substantially improved, low-cost surrogate for single FDS as the reference source for the bulk corpus. We stress the distinction between three tiers of "reference" used in this work: (i) FDS is the high-fidelity CFD reference, (ii) the FDS-calibrated CFAST is a calibrated surrogate that generates the large corpus, and it is this surrogate, not independent physical reality, that serves as the target for most in-distribution experiments, and (iii) the UL FSRI laboratory-fire experiments (Section 4) provide the only truly independent physical validation. Accordingly, we treat the in-distribution results as benchmarking on a physics-calibrated corpus and reserve claims of real-world generalization for the external laboratory-fire evaluation.

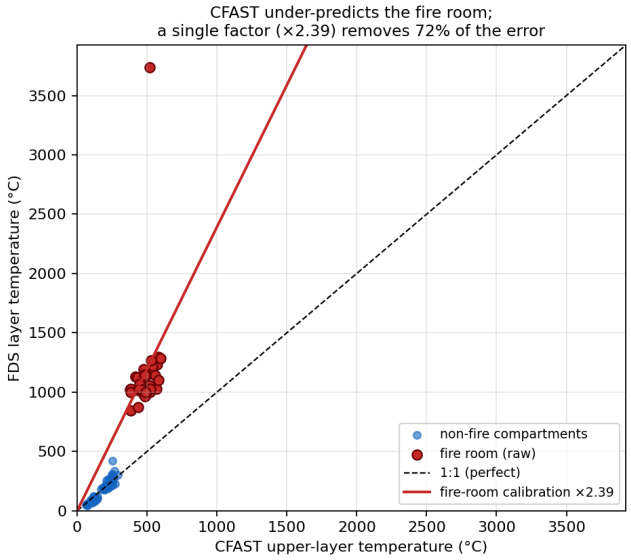


Figure 2. FDS vs. CFAST upper-layer temperature for 111 compartments. CFAST matches non-fire rooms ($r = 0.86$) but under-predicts the fire room (red). A single factor $k = 2.39$ aligns the fire room with FDS.

Table 1. FDS vs. FDS-calibrated CFAST (34 paired cases, 111 compartments).

Quantity	Value
Fire-room MAE, raw CFAST	676.6 °C
Fire-room MAE, calibrated ($k = 2.39$)	189.0 °C
Error reduction	**72.1%**
Overall FDS–CFAST correlation r	0.86

Bold marks the headline calibration result.

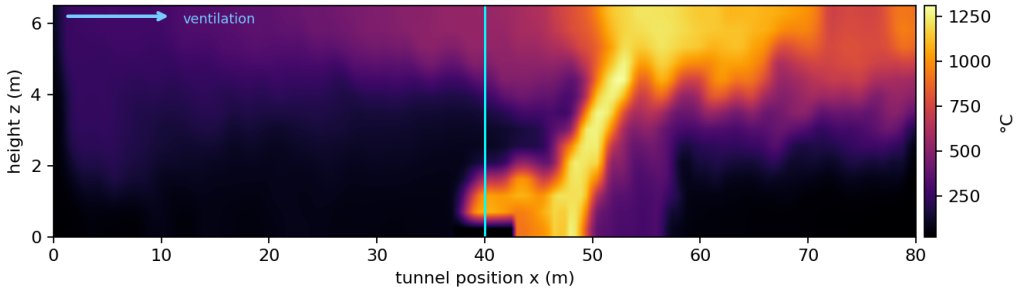


Figure 3. Example high-fidelity reference: FDS arched-tunnel fire ceiling-jet temperature field $T(x, z)$ under longitudinal ventilation. The vertical blue line marks the tunnel mid-span at $x = 40$ m, and the blue arrow indicates the longitudinal ventilation direction.

Table 2. Downstream value of calibration: fire-room damage-grade agreement with the FDS reference (34 cases).

Temperature Source	Exact Grade Agreement	Grade MAE
Raw CFAST	0.00	1.82
FDS-calibrated CFAST ($k = 2.39$)	**0.79**	**0.21**

Bold marks the best value in each column.

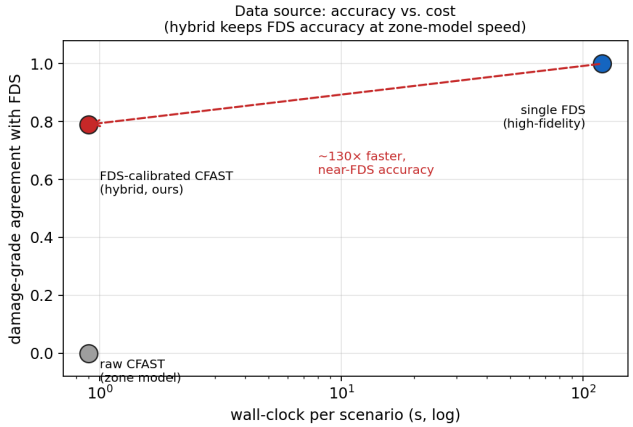


Figure 4. Three reference data sources in accuracy–cost space. The FDS-calibrated CFAST combination substantially improves damage-grade agreement with the FDS reference (79% out-of-sample vs. 0% for raw CFAST) at zone-model speed.

3.1.3. Governing Equations of the Data-Generating Simulators

FDS solves mass, momentum, energy, and species transport,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0, \tag{2}$$

$$\frac{\partial (\rho \mathbf{u})}{\partial t} + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) = -\nabla \tilde{p} + (\rho - \rho_\infty)\mathbf{g} + \nabla \cdot \boldsymbol{\tau}, \tag{3}$$

$$\frac{\partial (\rho h)}{\partial t} + \nabla \cdot (\rho h \mathbf{u}) = \frac{D\bar{p}}{Dt} + \dot{q}''' - \nabla \cdot \dot{\mathbf{q}}''_{\text{rad}} - \nabla \cdot \dot{\mathbf{q}}''_{\text{cond}}, \tag{4}$$

$$\frac{\partial (\rho Y_\alpha)}{\partial t} + \nabla \cdot (\rho Y_\alpha \mathbf{u}) = \nabla \cdot (\rho D_\alpha \nabla Y_\alpha) + \dot{m}'''_\alpha, \tag{5}$$

closed by $p_0 = \rho RT/\overline{W}$, a mixing-controlled combustion model, an LES sub-grid stress, and gray-gas radiation. Equation (4) yields the gas temperature used downstream, Equation (5) the soot field. Each CFAST compartment reduces to an upper-layer balance

$$\frac{\mathrm{d}T_u}{\mathrm{d}t} = \frac{1}{m_u c_p}\left(\dot{Q}_c - \dot{q}_{\mathrm{wall}} - \dot{m}_{\mathrm{vent}} c_p (T_u - T_\infty)\right), \tag{6}$$

bias-corrected by k. OpenSees solves static equilibrium by Newton iteration, $\mathbf{K}_T\,\Delta\mathbf{u} = \mathbf{R}$, with restrained thermal strain and temperature-dependent strength,

$$\varepsilon = \varepsilon_{\mathrm{mech}} + \alpha\Delta T, \qquad N = \int_A \sigma(\varepsilon, T)\,\mathrm{d}A, \quad P_{nT} = 0.85\,k_c(T) f_c' A_c + k_y(T) f_y A_s, \tag{7}$$

(Eurocode 2, EC2, concrete factor k_c, Eurocode 3, EC3, steel factor k_y), giving the interstory drift ratio IDR $= \Delta u/h$ and demand–capacity ratios N/P_{nT} and $V/(1.5V_c)$ that the HAZUS-style rule thresholds into the grade.

3.1.4. Damage Labels and Physics Constraint

We label each event with an ordinal structural-damage grade so that the forecasting target is a decision-relevant hazard state rather than a raw temperature. The grade combines a mechanical and a thermal criterion and takes the more severe of the two, reflecting that a member can fail either mechanically (excessive drift or over-demand) or thermally (loss of material strength when heated). Both criteria act on the peak (maximum-over-time) response. We emphasize that the specific numeric thresholds below are study-defined choices, informed by, but not prescribed by, the cited standards. They are meant to give a decision-relevant, ordinal target, not to reproduce a code-exact fragility model.

For the mechanical criterion, we adopt interstory-drift thresholds of 0.0033/0.0150/0.0400 and column demand–capacity thresholds of 0.8/1.0/1.3. These are study-defined constants, not values extracted from a specific fragility table: they are of the same order as typical HAZUS drift damage-state ratios [30], but those ratios are not universal, since HAZUS specifies them separately for each building class, height range, and seismic-design level, and defines four states (slight/moderate/extensive/complete). We do not tie our thresholds to a particular HAZUS class. We use three drift levels mapped to the slight/extensive/complete boundaries of our scale, folding the intermediate "moderate" state into the adjacent grades. The demand–capacity thresholds 0.8/1.0/1.3 are likewise study-specific engineering heuristics (approximate onset of yielding, nominal capacity, and substantial over-demand), not HAZUS-prescribed values.

The thermal criterion, a peak member temperature at 300/600/1000 °C, is likewise a study-defined classification threshold, motivated by, but not prescribed by, the Eurocode strength-reduction curves for heated members: the EC2 concrete factor $k_c(T)$ begins to drop above 300 °C (0.85 at 300 °C, 0.60 at 500 °C), and the EC3 steel factor $k_y(T)$ falls steeply through 500–600 °C (0.78 → 0.47) toward near-total loss by 1000 °C [28,29]. The worst of the two criteria sets the grade, mapping to {normal, damage, high_risk, collapse}.

Crucially, the temperature that drives damage is the member temperature, not the instantaneous gas temperature: a structural member responds to how long it is held hot, not merely to the gas peak. We therefore convert each gas-temperature history $T_g(t)$ to a member-temperature history $T_m(t)$ with the EC3 [29] lumped-capacitance heat balance for an unprotected section,

$$\Delta T_m = \frac{A_m/V}{c_a \rho_a}\left[\alpha_c\,(T_g - T_m) + \varepsilon\,\sigma\left((T_g + 273)^4 - (T_m + 273)^4\right)\right]\Delta t, \tag{8}$$

where A_m/V is the section factor and α_c, ε the convective and radiative coefficients. Because Equation (8) integrates the time–temperature exposure, a short hot spike and a sustained burn with the same gas peak yield very different member peaks and hence different grades (Figure 5). Over the corpus, the member peak is damped relative to the gas peak by 47% (bare steel) to 68% (rebar behind cover) (Figure 6), and an exposure-

based grade reclassifies 28–31% of cases relative to a naive gas-peak rule, almost always downward, because brief excursions do not heat the section. The gas peak is therefore a strong but imperfect proxy for the member demand: across the corpus, the two correlate at $r = 0.94$ and a peak-only fit explains $R^2 = 0.88$ of the member peak; yet, at a fixed gas peak (480–560 °C), the member peak still spans 78–397 °C purely through exposure, and adding the thermal dose raises the explained variance to $R^2 = 0.95$. The operational labels in this study are computed from the gas peak (a deliberately simple, cheap definition). The exposure analysis shows this is conservative and right about 70% of the time at grade level. Two consequences matter downstream (Section 4): the simple-vs-complex damage result is conditional on this peak-derived label, and, under the exposure-corrected label, where duration carries real signal, peak alone is no longer sufficient, a further instance of the label (not the model) setting the ceiling.

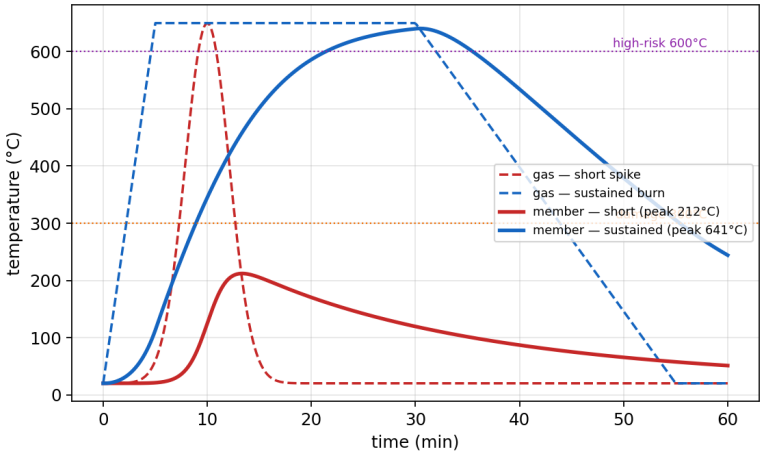


Figure 5. Same gas peak (650 °C), different exposure. A short spike leaves the member at 212 °C (normal). A sustained burn drives it to 641 °C (high-risk). Exposure time, not the instantaneous gas peak, sets the member temperature and the damage grade.

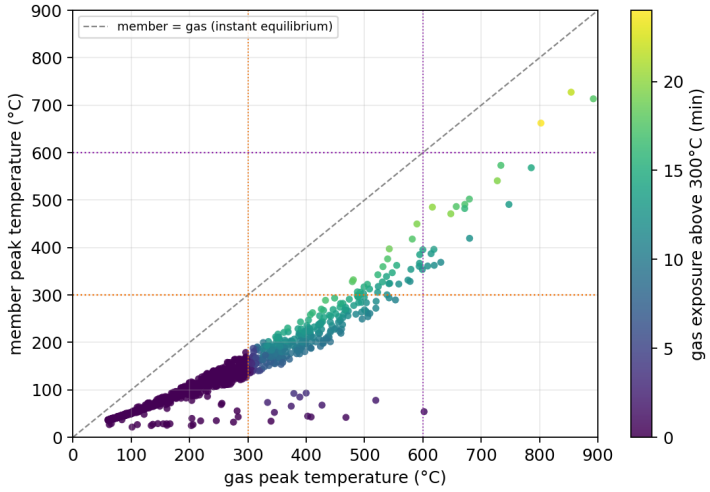


Figure 6. Gas peak overstates member demand across the corpus. The member peak (EC3 lumped-capacitance) is governed by the exposure time above 300 °C (color), not by the instantaneous gas peak. The orange and purple dotted lines mark the 300 and 600 °C grade thresholds on both the gas and member axes, partitioning the plane into gas/member grade combinations.

The physics-informed forecasters constrain the predicted temperature by the lumped energy balance

$$\frac{\mathrm{d}T}{\mathrm{d}t} = c_1\, q_{\mathrm{src}}(t) - c_2\,(T - T_{\mathrm{amb}}) \tag{9}$$

(net inflow minus cooling), added to the loss as a residual. The constants c_1, c_2 are calibrated once against the simulator reference.

3.1.5. Datasets, Protocol, and Implementation

The bulk building-fire corpus contains 2000 calibrated-CFAST scenarios (floors 2–12, floor area 150–1200 m^2, opening ratio 0.03–0.25, peak heat-release rate up to 16 MW), each coupled to OpenSees and labeled. The high-fidelity FDS subset comprises 34 compartment and 40 tunnel cases (20 rectangular, 20 arched, heat-release rate 5–112 MW, ventilation 0.5–4 m/s). All trajectories are resampled to $L = 40$. Validation sets and corpus pools are fixed across models with fixed seeds, and the sweeps reuse the same held-out scenarios. The physics-informed networks are per-event coordinate multilayer perceptrons (two hidden layers, 64 units, tanh) trained with the residual of Equation (9). The corpus forecasters are PatchTST, TimesNet, and DLinear. The damage models are a five-feature MLP/gradient-boosted-tree (tabular) and curve classifiers (ROCKET, 1D-CNN, InceptionTime).

3.2. Proposed Model: A Corpus-Distilled Physics-Informed Forecaster

The two axes above are usually treated as alternatives: either hand-specify a physics residual (per-event, corpus-free) or train a black box on the corpus. We propose a model that bridges them. We emphasize at the outset that the method does not discover physics de novo: sparse identification selects and weights terms within a physics-guided candidate library, so the recovered form necessarily resembles the hand-set energy balance (Equation (9)) and cannot capture dynamics absent from the library. What is new is threefold and distinguishes it from the closest prior work: unlike a standard PINN [14], it does not hand-set the residual coefficients but estimates them data-optimally from the forecasting corpus. Unlike a black-box corpus model, it retains an interpretable, reduced-order physics constraint, and unlike stand-alone equation identification [22], it transfers the estimated residual to a per-event coordinate network as a regularizing prior rather than using it as a simulator. In practice this yields an interpretable, corpus-light forecaster that attains corpus-model accuracy from a short observed window and degrades gracefully when the forcing is unknown at deployment. Concretely, rather than guessing the residual constants, we estimate a data-optimal reduced-order residual from the corpus. Offline, we start from a physics-guided candidate library of first- and second-order terms,

$$\Theta = \left\{\, q_{\mathrm{src}},\ (T - T_{\mathrm{amb}}),\ q_{\mathrm{src}}(T - T_{\mathrm{amb}}),\ (T - T_{\mathrm{amb}})^2,\ q_{\mathrm{src}}^2,\ 1 \,\right\}, \tag{10}$$

and regress the finite-difference $\dot{T}$ on it over all trajectories and time steps, a sparse-identification setup in the spirit of SINDy [22]. We emphasize an honest empirical fact: unconstrained sparse regression on this library does not by itself single out the energy balance. Because it fits the noisy finite-difference derivative, the interaction and quadratic terms carry comparable or larger normalized weight and are retained (in-sample $R^2 \approx 0.64$, Table 3). They lower the pointwise derivative error but overfit derivative noise and do not transfer. We therefore make the physical prior explicit and restrict the library to its interpretable first-order energy-balance sub-model $\{q_{\mathrm{src}},\ (T - T_{\mathrm{amb}}),\ 1\}$, and estimate only its coefficients data-optimally, giving

$$\frac{\mathrm{d}T}{\mathrm{d}t} = a\,q_{\mathrm{src}} + b\,(T - T_{\mathrm{amb}}) + c, \qquad (a,b,c) = (+0.040,\ -0.0088,\ +0.50),\ R^2 = 0.48, \tag{11}$$

on the pooled corpus. This is a deliberate reduced-order choice trading a little in-sample derivative fit for interpretability and transferability: the recovered form is the energy balance because we require it to be, and only its coefficients (a, b, c), not its functional structure, are learned from data. In the notation $\dot{T} = \Theta\,\Xi$, Ξ is thus a physics-selected sparse coefficient vector over Θ rather than one chosen purely by a data-driven threshold. This directly acknowledges that sparse regression can only select within the library and cannot recover physics absent from it. The value of the distilled residual is judged by the held-out forecast it produces (below), not by de novo discovery.

At deployment, a coordinate network $T_\theta(t)$ is fitted per event by minimizing the standard physics-informed objective

$$\mathcal{L} = \mathcal{L}_{\mathrm{data}} + \lambda\,\mathcal{L}_{\mathrm{phys}}, \quad \mathcal{L}_{\mathrm{data}} = \frac{1}{K}\sum_{i=1}^{K}\left\|T_\theta(t_i) - T_i\right\|^2, \quad \mathcal{L}_{\mathrm{phys}} = \left\|R\right\|^2, \tag{12}$$

where $\mathcal{L}_{\mathrm{data}}$ fits the K observed samples and the residual of the distilled law,

$$R = \frac{\mathrm{d}T_\theta}{\mathrm{d}t} - \left[a\,q_{\mathrm{src}} + b\,(T_\theta - T_{\mathrm{amb}}) + c\right], \tag{13}$$

regularizes the unobserved horizon. The derivative $\mathrm{d}T_\theta/\mathrm{d}t$ is obtained by automatic differentiation, q_{src} and T_{amb} are the (known or inferred) forcing and ambient inputs, and the coefficients (a, b, c) come from the offline distillation (11). The resulting corpus-distilled physics-informed forecaster (CD-PINN) is still per-event at deployment (it fits the observed window of a single event), but its physics prior carries the corpus's calibrated dynamics. When the deployment forcing is unknown (external laboratory fires, Section 4), only the distilled cooling coefficient b is transferred and the source is inferred as in Equation (14). Figure 7 shows the network.

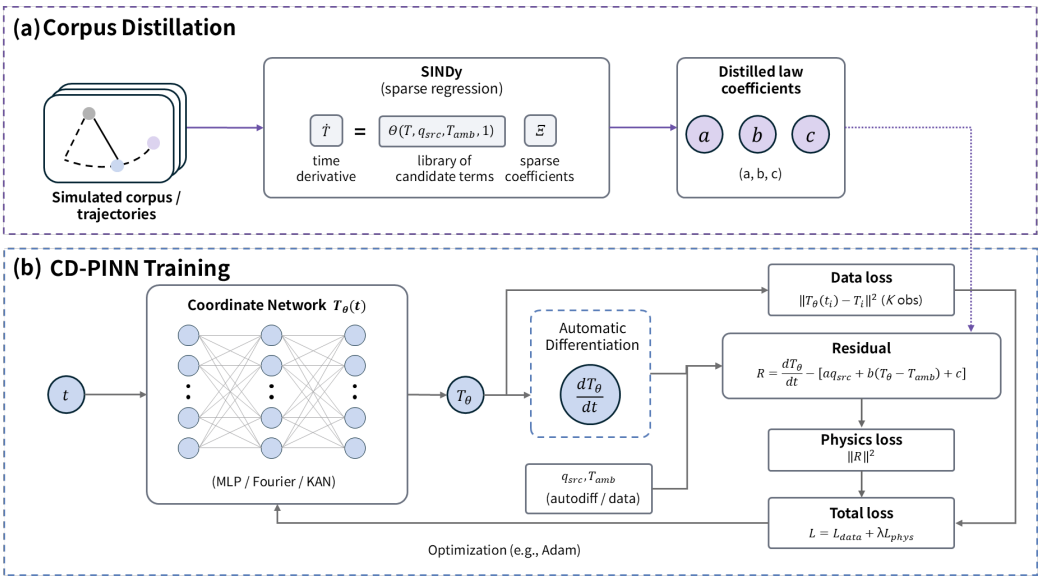


Figure 7. The proposed CD-PINN. (**a**) Offline, physics-guided regression over a candidate library (SINDy-style) estimates the energy-balance coefficients (a, b, c) from the simulated corpus. (**b**) At deployment, a per-event coordinate network $T_\theta(t)$ is trained on a data loss plus the distilled-equation residual, $\mathcal{L} = \mathcal{L}_{\mathrm{data}} + \lambda\mathcal{L}_{\mathrm{phys}}$.

Table 3. Physics-guided distillation over the candidate library Θ (Equation (10)), pooled over the training corpus. Normalized weight is the magnitude of the standardized (dimensionless) regression coefficient from the full-library fit. It measures how strongly a purely data-driven fit would use each term. Unconstrained sparse regression retains the interaction/quadratic terms (larger weight, $R^2 = 0.64$), but these overfit the noisy derivative and do not transfer. We instead retain the interpretable first-order energy-balance terms and estimate their coefficients data-optimally ($R^2 = 0.48$), the reduced-order residual used by CD-PINN.

Candidate Term	Normalized Weight $\lvert\Xi\rvert$	Reduced-Order Model
$q_{\mathrm{src}}(T - T_{\mathrm{amb}})$	2.58	discarded (non-transferable)
$(T - T_{\mathrm{amb}})^2$	1.63	discarded (non-transferable)
q_{src}	1.30	**retained**, $a = +0.040$
$(T - T_{\mathrm{amb}})$	1.25	**retained**, $b = -0.0088$
q_{src}^2	0.94	discarded (non-transferable)
1	0.74	**retained**, $c = +0.50$

Bold marks the terms retained in the distilled reduced-order model.

3.3. Benchmarked Forecasters

All models share the observation window, static features, and splits. Classical, structure-free (per-event): persistence (naive), mean, simple exponential smoothing (SES), linear/drift, Holt, and Theta. Physics-informed (per-event): corpus-free coordinate networks (PINN-MLP) and PIKAN with the energy-balance residual (9), plus the proposed corpus-distilled CD-PINN. Data-driven (corpus-trained): the linear DLinear and the deep PatchTST and TimesNet. For damage classification, we compare a five-feature tabular model (MLP and gradient-boosted trees) against curve classifiers (ROCKET, 1D-CNN, InceptionTime).

4. Results

4.1. Hazard-State Forecasting and the Cost of Complexity

Because the simplicity principle is usually demonstrated with classical univariate benchmarks, we first evaluate them as the simplest possible forecasters (per-event, no corpus): persistence (naive), the mean, linear/drift extrapolation, simple exponential smoothing (SES), Holt's linear trend, and the Theta method. Table 4 and Figure 8 show that on this task they fail: a fire trajectory rises to a peak and decays, but only its rising 20% is observed, so flat methods (Naive, Mean, SES) miss the peak and decay (nRMSE 0.43–0.58) and trend methods (Linear, Holt, Theta) diverge by extrapolating the steep rise (nRMSE 2.2–4.0). Structure-free benchmarks are thus strong for stationary series but inadequate for single-event, non-stationary hazard extrapolation: what succeeds is not complexity per se but the right minimal structure, either a physics constraint (per-event, nRMSE ≈ 0.14) or a small corpus that supplies the missing dynamics (data-driven, nRMSE ≈ 0.11). The rest of this section dissects which structure is needed and how much.

Table 4. Master comparison: all forecasters on the same 30 held-out fire events (20% observed, $K = 8$). Data-driven models are trained on a 30-scenario corpus. Classical and physics models are corpus-free (per-event). Values are nRMSE (mean $\pm$ s.d. over five seeds where applicable, lower is better).

Model	Family	nRMSE (↓)
PatchTST	data-driven, complex (corpus 30)	**0.103 ± 0.003**
TimesNet	data-driven, complex (corpus 30)	0.105 ± 0.003
DLinear	data-driven, simple (corpus 30)	0.117 ± 0.002

Table 4. *Cont.*

Model	Family	nRMSE (↓)
PINN-MLP	physics, per-event (corpus-free)	0.141 ± 0.001
PIKAN	physics, per-event (corpus-free)	0.142
Mean	classical, per-event	0.434
SES	classical, per-event	0.547
Naive	classical, per-event	0.583
Theta	classical, per-event	2.216
Holt	classical, per-event	3.302
Linear/Drift	classical, per-event	3.963

↓: lower is better. Bold marks the best (lowest) nRMSE.

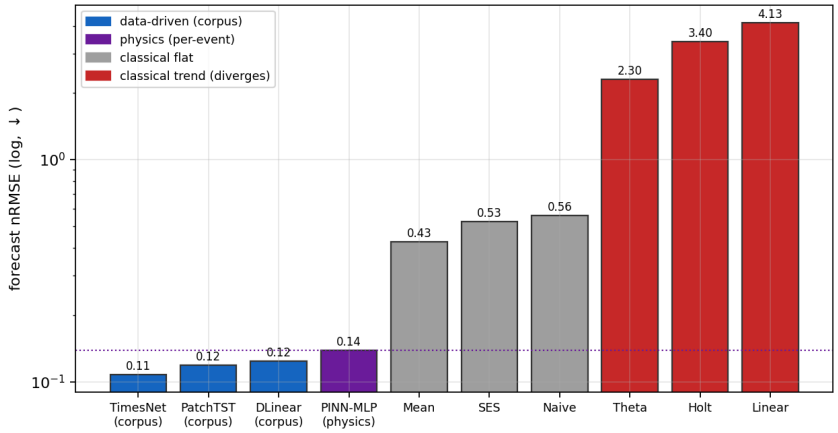


Figure 8. Structure-free classical benchmarks fail on single-event extrapolation. Trend methods (red) diverge. Minimal structure, physics or a small corpus, is needed (log scale). The purple dotted line marks the per-event physics baseline (PINN-MLP, nRMSE ≈ 0.14); the ↓ on the y-axis denotes lower is better.

4.1.1. Corpus-Size Crossover

With a fixed validation set and 20% observation, we vary the training corpus over five seeds (Table 5, Figure 9). We define the crossover as the corpus size at which the mean forecast nRMSE (averaged over the five seeds) of the corpus-trained forecasters first drops below the corpus-free physics baseline. It is thus a metric-based intersection of the two mean-nRMSE curves, not a visual estimate. Complex transformers need data: at ten scenarios they clearly trail the physics model (0.246–0.256 vs. 0.141, several standard deviations apart), whereas at twenty they have already dropped below it (0.118–0.122 vs. 0.141). Twenty scenarios is thus the first corpus size at which the corpus-trained models overtake the corpus-free physics baseline (the crossover), so the mean-nRMSE curves cross between ten and twenty scenarios. Parsimony is competitive throughout: the simple linear DLinear (0.117) stays close to the complex transformers (0.103–0.105) in the master comparison (Table 4). Consistent with the event-level bootstrap reported below (95% CI half-widths ≈ 0.02 nRMSE), we treat the crossover as a region rather than a sharp point and do not over-interpret sub-0.02 differences among the corpus-trained models. The effect we rely on is the reversal itself: the physics model is better than the corpus models by about 0.10 nRMSE at ten scenarios but is already beaten by them at twenty, a reversal far larger than that uncertainty.

Table 5. Fire forecast nRMSE (lower is better) vs. corpus size (20% observed), mean ± s.d. over five seeds. The physics-informed model is corpus-free.

Corpus	PatchTST	TimesNet	PINN-MLP (Corpus-Free)
10	0.256 ± 0.022	0.246 ± 0.025	**0.141 ± 0.001**
20	0.122 ± 0.009	0.118 ± 0.003	0.141 ± 0.001
30	0.112 ± 0.008	**0.107 ± 0.001**	0.141 ± 0.001

Bold marks the best (lowest) nRMSE.

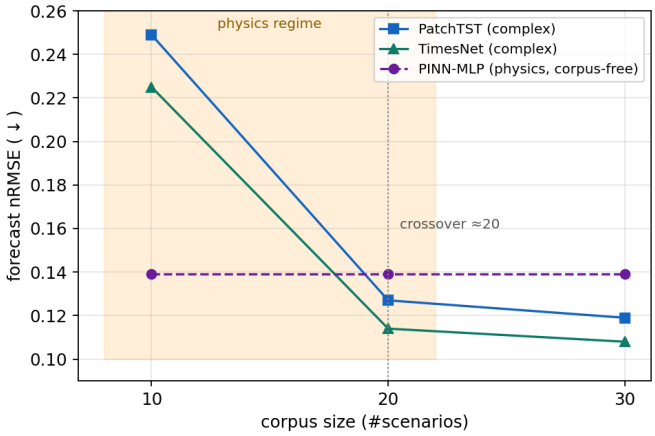


Figure 9. Corpus-size crossover. Complex deep forecasters trail the corpus-free physics model at ten scenarios but overtake it by twenty scenarios. Parsimonious models remain competitive throughout. ↓: lower is better.

4.1.2. Observation Window and Under-Determined Reconstruction

With only 10% observed and a ten-scenario corpus, the physics-informed model beats the transformer (Table 6, Figure 10). The gap widens as the problem becomes more under-determined: in a Raissi-style sparse-sensor reconstruction of the full diffusion field, the physics constraint cuts nRMSE from 0.146 to 0.026 with only four sensors (an 82% reduction), the gain shrinking monotonically as sensors are added (76% at eight, 26% at thirty-two) until, once the field is well sampled, the data-only model is adequate. Sparse, under-determined reconstruction is thus the clearest regime where physics complexity is worth it.

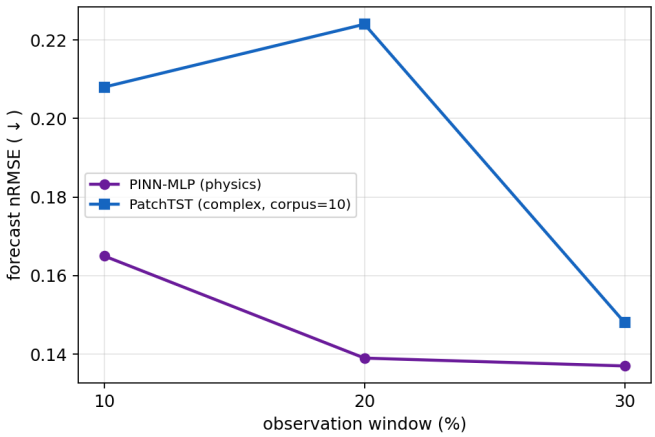


Figure 10. Short observation windows with a scarce corpus favor the physics-informed model. ↓: lower is better.

Table 6. Fire forecast nRMSE vs. observation window (corpus fixed at 10).

Observation	PINN-MLP (Physics)	PatchTST (Complex)
10%	**0.165**	0.208
20%	**0.139**	0.224
30%	0.137	0.148

Bold marks the best (lowest) nRMSE.

4.1.3. Reconstructing the Full Reference

With a 100-scenario corpus and 50 held-out events (same $K = 8$ protocol), the data-driven forecasters reach their best accuracy: PatchTST nRMSE 0.073, TimesNet 0.088, DLinear 0.098, while the corpus-free per-event physics, which cannot exploit the larger corpus, sits at PINN-MLP 0.163. Its accuracy is set by the per-event fit rather than the corpus, so, unlike the data-driven models it does not improve with more data (the absolute level here reflects this particular 50-event held-out set, which differs from the sets used in Tables 4 and 5). When a corpus exists, data-driven forecasting wins, and the simple linear DLinear stays within range of the transformers.

4.1.4. The Proposed CD-PINN Closes the Gap

This is where the proposed CD-PINN of Section 3.2 pays off. Restricting the candidate library (10) to its interpretable energy-balance terms and fitting them to the corpus yields the reduced-order law $\dot{T} = +0.040\, q_{\mathrm{src}} - 0.0088\,(T - T_{\mathrm{amb}}) + 0.50$: the same governing structure we had hand-specified (by construction, since we impose it rather than let an unconstrained fit choose, as shown in Table 3), but now with data-optimal coefficients. Its pointwise fit to the noisy finite-difference derivatives is modest ($R^2 = 0.48$ on the pooled corpus), as expected for a single global law over a diverse corpus. The law is not meant to be a high-accuracy point predictor but a regularizing prior, and its value is therefore judged by the held-out forecast it produces rather than by its own R^2. On that measure it is decisive: using the distilled law as the physics residual in the per-event fit cuts the in-distribution held-out error from the hand-set physics 0.153 ± 0.003 to 0.091, a 41% reduction, with a five-seed standard deviation below 0.001 (shown as 0.000 in Table 7). The residual regularizes the fit so consistently that it is almost initialization-independent (Table 7, Figure 11). A per-event physics model now matches the full corpus-trained models (0.076–0.096). This match survives an event-level bootstrap, under which CD-PINN is statistically indistinguishable from every corpus model (Table 8), while remaining interpretable and corpus-light at deployment. Distilling the physics from the corpus, rather than hand-tuning it or learning a black box, is thus the most effective use of the corpus for the physics-informed branch. The distilled coefficients are, however, tied to the corpus distribution: on the external laboratory fires CD-PINN degrades like the corpus models (0.356 ± 0.017; reported below), so its advantage, like theirs, is an in-distribution one. Because the distillation step is cheap (a sub-second regression), this sensitivity can be mitigated in deployment by (i) online re-distillation, re-estimating the coefficients on a sliding window of recent events, (ii) hybridizing the corpus-distilled model with the forcing-agnostic variant of Equation (14), which is shift-invariant, and selecting or averaging by an uncertainty estimate, or (iii) periodic recalibration when a drift monitor flags a distribution change. We leave a systematic study of these strategies to future work.

Table 7. In-distribution forecast nRMSE (mean ± s.d., five seeds): distilling the reduced-order physics residual from the corpus (CD-PINN) lifts per-event physics to corpus-model accuracy. CD-PINN's s.d. is below 0.001 (the distilled-physics fit is nearly initialization-independent).

Model	Type	nRMSE (↓)
Hand-set physics PINN	per-event, hand-set law	0.153 ± 0.003
CD-PINN (proposed)	per-event, corpus-distilled law	$\mathbf{0.091 \pm 0.000}$
DLinear	corpus, linear	0.096 ± 0.002
PatchTST	corpus, deep	0.076 ± 0.004
TimesNet	corpus, deep	0.085 ± 0.002

↓: lower is better. Bold marks the proposed model (CD-PINN).

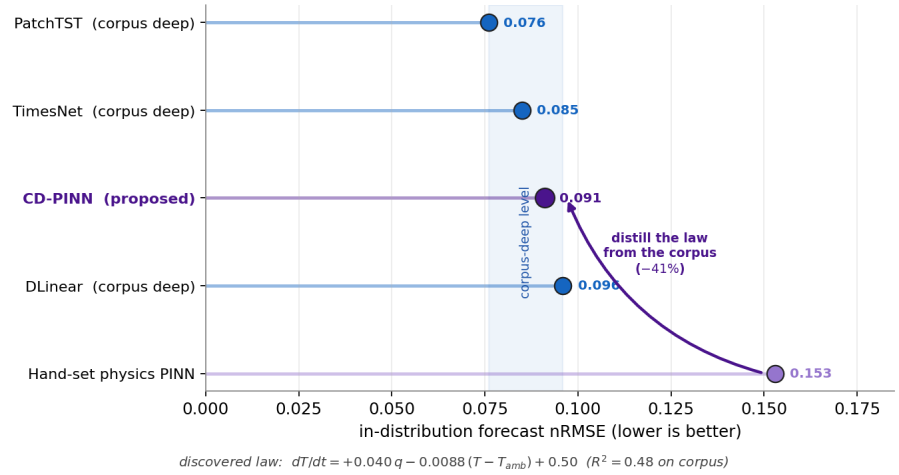


Figure 11. Corpus-distilled physics (CD-PINN). Estimating a data-optimal reduced-order physics residual from the corpus and using it as the physics constraint lifts the per-event physics forecaster (0.153 → 0.091) to the level of the corpus-trained models, while keeping an interpretable, reduced-order law.

Table 8. In-distribution uncertainty on the 30 held-out scenarios: nRMSE with 95% confidence intervals from an event-level bootstrap (10,000 resamples), and representative paired per-event differences (same procedure as the external-fire bootstrap). Distilling the physics residual gives a significant improvement over the hand-set law, after which CD-PINN is indistinguishable from the corpus-trained models.

Model	Mean nRMSE	95% CI (Event Bootstrap)
PatchTST	0.080	[0.059, 0.107]
TimesNet	0.088	[0.068, 0.111]
CD-PINN (proposed)	0.092	[0.078, 0.107]
DLinear	0.097	[0.078, 0.122]
Hand-set physics PINN	0.152	[0.125, 0.184]
Paired per-event differences (95% CI):		
CD-PINN − Hand-set physics	−0.060	[−0.096, −0.027]
CD-PINN − PatchTST	+0.012	[−0.018, +0.037]
CD-PINN − DLinear	−0.006	[−0.034, +0.019]
PatchTST − DLinear	−0.018	[−0.028, −0.007]

Bold marks the proposed model (CD-PINN).

4.1.5. Uncertainty on the In-Distribution Comparison

Our external evaluation used a fire-level bootstrap. To hold this in-distribution comparison to the same statistical standard, we apply the identical procedure at the event level: we resample the 30 held-out scenarios (10,000 times) to obtain per-model confidence intervals and form paired per-event differences between models. Table 8 reports

the result, computed on one representative seed's per-event errors, so its means are close to but not identical to the five-seed averages of Table 7. Two conclusions survive this stricter test. First, distilling the residual from the corpus yields a statistically significant improvement over the hand-set physics law: the CD-PINN − hand-set difference is −0.060 with a 95% CI of $[-0.096, -0.027]$ that excludes zero, and the hand-set model is separated from every corpus-trained model. The distilled physics is therefore not a within-noise re-parameterization. It genuinely lifts a per-event physics forecaster into the corpus-model band. Second, once lifted, CD-PINN is statistically indistinguishable from all three corpus-trained models (every CD-PINN − corpus CI includes zero), so a corpus-light, interpretable per-event model matches the black-box corpus forecasters within event-to-event variability. Among the corpus models themselves, the differences are small: PatchTST's edge over DLinear is significant but only 0.018 in nRMSE, while PatchTST, TimesNet, and CD-PINN are mutually indistinguishable. This mirrors, in distribution, the external finding: genuine differences exist but are modest, and no single model class dominates on complexity alone.

4.1.6. Robustness to the Physics Weight

The CD-PINN result is not an artifact of a tuned loss weight. Sweeping the physics weight λ in Equation (12) over a 50× range (0.1–5.0) changes the in-distribution nRMSE by at most 0.003 (Table 9): the deployed value $\lambda = 0.5$ gives 0.091 and the best value ($\lambda = 0.1$) only 0.090, so the physics prior helps consistently and λ was not chosen to flatter the model.

Table 9. CD-PINN in-distribution nRMSE vs. the physics-loss weight λ (30 held-out scenarios, mean over three seeds). The result is flat over a 50× range, so it is robust to λ.

λ	**0.1**	**0.2**	**0.5**	**1.0**	**2.0**	**5.0**
nRMSE	0.090	0.091	**0.091**	0.093	0.093	0.093

Bold marks the deployed setting.

4.2. Robustness and External Generalization

We next ask whether the physics-informed result depends on a dated multilayer perceptron and whether the rankings survive a distribution shift. Replacing the per-event backbone by modern coordinate networks, namely a Fourier-feature network, a sinusoidal SIREN, and a 2024 Kolmogorov–Arnold network (PIKAN), leaves the result essentially unchanged (Table 10): PIKAN (0.134) matches the plain MLP (0.132), the Fourier network is close (0.140), and only SIREN is worse (0.228). The physics constraint, not the backbone, governs accuracy. We then shift the test set from the corpus distribution (peak 121–156 °C) to higher-intensity fires (out-of-distribution, OOD, 230–253 °C). Over five seeds (Table 11), the complex transformers degrade most (+17%), the per-event physics most of all (+31%), whereas the simple linear DLinear barely degrades (0.078 → 0.079, +1%, within seed variability) and stays the most accurate model out of distribution. Under this milder, same-simulator shift, simplicity therefore buys robustness, exactly the property that motivates benchmark methods. (As shown next, this no longer holds under a severe shift to laboratory fires.)

Table 10. Modern physics-informed backbones (per-event, in-distribution, mean ± s.d.). The physics constraint dominates the backbone choice.

Backbone	**nRMSE (↓)**
PINN-MLP	0.132 ± 0.058
PIKAN (Kolmogorov–Arnold, 2024)	0.134 ± 0.060
PINN-Fourier features	0.140 ± 0.055
PINN-SIREN	0.228 ± 0.124

↓: lower is better.

Table 11. Robustness under a mild, same-simulator distribution shift (nRMSE on in-distribution vs. higher-intensity OOD test sets, corpus of 100, mean ± s.d. over five seeds).

Model	In-Dist.	OOD	Degradation
DLinear (simple)	0.078 ± 0.001	0.079 ± 0.002	**+1%**
TimesNet (complex)	0.072 ± 0.001	0.084 ± 0.002	+17%
PatchTST (complex)	0.074 ± 0.006	0.087 ± 0.006	+17%
PINN-MLP (physics)	0.138 ± 0.004	0.181 ± 0.004	+31%

Bold marks the smallest performance degradation.

External Validation on Laboratory-Fire Experiments

The shift above is still within our own simulator. The decisive test is whether the corpus advantage survives on fires we did not generate. We therefore evaluate the identical, frozen corpus models on 26 laboratory compartment-fire experiments from the UL FSRI 2018 program [33], an independent laboratory dataset (gas burner and furniture fuels, 100–500 kW, open/closed ventilation, measured ceiling-jet temperature at 1 Hz, peaks 153–982 °C). No laboratory-fire data ever enters training. Because a laboratory fire's heat release is unknown, the physics baseline is run in its hardest, forcing-agnostic form: we supply no forcing, keep only the energy-balance structure, and infer the source jointly with the temperature, constraining it to a single non-negative growth–decay heat pulse,

$$\frac{\mathrm{d}T}{\mathrm{d}t} = s_\phi(t) - k\,(T - T_{\mathrm{amb}}), \qquad s_\phi(t) \geq 0 \text{ (one pulse)}, \tag{14}$$

with s_ϕ and the cooling rate k fitted from the same first eight samples. This needs neither a corpus nor a known forcing and applies unchanged to any fire.

We stress that this evaluation changes the data source, the distribution, and the underlying physics jointly. It therefore probes transfer limitations of the complete simulator-trained forecasting pipeline, not an isolated, architecture-level difference in out-of-distribution generalization. Table 12 and Figure 12 report the result. In our simulator, the models span a 4–6× accuracy range (nRMSE 0.076–0.455). On the laboratory fires they all collapse into a narrow band of 0.31–0.41: the corpus-trained models, whose lab-fire nRMSE is 220–340% higher than in distribution (an absolute rise from about 0.08–0.10 to about 0.31–0.36), lose most of their in-distribution advantage. Two factors make this relative increase large. First, it is measured against a very small in-distribution baseline: a moderate absolute error on the laboratory fires (nRMSE about 0.3, i.e., roughly a third of each event's temperature range) becomes a several-fold percentage change when the in-distribution error is only about 0.08. Second, and substantively, the increase reflects the joint shift in data source, distribution, and physics: the corpus models were trained on FDS-calibrated CFAST dynamics and never saw the real combustion, fuels, and ventilation of the laboratory fires, so patterns that were predictive in the simulator do not carry over.

Because the seed standard deviations in Table 12 reflect training-run variability rather than the dominant case-to-case variability among the 26 fires, we additionally quantify uncertainty with a fire-level bootstrap (10,000 resamples of the 26 fires, which are all distinct configurations) and paired per-fire differences (Table 13). The bootstrap point estimates are computed on a single representative seed's per-fire errors, so they differ slightly from the five-seed means in Table 12. The shift is largest for the forcing-agnostic physics model, which has by far the highest per-fire and per-seed variance, but the qualitative conclusion below is insensitive to this choice. The 95% confidence intervals overlap substantially, and every pairwise per-fire difference has an interval that includes zero. On the laboratory fires, therefore, no model is statistically distinguishable from another: the large in-distribution differences are not significant out of sample. The practical lesson,

directly relevant to benchmarking, is cautionary: a model's position on an in-distribution leaderboard is a poor predictor of its accuracy on independent laboratory fires.

Table 12. Out-of-corpus generalization to laboratory fire experiments (UL FSRI 2018, 26 experiments, nRMSE mean ± s.d. over five seeds). Corpus models are frozen after training on 120 simulated scenarios. The physics model is given no forcing (Equation (14)). The 4–6× in-distribution spread collapses to a 0.31–0.41 band on the laboratory fires.

Model	Type	In-Dist.	Lab. Fires	Change
PatchTST	corpus	0.076 ± 0.004	0.331 ± 0.017	+336%
TimesNet	corpus	0.085 ± 0.002	0.360 ± 0.016	+324%
DLinear	corpus	0.096 ± 0.002	0.310 ± 0.004	+223%
CD-PINN (proposed)	corpus-distilled	0.091 ± 0.000	0.356 ± 0.017	+291%
Physics (forcing-agnostic)	corpus-free	0.455 ± 0.057	0.413 ± 0.048	−9%

Table 13. Fire-level uncertainty on the 26 laboratory fires: nRMSE with 95% confidence intervals from a bootstrap over fires (10,000 resamples), and representative paired per-fire differences. All intervals overlap and all paired differences include zero, so no model is statistically distinguishable on the laboratory fires.

Model	Mean nRMSE	95% CI (Fire Bootstrap)
DLinear	0.308	[0.276, 0.340]
PatchTST	0.326	[0.277, 0.380]
TimesNet	0.334	[0.285, 0.384]
CD-PINN (proposed)	0.359	[0.291, 0.433]
Physics (forcing-agnostic)	0.339	[0.258, 0.428]
Paired per-fire differences (95% CI):		
DLinear − PatchTST	−0.018	[−0.056, +0.017]
DLinear − TimesNet	−0.025	[−0.057, +0.005]
CD-PINN − Physics	+0.020	[−0.051, +0.083]

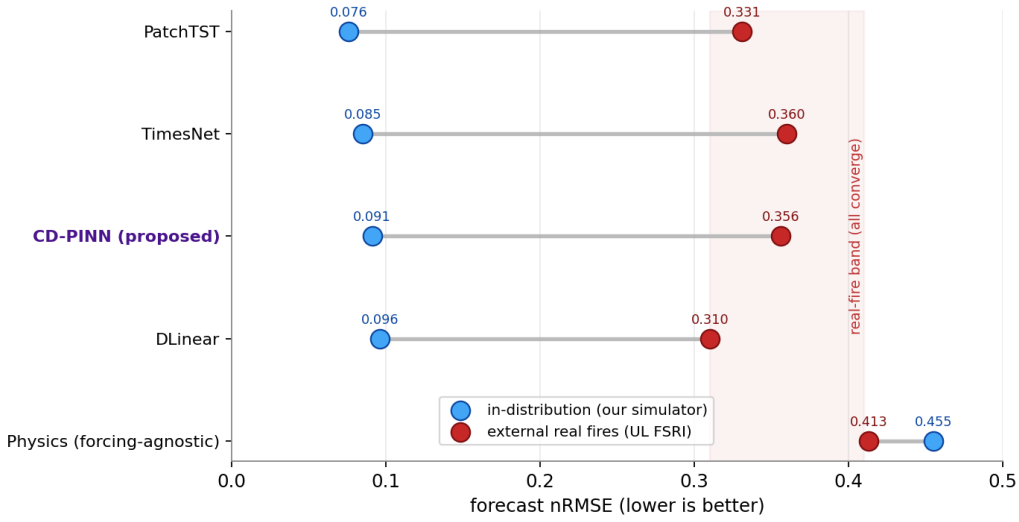


Figure 12. In-distribution vs. external laboratory fires (UL FSRI 2018, five seeds, error bars ± s.d.). The 4–6× in-distribution spread collapses to a narrow 0.31–0.41 band on the laboratory fires: in-distribution rankings do not transfer. The corpus-free physics model alone is shift-invariant, but it is the weakest in absolute terms.

4.3. Downstream Damage: The Label Sets the Ceiling

4.3.1. Simple Beats Complex

For the downstream damage grade, a five-feature tabular model dominates curve-based deep classifiers (Table 14, Figure 13): even ROCKET, a strong time-series classifier, does not beat a compact MLP on the peak feature, and InceptionTime collapses on the small, near-binary problem. The reason is definitional: the operational label is a peak-determined functional (Section 3.1.4), so the full curve carries little signal beyond the peak, and a peak-only fit explains 91% of IDR and 87% of the grade with MLP ≈ gradient-boosted trees ≈ random forest. We stress that this is conditional on the peak-based label rather than a universal property: under the exposure-corrected member-temperature label, where duration carries real signal (Section 3.1.4, R^2 rising 0.88 → 0.95 once thermal dose is added), peak features alone are no longer sufficient, and the time-resolved label below reverses the ranking outright. For the operational label, replacing the heavy OpenSees layer by a tabular surrogate (a distinct task from the curve-to-grade recovery of Table 14, here predicting the structural response from the scenario features) gives IDR $R^2 = 0.981$, DCR $R^2 = 0.983$, and grade accuracy 0.947 (macro-F1 0.939), so the heavy structural solver can be replaced by a small tabular model for fast inference.

Table 14. Damage-grade recovery (five-fold CV on reference curves): simple tabular vs. complex curve models.

Input	Model	Accuracy	Macro-F1
peak (tabular)	MLP	**0.952**	**0.947**
peak (tabular)	GBDT	0.912	0.904
curve	ROCKET	0.902	0.895
curve	Curve-MLP	0.888	0.880
curve	1D-CNN	0.863	0.854
curve	InceptionTime	0.542	0.509

Bold marks the best value in each column.

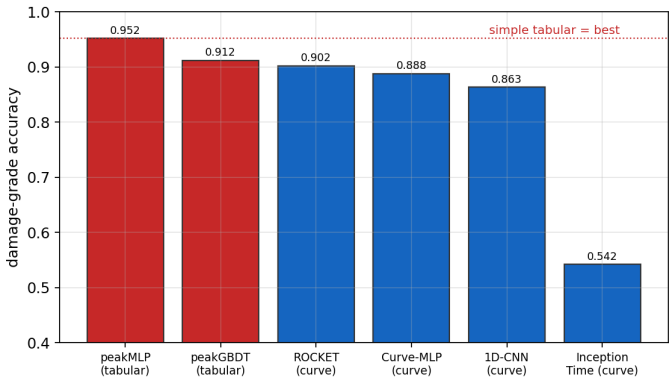


Figure 13. Damage-grade recovery: a five-feature tabular model (red) beats strong curve-based deep classifiers (blue) because the label is determined by the peak temperature, not by curve shape.

4.3.2. Accuracy Versus Computational Cost

The simplicity advantage also holds on the cost axis that matters for deployment. Figure 14 plots damage-grade accuracy against model size: the most accurate model (the five-feature tabular MLP, 2.8 k parameters) is also among the smallest, while the largest curve model (InceptionTime, 14.6 k parameters) is the least accurate. There is thus no accuracy–cost trade-off to exploit here. The parsimonious model dominates on both axes, reinforcing that complexity is not the useful axis for this label.

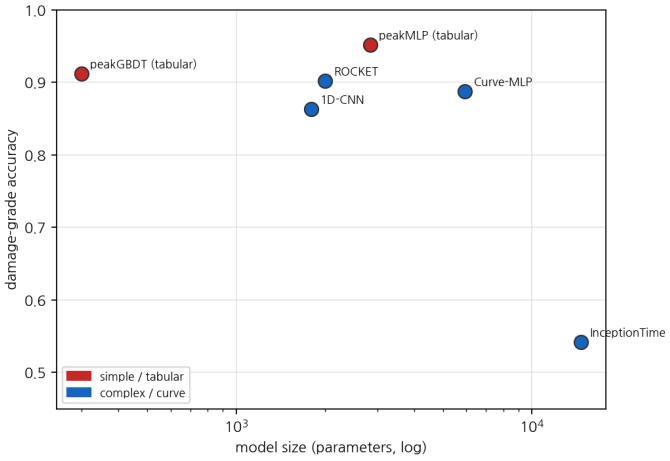


Figure 14. Accuracy vs. cost for damage classifiers. The smallest (tabular) model is the most accurate. Larger curve models do not buy accuracy.

The cost picture for the hazard-state forecasters is summarized in Table 15 (measured on CPU, corpus of 120 for training). The two model families trade cost differently: corpus-trained models pay a one-time training cost (from 3.9 s for the linear DLinear to 114.7 s for PatchTST) and then infer in under a millisecond per event, whereas the per-event physics models (including the proposed CD-PINN) require no training corpus but optimize per event at deployment (∼1.8–9.5 s/event). The distillation step of CD-PINN adds only a sub-second one-time cost. DLinear is the cheapest overall (0.8 k parameters), and CD-PINN's cost equals that of the hand-set PINN while being markedly more accurate. We therefore use "real-time" in the sense appropriate to this application: the corpus-trained models are truly real-time (sub-millisecond) at inference, whereas the per-event physics and CD-PINN are near-real-time, requiring a few seconds per forecast, which is fast relative to the minutes-long fire timescale and one to two orders of magnitude faster than running the simulator, but not sub-second. Model-parameter storage is not a distinguishing factor: every model listed has a sub-megabyte parameter footprint (the largest, PatchTST at 21.5 k parameters, requires less than 0.1 MB in single precision). Because peak runtime memory was not separately profiled and is hardware- and implementation-dependent, we report parameter count as a hardware-independent storage proxy, together with measured training and deployment time. Table 15 covers the hazard-state forecasters compared throughout. The classical closed-form baselines (Table 4) are omitted because their computational cost is negligible relative to the trainable forecasters, and the downstream damage classifiers (Table 14) address a separate prediction task and are characterized there by parameter count.

Table 15. Computational cost of the hazard-state forecasters (CPU, corpus of 120 for training, default epochs). Parameter count is reported as a hardware-independent proxy for model-storage requirements. Peak runtime memory was not separately profiled. The classical baselines (Table 4) and downstream damage classifiers (Table 14) are omitted here as they address separate settings and carry negligible trainable cost.

Model	Params	One-Time Training	Deployment Cost
DLinear (corpus, linear)	0.8 k	3.9 s	<1 ms/event
TimesNet (corpus, deep)	9.5 k	33.4 s	<1 ms/event
PatchTST (corpus, deep)	21.5 k	114.7 s	<1 ms/event
PIKAN (physics, per-event)	0.8 k	n/a	∼1.8 s/event
PINN-MLP (physics, per-event)	7.2 k	n/a	∼9.2 s/event
CD-PINN (proposed)	7.2 k	distillation < 1 s	∼9.5 s/event

Bold marks the proposed model (CD-PINN).

4.3.3. The Label, Not the Model, Sets the Ceiling

We extend the single per-event grade to a time-resolved grade $g(t)$ by mapping the temperature history through an OpenSees structural response curve. This transient-structural label reproduces the official per-event grade for 97.4% of scenarios (vs. 49.3% for a temperature-only label, Figure 15). Under this history-dependent, irreversible label, the ranking reverses (Table 16): an instantaneous tabular model drops to 0.854 accuracy, while a history-aware tabular model (0.954) and a causal CNN (0.896) lead. Thus whether simple or complex wins is decided by what the label depends on, not by the model class.

Table 16. Time-resolved (history-dependent) grade: the simple-vs-complex ranking reverses relative to Table 14.

Input Representation	Nowcast Accuracy	Macro-F1
history-summary tabular	**0.954**	**0.941**
causal 1D-CNN	0.896	0.866
instantaneous tabular	0.854	0.817

Bold marks the best value in each column.

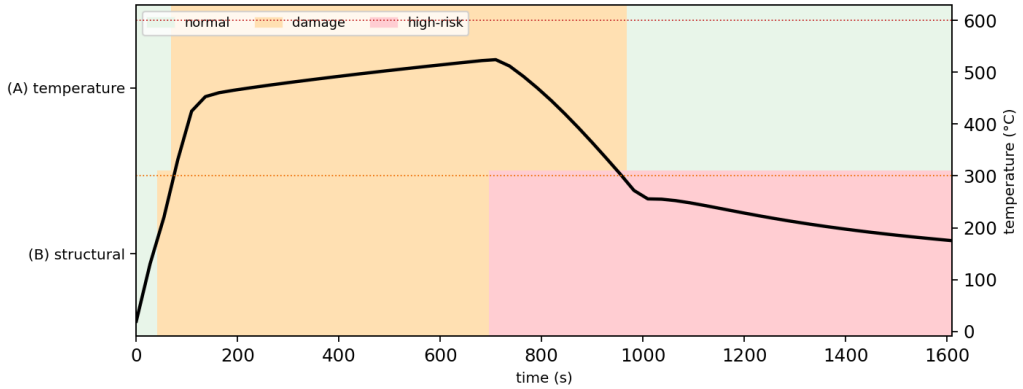


Figure 15. Time-resolved damage grading. Temperature history (top) is mapped to a transient structural response. The grade bands (bottom) progress normal → damage → high-risk, with structural damage onset earlier than the temperature threshold. The orange and red dotted lines mark the 300 and 600 °C grade thresholds, matching the shaded normal/damage/high-risk bands.

A feasibility demonstration that the same pipeline transfers unchanged to a different geometry (arched tunnel fires) is reported in Appendix A. It is a pipeline-feasibility check rather than a second benchmark.

5. Discussion

Table 17 summarizes the deployment rule implied by our results. Complex deep forecasters earn their cost only with a sufficient corpus and a forward forecasting task. Physics is worth its complexity only when the corpus is scarce/absent and the problem is under-determined, and for label recovery a parsimonious tabular model is the right default unless the label is genuinely history-dependent. These results both extend and qualify the long-standing benchmark-forecasting lesson: structure-free classical benchmarks, which are hard to beat for stationary series, fail on single-event hazard extrapolation, so the useful axis is not model complexity but the right minimal structure (physics or a small corpus). Once that structure is present, parsimony is competitive on accuracy (the remaining gains from added complexity are modest, Table 8) and wins decisively on computational cost, and the structure itself is best estimated as a reduced-order residual from the corpus (CD-PINN) rather than hand-set. Generalization, however, comes with a caution: under a mild shift within the same simulator (higher intensity), the simple corpus

model degrades least (Table 11), but under a severe shift to independent laboratory-fire experiments the in-distribution rankings do not transfer, every corpus model's nRMSE increasing by 220–340% and the models becoming statistically indistinguishable on a fire-level bootstrap (Table 13). The honest reading is not that one model class wins, but that forecast performance depends more on appropriate inductive structure, forcing information, and label construction than on raw model complexity: corpus-trained models are best in distribution, yet their advantage is not statistically distinguishable from the other models on the laboratory fires, so synthetic-benchmark rankings provide limited evidence of laboratory-fire generalization.

Table 17. When does forecasting complexity pay off?

Situation	Recommended Model
Corpus rich ($\gtrsim$20), in-distribution forecast	complex deep, or CD-PINN (proposed)
Corpus present, want interpretable physics	CD-PINN (corpus-distilled law)
Corpus scarce or absent (in-distribution)	physics-informed (PINN/PIKAN)
Severe shift (novel laboratory fires)	no single model; validate/calibrate externally
Sparse-sensor full-state reconstruction	physics-informed coordinate net
Damage from peak-determined label	simple tabular (MLP/GBDT)
Damage from time-resolved label	history-aware/sequence
Fast structural inference	tabular surrogate of OpenSees

The main limitations are that the structural analysis is quasi-static (a full dynamic time-history would raise the role of sequence models) and that the FDS reference subset is modest and uses an engineering-grade mesh (0.25 m cells) rather than a fine, validation-grade one. A mesh-convergence study and a larger FDS set are natural extensions.

6. Conclusions

On a reproducible, simulator-grounded benchmark for fire hazard-state and damage forecasting, spanning a single high-fidelity FDS model and a lightweight FDS-calibrated CFAST combination, and corpus-free (per-event) vs. corpus-trained forecasters, no single model class dominates: forecast performance depends more on appropriate inductive structure, forcing information, and label construction than on raw model complexity. Parsimonious models are competitive across most in-distribution regimes, physics helps most in the scarce/under-determined regime, and the damage label, not the model class, sets the ceiling. Beyond the benchmark, we propose CD-PINN, a corpus-distilled physics-informed forecaster that estimates a data-optimal reduced-order physics residual from the corpus and uses it as an interpretable physics constraint, raising a per-event physics model to corpus-model accuracy (0.153 $\rightarrow$ 0.091), and we show, on independent laboratory-fire experiments, that in-distribution rankings do not transfer, the models becoming statistically indistinguishable, a caution that synthetic-benchmark standing is a weak predictor of laboratory-fire generalization. The framework and the accompanying analysis provide a benchmark, a proposed model, and a concrete rule for when forecasting complexity is worth its cost, extensible from building compartments to tunnel fires. Figure 16 summarizes this deployment rule: with a sufficient corpus, a data-driven or corpus-distilled model suffices, but in the data-scarce regime a corpus-free physics-informed model is the better choice, and on genuinely novel fires no single model should be over-trusted.

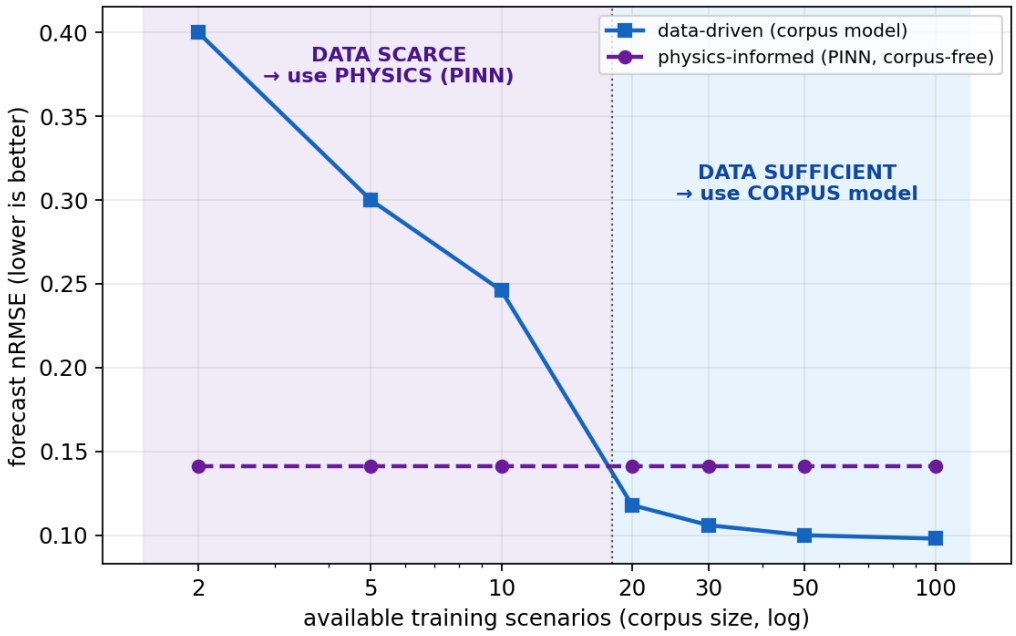


Figure 16. Deployment rule distilled from the benchmark. Below about twenty scenarios the corpus model is under-determined and the corpus-free physics-informed model wins. By twenty scenarios the data-driven model has overtaken and is both more accurate and cheaper to run.

Author Contributions: Conceptualization, D.K., S.L. and I.-H.L.; methodology, D.K.; software, D.K.; validation, D.K., S.L. and I.-H.L.; formal analysis, D.K.; investigation, S.L. and I.-H.L.; resources, D.K.; data curation, D.K.; writing—original draft preparation, D.K.; writing—review and editing, D.K., S.L. and I.-H.L.; visualization, D.K.; supervision, S.L. and I.-H.L.; project administration, S.L. and I.-H.L.; funding acquisition, S.L. and I.-H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (MOLIT), Republic of Korea [RS-2026-25543393], and by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 [RS-2024-00398310].

Institutional Review Board Statement: Not applicable. This study did not involve humans or animals. All data are simulator-generated or come from a publicly available laboratory-fire dataset.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and code supporting the findings of this study, comprising the FDS-calibrated fire hazard-state time series, scenario metadata with damage grades, the FDS/CFAST calibration data, the key experiment result tables, the author-written figure-generation scripts, and the simulator decks (FDS/CFAST), OpenSees models, and experiment scripts, are available from the corresponding author on reasonable request. The external laboratory-fire measurements used for validation are the publicly available UL FSRI compartment-fire dataset [33] (https://github.com/ulfsri/fsri-compartments-2018, accessed on 11 August 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CD-PINN	Corpus-distilled physics-informed neural network (proposed)
CFAST	Consolidated Fire and Smoke Transport (zone model)

CFD	Computational fluid dynamics
DCR	Demand–capacity ratio
EC2	Eurocode 2 (concrete structures)
EC3	Eurocode 3 (steel structures)
FDS	Fire Dynamics Simulator
HAZUS	Hazards US (FEMA loss-estimation methodology)
IDR	Interstory drift ratio
nRMSE	Normalized root-mean-square error
OOD	Out-of-distribution
PIKAN	Physics-informed Kolmogorov–Arnold network
PINN	Physics-informed neural network
SINDy	Sparse identification of nonlinear dynamics

Appendix A. Cross-Geometry Feasibility: Tunnel Fires

To test whether the framework is tied to building compartments, we applied the entire pipeline, unchanged, to fires of a completely different geometry. We generated 20 FDS arched (horseshoe) road-tunnel fires in an 80 m long, 8 m wide tunnel of 2.5 m wall height, meshed at 0.5 m cells, with heat-release rates spanning 5–112 MW and longitudinal ventilation velocities of 0.8–3.9 m/s (Figure 3 shows an example ceiling-jet field). Each fire is labeled with the same ordinal damage grade used for compartments, driven here by the peak ceiling temperature and the length of ceiling held above 600 °C.

All three layers of the pipeline apply without modification. The data-generation layer produces per-scenario temperature and hot-length histories, the labeling layer assigns grades by the same rule, and the forecasting layer ingests the resulting time series with no change to its inputs or architecture. The 20 fires span all four damage grades with a reasonably balanced distribution (Table A1), with peak ceiling temperatures ranging from about 110 to 1400 °C, so the tunnel set exercises the labeling rule across its full range rather than saturating at one extreme.

Table A1. Damage-grade distribution of the 20 FDS arched-tunnel fires (heat-release rate 5–112 MW, longitudinal ventilation 0.8–3.9 m/s).

Grade	Normal	Damage	High-Risk	Collapse	Total
Count	6	4	7	3	20

We deliberately present this as a pipeline-feasibility result rather than a second benchmark. With only 20 fires, the tunnel subset is far too small to re-estimate the corpus-size crossover, the robustness sweeps, or the damage-label analyses that anchor the main study, and the two geometries differ enough that their quantitative results should not be pooled. What the demonstration does establish is that the data-generation, labeling, and forecasting components are geometry-agnostic, requiring no tunnel-specific engineering. A full tunnel benchmark, with the larger corpus those analyses require, is a natural direction for future work.

References

1. Makridakis, S.; Spiliotis, E.; Assimakopoulos, V. The M4 Competition: 100,000 time series and 61 forecasting methods. *Int. J. Forecast.* **2020**, *36*, 54–74. [CrossRef]
2. Kyriazi, F.; Thomakos, D.D. Benchmarks in univariate time-series forecasting: A historical and methodological review. *IMA J. Manag. Math.* **2026**, *37*, 327–374. [CrossRef]
3. Hyndman, R.; Koehler, A.; Ord, K.; Snyder, R. *Forecasting with Exponential Smoothing: The State Space Approach*; Springer: Berlin/Heidelberg, Germany, 2008. [CrossRef]

4. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A time series is worth 64 words: Long-term forecasting with transformers. arXiv **2022**, arXiv:2211.14730.
5. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. TimesNet: Temporal 2D-variation modeling for general time series analysis. arXiv **2022**, arXiv:2210.02186.
6. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2021; Volume 35, pp. 11106–11115. [CrossRef]
7. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22419–22430.
8. Oreshkin, B.N.; Carpov, D.; Chapados, N.; Bengio, Y. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. arXiv **2019**, arXiv:1905.10437.
9. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2023; Volume 37, pp. 11121–11128. [CrossRef]
10. Das, A.; Kong, W.; Sen, R.; Zhou, Y. A decoder-only foundation model for time-series forecasting. In Proceedings of the 41st International Conference on Machine Learning (ICML), Vienna, Austria, 21–27 July 2024; Volume 235, pp. 10148–10167.
11. Ansari, A.F.; Stella, L.; Türkmen, A.C.; Zhang, X.; Mercado, P.; Shen, H.; Shchur, O.; Rangapuram, S.S.; Pineda Arango, S.; Kapoor, S.; et al. Chronos: Learning the language of time series. *arXiv* **2024**, arXiv:2403.07815.
12. Dempster, A.; Petitjean, F.; Webb, G.I. ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Min. Knowl. Discov.* **2020**, *34*, 1454–1495. [CrossRef]
13. Ismail Fawaz, H.; Lucas, B.; Forestier, G.; Pelletier, C.; Schmidt, D.F.; Weber, J.; Webb, G.I.; Idoumghar, L.; Muller, P.A.; Petitjean, F. Inceptiontime: Finding alexnet for time series classification. *Data Min. Knowl. Discov.* **2020**, *34*, 1936–1962. [CrossRef]
14. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. [CrossRef]
15. Karniadakis, G.E.; Kevrekidis, I.G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-informed machine learning. *Nat. Rev. Phys.* **2021**, *3*, 422–440. [CrossRef]
16. Tancik, M.; Srinivasan, P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J.; Ng, R. Fourier features let networks learn high frequency functions in low dimensional domains. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 7537–7547.
17. Sitzmann, V.; Martel, J.; Bergman, A.; Lindell, D.; Wetzstein, G. Implicit neural representations with periodic activation functions. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 7462–7473.
18. Liu, Z.; Wang, Y.; Vaidya, S.; Ruehle, F.; Halverson, J.; Soljacic, M.; Hou, T.Y.; Tegmark, M. KAN: Kolmogorov–Arnold networks. arXiv **2024**, arXiv:2404.19756.
19. Patra, S.; Panda, S.; Parida, B.K.; Arya, M.; Jacobs, K.; Bondar, D.I.; Sen, A. Physics Informed Kolmogorov-Arnold Neural Networks for Dynamical Analysis via Efficient-KAN and WAV-KAN. *J. Mach. Learn. Res.* **2025**, *26*, 1–39.
20. Shukla, K.; Toscano, J.D.; Wang, Z.; Zou, Z.; Karniadakis, G.E. A comprehensive and FAIR comparison between MLP and KAN representations for differential equations and operator networks. *Comput. Methods Appl. Mech. Eng.* **2024**, *431*, 117290. [CrossRef]
21. Lu, L.; Jin, P.; Pang, G.; Zhang, Z.; Karniadakis, G.E. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nat. Mach. Intell.* **2021**, *3*, 218–229. [CrossRef]
22. Brunton, S.L.; Proctor, J.L.; Kutz, J.N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 3932–3937. [CrossRef] [PubMed]
23. McGrattan, K.B.; Hostikka, S.; McDermott, R.; Floyd, J.; Weinschenk, C.; Overholt, K. Fire Dynamics Simulator Technical Reference Guide, Volume 1: Mathematical Model. In *Technical Report NIST Special Publication 1018*, 6th ed.; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2013. [CrossRef]
24. Peacock, R.D.; McGrattan, K.B.; Forney, G.P.; Reneke, P.A. CFAST—Consolidated Model of Fire Growth and Smoke Transport (Version 7), Volume 1: Technical Reference Guide. In *Technical Report NIST Technical Note 1889v1*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2015. [CrossRef]
25. Li, Y.Z.; Lei, B.; Ingason, H. The maximum temperature of buoyancy-driven smoke flow beneath the ceiling in tunnel fires. *Fire Saf. J.* **2011**, *46*, 204–210. [CrossRef]
26. McKenna, F. OpenSees: A framework for earthquake engineering simulation. *Comput. Sci. Eng.* **2011**, *13*, 58–66. [CrossRef]
27. Jiang, J.; Usmani, A. Modeling of steel frame structures in fire using OpenSees. *Comput. Struct.* **2013**, *118*, 90–99. [CrossRef]
28. European Committee for Standardization. *Eurocode 2: Design of Concrete Structures, Part 1-2: General Rules, Structural Fire Design (EN 1992-1-2)*; Technical Report; CEN: Brussels, Belgium, 2004.
29. European Committee for Standardization. *Eurocode 3: Design of Steel Structures, Part 1-2: General Rules, Structural Fire Design (EN 1993-1-2)*; Technical Report; CEN: Brussels, Belgium, 2005.

30. Federal Emergency Management Agency. *Multi-Hazard Loss Estimation Methodology: HAZUS-MH 2.1, Earthquake Model Technical Manual*; Technical Report; Federal Emergency Management Agency: Washington, DC, USA, 2012.
31. Peherstorfer, B.; Willcox, K.; Gunzburger, M. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *Siam Rev.* **2018**, *60*, 550–591. [CrossRef]
32. Kennedy, M.C.; O'Hagan, A. Predicting the output from a complex computer code when fast approximations are available. *Biometrika* **2000**, *87*, 1–13. [CrossRef]
33. Willi, J.M.; Madrzykowski, D.; McKinnon, M.; Dow, N. Experimental data from compartment fires with gas burner and upholstered furniture fuels. *Data Brief* **2023**, *47*, 108934. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.