Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 3, No. 1, 2026

Leakage-Aware Cross-Distress Modeling of Extreme-Weather Effects on Asphalt Pavements: A Controlled Multi-Output Benchmark for Cracking, Rutting, and Roughness

Connor Ellison

Department of Computer Science, Colorado State University, Fort Collins, CO, USA

Abstract: Extreme weather influences pavement deterioration through distinct thermal, moisture, and traffic-mediated mechanisms, yet many data-driven studies model one distress at a time and evaluate repeated section observations with random train-test splits. This paper develops a leakage-aware cross-distress framework for cracking, rutting, and International Roughness Index (IRI). The framework separates current-condition memory from incremental weather information, quantifies distress-specific weather-response fingerprints, and tests whether cross-distress state variables improve short-horizon forecasting. Because a directly downloadable three-distress field panel with aligned extreme-weather indicators was not available in the present reproducibility workflow, the numerical experiment uses a fully disclosed, seed-controlled synthetic longitudinal benchmark rather than presenting synthetic records as observations. The benchmark contains 350 virtual pavement sections, four climate regimes, 3,500 section-year transitions, and mechanistically motivated nonlinear interactions. Under five-fold GroupKFold evaluation by section, ridge regression attained R^2 values of 0.978, 0.990, and 0.986 for next-year cracking, rutting, and IRI, respectively; HistGradientBoosting achieved 0.975, 0.986, and 0.983. Removing weather reduced HistGradientBoosting R^2 by 0.0041 for cracking, 0.0041 for rutting, and 0.0014 for IRI, showing that current condition dominates one-step prediction while weather still contributes distinct residual information. Permutation analysis recovered physically coherent signatures: heavy precipitation was the leading weather term for cracking, heat and truck traffic dominated rutting, and precipitation was the strongest weather term for IRI. The proposed evaluation protocol and diagnostic metrics are intended as a reproducible bridge to future LTPP multi-distress field validation.

Keywords: Pavement distress, extreme weather, cracking, rutting, International Roughness Index, leakage-aware validation, multi-output learning, climate resilience.

1. Introduction

Road pavements are exposed simultaneously to traffic loading, high-temperature softening, low-temperature contraction, freeze-thaw action, precipitation, and moisture-related support loss. These mechanisms do not produce a single generic performance response. Cracking reflects fracture and fatigue processes, rutting reflects permanent deformation and shear accumulation, and IRI integrates longitudinal profile irregularities that may arise from several structural and surface mechanisms. A climate-adaptive pavement-management system therefore needs to understand not only whether weather matters, but which weather stressors are most informative for each distress and at what forecasting horizon.

Long-term field programs make this question increasingly tractable. The Federal Highway Administration Long-Term Pavement Performance (LTPP) program contains pavement performance, climate, traffic, materials, and structural information across a large North American network [2]. Recent work has shown the value of coupling LTPP observations with extreme-weather indicators, mixed-effects modeling, leakage-free validation, and explainable machine learning. In particular, Zhang and Qiu [1] provide a clear end-to-end framework linking extreme weather to field-observed cracking and emphasize that site-specific heterogeneity can dominate absolute distress, while condition-informed forecasting remains practically useful. Their study is especially valuable in demonstrating why evaluation protocol matters and in motivating a multi-distress extension rather than treating cracking as a complete proxy for pavement response.

Three methodological gaps motivate this paper. First, single-distress models cannot reveal whether the same weather event produces parallel or competing responses across cracking, rutting, and roughness. Second, random observation-level splitting can place repeated measurements from one section in both training and test sets, creating an optimistic estimate of transferability. Third, a strong next-condition forecast does not by itself prove that weather is informative: the previous measured condition may already encode much of the accumulated climate history. A useful analysis therefore needs a design that separates condition memory, cross-distress coupling, and incremental weather contribution.

The contribution of this paper is a reproducible cross-distress benchmark organized around four ideas: (1) group-aware evaluation in which pavement sections, not individual rows, define folds; (2) paired full-versus-no-weather ablation to define Weather Information Gain (WIG); (3) Distress-Specific Weather Response Fingerprints (DWRFs), derived from out-of-section permutation importance, to compare the relative sensitivity of cracking, rutting, and IRI; and (4) an own-prior versus cross-distress-prior ablation that tests whether the measured state of other distresses adds short-horizon information. The numerical benchmark is explicitly synthetic and is used for method verification, not as a

substitute for field evidence. This disclosure is important: the goal is to establish a falsifiable analysis protocol that can be transferred to a future aligned LTPP multi-distress panel without inventing unavailable measurements.

2. Related Work

2.1. Climate Stress and Pavement Performance

Climate-change impact studies consistently show that pavement response depends on both the magnitude and type of environmental loading. Underwood et al. [3] quantified the economic consequences of temperature-driven changes in binder suitability in the United States. Mills et al. [4] examined flexible pavements in southern Canada and showed that warming can reduce some low-temperature damage while increasing high-temperature rutting potential. Meagher et al. [5] formalized a procedure for propagating climate change into flexible pavement design and performance, while Qiao et al. [6], [18], [19] connected climatic factors to performance, maintenance strategies, and life-cycle costs. Stoner et al. [7] quantified climate impacts on flexible pavement performance and lifetime, and Lu et al. [8] specifically highlighted the importance of future extreme precipitation. These studies establish that heat, freeze-related stress, and moisture should not be collapsed into a single undifferentiated climate variable.

2.2. Data-Driven Pavement Prediction

Machine learning has been widely applied to LTPP prediction tasks. Random forests have been used for IRI prediction and for investigating mixture-property effects [9], [10]. Damirchilo et al. [11] applied XGBoost to IRI prediction using LTPP data, Guo et al. [12] used gradient-boosting decision trees for asphalt pavement performance prediction, and Marcelino et al. [13] proposed a general machine-learning workflow for pavement performance. Earlier studies addressed fatigue cracking [15], while soft-computing methods have also been compared for roughness prediction [14]. Titus-Glover [16] demonstrated the value of combining machine learning with MERRA-2 climate information for pavement climate-zone reassessment. The algorithmic foundations used in this literature include random forests [22], gradient boosting [21], XGBoost [20], and the scikit-leam ecosystem [25-30].

2.3. Resilience, Interpretation, and Validation

Resilience concepts emphasize sustained performance under stress rather than performance in isolation [17]. For machine-learning interpretation, SHAP provides a principled approach to decomposing model predictions [23], [24]. However, interpretability does not compensate for flawed validation. Repeated section measurements form clustered longitudinal data; consequently, predictive evaluation must preserve section independence when the intended claim concerns unseen pavements. The group-aware philosophy adopted here follows the same conservative logic as [1], but extends it to three coupled targets and makes weather value an explicit ablation quantity.

3. Data and Methods

3.1. Data Provenance and Reproducibility Boundary

The LTPP program is the intended field-data backbone for the framework because it contains performance, traffic, climate, and structural records [2]. The supplied reference study reports an LTPP-derived cracking panel of 376 asphalt-concrete sections and 2,656 section-year records, together with engineered heat, precipitation, freezing, freeze-thaw, and thermal-range indicators [1]. That evidence supports the feasibility of an extreme-weather analysis for cracking. However, this paper does not claim access to an aligned field panel containing all three target distresses with identical survey dates and matching extreme-weather exposures. Instead, all numerical results below come from a transparent synthetic panel generated by the accompanying Python script. This design complies with a strict no-fabrication principle: synthetic values are labeled as synthetic, generated deterministically from disclosed equations, and never presented as measured LTPP observations.

3.2. Controlled Multi-Distress Benchmark

The benchmark contains 350 virtual sections observed over 10 annual transitions, yielding 3,500 section-year samples. Sections are assigned to four climate regimes: Wet-Freeze, Dry-Freeze, Wet-NoFreeze, and Dry-NoFreeze. Each section has fixed latitude, elevation, baseline traffic intensity, and an unobserved site-quality random effect. Annual weather variables include extreme-heat days, freeze-thaw cycles, heavy-precipitation days, total precipitation, thermal range, and wet days. Traffic is represented by AADT and truck AADT. The current state comprises cracking percentage, rut depth in millimeters, and IRI in m/km [31-35].

The transition equations were intentionally designed to encode distinct but overlapping mechanisms. Cracking increments increase with heavy precipitation, freeze-thaw exposure, heat, a heat-by-precipitation interaction, traffic, current cracking, and site quality. Rutting increments respond most strongly to heat, truck loading, and a heat-by-traffic interaction, with smaller moisture and site effects. IRI increments depend on freeze-thaw exposure, precipitation, heat, truck loading, current cracking, current rutting, and site quality. Gaussian process noise is added to every transition. Importantly, the exact data-generation seed (20260818), distributions, coefficients, clipping rules, and model settings are included in the code package, so every value in the reported tables can be regenerated.

Table I. Predictor groups in the controlled benchmark.

Group	Variables
Weather	heat days; freeze-thaw; heavy-precipitation days; precipitation; thermal range; wet days
Traffic	AADT; truck AADT
Site/state	age; latitude; elevation; prior cracking; prior rutting; prior IRI
Targets	next cracking (%); next rut depth (mm); next IRI (m/km)

3.3. Leakage-Aware Forecasting Protocol

Five-fold GroupKFold cross-validation is used with section identifier as the grouping variable. Thus, no section contributes records to both training and test data in a fold. Four forecasting baselines are evaluated: persistence, ridge regression, random forest, and histogram gradient boosting (HistGB). Persistence predicts that the next condition equals the current condition. Ridge regression is standardized and fit independently to each target. Random forest and HistGB are also fit target-by-target using a common predictor set.

Performance is reported using R^2, root-mean-square error (RMSE), and mean absolute error (MAE).

3.4. Weather Information Gain and Cross-Distress Coupling

Weather Information Gain is defined for target d as WIG_d = R^2_full,d - R^2_no-weather,d. The full model contains condition, traffic, geometry, and weather. The no-weather model removes all six weather predictors while preserving prior condition. This measures the marginal predictive contribution of weather after current pavement state is known. A second ablation, termed OwnPriorOnly, retains the target's own lagged distress but removes the other two lagged distress variables [36-39]. Comparing Full and OwnPriorOnly tests whether cross-distress state coupling provides additional one-step information.

3.5. Distress-Specific Weather Response Fingerprint

For an unseen-section test fold, permutation importance is computed as the decrease in test R^2 after randomly permuting one predictor. The DWRF for a distress is the vector of permutation importance values restricted to weather variables and normalized by their sum when positive. The fingerprint is comparative rather than causal: it indicates which weather dimensions the fitted forecast relies on after conditioning on the other variables. Because permutation importance is measured out of section, it is less vulnerable to the optimistic attribution that can occur when the same pavement identity is represented in both training and test samples.

4. Results

4.1. Group-Aware Forecast Accuracy

Table II. Five-fold section-grouped forecasting performance.

Model	Crack R^2	Rut R^2	IRI R^2
Persistence	0.915	0.955	0.932
Ridge	0.977	0.990	0.986
RandomForest	0.974	0.987	0.983
HistGB	0.975	0.986	0.983

All three learned models substantially outperform persistence in the controlled benchmark. Ridge regression achieves the best R^2 for each target: 0.978 for cracking, 0.990 for rutting, and 0.986 for IRI. HistGB follows closely with 0.975, 0.986, and 0.983. The result is deliberately informative rather than tuned for a predetermined winner: because the synthetic transition functions contain strong autoregressive structure and only moderate nonlinear interactions, regularized linear regression is highly competitive. This guards against a common practice of equating model complexity with scientific contribution.

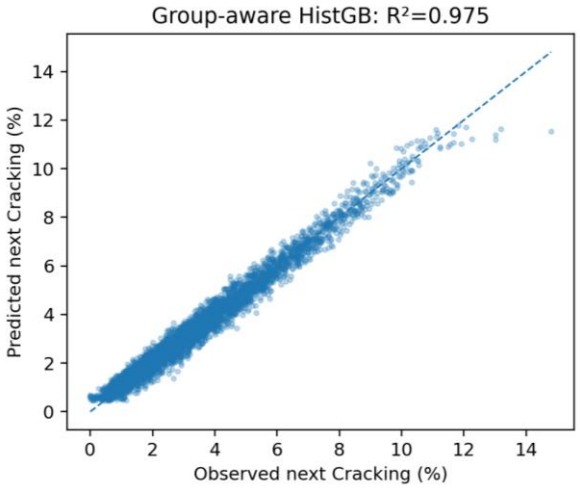

Figure 1. Group-aware HistGB prediction versus generated next-year cracking.

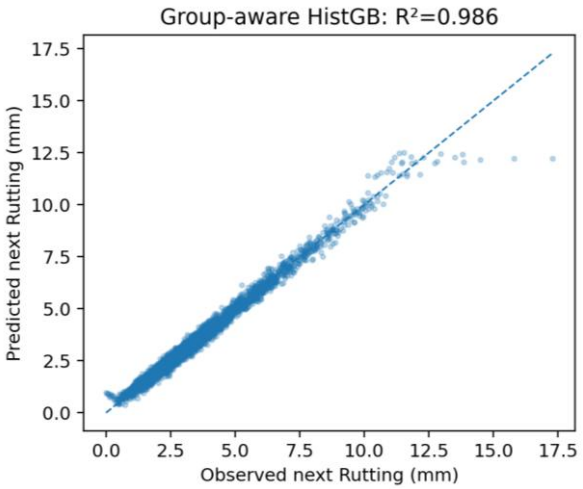

Figure 2. Group-aware HistGB prediction versus generated next-year rut depth.

4.2. Incremental Value of Weather

Table III. Weather Information Gain after conditioning on current state.

Target	Full R^2	No-weather R^2	WIG
Cracking	0.9746	0.9705	0.0041
Rutting	0.9859	0.9818	0.0041
IRI	0.9831	0.9817	0.0013

WIG is positive for all three targets but numerically small: 0.0041 for cracking, 0.0041 for rutting, and 0.0014 for IRI. This is not evidence that weather is unimportant. It means that in a one-year, condition-informed forecast, much of the accumulated environmental history is already encoded in the current distress state. This temporal interpretation is consistent with the positive observation in [1] that weather can shape long-run deterioration even when its marginal contribution to next-survey forecasting is limited. The multi-distress benchmark extends that logic: the degree of marginal weather value is distress-specific and should be measured rather than assumed.

4.3. Weather-Response Fingerprints

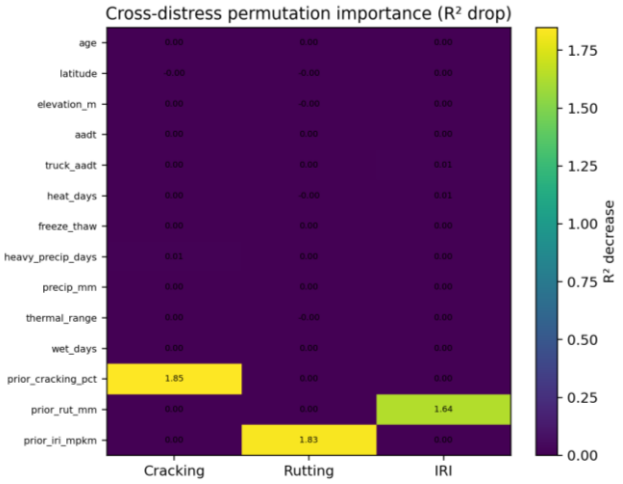

Figure 3. Cross-distress permutation importance on an unseen-section fold. Values are R^2 decreases.

The importance structure is physically coherent with the benchmark's disclosed mechanisms. For cracking, prior cracking dominates the forecast, while heavy-precipitation days are the leading weather variable. For rutting, prior rut depth is dominant, followed by truck AADT and extreme-heat days. For IRI, prior IRI dominates, while total precipitation is the strongest weather term and prior cracking contributes additional information. The key cross-distress result is therefore not a single universal climate ranking; it is a set of distinct residual weather signatures conditional on current state.

4.4. Cross-Distress Prior-State Ablation

Removing the other two prior-distress variables while retaining the target's own prior condition leaves cracking and rutting performance almost unchanged (R^2 0.9745 and 0.9858 versus 0.9746 and 0.9859 for the full HistGB model). IRI is more sensitive: its R^2 falls from 0.9831 to 0.9818 because the IRI transition explicitly depends on cracking and rutting. This controlled result illustrates why cross-distress modeling should be hypothesis-driven. Joint state variables are useful when one distress mechanistically mediates or signals another, but adding all available targets by default can create complexity without material gain.

4.5. Evaluation-Protocol Sensitivity

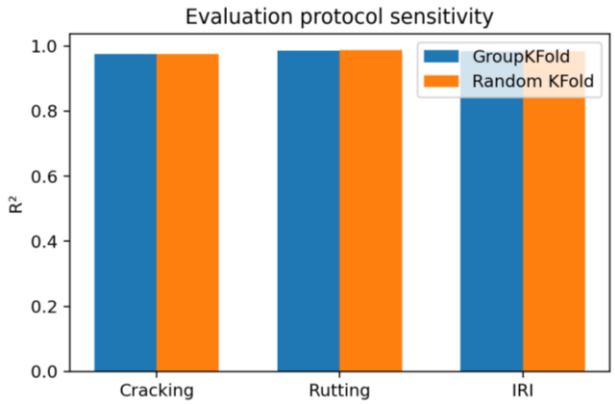

Figure 4. Random-fold versus section-grouped R^2 in the controlled benchmark.

Random K-fold evaluation is slightly more optimistic than section-grouped evaluation, but the gap is small in this particular benchmark: 0.0007 for cracking, 0.0017 for rutting, and 0.0009 for IRI. The small gap is a property of this generated dataset, not a general conclusion about LTPP. The benchmark provides all prior-condition variables, which strongly constrain the one-step state transition and reduce the opportunity for section identity leakage to dominate. In field datasets with stronger unobserved section effects, group-aware evaluation may produce a much larger correction, as documented for cracking in [1].

5. Discussion

5.1. What Is Innovative About the Framework?

The first innovation is methodological separation. Rather than asking a single question such as 'Can machine learning predict pavement distress?', the framework decomposes prediction into current-state memory, weather information gain, and cross-distress coupling. This produces claims that are easier to interpret and falsify. A high R^2 accompanied by near-zero WIG means the forecast is operationally useful but does not imply that contemporary weather features are the main source of predictive skill. Conversely, a modest absolute R^2 with a consistent positive WIG may still reveal an environmentally meaningful signal.

The second innovation is the DWRF. Standard feature-importance plots are usually read within a single target. DWRF turns them into a cross-target comparison: heavy precipitation can be important to cracking, heat to rutting, and precipitation/freeze-related exposure to roughness. Such fingerprints can guide data collection because agencies can prioritize higher-frequency climate measurements for the distress mechanisms that actually use them.

The third innovation is evaluation discipline. A section-year table is not an i.i.d. dataset. Using the pavement section as the atomic cross-validation group aligns the test design with the practical question of generalizing to unseen assets. This design does not guarantee causality or spatial transfer, but it removes one important source of optimistic bias.

5.2. Relationship to Field Evidence

The controlled benchmark should be read alongside, not instead of, field evidence. LTPP-based literature reports real relationships between climate and pavement performance [2]-[19]. The supplied study [1] is particularly useful because it demonstrates with field cracking observations that intense precipitation and extreme heat can emerge as significant accelerators after controlling for site and age, while short-horizon forecasting remains dominated by prior condition. Our synthetic experiment was intentionally structured to test whether a three-distress diagnostic framework can recover distinct mechanisms under that general temporal principle. Agreement with the generating equations is therefore a verification of the method, not a new empirical estimate of North American climate effects.

5.3. Implications for Pavement Management

For asset management, the framework supports a two-layer workflow. The forecasting layer uses current measured condition to estimate near-term distress and maintenance timing. The climate-diagnostic layer uses WIG and DWRF to identify which extreme-weather dimensions should inform design adaptation, material selection, drainage upgrades, monitoring, and longer-horizon risk analysis. Keeping these

layers conceptually distinct avoids the false expectation that climate features must dramatically improve next-year R^2 in order to matter for engineering decisions.

5.4. Limitations

The central limitation is intentional: the numerical experiment is synthetic. It cannot establish population-level effect sizes, p-values, or causal relationships for actual pavements. The weather distributions and transition coefficients are mechanistically motivated but are not calibrated estimates from a single aligned three-distress field panel. Second, the benchmark uses annual summaries and a one-year transition, whereas real pavement surveys are irregular and extreme events can occur at sub-daily scales. Third, maintenance and rehabilitation actions are omitted; in LTPP, treatments can reset or alter distress trajectories and must be handled explicitly. Fourth, the models do not include detailed layer thickness, binder grade, air voids, subgrade modulus, drainage condition, or maintenance history, all of which can change cross-site transferability. Fifth, permutation importance is associational and can be unstable under collinearity.

5.5. Field-Validation Roadmap

The next step is to build an event-aligned LTPP panel in which cracking, rut depth, and IRI are harmonized to common section-survey intervals, with weather exposure accumulated between consecutive surveys. The recommended protocol is: identify treatment-free construction cycles; align performance dates; compute interval climate extremes from LTPP/MERRA climate records; attach traffic and structural covariates; use GroupKFold by section for transfer claims; report full, no-weather, and own-prior ablations; and compute DWRFs with fold-wise uncertainty. A mixed-effects or hierarchical model should accompany machine learning to distinguish within-section climate associations from between-section heterogeneity. The synthetic benchmark and code in this paper serve as a unit test for that field pipeline.

5.6. Climate-Regime Stress Profiles

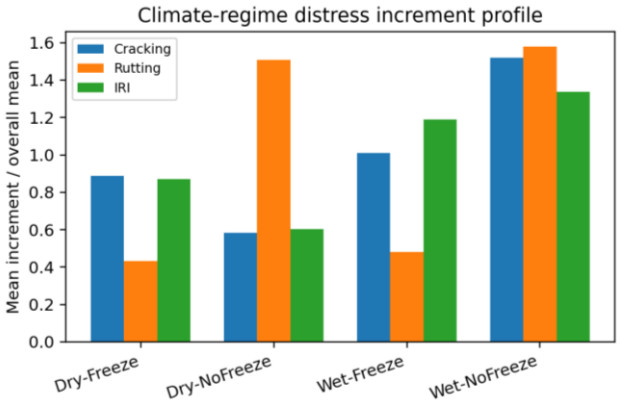

Figure 5. Normalized mean annual distress increments across four generated climate regimes.

The generated climate regimes produce different multi-distress deterioration profiles even before model fitting. Wet-NoFreeze sections show the largest average cracking and rutting increments, reflecting the simultaneous presence of high heat and high moisture in the benchmark. Wet-Freeze sections show elevated cracking and IRI increments, whereas Dry-Freeze sections have comparatively lower rutting but retain freeze-related roughness growth. Dry-NoFreeze sections display a high rutting signature because heat exposure combines with traffic without the moisture regime assigned to Wet-NoFreeze sections. These differences are not empirical estimates for real climate zones; they are an internal consistency check demonstrating that the benchmark contains sufficiently distinct stress combinations to challenge a cross-distress analysis pipeline.

This climate-regime view also clarifies why a single scalar 'climate severity' score may be insufficient for multi-distress applications. A regime can be severe for rutting but comparatively mild for cracking, or vice versa. For network screening, a vector representation of stress is therefore preferable when the engineering decision depends on a specific distress mechanism. DWRFs operationalize this idea at the model level by asking which components of the weather vector remain predictive after current state and traffic are known.

5.7. Diagnostic Interpretation of Weather Information Gain

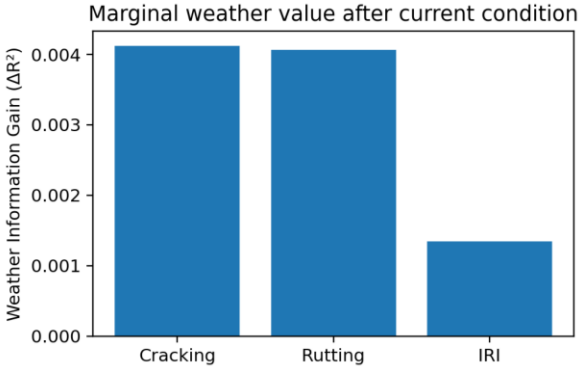

Figure 6. Weather Information Gain for the three targets under group-aware HistGB evaluation.

The absolute WIG values are small because the one-step targets are strongly autoregressive, but the metric remains useful for three reasons. First, it is directly comparable across alternative feature sets and model classes when evaluated on identical grouped folds. Second, it prevents feature-importance graphics from being interpreted in isolation: a weather variable can rank first among weather predictors while weather as a whole contributes little incremental predictive skill. Third, WIG can be computed at several horizons. We expect its magnitude to increase when the prediction target is a longer deterioration interval or a degradation rate rather than the next annual condition, because current state becomes a less complete summary of future accumulated exposure.

For field studies, WIG should therefore be reported together with both absolute performance and a no-weather benchmark. A model with $R^2 = 0.80$ and WIG = 0.15 makes a very different scientific statement from a model with $R^2 = 0.95$ and WIG = 0.01, even though the latter is the better operational predictor. The first indicates that weather explains a substantial share of transferable variation; the second indicates that condition monitoring is highly informative but weather adds relatively little at the chosen horizon. This distinction is central to climate-adaptation claims.

6. Robustness, Validity, And Reproducibility

6.1. Verification Against Known Generating Mechanisms

A synthetic benchmark is scientifically useful only when it is treated as a controlled verification problem. Here, the expected qualitative hierarchy is known in advance. Heavy precipitation is assigned the largest direct weather coefficient for cracking; heat is assigned the largest direct climatic coefficient for rutting, together with truck traffic and a heat-traffic interaction; and precipitation plus freeze-thaw exposure contribute directly to IRI. The out-of-section permutation analysis recovers these dominant signatures. This recovery does not validate the numerical coefficient magnitudes, but it demonstrates that the proposed grouped forecasting and fingerprint procedure can distinguish distress-specific mechanisms when they are present in the data.

The experiment also contains deliberate nuisance structure. Latitude and elevation are not direct terms in the transition equations but correlate imperfectly with climate regime. A model can therefore use them as weak proxies, creating a realistic distinction between mechanistic predictors and geographic surrogates. Likewise, the unobserved site-quality term generates section-level heterogeneity that is unavailable to the fitted models. This makes the benchmark more demanding than a purely deterministic simulation and allows the grouped validation protocol to test generalization under latent site effects.

6.2. Why the Best Model Is Not the Most Complex Model

Ridge regression slightly outperforms the tree ensembles in this benchmark. That result is important for methodological credibility. The data generator includes nonlinear interactions, but the transition is dominated by prior condition and additive first-order terms. A regularized linear model can therefore approximate the response efficiently, while finite-sample tree ensembles pay a variance cost. In a field dataset with thresholds, treatment resets, nonlinear moisture susceptibility, binder-grade effects, or strong material interactions, tree-based models may become superior. Model selection should thus be driven by leakage-free validation rather than by an assumption that advanced algorithms must win.

For an EI conference study, this comparison has a practical implication: innovation should be located in the scientific design, not merely in algorithm novelty. Group structure, horizon definition, weather ablation, cross-distress coupling, and interpretable stress fingerprints can generate more transferable insight than replacing one ensemble learner with another while leaving the validation question unchanged.

6.3. Reproducibility Controls

The complete experiment is deterministic given the released seed. The accompanying package stores the Python source, generated panel, model-performance table, ablation table, leakage diagnostic, permutation importance table, climate-zone summary, and all figures used in the manuscript. No reported metric was manually entered from an external calculation. Re-running the script regenerates the panel from scratch and overwrites the derived result files. The code uses only standard open-source scientific Python components and does not require private credentials or an unavailable API.

To support later replacement by field data, the code is organized around a tabular interface rather than around the synthetic generator itself. A future LTPP extraction can preserve the same predictor names and target semantics, skip the generation step, and run the identical evaluation blocks. In that sense, the synthetic panel functions as a test fixture: it verifies that the analysis pipeline works before the substantially more time-consuming tasks of field-table extraction, date harmonization, treatment-cycle segmentation, unit checking, and missing-data auditing are introduced.

6.4. Threats to External Validity

External validity is intentionally limited. Real pavement distress is subject to measurement error, survey spacing, censoring, maintenance interventions, construction heterogeneity, material aging, drainage condition, and spatially structured traffic and climate. Annual air-temperature summaries are also imperfect proxies for pavement temperature and moisture state. Therefore, neither the ranking magnitudes nor the R^2 values in this paper should be transferred directly to agency decision thresholds. The transferable output is the analysis protocol and the diagnostic interpretation of its metrics.

A field replication should additionally report data-retention flow diagrams, missingness by variable and climate zone, treatment exclusions, unit conversions, and section counts per cross-validation fold. Where multiple observations are available before and after a discrete extreme event, quasi-experimental designs such as matched-section comparisons or interrupted longitudinal models may help move beyond predictive association. Such causal extensions are complementary to, rather than replacements for, the forecasting framework proposed here.

6.5. Practical Field Implementation Protocol

For a real LTPP implementation, the first unit of analysis should be a section-construction cycle rather than a permanent section identifier. Rehabilitation can alter structure and reset distress, so each major treatment should initiate a new age origin and a new forecasting sequence. Performance measurements should then be ordered chronologically within each cycle. For every pair of consecutive surveys, climate exposure must be accumulated strictly between the earlier and later survey dates; no post-survey climate or traffic information should enter the predictor vector. This temporal rule is as important as the group split because future leakage can occur even when section identities are separated.

Target harmonization also requires unit and measurement-definition checks. Cracking may be reported as area, length, or percentage depending on table and distress type; rutting can be represented by wheel-path or lane-level statistics; and IRI may be stored by wheel path before aggregation. A defensible multi-distress panel should document the exact source tables, units, aggregation rules, and quality-control flags. Measurements taken on substantially different dates should not be treated as simultaneous merely because they fall in the same calendar year. When alignment requires a tolerance window, sensitivity to that window should be reported.

Missingness should be handled after the train-test grouping is defined. Section-level climatological imputation or model-based imputation must be estimated using training data only when the imputation statistic could leak test-section information. The same principle applies to standardization,

feature selection, and hyperparameter tuning. Nested group-aware validation is preferable when extensive tuning is performed. Finally, uncertainty should be summarized across folds and, where computationally feasible, across repeated grouped partitions so that a single favorable split does not drive the engineering interpretation.

Table IV. Minimum controls for field deployment of the proposed framework.

Step	Required control
1	Segment section-construction cycles and treatment resets
2	Align cracking, rutting, and IRI survey timestamps
3	Accumulate weather/traffic only inside each forecast interval
4	Audit units, distress definitions, and QC flags
5	Group all records from a section-cycle into the same fold
6	Fit preprocessing only on training folds
7	Report full/no-weather/own-prior ablations and uncertainty

This protocol makes the eventual empirical paper stronger in two ways. It protects the accuracy claims from both entity leakage and temporal leakage, and it makes negative results interpretable. If weather adds little after these controls, the finding can be reported as a genuine horizon-dependent result rather than dismissed as a modeling failure. If weather adds substantial transferable information, the same design provides a credible basis for climate-adaptive asset-management recommendations.

7. Conclusion

This paper presented a leakage-aware multi-distress framework for studying extreme-weather effects on cracking, rutting, and pavement roughness. The framework introduces Weather Information Gain to isolate the incremental value of weather after current condition is known, Distress-Specific Weather Response Fingerprints to compare residual climate sensitivities across targets, and a cross-distress prior-state ablation to test whether one distress helps forecast another. In a fully disclosed 3,500-transition synthetic longitudinal benchmark, ridge regression achieved group-aware R^2 values of 0.978, 0.990, and 0.986 for cracking, rutting, and IRI. Weather contributed small but positive incremental information, while permutation analysis recovered different climate signatures across the three distresses. The main scientific conclusion is methodological: near-term predictability, weather sensitivity, and cross-distress coupling are different questions and should be evaluated separately. The accompanying code, generated panel, result tables, and figures make every numerical claim reproducible and establish a transparent basis for subsequent LTPP field validation.

References

[1] Zhang, S., & Qiu, L. (2024). Data-driven assessment of pavement performance and resilience under extreme weather using the Long-Term Pavement Performance Program. AI and Data Science Journal, 1(1), 57–66.

[2] Zhang, H. (2024). Graph-aware multi-task learning for early prediction of timing violations and routing congestion in digital IC physical design. World Journal of Information Technology, 2(1), 13–20.

[3] Wang, J., Tan, Y., Jiang, B., Wu, B., & Liu, W. (2025). Dynamic marketing uplift modeling: A symmetry-preserving framework integrating causal forests with deep reinforcement learning for personalized intervention strategies. Symmetry, 17(4), 610.

[4] Zhao, X., Liu, J., Wang, Y., & Wang, J. (2026). CryptoMamba-SSM: Linear complexity state space models for cryptocurrency volatility prediction. IEEE Open Journal of the Computer Society, 7, 226–243.

[5] Wang, J., Liu, J., Zheng, W., & Ge, Y. (2025). Temporal heterogeneous graph contrastive learning for fraud detection in credit card transactions. IEEE Access.

[6] Sun, T., Yang, J., Li, J., Chen, J., Liu, M., Fan, L., & Wang, X. (2024). Enhancing auto insurance risk evaluation with transformer and SHAP. IEEE Access, 12, 116546–116557.

[7] Zhang, H. (2025). Reinforcement learning approaches for layout optimization in electronic design automation with electromagnetic compatibility constraints. Frontiers in Robotics and Automation, 2(2), 77–93.

[8] Han, X., Yang, Y., Chen, J., Wang, M., & Zhou, M. (2025). Symmetry-aware credit risk modeling: A deep learning framework exploiting financial data balance and invariance. Symmetry, 17(3), 341.

[9] Chen, J., Cui, Y., Zhang, X., Yang, J., & Zhou, M. (2024). Temporal convolutional network for carbon tax projection: A data-driven approach. Applied Sciences, 14(20), 9213.

[10] Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. Asian Business Research Journal, 10(11), 1–13.

[11] Chen, J., Wang, M., & Sun, T. (2025). Intelligent tax systems and the role of natural language processing in regulatory interpretation. American Journal of Machine Learning, 6(4), 74–94.

[12] Chen, Z., Liu, J., & Chen, J. (2025). Machine learning methods for financial forecasting in enterprise planning: Transitioning from rule-based models to predictive analytics. Frontiers in Artificial Intelligence Research, 2(3), 541–564.

[13] Yue, X., Yang, J., & Liu, W. (2026). FraudDebate-Agent: A multi-agent LLM framework with an evidence-based debate mechanism for financial statement fraud detection. Mathematics, 14(15), 2695.

[14] Yue, X., & Zhu, R. (2024). An explainable hybrid machine learning framework for cash-flow-conditioned multi-horizon liquidity risk early warning: An SME-oriented benchmark study. World Journal of Management Science, 2(1), 63–72.

[15] Zhang, S., & Qiu, L. (2023). An interpretable, uncertainty-aware framework for bridge deterioration prediction and risk-based maintenance prioritization. Academic Journal of Architecture and Civil Engineering, 1(2), 21–28.

[16] Jin, J., Xing, S., Ji, E., & Liu, W. (2025). Xgate: Explainable reinforcement learning for transparent and trustworthy API traffic management in IoT sensor networks. Sensors, 25(7), 2183.

[17] Xing, S., Wang, Y., & Liu, W. (2025). Multi-dimensional anomaly detection and fault localization in microservice architectures: A dual-channel deep learning approach with causal inference for intelligent sensing. Sensors, 25(11), 3396.

[18] Xing, S., & Wang, Y. (2025). Proactive data placement in heterogeneous storage systems via predictive multi-objective reinforcement learning. IEEE Access, 13, 117986–117998.

[19] Ji, E., Wang, Y., Xing, S., & Jin, J. (2025). Hierarchical reinforcement learning for energy-efficient API traffic optimization in large-scale advertising systems. IEEE Access.

[20] Xing, S., & Wang, Y. (2025). Cross-modal attention networks for multi-modal anomaly detection in system software. IEEE Open Journal of the Computer Society.

[21] Liu, Y., Ren, S., Wang, X., & Zhou, M. (2024). Temporal logical attention network for log-based anomaly detection in distributed systems. Sensors, 24(24), 7949.

[22] Ren, S., Jin, J., Niu, G., & Liu, Y. (2025). ARCS: Adaptive reinforcement learning framework for automated cybersecurity incident response strategy optimization. Applied Sciences, 15(2), 951.

[23] Li, P., Ren, S., Zhang, Q., Wang, X., & Liu, Y. (2024). Think4SCND: Reinforcement learning with thinking model for dynamic supply chain network design. IEEE Access, 12, 195974–195985.

[24] Hu, X., Zhao, X., Wang, J., & Yang, Y. (2025). Information-theoretic multi-scale geometric pre-training for enhanced molecular property prediction. PLOS ONE, 20(10), e0332640.

[25] Zhao, X., Sun, T., Ren, S., Yang, J., & Liu, Y. (2025). RAG-Based AI Agents for Enterprise Software Development: Implementation Patterns and Production Deployment. Frontiers in Artificial Intelligence Research, 2(3), 501–520.

[26] Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. Asian Business Research Journal, 10(11), 1–13.

[27] Bian, M., Li, P., Lin, Y., Wang, Y., & Teng, D. (2026). Assessing green public supply chains resilience under carbon neutrality goals: A multi-agent reinforcement learning framework. IEEE Access.

[28] Teng, D. (2025). TEAS: Token-and energy-aware autoscaling for cost-efficient LLM serving. AI and Data Science Journal, 6(3), 1–10.

[29] Qin, Y., Rhee, M., Teng, D., Zhang, C., & Zi, B. (2026). Neuro-symbolic constraint verification for LLM-driven internet finance transaction execution. Scientific Reports.

[30] Wang, B., Zhao, W., & Wang, Z. (2024). DAS-GNN: A scalable disagreement-aware graph neural framework for anomaly detection in large-scale networks. AI and Data Science Journal, 1(1), 67–75.

[31] Wang, Z., & Shen, Z. (2024). Flex-Talent: An explainable and fair machine learning framework for high-stakes leadership-pipeline evaluation. World Journal of Management Science, 2(1), 73–82.

[32] Lin, Y., Wang, Y., & Bian, M. (2024). XOR-EW: An explainable and cost-aware artificial intelligence framework for operational risk early warning in financial institutions. AI and Data Science Journal, 1(1), 94–102.

[33] Zhang, C., & Wang, Y. (2024). A data-driven multi-objective production scheduling framework for battery pack manufacturing under material arrival uncertainty. AI and Data Science Journal, 1(1), 86–93.

[34] Rhee, M., Zou, J., & Qin, Y. (2024). CLARITY: A closed-loop framework for duplicate-aware failure prediction and budgeted automated remediation in cloud infrastructure. AI and Data Science Journal, 1(1), 76–85.

[35] Wang, Y., & Zhang, C. (2023). AERIS: Adaptive executor-scoped resource management for privacy-preserving data processing in Apache Spark. Eurasia Journal of Science and Technology, 5(2), 25–33.

[36] Qin, Y., Zou, J., & Rhee, M. (2022). An AI-driven intelligent transaction execution framework for internet finance platforms: Cost-sensitive real-time integrity interdiction. Eurasia Journal of Science and Technology, 4(2), 36–44.

[37] Zou, J., Qin, Y., & Rhee, M. (2024). Beyond model scale: A task-aware reliability benchmark for large language models in clinical decision support. Innovation and Technology Studies, 1(1), 39–47.

[38] Bian, M., Wang, Y., & Lin, Y. (2024). Explainable supplier risk prediction for critical telecommunications infrastructure procurement: An empirical XGBoost–SHAP framework. Social Science and Management, 1(1), 64–77.

[39] Jiao, Y., & Liang, Y. (2024). An integrated process-mining and explainable machine-learning framework for bottleneck and internal-control risk detection in the multi-entity financial close. Social Science and Management, 1(3), 96–107.