BER the international journal of bridge engineering management and research

Transformer-Based Sequence Learning for Multi-Horizon Corrosion Initiation Probability Prediction in Reinforced Concrete Bridges

Yuzhong Huang and Wei Zheng*

Submitted: 04 May 2026 **Accepted:** 01 July 2026 **Publication date:** 10 July 2026

DOI: 10.70465/ber.v3i3.94

Abstract: Chloride-induced corrosion initiation is a major durability concern in reinforced concrete bridges. This study frames corrosion-initiation probability prediction as a multi-horizon sequence-learning problem motivated by the future use of deterioration records, laboratory studies, sensor histories, exposure variables, and published deterioration sequences for infrastructure forecasting.

Because compatible real-world deterioration sequence data are not yet fully organized for this task, a diffusion-based stochastic reliability framework is used to generate controlled simulator-derived corrosion-initiation probability trajectories. These trajectories provide a first-stage proof-of-concept testbed for evaluating whether a temporal fusion transformer (TFT) can organize scenario variables and historical sequence information to predict future probability trajectories.

The results show that the sequence-learning model closely follows the simulator-derived corrosion-initiation probability trajectory within the sampled scenario space and preserves the expected cover-depth ordering. Benchmark comparison under a common sliding-window task shows that gated recurrent units achieved the lowest numerical error in this smooth simulation setting, while TFT remained competitive and provided an interpretable and extensible multi-horizon forecasting structure. MC Dropout further provides an approximate model-level uncertainty estimate for the trained TFT predictor.

These findings support the feasibility of simulation-trained sequence learning for controlled corrosion-initiation probability trajectory prediction, but they do not constitute direct field validation. The contribution is a structured proof-of-concept workflow for organizing historical observations, covariates, future or scenario-defined inputs, benchmark evaluation, and approximate uncertainty quantification. Field validation using bridge inspection records, laboratory deterioration sequences, sensor histories, exposure histories, and maintenance records remains necessary before practical deployment.

Author keywords: Corrosion initiation; chloride ingress; reinforced concrete bridges; sequence learning; temporal fusion transformer; durability assessment

Introduction

Reinforced concrete (RC) bridge infrastructure forms a critical component of modern transportation systems. However, chloride-induced corrosion of embedded steel reinforcement remains one of the primary drivers of structural deterioration and life-cycle cost escalation. Long-term exposure to marine environments and de-icing salts leads to progressive chloride ingress through the concrete cover, eventually initiating corrosion, cracking, and loss of structural capacity.[1–3] Environmental factors such as atmospheric chloride deposition, aerosol transport, and long-term accumulation further influence chloride transport and corrosion development.[4–6] As a result, reliable estimation of corrosion-initiation probability is essential for durability assessment and maintenance planning of RC bridge structures.

Chloride transport in concrete is commonly described using diffusion-based models derived from Fick's second law.[7–9] While these models provide a useful first-order approximation, field observations indicate that long-term deterioration involves more complex interactions, including time-dependent diffusivity, microstructural changes, cracking, and environmental variability.[10,11] In addition, data-driven probabilistic approaches have been increasingly explored to capture environmental influences on deterioration processes.[12] These considerations highlight the limitations of purely diffusion-based formulations and motivate the need for modeling approaches that can account

*Corresponding Author: Wei Zheng. Email: j00411035@jsums.edu
Department of Civil and Environmental Engineering, Jackson State University, Jackson, MS, USA

for multiple sources of uncertainty. Recent developments in chemo-mechanical modeling further emphasize the coupled nature of corrosion-induced damage and transport processes.[13] For bridge-scale and network-level applications, repeated probabilistic evaluation under varying exposure and design conditions can become computationally demanding.

To address uncertainty in deterioration evolution, probabilistic reliability methods have been widely adopted to estimate time-dependent corrosion-initiation probability and structural performance.[14–17] These approaches incorporate uncertainties in material properties, environmental exposure, and structural geometry, and have been extended to include data fusion from inspection and monitoring systems.[18,19] Despite these advances, traditional reliability analyses often rely on repeated stochastic simulations for each scenario, which can limit their scalability for large-scale or network-level infrastructure assessments.[20,21]

Recent studies have emphasized the importance of integrating environmental variables and physical degradation mechanisms into predictive models for improved reliability forecasting.[22] Long-term exposure experiments and durability studies further demonstrate the need to explicitly account for environmental conditions and material behavior in deterioration modeling.[23] In infrastructure asset management, deterioration models play an important role in life-cycle cost analysis and maintenance planning, and system-level modeling frameworks have shown that incorporating probabilistic uncertainty propagation can improve prediction consistency and support decision-making processes.[24–27]

In parallel, data-driven and machine-learning approaches have been increasingly applied to infrastructure deterioration prediction. Neural network-based models and data-driven frameworks have demonstrated promising performance in predicting structural condition and long-term degradation trends.[28–30] Large-scale data analytics and multistate deterioration models further enable improved prediction of infrastructure performance under uncertainty.[31,32] However, many existing machine-learning studies focus on predicting condition states or degradation indicators, rather than directly representing reliability quantities such as corrosion-initiation probability, and they often face challenges related to data quality, interpretability, and generalization across varying environmental conditions.[33]

Advances in probabilistic machine learning and sequence modeling have created new opportunities for representing time-dependent deterioration processes. Techniques such as autoregressive probabilistic forecasting and recurrent neural networks enable distributional prediction under uncertainty.[34,35] More recently, Temporal Fusion Transformers (TFTs) have demonstrated strong capability in multi-horizon forecasting through attention mechanisms and interpretable feature selection.[36] These characteristics make TFT suitable for problems in which the target evolves over time and depends on both static and time-varying inputs.

Real deterioration-related sequence data already exist in bridge engineering practice and research, including inspection histories, component condition ratings, field chloride measurements, corrosion monitoring, structural-health-monitoring records, traffic records, weather and exposure histories, maintenance histories, long-term laboratory corrosion tests, and published experimental deterioration datasets. However, these sources are fragmented, heterogeneous, and not yet organized into a machine-learning-ready multi-horizon sequence format that separates historical deterioration states, static bridge/material/environmental covariates, observed historical external inputs, known or scenario-defined future inputs, and future deterioration targets.

The broader prediction problem motivating this work is long-term bridge deterioration forecasting from accumulated deterioration-status sequences and external impact variables. In future field applications, annual bridge inspection histories, component condition-state records, maintenance records, structural health monitoring histories, weather and climate histories, traffic histories, chloride or marine exposure records, and laboratory or literature-derived deterioration sequences can be organized as longitudinal training data. The present simulator-derived corrosion-initiation dataset is used as a controlled first-stage testbed for developing the sequence-learning structure before such heterogeneous real-world data are cleaned, aligned, and validated for model training.

Accordingly, the prediction accuracy reported in this study represents agreement between the trained sequence-learning model and the stochastic simulator, rather than validation against observed corrosion initiation in real bridges. Direct field validation requires structured data that include bridge identity, inspection dates, deterioration states, material and geometric properties, cover depth, exposure environment, chloride measurements or proxies, traffic and loading histories, maintenance histories, and observed corrosion or deterioration indicators.

Although diffusion-threshold reliability models can generate time-dependent corrosion-initiation probability trajectories, they do not by themselves provide a structured sequence-learning framework that separates historical observations, static bridge/material/environmental variables, known or scenario-defined future inputs, and future deterioration targets. Existing bridge deterioration studies have used inspection data, experimental data, physics-based models, Markov or semi-Markov models, and machine-learning methods, but much of this work focuses on condition ratings, corrosion rates, or state-transition probabilities. The remaining gap addressed here is the lack of a structured, interpretable, and uncertainty-aware multi-horizon sequence-learning framework for future corrosion-initiation probability trajectory prediction.

The TFT was selected because its architecture aligns with this forecasting structure. It separates static covariates, historical observed inputs, known or scenario-defined future inputs, variable-selection mechanisms, temporal attention, and multi-horizon outputs. This organization is useful for future bridge deterioration forecasting, where historical deterioration states, material properties, environmental exposure, traffic loading, and maintenance scenarios may need to be represented jointly. The value of TFT in this

study is architectural alignment with the intended real-data forecasting problem, not a claim of universal numerical superiority over gated recurrent units (GRU) or other sequence models.

Recent studies also show why a TFT-based deterioration forecasting framework requires careful positioning relative to related civil-engineering time-series work. TFT has been used for structural health monitoring and anomaly detection in heritage structures,[37] while Transformer or attention-based models have been explored for pavement skid, texture, and distress deterioration prediction.[38,39] These studies demonstrate growing interest in attention-based civil-infrastructure time-series modeling, but use of a TFT-style framework for scenario-based, multi-horizon bridge corrosion or deterioration probability forecasting remains limited. To the authors' knowledge, prior bridge deterioration studies have not yet widely developed a forecasting structure that explicitly separates long-term historical deterioration states, observed historical external inputs, static bridge or material attributes, known or expected future external inputs, and future deterioration probability outputs within one interpretable multi-horizon model.

Accordingly, this study is positioned as a controlled proof-of-concept sequence-learning framework. Beyond reproducing simulator-derived trajectories, the study evaluates whether a sequence-learning model can learn population-level corrosion-initiation probability trajectories, compares TFT with simpler baseline models under a common forecasting task, identifies dominant physical parameters through Sobol sensitivity analysis, and assesses approximate model-level uncertainty using MC Dropout. These additions clarify the methodological contribution while preserving the limitation that field validation remains future work.

Methodology

Chloride diffusion and stochastic reliability modeling

Chloride-induced corrosion initiation in RC can be formulated as a transport-driven limit-state problem, in which corrosion onset occurs when the chloride concentration at the reinforcement depth exceeds a critical threshold. The overall simulator-to-surrogate workflow used in this study is summarized in Fig. 1. Random material, environmental, and geometric inputs were first sampled to drive a chloride diffusion simulator. corrosion-initiation labels were then generated using a threshold-crossing criterion at the reinforcement depth, and the resulting realization-level labels were aggregated into population-level corrosion-initiation probability trajectories, $P_f(t)$, for surrogate-model training and evaluation.

Diffusion-based models derived from Fick's second law remain a widely used baseline for representing chloride ingress in concrete structures.[8] However, their direct applicability is limited because concrete is a reactive and heterogeneous medium, and the assumption of constant diffusivity is rarely satisfied over depth and time. Experimental and analytical studies further indicate that chloride transport is influenced by ionic interactions, binding effects, and electrochemical processes, motivating the use of effective diffusion parameters rather than intrinsic material constants.[9]

Let $C(x, t)$ denote the chloride concentration at depth x and time t. Corrosion initiation at the reinforcement depth x_{rebar} is described by the limit-state function

$$g(t) = C_{crit} - C(x_{rebar}, t), \tag{1}$$

where C_{crit} is the critical chloride concentration. Corrosion initiation is assumed to occur when

$$g(t) \leq 0. \tag{2}$$

The population-level corrosion-initiation probability is then defined as

$$P_f(t) = P(T_{init} \leq t) = P(g(t) \leq 0), \tag{3}$$

where T_{init} denotes the corrosion-initiation time. This formulation is consistent with probabilistic reliability frameworks for infrastructure systems, in which the temporal evolution of failure probability provides a more informative basis for engineering interpretation than a single deterministic estimate.

The present simulator focuses on chloride-induced corrosion initiation, defined as a threshold-crossing event when the chloride concentration at the reinforcement depth exceeds the critical chloride concentration. Therefore, post-initiation corrosion propagation, rust expansion, cracking, and spalling are not explicitly modeled. Chloride binding, wetting/drying cycles, and temperature-dependent transport are also not resolved explicitly; their potential influence is represented only indirectly through the apparent diffusion coefficient, surface chloride concentration, aging exponent, and stochastic variability of the input parameters.

In practice, chloride transport and corrosion initiation are influenced by multiple uncertain factors, including surface chloride concentration C_s, diffusion coefficient D, concrete cover depth x, and critical chloride threshold C_{crit}. These parameters may vary due to differences in material properties and environmental exposure conditions.

Model-form limitations and measurement uncertainty are acknowledged when interpreting the simulator-derived trajectories. Cracking and variability in material and exposure conditions can significantly affect initiation behavior relative to idealized uncracked assumptions.[11] In addition, uncertainty in chloride content assessment, including variability associated with semi-destructive testing and data processing, influences the calibration and interpretation of transport parameters.[19] These considerations motivate the use of stochastic reliability modeling as a physically grounded approach for generating consistent corrosion-initiation probability trajectories that serve as reference outputs for subsequent data-driven approximation.

Importantly, in the proposed framework, the simulated chloride concentration field $C(x, t)$ is used solely to generate

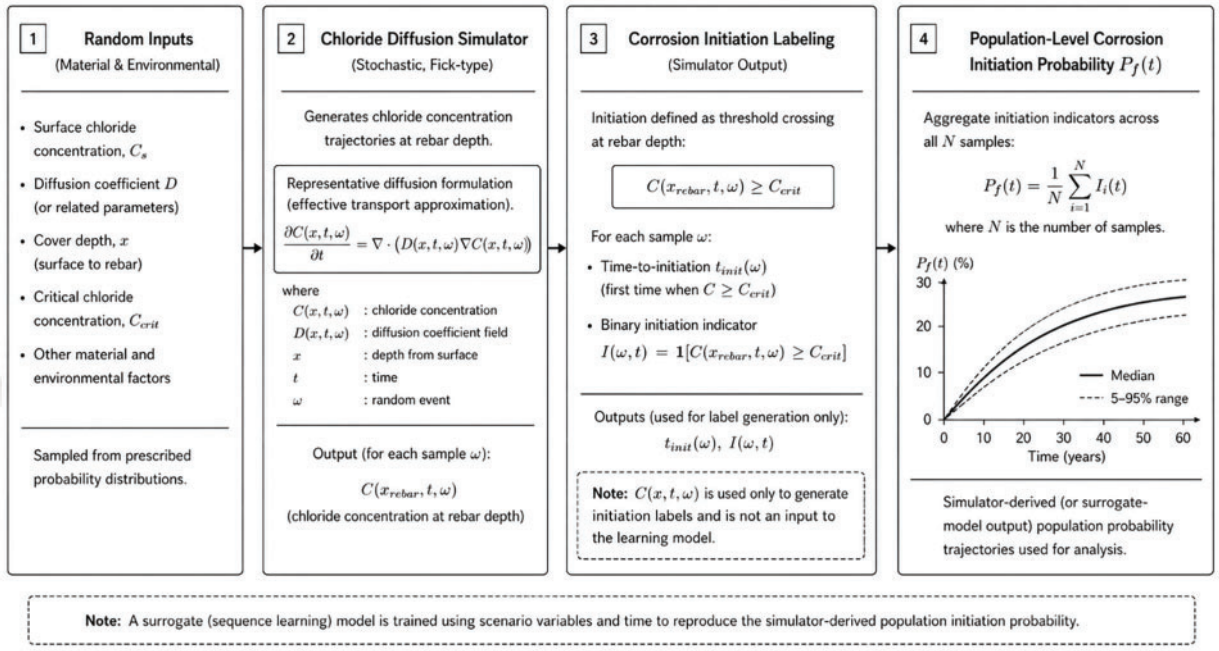

Figure 1. Schematic framework for generating simulator-derived corrosion-initiation probability trajectories $P_f(t)$

corrosion-initiation labels through the threshold condition in Eqs. (1) and (2) and is not provided as an input to the learning model. This design explicitly prevents circularity and ensures that the sequence-learning model learns to approximate population-level probability trajectories from scenario variables rather than reproducing simulator outputs through direct leakage.

Fig. 1 provides an overview of the modeling framework, illustrating the progression from stochastic input sampling to diffusion-based transport, corrosion-initiation labeling, and aggregation into population-level corrosion-initiation probability, $P_f(t)$. Within this framework, chloride concentration fields are generated by a stochastic transport simulator and used only to determine initiation events through the threshold criterion. The resulting simulator-derived probability trajectories define the reference outputs for training.

Any subsequent learning model is trained using scenario variables and time to reproduce this probability evolution, without directly using concentration fields as inputs. This separation between simulator outputs and learning inputs ensures consistency with the underlying reliability formulation and enables the development of a reusable sequence-learning approximation while preserving physically meaningful relationships (Table 1).

Stochastic population simulation

The population simulator generates ensembles of deterioration realizations by sampling uncertain input parameters and mapping each realization to a corrosion-initiation outcome over time. For each simulated element, a diffusion-based transport model produces $C(x_{\text{rebar}}, t)$ over the service-life horizon, and corrosion initiation is defined as the first time at which $C(x_{\text{rebar}}, t) \geq C_{\text{crit}}$. Aggregation across realizations yields an empirical estimate of the population-level corrosion-initiation probability, $P_f(t)$, expressed as the fraction of realizations that have initiated corrosion by time t.

In this study, the stochastic simulator is used as a controlled data-generation mechanism to construct reference probability trajectories for sequence-learning model development, rather than as a direct representation of field-observed corrosion behavior. The simulator-derived trajectories provide a consistent and physically grounded basis for evaluating whether a sequence-learning model can reproduce time-dependent corrosion-initiation probability under uncertainty.

The simulation inputs are sampled from prescribed probability distributions representing material properties, environmental exposure, and structural configuration. The simulation setup includes the service-life horizon, time discretization, number of realizations, and scenario configurations. These parameters are selected to ensure sufficient coverage of the input space while maintaining computational tractability. The full configuration of the simulation and dataset generation is summarized in Table 2.

In addition to representing uncertainty propagation, the simulation framework supports conditional evaluation across representative structural or exposure conditions. For example, stratification by concrete cover depth produces distinct reliability trajectories that reflect the sensitivity of corrosion-initiation probability to transport path length, a dependence that is well documented in chloride ingress design and durability assessment.[14] These stratified responses are used in subsequent sections to assess whether the sequence-learning model preserves physically meaningful relationships.

Table 1. Input variables used in the corrosion-initiation model

Variable	Description	Unit	Distribution	Range
C_s	Surface chloride concentration	kg/m^3	Lognormal	2–6 kg/m^3
D_{28}	Reference diffusion coefficient (28 days)	m^2/s	Lognormal	$1 \times 10^{-12} - 5 \times 10^{-12}$ m^2/s
m	Aging exponent	–	Normal	0.2–0.6
x	Concrete cover depth	mm	Uniform	40–110 mm
C_{crit}	Critical chloride concentration	kg/m^3	Lognormal	0.6–1.2 kg/m^3

Table 2. Simulation and dataset configuration

Item	Value
Service-life horizon	0–60 years
Time step	4 weeks (~0.077 years)
Number of scenario configurations	1000
Number of time steps per series	783
Total dataset size	783,000 rows
Train/validation/test split	700/150/150 series
Approximate row split	548, 100/117, 450/117, 450 rows
Cover-depth range	40–110 mm
Stratified groups	40–60, 60–80, 80–110 mm
Random seed	20250111

To ensure methodological consistency and avoid circularity, the simulator outputs are used only to generate corrosion-initiation labels and aggregated probability trajectories. The underlying chloride concentration field $C(x, t)$ is not used as an input to the learning model. Instead, the sequence-learning model is trained using scenario variables and time to approximate the resulting population-level probability evolution. This separation ensures that the learning task remains aligned with the reliability formulation and prevents information leakage from intermediate simulator states.

The simulator is formulated to remain consistent with both simplified reliability methods and full stochastic simulation-based approaches, which have been compared in corrosion modeling studies and motivate the use of learned approximations for repeated scenario evaluation.[20] By generating large ensembles of physically consistent trajectories, the simulation framework provides a reproducible and computationally tractable basis for training and evaluating sequence-learning models.

The dataset generated in this study consists of population-level corrosion-initiation probability trajectories, $P_f(t)$, associated with sampled input configurations. These trajectories are derived from simulator outputs and serve as reference responses for subsequent sequence-learning model development under uncertainty.

TFT sequence learning model

The learning task is formulated as multi-horizon prediction of simulator-derived population-level corrosion-initiation probability trajectories from historical sequence information and associated covariates. In the present proof-of-concept setting, the sequences are generated by a stochastic diffusion-threshold simulator. In the intended future real-data setting, the same forecasting structure could organize accumulated inspection histories, laboratory deterioration sequences, sensor histories, exposure variables, and maintenance or scenario-defined future inputs for predicting $P_f(t)$.

Within this formulation, the input–output structure distinguishes static variables, historical observed variables, known or scenario-defined future inputs, and future prediction targets. This distinction matters because infrastructure forecasting may require evaluating how future deterioration trajectories change under different exposure, traffic, weather, chloride, or maintenance scenarios. TFT is used because it is directly structured for this kind of multi-horizon forecasting task, although other architectures may achieve lower numerical error in simplified datasets.

This input-role separation is also important when comparing TFT with other sequence models. Markov and semi-Markov models remain useful for condition-state transition modeling, but they are less directly structured for preserving a full historical deterioration path together with time-varying historical and future external inputs. Recurrent models such as recurrent neural networks, GRUs, and long short-term memories can process ordered deterioration histories and remain useful benchmark architectures[40–42]; however, in very long deterioration records they compress past information into hidden states, which may limit retention, interpretation, or selective use of distant time steps.

General Transformer models improve long-range dependency modeling through attention mechanisms[43] and can be adapted to include external covariates. The advantage of TFT for the present research direction is not that other models cannot use such variables, but that TFT provides a forecasting-specific organization for static covariates, observed historical time-varying inputs, known future inputs, variable selection, gating, temporal attention, and multi-horizon outputs.

The model is trained to reproduce the simulator-derived trajectories $P_f(t)$ directly, rather than intermediate state variables such as chloride concentration. Consistent with the data-generation framework described in the previous section, the underlying chloride concentration field $C(x, t)$ is not used as an input to the model, thereby preventing information leakage and ensuring that the learning task remains aligned with the reliability formulation.

Although TFT provides a flexible architecture for interpretable and extensible multi-horizon reliability-trajectory learning, its use is not assumed to imply universal superiority over simpler models. Therefore, additional benchmark models, including pointwise Logistic Regression, Windowed multi-output linear/sigmoid regression, Windowed MLP, and GRU, were included to contextualize predictive performance under the present simulation setting. The benchmark is intended to evaluate the credibility of the forecasting formulation, not to establish that TFT is the best model for every deterioration dataset.

TFT architecture and temporal fusion rationale

The TFT architecture used in this study is based on the TFT framework.[36] TFT separates the forecasting problem into static covariates, historically observed inputs, known future inputs, and future prediction targets. This separation is important for bridge deterioration forecasting because future deterioration probability is not determined only by the past target trajectory; it may also depend on observed historical external conditions and known or scenario-defined future external conditions.

In the present forecasting formulation, the historical observed input window contains two distinct components. The first component is the historical state sequence, which represents the past evolution of the target deterioration or corrosion-initiation state. The second component is the observed historical external input sequence, which represents exogenous variables observed over the same historical period. In addition, the prediction horizon is associated with a known or scenario-defined future external input window. These future external inputs are conditioning variables over the prediction horizon and are not prediction targets. The output of the model is the future corrosion-initiation probability sequence over the forecast horizon.

The term temporal fusion refers to the model's ability to combine multiple categories of time-related information within a unified forecasting architecture. TFT separates static covariates, historical state information, observed historical external inputs, and known future external inputs according to their forecasting roles, and then fuses them through variable-selection networks, static covariate encoders, local temporal processing, static enrichment, interpretable temporal attention, and gated residual components. This design supports future bridge-deterioration applications in which past deterioration states, past external conditions, static attributes, and expected future weather, traffic, exposure, or maintenance conditions may jointly influence future deterioration probability trajectories.

Compared with a general Transformer, TFT is more directly designed for structured multi-horizon forecasting. A general Transformer can be adapted to include external covariates, but the distinction among static covariates, historical observed inputs, known future inputs, and future prediction targets must usually be imposed by the model developer. In contrast, TFT provides built-in mechanisms for this separation through variable-selection networks, static context encoding, recurrent local temporal processing, interpretable self-attention, and gated residual networks. Therefore, the use of TFT in this study is motivated by its forecasting-specific organization of state sequences and external input sequences, not by a claim that TFT is always numerically superior to other sequence models.

The resulting TFT input–output organization is illustrated in Fig. 2. In this study, each supervised sample uses a 52-step historical window and a 13-step prediction horizon. The 52-step historical window includes both the historical target-state sequence and the observed historical external input sequence. The 13-step future horizon includes known or scenario-defined external inputs that condition the forecast. The model then produces a 13-step future target-state sequence representing the predicted corrosion-initiation probability trajectory.

To ensure a fair comparison among the windowed and sequence-learning baselines, Windowed multi-output linear/sigmoid regression, Windowed MLP, GRU, and TFT were evaluated using the same sliding-window forecasting task. A common windowed dataset was constructed using a 52-step historical input window, a 13-step future prediction horizon, and a sliding stride of one time step. These models used the same physical covariates, cumulative corrosion-initiation target, scenario-level train/validation/test split, overlap-averaged inference procedure, population-level aggregation method, and mean absolute error/root mean square error (MAE/RMSE) evaluation metrics.

For benchmark interpretation, Pointwise Logistic Regression was retained only as a simple pointwise linear reference; it does not use a 52-step historical window to predict a 13-step future horizon. The Windowed multi-output linear/sigmoid regression baseline uses the same 52-step historical window and 13-step horizon but remains a fixed-window linear baseline. The Windowed MLP uses the same 52-step history after flattening it into fixed features and can capture nonlinear feature interactions, but it does not explicitly model temporal order through recurrence or attention. GRU processes ordered historical sequences through gated recurrent hidden states and can perform strongly on smooth, low-dimensional simulator trajectories. TFT explicitly separates static covariates, observed historical inputs, known

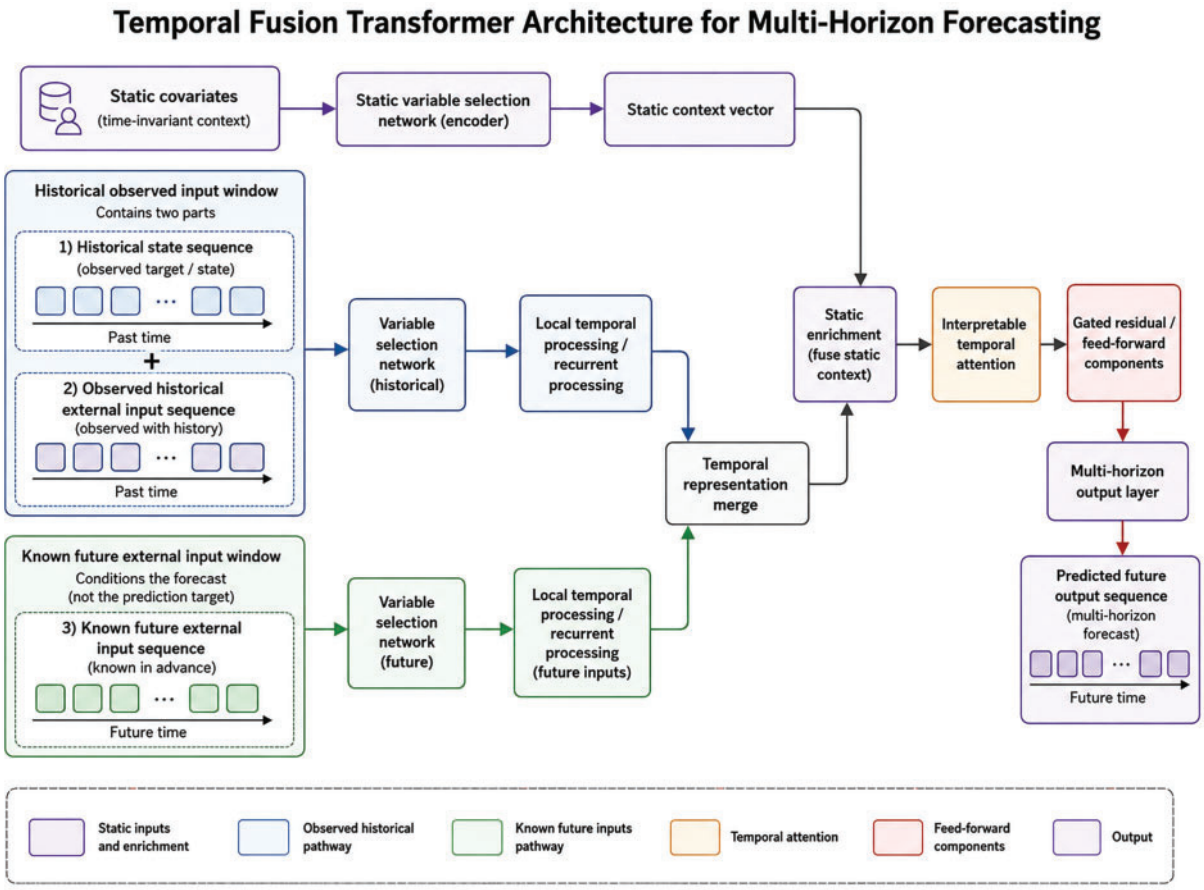

Figure 2. Temporal fusion transformer architecture for multi-horizon corrosion-initiation forecasting. The framework separates static covariates, historical observed inputs, known or scenario-defined future inputs, and multi-horizon probability outputs. Variable selection, temporal processing, static enrichment, attention, and gated residual components support structured sequence-to-sequence prediction

or scenario-defined future inputs, attention, gating, and multi-horizon outputs, which is why it remains structurally important for the intended real-data forecasting problem even when it is not the lowest-error model in this simulation dataset.

The model was trained using an independent series-level split of the simulated dataset, with early stopping based on validation loss to prevent overfitting. The detailed TFT training configuration, including encoder length, prediction horizon, hidden dimension, attention heads, dropout rate, batch size, optimizer, learning rate, number of epochs, and early stopping criterion, as well as train/validation split, is summarized in Table 3.

This setup ensures that the model learns generalizable temporal patterns rather than memorizing simulator outputs. The approach is consistent with probabilistic sequence modeling frameworks that learn conditional distributions for time series under uncertainty; including distribution-oriented forecasting approaches based on quantile representations.[34,35]

The role of physical modeling in this framework is reflected using simulator-generated targets derived from the diffusion-threshold reliability formulation, rather than through explicit enforcement of governing equations during training. In this sense, the sequence model acts as a data-driven approximation of simulator outputs, while remaining consistent with the underlying physical assumptions embedded in the transport-based reliability model. This distinction differentiates the present formulation from physics-informed neural network approaches that aim to infer transport fields directly.[44]

Training and loss design

The training data are constructed from simulated samples of the form $\left(\mathbf{x}, \mathbf{z}_{1:T}, P_f(1:T)\right)$, where $\mathbf{x}$ denotes static covariates, $\mathbf{z}_{1:T}$ represents time-varying covariates, and $P_f(1:T)$ corresponds to the simulator-derived population-level corrosion-initiation probability trajectory over the prediction horizon. Each sample therefore represents a complete trajectory associated with a given scenario configuration, rather than independent point-wise observations, which is consistent with the objective of learning time-dependent reliability evolution.

To explicitly define the learning task and avoid ambiguity regarding variable usage, the input-output structure of the sequence-learning model is summarized in Table 4. In particular, simulator-generated chloride concentration $C(x, t)$

Table 3. TFT model training configuration

Item	Value
Encoder length	52 time steps (∼4.0 years)
Decoder/prediction horizon	13 time steps (∼1.0 year)
Hidden dimension	32
Number of attention heads	4
Dropout rate	0.1
Batch size	32 training/64 inference
Optimizer	Adam
Learning rate	3×10^{-4}
Maximum epochs	10
Early stopping	Validation loss monitored; patience = 3
Data split	Independent series split: 700/150/150
Random seeds	20250111, 20250112, 20250113

Table 4. Input–output definition for the surrogate learning task

Component	Role in study	Used as TFT input?
Surface chloride concentration (C_s)	Environmental variable	Yes
Diffusion coefficient (D_{28})	Material parameter	Yes
Aging exponent (m)	Time-dependent parameter	Yes
Cover depth (x)	Structural variable	Yes
Time (t)	Sequence index	Yes
Chloride concentration $C(x, t)$	Simulator intermediate variable	No (label generation only)
Critical threshold C_{crit}	Initiation threshold	Yes
Population-level $P_f(t)$	Target variable	Predicted output

is used only for label generation and is not included as a model input, thereby preventing information leakage and ensuring that the model does not directly access intermediate simulator states.

The target variable $P_f(t)$ is bounded within [0, 1] and typically evolves smoothly and monotonically over time under fixed exposure conditions. Accordingly, the learning objective emphasizes reproducing stable trajectory patterns while reducing sensitivity to high-frequency fluctuations arising from finite-sample stochastic simulation, which are not directly relevant for population-level reliability assessment.

Model performance is evaluated using trajectory-level error measures, including MAE and RMSE, which quantify deviations between predicted and simulator-derived $P_f(t)$ trajectories. These metrics are adopted as first-stage fidelity measures because the primary objective is to approximate the simulator-defined reliability evolution rather than to perform full probabilistic calibration.

Although quantile-based loss formulations are widely used in probabilistic sequence modeling frameworks,[34,35] they are not explicitly implemented in the present study. Instead, such formulations are referenced only to indicate methodological compatibility, while the current implementation prioritizes deterministic trajectory approximation consistent with the simulator-fidelity objective.

The resulting trained model serves as a learned representation of the stochastic simulation process, providing a reusable approximation for repeated scenario evaluation while maintaining consistency with the probabilistic interpretation of corrosion-initiation trajectories derived from established durability modeling frameworks.[14,15] This formulation enables efficient evaluation across multiple exposure scenarios and structural configurations within the sampled parameter space and supports large-scale screening applications in bridge deterioration assessment.[20]

Experimental Design

Simulation scenario generation

The experimental design is based on a controlled generator of stochastic chloride-ingress scenarios intended to represent dominant sources of uncertainty in chloride-induced corrosion initiation in RC under representative marine aerosol and de-icing exposure conditions. In this study, the simulation framework is used as a controlled environment for evaluating the ability of a data-driven sequence-learning model to reproduce time-dependent reliability trajectories, rather than as a direct representation of field-calibrated deterioration

behavior. Corrosion initiation is formulated as a threshold-crossing event at the reinforcement depth, governed by the temporal evolution of chloride concentration at the bar location. Diffusion-based formulations derived from Fick's second law are used as a baseline representation; however, their underlying assumptions, including constant diffusivity and homogeneous transport, are not strictly satisfied in concrete. Accordingly, diffusion parameters are interpreted as effective transport quantities rather than intrinsic material constants.[8,9]

Each realization in the simulation is defined by a sampled set of material, environmental, and geometric variables. The diffusion model is used to generate chloride concentration histories at the reinforcement depth, which are subsequently used to determine corrosion initiation. In implementation, the scenario generator samples surface chloride concentration C_s, reference diffusivity D_{28} defined at 28 days, an aging exponent m controlling time-dependent diffusivity, concrete cover depth x, and a critical chloride threshold C_{crit}.

The effective diffusivity is modeled as a time-dependent quantity using a power-law aging function,

$$D(t) = D_{28}\left(\frac{t}{t_{\text{ref}}}\right)^{-m}, \tag{4}$$

which reflects the empirically observed reduction in transport rate over time under field conditions and is widely adopted in durability-oriented diffusion modeling.[8,9]

The chloride concentration at the reinforcement depth is evaluated over a discretized service horizon, 0–60 years, using the complementary error-function solution associated with one-dimensional diffusion under constant surface boundary conditions, with $D(t)$ treated as a time-dependent effective coefficient. This formulation provides an analytically tractable approximation while acknowledging model-form limitations and unresolved transport complexities in heterogeneous concrete systems.

Corrosion initiation is defined as the first time at which the computed chloride concentration at the reinforcement depth satisfies

$$C(x_{\text{rebar}}, t) \geq C_{\text{crit}}, \tag{5}$$

and for each realization the corresponding initiation time T_{init} is recorded. A time-dependent binary indicator is then constructed as

$$I(t) = 1(t \geq T_{\text{init}}), \tag{6}$$

where $1(\cdot)$ denotes the indicator function. The population-level corrosion-initiation probability $P_f(t)$ is estimated as the ensemble average,

$$P_f(t) = \mathbb{E}[I(t)], \tag{7}$$

which represents the fraction of realizations that have initiated corrosion by time t.

This probabilistic representation provides a consistent measure of time-dependent reliability for durability-oriented assessment, capturing uncertainty propagation across realizations rather than relying on a single deterministic estimate.[14,15] The formulation also reflects variability arising from heterogeneous transport conditions, cracking effects, and measurement uncertainty, which can influence apparent initiation behavior relative to idealized assumptions.[11] Importantly, the simulation framework preserves the expected monotonic relationship between cover depth and corrosion-initiation probability, providing a physically interpretable reference for subsequent sequence-learning evaluation.

Dataset construction

The learning task is formulated at the population level, while model training is performed using realization-level time series that collectively represent the underlying stochastic behavior of corrosion initiation. Each stochastic realization corresponds to one simulated structural element evaluated over a common discrete time grid associated with physical time t. This formulation allows the sequence-learning model to learn the evolution of population-level corrosion-initiation probability while remaining consistent with probabilistic reliability-based durability assessment frameworks.

For each realization, the dataset is constructed from sampled material, environmental, and geometric parameters, including surface chloride concentration C_s, reference diffusivity D_{28}, aging exponent m, concrete cover depth x, and the critical chloride threshold C_{crit}. These variables define the scenario configuration and are used as model inputs. The chloride concentration field $C(x_{\text{rebar}}, t)$ is generated by the diffusion-based simulator and is used solely to determine corrosion-initiation labels and reference trajectories; it is not included as an input to the learning model, thereby avoiding circularity in the mapping from inputs to outputs and preventing information leakage between the simulator and the sequence-learning model.

Corrosion initiation is represented through a binary indicator that is equal to zero prior to initiation and unity thereafter, and the corresponding population-level initiation probability $P_f(t)$ is obtained as the ensemble mean of this indicator across realizations at each time step, consistent with established probabilistic durability assessment practice.[14,15] The resulting dataset therefore consists of realization-level input variables and associated time-indexed initiation indicators, together with the aggregated population-level probability trajectory used as the reference target.

To examine whether the learned mapping preserves physically meaningful behavior, the dataset is interpreted through both aggregated and covariate-conditioned views. Aggregation across all realizations provides the overall population trajectory. Cover-depth conditioning is examined using finer 10-mm bands in Fig. 3 and summarized using broader intervals, 40–60, 60–80, and 80–110 mm, in Table 6. Because cover depth directly influences chloride transport path length, shallower cover is expected to lead to earlier corrosion initiation and higher values of $P_f(t)$ over a fixed horizon.[14] Preservation of this ordering is treated as a physical consistency check complementary to numerical error metrics.

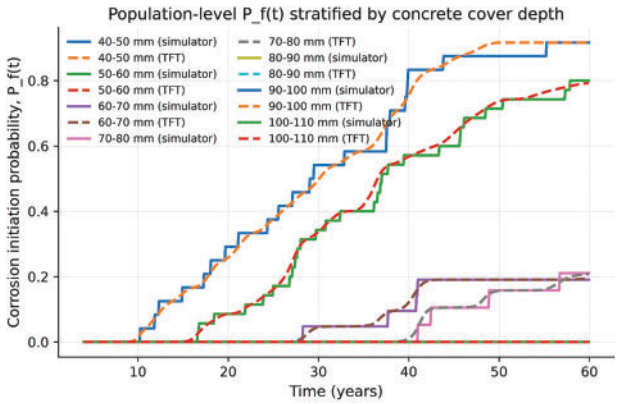

Figure 3. Cover-depth-stratified comparison between simulator-derived and TFT-predicted corrosion-initiation probability $P_f(t)$

Table 5. Population-level prediction accuracy metrics for $P_f(t)$

Metric	Value
Mean absolute error (MAE)	0.004543
Root mean squared error (RMSE)	0.006382
Maximum absolute error	0.024123
Year of maximum error	36.80 years
Final-year error (∼60 years)	0.001347

During inference, the sequence-learning model produces the probability of corrosion initiation at each time step for each realization. Because sequence models operate using sliding decoder windows, multiple predictions may be generated for the same realization-time pair. These overlapping predictions are combined through averaging to obtain a single pointwise estimate. The population-level predicted trajectory $P_{f,\text{pred}}(t)$ is then computed as the mean of these pointwise probabilities across realizations at each time step, mirroring the Monte Carlo estimator used to define the reference trajectory $P_{f,\text{true}}(t)$. This construction ensures that comparisons between predicted and reference curves are performed consistently at the level of the population-level reliability quantity, rather than being influenced by sequence-level prediction artifacts.

Evaluation metrics

Evaluation focuses on how accurately the sequence-learning model reproduces the simulator-defined population-level corrosion-initiation probability trajectory under the controlled proof-of-concept setting. Model accuracy is assessed by comparing the predicted trajectory with the reference trajectory over the full service-life horizon, with errors evaluated on the probability scale. These metrics should be interpreted as simulator-to-model fidelity measures, not as direct field-validation metrics.

MAE and RMSE are adopted as primary performance metrics to quantify trajectory-level agreement between predicted and reference probability curves. These metrics summarize deviations over the full time horizon and are consistent with standard practice when the quantity of interest is a time-indexed reliability function rather than a single endpoint estimate.[15]

To characterize the temporal distribution of discrepancies, the pointwise absolute error is defined as

$$e(t) = \mid P_{f,\text{pred}}(t) - P_{f,\text{true}}(t) \mid, \tag{8}$$

where $P_{f,\text{pred}}(t)$ denotes the model-predicted population-level corrosion-initiation probability and $P_{f,\text{true}}(t)$ denotes the simulator-derived reference trajectory. The temporal evolution of $e(t)$ is analyzed to identify regions where deviations are relatively large. Such localized discrepancies may arise from finite-sample Monte Carlo variability in the reference trajectory and increased sensitivity of $P_f(t)$ during periods of rapid change.

In addition to accuracy, computational performance is evaluated by comparing inference-time runtime between two evaluation procedures defined within the same simulation framework. The first corresponds to repeated stochastic simulation, in which corrosion-initiation probability is estimated through Monte Carlo sampling for each scenario configuration. The second corresponds to learned-model inference, where the trained model directly maps input variables and time to the probability trajectory without repeated simulation. The comparison therefore isolates differences between simulation-based evaluation and learned-model inference, rather than post-processing steps.

The computational comparison is interpreted as an inference-stage advantage, recognizing that model training incurs an upfront computational cost. Once trained, the trained model provides a consistent approximation to the simulator-defined mapping while enabling efficient evaluation across a large number of scenarios. This trade-off is particularly relevant for network-level screening and reliability assessment tasks that require repeated evaluation under diverse configurations.[20,21]

Taken together, the evaluation framework assesses trajectory fidelity, physical consistency, benchmark context, approximate model-level uncertainty, and conditional computational reuse. It does not establish universal acceleration over physics-based simulation or field-ready bridge deterioration prediction.

Global sensitivity analysis

To quantify the influence of major physical input parameters, a variance-based Sobol global sensitivity analysis was conducted using the same diffusion-threshold corrosion-initiation model and locked input distributions used for dataset generation. The analyzed parameters were C_s, D_{28}, m, concrete cover depth x, and C_{crit}. Saltelli sampling with a base sample size of 2,048 generated 14,336 model evaluations. Sensitivity indices were computed at 20, 40, and 60 years using the smooth limit-state margin $C_{rebar}(t) - C_{crit}$ as the primary response. A binary corrosion-initiation response

was also evaluated as a supplemental check. First-order S_1 and total-order S_T indices were reported.

Uncertainty quantification using MC Dropout

To quantify the predictive uncertainty of the trained TFT model, MC Dropout was used as a practical approximation for uncertainty quantification. Dropout was originally introduced as a regularization method for reducing overfitting in neural networks.[45] The theoretical foundation of MC Dropout is provided by Gal and Ghahramani, who showed that dropout can be interpreted as approximate Bayesian inference in deep neural networks.[46] During inference, dropout layers were kept active while the remaining model components were retained in evaluation mode. Under identical input conditions, repeated stochastic forward passes were performed using the trained TFT checkpoint. Because each forward pass uses a different dropout mask, the model produces a slightly different corrosion-initiation probability trajectory. The resulting ensemble of predictions was treated as an empirical predictive distribution of $P_f(t)$.

From this predictive distribution, the predictive mean, predictive standard deviation, empirical 2.5th percentile, empirical 97.5th percentile, approximate 95% predictive interval, and corresponding lower and upper predictive bounds were computed at each time step. In the present study, 50 stochastic forward passes were used for the main uncertainty estimation, and an additional 100-pass run was conducted to assess the of the predictive mean and standard deviation. The MC Dropout results should be interpreted as approximate model-level epistemic uncertainty within the simulation setting, not as exact Bayesian inference or fully calibrated physical, measurement, environmental, or field uncertainty.

Results

Population $P_f(t)$ prediction accuracy

The sequence-learning model reproduced the simulator-derived population-level trajectory with small error relative to the probability scale. As summarized in Table 5, the representative TFT checkpoint achieved an MAE of 0.004543 and an RMSE of 0.006382, with a maximum absolute error of 0.024123 and a final-year error of 0.001347. These results indicate close simulator-to-model fidelity within the sampled scenario space.

The results should be interpreted in the context of model-form uncertainty. Diffusion-based transport provides an effective representation of chloride ingress in concrete; however, it does not fully account for multi-ionic transport, binding effects, wetting and drying cycles, temperature effects, electrochemical coupling, corrosion propagation, cracking, or spalling. Within this context, the TFT is a learned approximation of the simulator-defined $P(t)$ trajectory, and its fidelity remains tied to the assumptions and sampled parameter ranges of the underlying simulator.

Cover depth reliability prediction

When population trajectories are stratified by concrete cover depth, the TFT preserves the expected ordering and separation of reliability curves. As shown in Fig. 3, shallower cover is associated with earlier and higher initiation probability, whereas deeper cover corresponds to delayed initiation and lower $P_f(t)$ over the same time horizon. This behavior reflects the fundamental role of transport path length in chloride ingress, where shorter diffusion distances accelerate the time required for chloride concentration to reach the reinforcement level.

The stratified trajectories in Fig. 3 demonstrate that the sequence-learning model captures conditional variation associated with structural covariates, rather than representing the population using a single aggregated trajectory. To further quantify this agreement, Table 6 reports the number of test series, final-year reference and predicted probabilities, and corresponding absolute errors for each cover-depth group.

The results indicate that prediction errors remain small across all cover-depth groups and that the monotonic ordering of initiation probability is consistently preserved. This ordering is a physically meaningful feature of diffusion-driven corrosion initiation and serves as an important physical consistency check for the sequence-learning model. The ability to reproduce such stratification supports the interpretation that the learned mapping respects the dominant transport-controlled mechanism embedded in the diffusion-threshold formulation, rather than merely fitting aggregate trends.

It is important to note that this agreement should be interpreted as consistency with the assumed physical model, rather than independent validation of corrosion mechanisms. The diffusion-threshold formulation provides a simplified representation of chloride ingress and does not explicitly account for additional processes such as multi-species transport or chemical binding. Nevertheless,

Table 6. Final-year population-level corrosion-initiation probability by broad cover-depth group on the held-out test set

Cover-depth group	Test series (N)	Final reference probability	Final predicted probability	Final absolute error
40–60 mm	59	0.847	0.843	0.004
60–80 mm	40	0.200	0.201	0.001
80–110 mm	51	0.000	0.000	0.000

Table 7. Sobol first-order and total-order sensitivity indices for the smooth limit-state margin $C_{rebar}(t) - C_{crit}$

Parameter	(S_1), 20 yr	(S_T), 20 yr	(S_1), 40 yr	(S_T), 40 yr	(S_1), 60 yr	(S_T), 60 yr
C_S	0.005	0.021	0.013	0.028	0.020	0.033
D_{28}	0.018	0.032	0.019	0.026	0.020	0.024
m	0.101	0.183	0.143	0.193	0.170	0.204
Cover depth	0.671	0.781	0.717	0.783	0.716	0.760
C_{crit}	0.092	0.092	0.037	0.037	0.025	0.025

preservation of the expected cover-depth dependence indicates that the model maintains coherence with the governing physical assumptions underlying the simulator.

More broadly, the observed stratified behavior aligns with the intended role of sequence models in learning covariate-conditioned temporal patterns. The TFT is designed for multi-horizon prediction with mixed covariates and includes mechanisms for covariate-dependent temporal modeling.[36] In this context, the ability to recover cover-depth-dependent trajectories can be interpreted as a learned approximation of simulator-defined responses across heterogeneous scenario groups. Such behavior is particularly relevant in civil infrastructure applications, where deterioration processes are governed by heterogeneous material, geometric, and environmental conditions, and stratified prediction provides a structured representation of variation across related groups.[32]

Sensitivity analysis of input parameters

The total-order index S_T of cover depth remained high across the three service times, with values of 0.781, 0.783, and 0.760 at 20, 40, and 60 years, respectively. The aging exponent was consistently the second most influential parameter, indicating that long-term diffusivity evolution also contributes to uncertainty in corrosion initiation. The remaining parameters, including C_s, D_{28}, and C_{crit}, had smaller but non-negligible effects depending on service time. The corresponding Sobol first-order and total-order sensitivity indices are summarized in Table 7.

These results indicate that, within the assumed diffusion-threshold framework, geometric protection provided by concrete cover has the strongest influence on corrosion-initiation risk. The increasing contribution of the aging exponent from 20 to 60 years further suggests that long-term diffusivity evolution becomes more influential as service time increases.

Time-dependent prediction error

The absolute error over time remains small relative to the probability scale and exhibits limited variation across the prediction horizon. As shown in Fig. 4, the error profile is not uniform and localized fluctuations are present; however, these variations do not introduce sustained bias in the predicted trajectory. This behavior is consistent with the nature of the learning target, which is derived from a finite-sample Monte Carlo estimate of a probability function. Even when the underlying transport process is deterministic under fixed parameters, the population-level reliability $P_f(t)$ reflects sampling variability, and localized irregularities may arise from stochastic labeling near the threshold-crossing region.

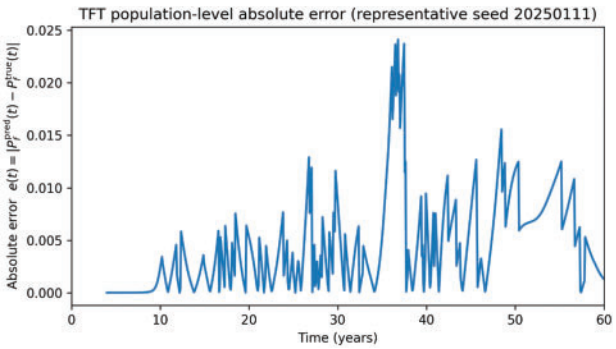

Figure 4. Time-dependent absolute error between TFT-predicted and simulator-derived population-level corrosion-initiation probability using the representative TFT checkpoint

The maximum absolute error over the full service-life horizon was 0.024123, occurring at approximately 36.80 years. The final-year absolute error was 0.001347, indicating that the long-term endpoint remained closely reproduced.

Such effects tend to be more pronounced in time intervals where initiation probability increases rapidly, as small differences in the implied initiation-time distribution can lead to larger instantaneous deviations in $P_f(t)$. As illustrated in Fig. 4, these discrepancies remain bounded and do not propagate into systematic divergence between predicted and reference trajectories.

Time-dependent reliability analysis typically relies on the evolution of $P_f(t)$ to support inspection scheduling and maintenance planning; therefore, preserving trajectory shape and avoiding sustained bias are more important than matching every localized finite-sample fluctuation exactly.

The observed error characteristics indicate that the TFT provides a stable learned approximation of the simulator-defined reliability trajectory for scenario-level evaluation within the sampled parameter space.

Computational efficiency

The computational comparison is interpreted cautiously. A single vectorized Monte Carlo simulation can be faster

Table 8. MC Dropout uncertainty summary

Metric	Value	Purpose/interpretation
Number of MC Dropout forward passes	50	Main stochastic inference setting
Predictive mean MAE	0.004616	Accuracy of MC Dropout mean relative to simulator-derived trajectory
Predictive mean RMSE	0.006503	Trajectory-level deviation of MC Dropout mean
Average predictive standard deviation over time	0.000461	Average spread among stochastic dropout predictions
Maximum predictive standard deviation over time	0.001682	Largest time-localized model-level uncertainty
Average 95% predictive interval width	0.001668	Average uncertainty-band width across service life
Maximum 95% predictive interval width	0.006213	Largest uncertainty-band width across service life
Year of maximum predictive interval width	59.95 years	When model-level uncertainty is largest
Final-year predictive mean	0.38564	Predicted corrosion-initiation probability at final service year
Final-year lower predictive bound	0.38309	Lower 2.5th percentile at final service year
Final-year upper predictive bound	0.38930	Upper 97.5th percentile at final service year
Average absolute change in predictive mean, 50 vs. 100 passes	3.77×10^{-5}	Stability check for predictive mean
Average absolute change in predictive standard deviation, 50 vs. 100 passes	2.73×10^{-5}	Stability check for uncertainty estimate
Maximum absolute change in predictive mean, 50 vs. 100 passes	2.47×10^{-4}	Worst-case convergence difference for the mean
Maximum absolute change in predictive standard deviation, 50 vs. 100 passes	2.40×10^{-4}	Worst-case convergence difference for the standard deviation
Prediction interval coverage probability, 50 passes	0.115	Empirical coverage of MC Dropout model-level prediction interval; not a calibrated field confidence interval.
Prediction interval coverage probability, 100 passes	0.114	Empirical coverage of MC Dropout model-level prediction interval; not a calibrated field confidence interval.

than neural-model inference in the present implementation, and model training requires an upfront cost. The practical value of the trained sequence model is therefore conditional: after training, it may support repeated scenario screening or future real-data forecasting workflows where many bridge, exposure, traffic, maintenance, or climate scenarios must be evaluated under a consistent learned mapping. Therefore, the reported runtime results should not be interpreted as a universal acceleration claim.

For a single simulation run, Monte Carlo evaluation is faster than TFT inference in the present implementation. The trained model becomes advantageous when it is reused across many scenario evaluations or embedded in larger screening workflows. Excluding training cost, the inference-time break-even point is approximately $612/5.35 \approx 115$ simulation-equivalent evaluations. Including the one-time training cost, the break-even point is approximately $(49,161 + 612)/5.35 \approx 9,303$ simulation-equivalent evaluations. Therefore, the reported computational benefit should be interpreted as a reusable inference-stage advantage for repeated large-scale scenario screening rather than as acceleration of a single simulation run.

Uncertainty quantification

From this predictive distribution, the predictive mean, predictive standard deviation, empirical 2.5th percentile, empirical 97.5th percentile, approximate 95% predictive interval, and corresponding lower and upper predictive bounds for $P_f(t)$ were computed at each time step. The resulting MC Dropout uncertainty metrics are summarized in Table 8.

Increasing the number of stochastic forward passes from 50 to 100 changed the predictive mean by 3.77×10^{-5} on average and the predictive standard deviation by 2.73×10^{-5} on average; the maximum absolute changes were 2.47×10^{-4} and 2.40×10^{-4}, respectively.

To clarify the benchmark comparison basis, Table 9 summarizes the input representation, historical input window, prediction horizon, target, and evaluation role of each surrogate model. Pointwise Logistic Regression was retained as a simple pointwise linear reference, whereas the Windowed multi-output linear/sigmoid regression, Windowed MLP, GRU, and TFT were evaluated using the same 52-step historical input window and 13-step prediction horizon.

Table 9. Benchmark design for surrogate model comparison

Model	Input representation	Historical input window	Prediction horizon	Sliding stride	Target	Evaluation role
Pointwise logistic regression	Pointwise covariates at each evaluation time	Not sequence-to-sequence	Pointwise prediction	Not applicable	Cumulative corrosion-initiation probability	Simple pointwise linear reference only
Windowed multi-output linear/sigmoid regression	Flattened 52 × 7 input window	52 time steps	13 time steps	1 time step	Same cumulative corrosion-initiation target	Fairer fixed-window linear baseline
Windowed MLP	Flattened 52 × 7 input window	52 time steps	13 time steps	1 time step	Same cumulative corrosion-initiation target	Fixed-window nonlinear baseline
GRU	Ordered 52-step sequence	52 time steps	13 time steps	1 time step	Same cumulative corrosion-initiation target	Recurrent sequence-learning baseline
TFT	Ordered 52-step sequence with structured covariates	52 time steps	13 time steps	1 time step	Same cumulative corrosion-initiation target	Interpretable multi-horizon sequence-learning model

Benchmark model comparison

To address the need for broader benchmark evaluation, Logistic Regression, Windowed multi-output linear/sigmoid regression, Windowed MLP, GRU, and TFT were compared using consistent target definitions, data splits, and population-level evaluation metrics. To improve benchmark fairness among fixed-window, nonlinear, and sequence-learning baselines, Windowed multi-output linear/sigmoid regression, Windowed MLP, GRU, and TFT were evaluated using the same sliding-window forecasting task. A common windowed dataset was constructed with a 52-step historical input window, a 13-step future prediction horizon, and a sliding stride of one time step. The Windowed multi-output linear/sigmoid regression and Windowed MLP baselines flattened the 52 × 7 covariate window into a fixed 364-dimensional input vector and used 13 future outputs, whereas GRU and TFT used ordered sequence inputs to predict the same 13-step horizon. All windowed models used the same physical covariates, cumulative corrosion-initiation target, scenario-level train/validation/test split, overlap-averaged inference procedure, population-level aggregation method, and MAE/RMSE evaluation metrics. Pointwise Logistic Regression was retained as a simpler linear reference using the same covariates and target at each evaluation time, but it was not treated as a fully equivalent sequence-to-sequence benchmark.

As summarized in Table 10, GRU achieved the lowest prediction error in the present low-dimensional simulation task, with an MAE of 0.001934 and an RMSE of 0.002931. TFT achieved the second-lowest error, with an MAE of 0.004542 and an RMSE of 0.006373, and outperformed the Windowed MLP, Windowed multi-output linear/sigmoid regression, and pointwise Logistic Regression baselines. Windowed MLP produced an MAE of 0.006975 and an RMSE of 0.009964. The Windowed multi-output linear/sigmoid regression baseline slightly improved over Pointwise Logistic Regression, reducing MAE from 0.020652 to 0.020212 and RMSE from 0.024271 to 0.023823, but remained much less accurate than Windowed MLP, GRU, and TFT. These results indicate that adding a historical window provides limited benefit to a linear model, while nonlinear and sequence-learning architectures provide additional predictive improvement. TFT should not be interpreted as universally superior, because GRU remained the most accurate model in this smooth simulator-derived dataset. Instead, TFT is retained as an interpretable and extensible multi-horizon architecture that is structurally aligned with future real-data deterioration forecasting, where historical deterioration states, static attributes, observed external inputs, and known or scenario-defined future external inputs may need to be modeled jointly.

Discussion and Guidelines for Applications

The results demonstrate simulator-to-model fidelity, physical consistency with the assumed diffusion-threshold

Table 10. Benchmark comparison of models under the common forecasting setup

Model	MAE	RMSE	Interpretation
Pointwise Logistic Regression	0.020652	0.024271	Pointwise linear reference; highest error
Windowed multi-output linear/sigmoid regression	0.020212	0.023823	Fairer fixed-window linear baseline
Windowed MLP	0.006975 ± 0.000843	0.009964 ± 0.001328	Fixed-window nonlinear baseline
GRU	0.001934 ± 0.000169	0.002931 ± 0.000207	Best numerical accuracy in the present simulation task
TFT	0.004542 ± 0.000261	0.006373 ± 0.000649	Interpretable multi-horizon sequence-learning model

Note: Pointwise Logistic Regression and Windowed multi-output linear/sigmoid regression are deterministic linear baselines and are reported without random-seed variability. Values for Windowed MLP, GRU, and TFT are reported as mean ± standard deviation across three random seeds: 20250111, 20250112, and 20250113.

mechanism, benchmark context, and approximate model-level uncertainty; they do not demonstrate universal TFT superiority, field validation, or complete physical uncertainty quantification. GRU achieved the lowest numerical error for the present smooth, low-dimensional simulation dataset, while TFT remains useful as an interpretable multi-horizon framework for future datasets that must separate static covariates, observed histories, known future or scenario-defined inputs, and future probability trajectories.

The MC Dropout analysis extends the TFT model from a deterministic predictor to an uncertainty-aware predictor. By retaining dropout during inference and repeatedly evaluating the trained model under identical input conditions, this procedure generates an approximate posterior predictive distribution for $P_f(t)$. The resulting distribution allows the model to report the predictive mean, predictive dispersion, and approximate 95% predictive interval, including lower and upper predictive bounds for $P_f(t)$.

However, this uncertainty estimate has important limitations. MC Dropout captures only approximate epistemic uncertainty associated with the trained sequence-learning model. It does not account for field measurement errors, environmental variability, chloride transport model assumptions, corrosion threshold uncertainty, or long-term deterioration mechanisms not included in the simulator. Therefore, the uncertainty estimates reported in this proof-of-concept study should be interpreted as model-level uncertainty indicators within the adopted simulation setting, rather than fully calibrated prediction intervals for real bridge deterioration.

The preceding results support a workflow in which physics-based stochastic reliability modeling provides traceable proof-of-concept targets, while sequence learning provides a structured forecasting formulation for organizing historical observations, covariates, and future probability trajectories. In the present study, this role remains simulation-based. Future extensions should test whether the same structure can be trained and validated using inspection records, laboratory deterioration sequences, sensor histories, exposure histories, maintenance records, and published deterioration datasets.

The framework may also be extended to settings where heterogeneous data sources are available and inspection or measurement uncertainty needs to be incorporated. Although the present results are derived from simulation-based reference trajectories, the underlying reliability formulation could be adapted to data-fusion applications in which monitoring or inspection data help constrain uncertainty in deterioration states and model parameters.[18] In practical applications, uncertainties in chloride measurements and semi-destructive testing may influence parameter calibration and should therefore be considered when transferring the sequence-learning model to field-oriented workflows.[19]

At the same time, the findings highlight that diffusion-based formulations should be interpreted with appropriate caution. Previous studies have documented the limitations of strict Fickian assumptions in concrete, particularly their inability to fully represent depth- and time-dependent transport behavior.[8,9] The sequence-learning model does not address these limitations directly; instead, it provides a computationally efficient approximation of the chosen reliability framework. From an engineering perspective, this distinction is important: while the approach improves scalability and facilitates systematic sensitivity analysis, the fidelity of the results remains tied to the adequacy of the underlying transport-initiation model and the representativeness of the sampled scenario space. Therefore, the applicability of the sequence-learning model should be interpreted within the scope of the adopted physical assumptions and parameter ranges.

Limitations

The present study should be interpreted as a first-stage concept demonstration of a physics-guided, simulation-trained sequence-learning framework for corrosion-initiation probability trajectory prediction. The reported agreement reflects simulator-to-model fidelity under the assumed stochastic diffusion-threshold reliability formulation, rather than

direct validation against field-observed bridge corrosion behavior.

Although field or laboratory validation is essential for practical deployment, a single external data point was not used as formal validation in this revision. Such a comparison would require compatible calibration of exposure conditions, cover depth, diffusion parameters, chloride measurement protocols, critical chloride threshold, and the definition of corrosion initiation. Without this compatibility, an isolated point comparison could be misleading and could overstate the level of validation. Future validation should therefore rely on structured datasets that support calibration and independent testing.

Corrosion initiation is represented as a threshold-crossing event at the reinforcement depth, which enables efficient population-level reliability evaluation. This simplified formulation was adopted to maintain a controlled proof-of-concept setting for simulator-to-model reliability-trajectory learning. It does not explicitly resolve chloride binding, wetting/drying cycles, temperature-dependent transport, corrosion propagation, rust expansion, or crack formation. Fully coupled treatment of these transport, electrochemical, and damage-evolution mechanisms would require additional material-specific, environmental, and field- or laboratory-calibrated parameters. These mechanisms should therefore be incorporated in future calibrated extensions rather than treated as fully resolved processes in the present simulation-based study.

The model was trained and tested within predefined parameter ranges; therefore, its predictions should be interpreted as interpolation within the sampled scenario space rather than reliable extrapolation beyond it. Outside these ranges, including unusually high surface chloride exposure, very low cover depth, extreme diffusivity, atypical critical chloride thresholds, or nonstationary environmental conditions, the learned mapping may become unreliable and the prediction error may increase. Extreme conditions may also change the governing deterioration mechanisms, making the diffusion-threshold assumptions less representative. Therefore, predictions outside the sampled domain should be treated as exploratory and verified through additional physics-based simulation, field calibration, or retraining with expanded parameter ranges.

In addition, the population-level corrosion-initiation probability is estimated from finite Monte Carlo realizations and therefore inherently contains sampling variability. This variability propagates into the training targets and may contribute to localized fluctuations in predicted probability trajectories, especially at later service times when uncertainty in initiation timing increases. Such behavior is consistent with the finite-ensemble nature of the reference data and should be interpreted within that context.

From a computational perspective, the efficiency advantage of the sequence-learning approach is primarily realized during inference. Model training requires non-negligible computational resources and depends on the quality and coverage of the training dataset. Furthermore, predictive stability can be influenced by hyperparameter selection and architectural design, which may require iterative calibration in practical implementations.

Although additional benchmark models were included in this revision, the comparison remains limited to a simulation-derived dataset. Future work should extend the benchmark evaluation to tree-based models, Gaussian process regression, larger datasets, and field-calibrated deterioration scenarios.

Finally, the current implementation focuses on corrosion-initiation probability rather than full deterioration progression or structural capacity loss. While initiation probability is a key durability indicator, its direct use in engineering decision-making requires integration with downstream models of damage evolution, structural performance, and life-cycle cost. Therefore, the applicability of the proposed sequence-learning framework should be interpreted within the scope of the adopted physical assumptions and the sampled scenario space.

Conclusion

This study presents a physics-guided, simulation-trained sequence-learning framework for multi-horizon forecasting of population-level corrosion-initiation probability trajectories in RC bridge systems. A stochastic diffusion-based simulator is used to generate controlled time-dependent reliability targets under uncertainty, and a TFT is trained to approximate the mapping from scenario variables, historical sequence structure, and time to the corresponding future probability evolution $P_f(t)$.

Within the controlled simulation setting, the results indicate that the sequence-learning model closely reproduces simulator-defined population-level reliability trajectories, preserves physically meaningful stratification across cover-depth groups, and maintains small bounded prediction deviations over the service-life horizon. The benchmark results further show that TFT should not be interpreted as universally superior, because GRU achieved the lowest numerical error in this smooth simulation dataset.

The contribution of this study lies in establishing a structured proof-of-concept workflow for sequence-based reliability trajectory prediction, benchmark comparison, sensitivity interpretation, and approximate model-level uncertainty quantification. The present simulation-generated dataset is a controlled first-stage substitute until real inspection, laboratory, sensor, exposure, maintenance, and published deterioration sequence data are structured for model training and field validation.

From an engineering perspective, the framework may support future scenario screening and deterioration forecasting once it is calibrated and validated with compatible real-world data. The current results demonstrate simulator-to-model fidelity only and should not be interpreted as readiness for direct bridge maintenance decision-making. Practical implementation requires future validation using bridge inspection records, laboratory deterioration sequences, field chloride profiles, sensor histories, exposure and traffic records, maintenance histories, and published deterioration datasets.

Recommendations for Future Work

Future work should extend the proposed framework beyond the current simulation-based setting. First, bridge inspection records, long-term laboratory corrosion measurements, field chloride profiles, structural health monitoring histories, exposure and traffic records, maintenance records, and published deterioration sequences should be cleaned, aligned, calibrated, and organized into machine-learning-ready sequence datasets. These data will support calibration of physical input distributions, external validation of prediction performance, and the transition from a simulator-based proof-of-concept to data-informed bridge deterioration forecasting.

Second, future work should extend the benchmark evaluation to additional model classes, larger datasets, and field-calibrated deterioration scenarios to determine whether the relative performance observed in the present simulation setting remains consistent under more complex conditions.

Third, the current formulation, which focuses on corrosion initiation, should be extended to incorporate subsequent deterioration stages such as corrosion propagation, cracking, section loss, and structural capacity degradation. Finally, the framework should be integrated with bridge management and life-cycle decision models so that predicted reliability trajectories can support inspection prioritization, maintenance planning, and scenario-based infrastructure assessment. These developments would enable the transition from a controlled simulation framework to data-informed and decision-oriented applications.

Acknowledgments

This work was supported in part by the Maritime Transportation Research and Education Center (MarTREC), a U.S. Department of Transportation University Transportation Center, under Contract No. 69A3552348331. The authors gratefully acknowledge this support.

Data Availability Statement

The data supporting the findings of this study are generated through simulation and are available from the corresponding author upon reasonable request. The code used to generate the simulation data and train the sequence-learning model can also be provided for research purposes.

Disclaimer

The contents of this paper reflect the views of the authors, who are responsible for the facts and accuracy of the information presented. The contents do not necessarily reflect the official views or policies of MarTREC, the U.S. Department of Transportation, or any affiliated institutions. This paper does not constitute a standard, specification, or regulation.

References

[1] Tang L, Boubitsas D, Huang L. Long-term performance of reinforced concrete under a de-icing road environment. *Cem Concr Res*. 2023;164:107039. doi:10.1016/j.cemconres.2022.107039.

[2] Khani S, Conciatori D, Chouinard L, et al. Chloride ingress in de-icing salt-exposed bridge: numerical modeling and field investigations. *Case Stud Constr Mater*. 2025:e05003. doi:10.1016/j.cscm.2024.e05003.

[3] Angst UM, Rossi E, Boschmann Käthler C, et al. Chloride-induced corrosion of steel in concrete. *Mater Struct*. 2024;57:56. doi:10.1617/s11527-024-02337-7.

[4] Liu J, Ou G, Qiu Q, Xing F, Tang K, Zeng J. Atmospheric chloride deposition in field concrete at coastal region. *Constr Build Mater*. 2018;190:1015–1022. doi:10.1016/j.conbuildmat.2018.09.118.

[5] Meira GR, Pinto WTA, Lima EEP, Andrade C. Vertical distribution of marine aerosol salinity in coastal area. *Constr Build Mater*. 2017;135:287–296. doi:10.1016/j.conbuildmat.2016.12.156.

[6] Meira GR, Ferreira PR, Andrade C. Long-term chloride accumulation on concrete surface in marine atmosphere zone. *Corros Mater Degrad*. 2022;3:349–362. doi:10.3390/cmd3030019.

[7] Angst UM, Geiker MR, Michel A, et al. The steel-concrete interface. *Mater Struct*. 2017;50:143. doi:10.1617/s11527-016-0907-5.

[8] Chatterji S. On the applicability of Fick's second law to chloride ion migration through Portland cement concrete. *Cem Concr Res*. 1995;25:299–303. doi:10.1016/0008-8846(94)00166-P.

[9] Zhang T, Gjørv OE. Diffusion behavior of chloride ions in concrete. *Cem Concr Res*. 1996;26:907–917. doi:10.1016/0008-8846(96)00073-5.

[10] Geiker M, Danner T, Polder R, Antonsen R, Hornbostel K. Long-term phase changes in cathodically protected marine reinforced concrete bridge. *Mater Struct*. 2025;58:168. doi:10.1617/s11527-024-02345-7.

[11] Konecny P, Lehner P. Effect of cracking and randomness of inputs on corrosion initiation of reinforced concrete bridge decks exposed to chlorides. *Fract Struct Integr*. 2017;11:29–37. doi:10.3221/IGF-ESIS.39.04.

[12] Zheng W, Qian F, Shen J, Xiao F. Probabilistic-based machine learning for bridge scour detection. *J Civ Struct Health Monit*. 2020;10:957–972. doi:10.1007/s13349-020-00424-5.

[13] Korec E, Jirásek M, Wong HS, Martínez-Pañeda E. Phase-field chemo-mechanical modelling of corrosion-induced cracking. *Theor Appl Fract Mech*. 2024;129:104233. doi:10.1016/j.tafmec.2023.104233.

[14] Vu KAT, Stewart MG. Structural reliability of concrete bridges including improved chloride-induced corrosion models. *Struct Saf*. 2000;22:313–333. doi:10.1016/S0167-4730(00)00018-7.

[15] Stewart MG, Mullard JA. Spatial time-dependent reliability analysis of corrosion damage and timing of first repair. *Eng Struct*. 2007;29:1457–1464. doi:10.1016/j.engstruct.2006.07.004.

[16] Li J, Guo X, Zhang X, Wu Z. Time-dependent reliability assessment and optimal design of corroded reinforced concrete beams. *Adv Struct Eng*. 2024;27:1313–1327. doi:10.1177/13694332231123456.

[17] Boonintra N, Tapsuphonkul K, Pheinsusom P, et al. Reliability assessment of RC bridge piers. *Case Stud Constr Mater*. 2025;23:e05610. doi:10.1016/j.cscm.2025.e05610.
[18] Wang B, Chen K, Wang B. Hierarchical Bayesian fusion of inspection and monitoring data for probabilistic bridge deterioration assessment. *Sci Rep*. 2026;16:5965. doi:10.1038/s41598-026-36808-4.
[19] Schoefs F, Awa Zahui Raissa K, Bonnet S, O'conor AJ. Uncertainty quantification of semi-destructive testing for chloride content assessment. *Front Built Environ*. 2023;9:1130066. doi:10.3389/fbuil.2023.1130066.
[20] Skrzypczak I, Halicka A, Słowik M. Modeling of chloride-induced corrosion in concrete bridge using simplified and full probabilistic methods. *Inż Bezp Obiektów Antropogenicznych*. 2023:35–45. doi:10.37105/iboa.194.
[21] Yehia A, Swei O. Probabilistic infrastructure performance models. *Transp Res Part C*. 2020;111:245–254. doi:10.1016/j.trc.2019.12.013.
[22] Qi L, Peng X, Yang Q, Xia K, Xu B. Prediction models for residual life of concrete structures. *Coatings*. 2025;15:693. doi:10.3390/coatings15060693.
[23] Bouteiller V, Marie-Victoire E, Bonnet A, et al. Reinforced concretes of tomorrow: corrosion behaviour according to exposure classes. *Proc DBMC*. 2023. doi:10.23967/c.dbmc.2023.027.
[24] Shen Y, Goodall JL, Chase SB. Condition state-based deterioration model at structure system level. *J Infrastruct Syst*. 2019;25:04018042. doi:10.1061/(ASCE)IS.1943-555X.0000482.
[25] Ibrahim A, Abdelkhalek S, Zayed T, Qureshi AH, Abdelkader EM. Key deterioration factors of concrete bridge decks: a review. *Buildings*. 2024;14:3425. doi:10.3390/buildings14113425.
[26] Anwar GA, Akber MZ, Ahmed HA, et al. Life-cycle performance modeling for sustainable and resilient structures. *Buildings*. 2024;14:3053. doi:10.3390/buildings14103053.
[27] Han D, Lee JH, Park KT. Deterioration models for bridge pavement materials for life cycle cost analysis. *Sustainability*. 2022;14:11435. doi:10.3390/su141811435.
[28] Medina PA, Gonzalez FJL, Todisco L. Data-driven prediction of long-term deterioration of RC bridges. *Constr Build Mater*. 2022;317:125790. doi:10.1016/j.conbuildmat.2021.125790.
[29] Althaqafi E, Chou E. Developing bridge deterioration models using artificial neural networks. *Infrastructures*. 2022;7:101. doi:10.3390/infrastructures7080101.
[30] Mia MM, Kameshwar S. Machine learning approach for predicting bridge component condition ratings. *Front Built Environ*. 2023;9:1254269. doi:10.3389/fbuil.2023.1254269.
[31] Chen CL, Hung CC, Zhang WY. Severe deterioration and maintenance rules of bridge components. *Case Stud Constr Mater*. 2025;22:e04907. doi:10.1016/j.cscm.2025.e04907.
[32] Zhang H, Marsh DWR. Multi-state deterioration prediction for infrastructure asset. *Inf Sci*. 2020;529:197–213. doi:10.1016/j.ins.2020.04.021.
[33] Faris N, Zayed T, Fares A. Review of condition rating and deterioration modeling approaches. *Buildings*. 2025;15:219. doi:10.3390/buildings15020219.
[34] Salinas D, Flunkert V, Gasthaus J, Januschowski T. DeepAR: probabilistic forecasting with autoregressive recurrent networks. *Int J Forecast*. 2020;36:1181–1191. doi:10.1016/j.ijforecast.2019.07.001.
[35] Gasthaus J, Benidis K, Wang Y, et al. Probabilistic forecasting with spline quantile function RNNs. *Proc AISTATS*. 2019.
[36] Lim B, Arik SO, Loeff N, Pfister T. Temporal fusion transformers for interpretable multi-horizon forecasting. *Int J Forecast*. 2021;37:1748–1764. doi:10.1016/j.ijforecast.2021.03.036.
[37] Carrara F, Falchi F, Girardi M, Messina N, Padovani C, Pellegrini D. Deep learning for structural health monitoring: an application to heritage structures. arXiv:2211.10351. 2022.
[38] Gao L, Din Z, Kim K, Senouci A. Time series transformer-based modeling of pavement skid and texture deterioration. arXiv:2507.01842. 2025.
[39] Tong S, Wu D, Liu X, Zheng L, Du Y, Zou D. STGAN: spatial-temporal Graph Autoregression Network for pavement distress deterioration prediction. arXiv:2503.01152. 2025.
[40] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9:1735–1780.
[41] Cho K, van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078. 2014.
[42] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans Neural Netw*. 1994;5:157–166. doi:10.1109/72.279181.
[43] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:5998–6008.
[44] Wu RJ, Xia Y, Xia J. Long-term prediction of surface chloride content using physics-informed neural networks. *Eng Struct*. 2025;329:119752. doi:10.1016/j.engstruct.2024.119752.
[45] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res*. 2014;15(56):1929–1958.
[46] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. *Proc 33rd Int Conf Mach Learn. PMLR*. 2016;48:1050–1059.