Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Feature Fusion and Selective Ensemble Learning for Flood Insurance Pure Premium Stress Testing

Nassamon Bootwisas
Rajamangala University of Technology Suvarnabhumi
Pasin Marupanthorn
quantfilab@gmail.com
Maejo University

Research Article

Keywords: NFIP, flood insurance, pure premium adequacy, ensemble learning, storm-event stress testing

Posted Date: June 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9943063/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Feature Fusion and Selective Ensemble Learning for Flood Insurance Pure Premium Stress Testing

Nassamon Bootwisas[1] and Pasin Marupanthorn[2*]

[1]Computation Science, Science and Technology, Rajamangala University of Technology Suvarnabhumi, Thailand.

[2*]Department of Mathematics, Faculty of Science, Maejo University, 63 Moo 4, Nong Han, San Sai, Chiang Mai, 50290, Thailand.

Abstract

We develop a provenance-controlled public-data benchmark for retrospective stress testing of National Flood Insurance Program (NFIP) pure premiums in a targeted Florida–Texas county-year panel. The study addresses a fundamental data-governance issue in public flood-insurance research: incomplete claims or policy retrievals can bias adequacy analysis when missing files are inadvertently treated as observed zero-loss records. The benchmark therefore imposes filename validation, state–county consistency checks, county FIPS verification, and a complete-history rule before any modeling is conducted. The retained sample consists of high-risk counties with observed policy and claims histories over 2020–2024, and the empirical target is realized county-year pure premium conditional on the corresponding insured portfolio and storm-event summaries. We compare lower-information insurance-only predictors with a depth-aware storm-event-fused selective ensemble that integrates policy exposure, county loss memory, NOAA Storm Events indicators, premium anchoring, signed residual correction, and a structural tail expert that is activated only when sufficient pre-period history is available. In the 2024 stress year, the fused benchmark substantially improves both dollar-scale and logarithmic error metrics. Rolling-origin checks indicate that the fused design outperforms the insurance-only ensemble in every fold, although premium-scaled rules remain competitive in calmer years. The findings should be interpreted as hypothesis-generating stress-test evidence rather than as confirmatory validation of a deployable pricing model. The main contribution is a transparent sample-governance protocol and a leakage-disciplined benchmark that clarifies the opportunities and limitations of public NFIP county-year analysis.

Keywords: NFIP, flood insurance, pure premium adequacy, ensemble learning, storm-event stress testing

1 Introduction

Flood-insurance adequacy lies at the intersection of actuarial loss measurement, public disaster finance, climate adaptation, and insurance-affordability policy. In the United States, this issue is especially salient for the National Flood Insurance Program (NFIP), where premiums are expected to communicate flood risk while maintaining participation in a public insurance program. The policy environment has changed materially since the introduction of Risk Rating 2.0. FEMA reports that the new NFIP pricing approach was fully implemented on April 1, 2023, with the aim of reflecting property-specific flood risk more accurately [1]. The U.S. Government Accountability Office similarly characterizes Risk Rating 2.0 as a response to the persistent tension among affordability, risk-based pricing, and fiscal solvency in the NFIP [2]. Recent empirical evidence indicates that the reform has nontrivial demand consequences: Ortega and Petkov document increased exits and reduced new entry under the new rate-setting regime [3], while Gourevitch, Snyder, and Kousky estimate that larger Risk Rating 2.0 premium increases reduced both new and existing policies-in-force, with stronger effects in lower-income zip codes [4]. These findings imply that empirical work on NFIP premium adequacy must clearly distinguish among loss adequacy, affordability, demand response, solvency, and policy design.

The actuarial challenge is amplified by the changing distribution of flood losses. National flood-risk studies indicate that exposure and expected losses are not stationary under climate change, sea-level rise, and continued development in flood-prone areas. Wing et al. estimate substantial increases in U.S. flood risk by mid-century and emphasize that the burden of future flood risk is unevenly distributed across communities [5]. Poudel et al. use socio-environmental modeling to project future NFIP claims, policy participation, and housing-market trajectories under coastal flooding scenarios [6]. Nayak et al. argue that NFIP insolvency is linked not only to isolated extreme events, but also to catastrophic spatiotemporal clustering and recurrent losses [7]. In parallel, the compound-flooding literature shows that coastal losses may arise from interacting drivers, including storm surge, extreme rainfall, fluvial flow, groundwater, and antecedent conditions, rather than from a single hazard mechanism [8]. These findings motivate stress testing and temporally disciplined evaluation, rather than sole reliance on pooled average-loss models.

A growing empirical literature uses NFIP claims, hazard data, and open socioeconomic information to model flood losses and flood-insurance outcomes. Salas, Saha, and Ravela show that NFIP claims, when combined with rainfall information, can support interannual flood-loss prediction and that machine-learning models can extract useful structure from historical claims and physical covariates [9]. Bhattacharyya and Hastak estimate county-level relationships between flood-insurance payouts and exposure, infrastructure vulnerability, social vulnerability, and resilience indicators [10]. Bowers, Serafin, and Baker demonstrate the value of interpretable machine learning for identifying hazard, exposure, and vulnerability drivers of atmospheric-river flood damage using NFIP claims [11]. Liu and Mostafavi develop the FloodGenome framework to characterize built-environment, hydrological, and topographical features shaping urban property flood-risk disposition [12]. Collectively, these studies show that public flood-loss data can be analytically powerful when combined with hazard

and vulnerability variables. They also highlight a more fundamental methodological requirement: the empirical panel must be defensible before predictive performance can be interpreted.

The actuarial modeling literature provides a second motivation for careful benchmark design. Pure premium is naturally connected to claim frequency and claim severity, and recent work in insurance analytics increasingly combines classical frequency–severity reasoning with modern machine learning and distribution-free uncertainty quantification. Graziadei et al. extend conformal prediction to frequency–severity modeling and emphasize finite-sample predictive guarantees for insurance claims [13]. Hong studies conformal prediction for future insurance claims in regression settings, stressing model-free validity for practical insurance applications [14]. These developments are directly relevant to flood-insurance stress testing because county-year losses are sparse, heavy-tailed, spatially dependent, and sensitive to extreme events. In such settings, a high-capacity model can appear successful while remaining fragile if the training sample is small, temporal availability assumptions are unclear, or tail adjustments are tuned after the extreme rows have been observed.

This paper is motivated by the gap between public data availability and defensible benchmark construction. The core data sources used here—FEMA NFIP redacted claims, FEMA NFIP redacted policies, NOAA Storm Events, and ACS county indicators—are public and widely useful [15–18]. Public availability, however, does not by itself establish sample validity. NOAA Storm Events provides a long official record of severe-weather events, but the database must be interpreted with attention to reporting conventions, damage-estimate quality, and missing or inconsistent records [19]. The same principle applies to public NFIP retrievals: a missing state-year claims archive is not an observed zero-loss year. If incomplete claims files, mismatched state–county identifiers, or legacy non-targeted policy files are silently retained, the resulting pure-premium benchmark can be distorted before any statistical model is fitted.

Accordingly, this paper addresses sample governance before predictive modeling. The primary contribution is a reproducible provenance-control protocol for public NFIP county-year research based on filename validation, state–county FIPS consistency checks, and complete-history restrictions over 2020–2024. The predictive analysis is deliberately narrower. It evaluates whether storm-event summaries and county-history information improve realized county-year pure-premium estimation in a selected high-risk Florida–Texas stress portfolio. The estimand is not national NFIP pricing adequacy; it is realized county-year pure premium for counties that satisfy the paper's complete observed claims and policy history rule.

The task definition is equally limited. The paper studies retrospective, event-conditioned adequacy measurement after year-t policy composition and year-t NOAA Storm Events summaries are observed. It is therefore not a pre-season pricing model, a prospective underwriting rule, a household-level replication of Risk Rating 2.0, or a climate-attribution exercise. At the same time, the contribution is intended to be visible from an AI and ensemble-learning perspective as well as from an actuarial one. Specifically, the manuscript contributes a leakage-controlled public-data benchmark, a multi-source feature-fusion design that combines policy composition, county-loss memory, storm-event summaries, and county context, a depth-aware selective ensemble for

heavy-tailed regression under limited historical support, a signed residual correction mechanism, a structural tail expert for storm-event stress years, and a temporally disciplined rolling-origin validation framework based on expanding-origin train/test splits. These elements define a generalizable ensemble-learning architecture for small-sample, heavy-tailed, multi-source risk modeling problems in which tail behavior and temporal leakage are central concerns. The value of the manuscript lies in clarifying what can be learned from a provenance-controlled public stress-test benchmark, and what cannot yet be claimed without broader samples, locked model specification, stronger actuarial baselines, dependence-aware uncertainty analysis, and external validation.

The remainder of the paper is organized as follows. Section 2 presents the mathematical structure of the exploratory benchmark and defines the leakage-safe notation. Section 3 describes the data architecture and provenance-based sample construction. Section 4 defines the empirical task, feature-availability timing, and model-comparison design. Section 5 reports the stress-year and rolling-origin evidence with a deliberately bounded interpretation. Section 6 concludes by summarizing the inferential scope and the validation work required for stronger actuarial or policy claims.

2 Mathematical Description of the Exploratory Benchmark

This section specifies the mathematical form of the exploratory benchmark used in the empirical analysis. The construction should be read as a leakage-controlled stress-testing architecture rather than as a deployable actuarial tariff. It is governed by three design constraints. First, each evaluation year is predicted using parameters fitted only on earlier years. Second, all county-memory variables are computed from histories that end before the evaluated year. Third, contemporaneous NOAA Storm Events variables are admitted only because the target is retrospective, event-conditioned pure-premium measurement after the storm year has been observed.

2.1 Target, information sets, and historical summaries

For county c and calendar year t, let $L_{c,t}$ denote total paid loss, $N_{c,t}$ denote policy count, and $W_{c,t}$ denote written premium. The county-year response is realized pure premium,

$$y_{c,t} = \frac{L_{c,t}}{N_{c,t}}, \qquad p_{c,t} = \frac{W_{c,t}}{N_{c,t}}, \tag{1}$$

where $p_{c,t}$ is the observed premium per policy used only as a scaling anchor. Let

$$x_{c,t} = \left(x_{c,t}^{\mathrm{pol}}, x_{c,t}^{\mathrm{hist}}, x_{c,t}^{\mathrm{storm}}, x_{c}^{\mathrm{ctx}}\right) \in \mathbb{R}^{p} \tag{2}$$

denote the feature vector at the benchmark evaluation time. The four blocks correspond to policy-composition variables, lagged county-loss memory, year-t storm-event summaries, and static county context. For an evaluation year e, the expanding training and test sets are

$$\mathcal{D}_{\mathrm{train}}(e) = \{(c,t) : t < e\}, \qquad \mathcal{D}_{\mathrm{test}}(e) = \{(c,e)\}. \tag{3}$$

The main stress split is obtained by setting $e = 2024$, so that $\mathcal{D}_{\text{train}}(2024)$ contains 2020–2023 and $\mathcal{D}_{\text{test}}(2024)$ contains only 2024 county-years.

For any county-year quantity $v_{c,t}$, define the one-period lag, expanding maximum, and expanding mean by

$$\ell^{v}_{c,t} = v_{c,t-1}, \qquad m^{v}_{c,t} = \max_{u<t} v_{c,u}, \qquad \bar{v}_{c,t} = \frac{1}{|\{u : u < t\}|} \sum_{u<t} v_{c,u}. \tag{4}$$

These summaries are used for pure premium, claim count, claim rate, paid-loss severity, and related loss-history variables. Equation (4) is the main temporal safeguard. The benchmark may use current-year policy and storm information because the task is event-conditioned, but it never uses future loss history when forming county-memory features.

2.2 Insurance ensemble and premium-scaled anchor

The first layer is an insurance-information ensemble fitted only on policy-composition and pre-period county-history variables. If $\widehat{f}^{(e)}_m$ denotes the mth learner trained on $\mathcal{D}_{\text{train}}(e)$ using the lower-information view $\mathcal{X}^{\text{ins}}_{c,e}$, then

$$\widehat{y}^{\text{ins}}_{c,e} = \sum_{m=1}^{M} \omega_m \widehat{f}^{(e)}_m(x_{c,e}), \qquad \omega_m \geq 0, \qquad \sum_{m=1}^{M} \omega_m = 1. \tag{5}$$

The implemented benchmark uses three learners with equal weights, but equation (5) records the more general form. Because flood losses are heavy-tailed and the number of retained counties is small, this lower-information ensemble is stabilized by a premium-scaled anchor. Let

$$\rho_{\text{train}}(e) = \underset{(i,u)\in\mathcal{D}_{\text{train}}(e)}{\text{median}} \left\{ \frac{y_{i,u}}{p_{i,u}} : p_{i,u} > 0 \right\}, \qquad \widehat{y}^{\text{anc}}_{c,e} = \rho_{\text{train}}(e) p_{c,e}. \tag{6}$$

The anchor is not treated as a theoretically correct premium rule; rather, it is a robust scale control that ties the prediction to the observed premium base. Let $s_{c,e}$ and $r_{c,e}$ denote the storm-shock and recurrence scores defined below, and let

$$u_{c,e} = \psi(s_{c,e}, r_{c,e}) \in [0, 1] \tag{7}$$

be the combined structural-stress index. The anchor is released more strongly as this index rises. Specifically,

$$a_{c,e} = a(u_{c,e}) \in [\underline{\omega}, 1], \qquad b_{c,e} = a_{c,e}\widehat{y}^{\text{ins}}_{c,e} + \big(1 - a_{c,e}\big)\widehat{y}^{\text{anc}}_{c,e}. \tag{8}$$

The quantity $b_{c,e}$ is the anchored central prediction. When structural stress is low, the benchmark remains close to the robust premium-scaled anchor. When structural stress is high, the benchmark places greater weight on the insurance ensemble.

2.3 Event-pressure scores and signed residual correction

The benchmark summarizes two stress mechanisms. Let $d_{c,e}$ denote NOAA storm property damage per policy, and let $f_{c,e}$ denote the flood-related event count. Let $\phi_d, \phi_f, \phi_{pp}, \phi_{cr}$ be monotone rescaling maps into $[0,1]$, and let $\mathrm{clip}_{[0,1]}(z) = \min\{1, \max\{0, z\}\}$. The contemporaneous storm-shock score and historical recurrence-cascade score are

$$s_{c,e} = \mathrm{clip}_{[0,1]}\big(\alpha_s \phi_d(d_{c,e}) + (1-\alpha_s)\phi_f(f_{c,e})\big), \tag{9}$$

$$r_{c,e} = \mathrm{clip}_{[0,1]}\big(\alpha_r \phi_{pp}\big(\max\{m^{pp}_{c,e}, \ell^{pp}_{c,e}\}\big) + (1-\alpha_r)\phi_{cr}\big(\max\{m^{cr}_{c,e}, \ell^{cr}_{c,e}\}\big)\big). \tag{10}$$

The score $s_{c,e}$ measures current event pressure, while $r_{c,e}$ measures whether the county has a recent or historically recurrent loss profile. These scores govern both anchor release and the selective tail logic.

The second prediction layer is a signed residual correction. Let $b^{(-t)}_{c,t}$ denote the expanding-origin version of equation (8), computed for training row (c,t) using only earlier years. The time-safe residual target is

$$\epsilon_{c,t} = y_{c,t} - b^{(-t)}_{c,t}, \qquad (c,t) \in \mathcal{D}_{\mathrm{train}}(e). \tag{11}$$

An expert model fitted on the fused feature view estimates $\widehat{h}_{c,e} \approx \mathbb{E}[\epsilon_{c,e} \mid \mathcal{X}^{\mathrm{fused}}_{c,e}]$. A classification gate estimates the probability that a large correction is needed:

$$z^{\mathrm{res}}_{c,t} = \mathbf{1}\left\{|\epsilon_{c,t}| \geq \eta_\epsilon,\ u_{c,t} \geq \eta_u\right\}, \qquad \widehat{g}_{c,e} = \Pr\left\{z^{\mathrm{res}}_{c,e} = 1 \mid \mathcal{X}^{\mathrm{fused}}_{c,e}\right\}. \tag{12}$$

The short-history conservative predictor is then

$$\widehat{y}^{\mathrm{res}}_{c,e} = b_{c,e} + \widehat{g}_{c,e}\widehat{h}_{c,e}. \tag{13}$$

In contrast to a one-sided uplift rule, equation (13) allows the fused layer to adjust the anchored baseline upward or downward. This signed correction is important for avoiding systematic over-reaction in calmer years.

2.4 History-depth selective tail expert and final predictor

When sufficient pre-period history is available, the benchmark activates a stronger tail expert while restricting that mechanism to the deep-history regime. Let

$$\widehat{y}^{\mathrm{tail}}_{c,e} = \mathcal{T}\big(\mathcal{X}^{\mathrm{fused}}_{c,e}; \theta_e\big) \tag{14}$$

denote the structural tail expert, where $\mathcal{T}$ includes the exceedance-uplift layer and the regime-specific structural floors used to prevent catastrophic underprediction in deep-history stress years. The activation of this stronger expert is controlled by historical depth rather than being imposed in every fold. Let

$$D_e = \mathbf{1}\left\{|\{u : u < e\}| \geq \kappa_H\right\}, \tag{15}$$

Algorithm 1 Leakage-controlled construction of the exploratory benchmark for evaluation year e

1: Construct $\mathcal{D}_{\text{train}}(e)$ and $\mathcal{D}_{\text{test}}(e)$ using equation (3).
2: Compute all lagged and expanding county histories using only $u < e$.
3: Fit the insurance-information ensemble and compute the premium-scaled anchor on $\mathcal{D}_{\text{train}}(e)$.
4: Form the anchored central prediction $b_{c,e}$ and the expanding-origin residual targets $\epsilon_{c,t}$.
5: Fit the fused residual expert and residual gate, and compute $\widehat{y}^{\text{res}}_{c,e}$ for the test counties.
6: If equation (15) is satisfied, fit the structural tail expert $\widehat{y}^{\text{tail}}_{c,e}$; otherwise retain the conservative residual regime.
7: Produce the final prediction from equation (16).

where κ_H is the minimum number of prior yearly histories required before the structural tail expert is allowed to govern the prediction. The final predictor is

$$\widehat{y}_{c,e} = (1 - D_e)\widehat{y}^{\text{res}}_{c,e} + D_e\widehat{y}^{\text{tail}}_{c,e}. \tag{16}$$

Equation (16) formalizes the central empirical idea of the benchmark: short-history folds are handled conservatively by a signed anchored-residual correction, whereas the more aggressive tail machinery is reserved for deep-history evaluation years in which recurrence structure is sufficiently observed.

The parameters and thresholds appearing in this section are reported in Section 4. They are reproducibility constants for the exploratory benchmark and should not be interpreted as statistically validated universal tuning values.

3 Data Architecture and Calibration Scope

The empirical benchmark is built from public, file-level data sources that can be independently retrieved and audited: FEMA NFIP redacted claims, FEMA NFIP redacted policies, NOAA Storm Events, and county-level ACS controls [15–18]. The data architecture is intentionally narrower than the full variable set that could be used in a flood-risk study. Climate normals, National Risk Index variables, proprietary insurer-market overlays, and licensed financial-market variables are excluded from the core evidence. This restriction makes the benchmark reproducible from public sources and prevents the reported results from depending on private data layers that other researchers cannot verify.

Let c denote a county and let $t \in \{2020, 2021, 2022, 2023, 2024\}$ denote the calendar year. The unit of analysis is the county-year (c, t). The claims file supplies realized paid losses and claim counts; the policy file supplies the contemporaneous NFIP portfolio, including policy count, written premium, coverage, and composition variables; NOAA Storm Events supplies county-year summaries of weather-event frequency and reported property damage; and ACS indicators supply slowly moving demographic and housing context. Table 1 summarizes the active architecture.

Table 1 Active public data architecture used in the core benchmark

Source	Role in the benchmark	Status
FEMA FIMA NFIP Redacted Claims v2	County-year paid loss, claim count, lagged-loss history, claim severity, claim rate, and historical pure-premium construction	Used in core benchmark
FEMA FIMA NFIP Redacted Policies v2	County-year written premium, policy count, coverage, policy-cost scale, and portfolio-composition variables	Used in core benchmark
NOAA Storm Events	County-year event counts, flood-related event summaries, and reported property-damage stress indicators	Used in core benchmark
ACS county indicators	Static or slowly varying demographic, housing, and socioeconomic controls joined by county FIPS	Used in core benchmark

3.1 Provenance-controlled sample construction

The central data problem is not merely missingness; it is the possibility of confusing an unretrieved file with an observed zero-loss outcome. The retrieval workspace contains targeted county-year policy files, legacy or non-targeted policy files, and incomplete state-year claims coverage. If these files are pooled mechanically, the resulting panel may include incorrect state–county assignments, duplicated legacy files, or unobserved claims years coded as zero. The benchmark therefore applies a file-level provenance filter before any statistical modeling is performed.

The governance rule has five steps. First, each policy file must follow the expected state–county–year naming convention. Second, the internal records of the file must agree with the state and county implied by the filename. Third, the county FIPS prefix must be consistent with the state FIPS code. Fourth, claims information is admitted only from observed state-year claims archives; missing state-year claims files are not interpreted as zero claims. Fifth, a county is retained only if both policy and claims histories are observed for every year from 2020 through 2024. These checks define the empirical estimand before model fitting and avoid contaminating pure-premium calculations with artificial zero-loss rows.

3.2 County-year aggregation and derived variables

After provenance filtering, all variables are aggregated to the county-year level. Let $\mathcal{C}_{c,t}$ be the set of retained claim records for county c in year t, and let $\mathcal{P}_{c,t}$ be the corresponding retained policy records. The realized paid loss, claim count, policy count,

Table 2 File-level sample construction summary

Stage	Count
Raw FEMA policy files in the six-state retrieval workspace	332
Targeted county-year policy files with valid naming pattern	324
Legacy or non-targeted policy files excluded from the benchmark workflow	8
Raw FEMA claims state-year files available in the six-state workspace	17
Counties satisfying complete observed claims and policy histories for 2020–2024	19
County-year observations retained in the final balanced panel	95

and written premium are constructed as

$$\begin{aligned} L_{c,t} &= \sum_{i \in \mathcal{C}_{c,t}} \text{PaidLoss}_i, \\ N^{\text{clm}}_{c,t} &= |\mathcal{C}_{c,t}|, \\ P_{c,t} &= \sum_{j \in \mathcal{P}_{c,t}} \text{PolicyCount}_j, \\ W_{c,t} &= \sum_{j \in \mathcal{P}_{c,t}} \text{WrittenPremium}_j. \end{aligned} \tag{17}$$

These quantities generate the two central adequacy measures used in the paper: realized pure premium $L_{c,t}/P_{c,t}$ and realized loss ratio $L_{c,t}/W_{c,t}$. Claim severity is computed on the positive-claim scale, whereas claim rate is computed per policy-year. Undefined ratios caused by zero denominators are not filled with structural zeros; they are treated as unavailable for the affected derived feature. This convention is important because the benchmark studies realized loss adequacy rather than claim occurrence alone.

Storm-event variables are joined by county and year. Let $\mathcal{S}_{c,t}$ denote the set of NOAA Storm Events records assigned to county c in year t. The core storm summaries are

$$\begin{aligned} F_{c,t} &= \sum_{k \in \mathcal{S}_{c,t}} \mathbf{1}\{k \text{ is flood related in the benchmark classification}\}, \\ D_{c,t} &= \sum_{k \in \mathcal{S}_{c,t}} \text{PropertyDamage}_k, \\ d_{c,t} &= \frac{D_{c,t}}{\max\{P_{c,t}, 1\}}. \end{aligned} \tag{18}$$

The variable $F_{c,t}$ captures event frequency, $D_{c,t}$ captures reported county-year property-damage pressure, and $d_{c,t}$ converts storm damage into a policy-scaled stress measure. Because NOAA Storm Events damage values are administrative estimates rather than audited insurance losses, they are used as stress indicators rather than as

substitutes for NFIP claims. This interpretation is consistent with the need for caution when using the Storm Events database for quantitative loss modeling [19].

Historical insurance-memory variables are computed only from prior years. For any realized county-year loss quantity $v_{c,t}$, the benchmark forms one-year lags $v_{c,t-1}$ and expanding histories such as $\max_{u<t} v_{c,u}$ and $|\{u < t : v_{c,u} > 0\}|$. These features summarize recurrence and tail memory without using the evaluated year's realized loss. ACS controls are joined through county FIPS and treated as contextual variables. They are not interpreted as causal drivers in this paper; instead, they adjust for persistent differences in housing, demographic, and socioeconomic composition across retained counties.

3.3 Retained panel and calibration scope

The provenance rules leave a balanced panel of 95 county-year observations: 19 counties observed for five years. The retained set consists of eight Florida counties and eleven Texas counties. The panel is balanced in time but highly imbalanced in loss contribution. Florida contributes most policy-years, claims, and paid losses, while the Texas counties provide lower-loss observations that help define the non-tail region of the benchmark. Table 3 reports the resulting state-level summary.

Table 3 Balanced provenance-controlled county-year panel used in the final benchmark

State	Counties	County years	Policy years	Claims	Paid loss (USD bn)	Written premium (USD bn)	Loss ratio
Florida	8	40	3,427,577	248,473	10.469	2.610	4.011
Texas	11	55	1,104,325	5,644	0.149	0.786	0.189

This retained panel should be interpreted as a selected Florida–Texas stress portfolio, not as a nationally representative NFIP sample. The complete-history rule increases internal validity by preventing missing claims archives from entering as artificial zero-loss years, but it also narrows external validity by excluding counties and states without complete observed histories. Consequently, the estimand is realized county-year pure premium for the retained high-risk counties under the stated public-data construction rule. The empirical results do not estimate national NFIP premium adequacy, validate Risk Rating 2.0, perform climate attribution, or justify prospective pre-season pricing. They instead provide a transparent calibration scope for the leakage-controlled stress-test benchmark developed in Sections 2 and 4.

4 Framework and Empirical Design

This section defines the empirical task used to evaluate the exploratory benchmark. The objective is not to estimate a national NFIP tariff, reproduce FEMA's property-level rating system, or make a pre-season underwriting decision. The objective is narrower: for each retained county c and evaluation year t, estimate realized county-year pure premium after the year-t policy portfolio and year-t NOAA Storm Events

summaries are available. The design is therefore a retrospective, event-conditioned stress test of public county-year data. This distinction is central because contemporaneous storm variables are informative for realized-loss measurement but would be unavailable in a genuine pre-year pricing exercise.

4.1 Response variables and adequacy interpretation

The main response is county-year pure premium,

$$\mathrm{PP}_{c,t} = \frac{\mathrm{TotalPaidLoss}_{c,t}}{\mathrm{PolicyCount}_{c,t}}, \tag{19}$$

measured in U.S. dollars per policy-year. The observed loss ratio is used as a secondary adequacy diagnostic,

$$\mathrm{LR}_{c,t} = \frac{\mathrm{TotalPaidLoss}_{c,t}}{\mathrm{WrittenPremium}_{c,t}}, \tag{20}$$

which compares realized paid losses with the written premium base. The modeling target in the reported experiments is $\mathrm{PP}_{c,t}$ rather than $\mathrm{LR}_{c,t}$ because pure premium is directly expressed on the loss-per-policy scale and avoids dividing the response by premium when premium is also used as an exposure and anchoring variable. Loss ratios are retained for descriptive interpretation of stress concentration and premium adequacy.

The benchmark reports equal-county errors; hence, each retained county receives the same weight in the headline MAE and RMSLE calculations. This choice makes the experiment a geographic stress benchmark rather than an aggregate-dollar actuarial scorecard. An exposure-weighted evaluation would answer a different question, namely how well the model explains aggregate loss dollars in the retained portfolio. That alternative is left for the next validation stage because the present sample is financially dominated by a small number of Florida tail counties.

4.2 Feature timing and leakage audit

The feature design follows the information structure defined in Section 2. Four blocks are used: within-year policy exposure, lagged county-loss history, contemporaneous storm-event summaries, and static county context. The historical block contains only pre-t summaries such as lagged pure premium, lagged claim rate, lagged severity, expanding county maxima, and expanding county means. The policy block contains year-t portfolio quantities such as policy count, written premium, coverage totals, premium per policy, coverage composition, and rating-zone composition. The storm block contains year-t NOAA event counts and damage summaries aggregated to the county-year level. The context block contains static or slowly moving county controls. Table 4 records the resulting availability interpretation.

All temporal transformations are recomputed inside each split. Training means, medians, scaling constants derived from empirical distributions, anchor ratios, and fitted model parameters are estimated only on the training window. County-history features for a test year use the same formula as training features but are accumulated

Table 4 Feature-availability audit for the benchmark task

Predictor block	Availability time	Interpretation in this paper
Lagged claims, losses, pure premium, claim rate, severity, and expanding county histories	Pre-year or start-of-year	Leakage-safe county memory variables computed only from years earlier than the evaluation year
ACS county controls	Static or pre-year	Contextual variables used to describe persistent county heterogeneity rather than current-year loss realization
Year-t policy count, written premium, coverage, premium per policy, and portfolio mix	Within-year portfolio observed	Retrospective adequacy measurement conditional on the observed insured portfolio, not a pre-season pricing rule
Year-t NOAA Storm Events counts and property-damage totals	Within-year or post-event	Event-conditioned loss-scaling information, not ex ante climate prediction or climate attribution

only up to year $t-1$. The availability audit therefore separates two issues that are often conflated: using current-year storm information is allowed because of the retrospective target, whereas using future loss history or test-year outcomes during feature construction is not allowed.

4.3 Model views and benchmark rules

Two feature views are compared. The insurance-only view is

$$\mathcal{X}_{c,t}^{\text{ins}} = \left(x_{c,t}^{\text{pol}}, x_{c,t}^{\text{hist}}\right), \tag{21}$$

where $x_{c,t}^{\text{pol}}$ denotes observed policy-composition variables and $x_{c,t}^{\text{hist}}$ denotes pre-t county-loss memory. The storm-event-fused view is

$$\mathcal{X}_{c,t}^{\text{fused}} = \left(x_{c,t}^{\text{pol}}, x_{c,t}^{\text{hist}}, x_{c,t}^{\text{storm}}, x_{c}^{\text{ctx}}, s_{c,t}, r_{c,t}\right), \tag{22}$$

where $x_{c,t}^{\text{storm}}$ contains NOAA storm-event summaries, x_{c}^{ctx} contains county context, and $s_{c,t}$ and $r_{c,t}$ are the shock and recurrence scores defined in Section 2. The insurance-only model is the main lower-information comparison, whereas the fused benchmark is the depth-aware selective architecture containing the insurance ensemble, premium anchor, signed residual correction, and a structural tail expert that is activated only when sufficient yearly history is available. Performance differences between the two views should therefore not be interpreted as the isolated marginal value of NOAA variables. They measure the combined contribution of storm information, county memory, premium anchoring, residual correction, and depth-aware tail selection.

Three simple actuarial-style rules are also reported so that the model comparison does not rely exclusively on another machine-learning benchmark. For evaluation year e, let $\mathcal{D}_{\text{train}}(e)$ denote all retained county-years with $t < e$. The training-mean rule is

$$\widehat{y}_{c,e}^{\text{mean}} = \frac{1}{|\mathcal{D}_{\text{train}}(e)|} \sum_{(i,u)\in\mathcal{D}_{\text{train}}(e)} y_{i,u}. \tag{23}$$

The lagged-pure-premium rule is

$$\widehat{y}_{c,e}^{\text{lag}} = y_{c,e-1}, \tag{24}$$

which tests whether the previous observed county-year pure premium already explains most of the next year's stress level. Finally, the premium-scaled rule is

$$\widehat{y}_{c,e}^{\text{prem}} = \widetilde{\rho}_e \, p_{c,e}, \qquad \widetilde{\rho}_e = \underset{(i,u)\in\mathcal{D}_{\text{train}}(e)}{\text{median}} \left\{ \frac{y_{i,u}}{p_{i,u}} : p_{i,u} > 0 \right\}, \tag{25}$$

where $p_{c,e}$ is observed premium per policy. This rule is deliberately simple and robust; it asks whether the written-premium scale alone provides a competitive stress benchmark.

4.4 Temporal evaluation design and metrics

The main split uses 2020–2023 as the training period and 2024 as the stress-year evaluation. This split is important because 2024 contains the most severe tail observations in the retained panel and tests whether the model can remain numerically credible under a concentrated stress year. As an auxiliary temporal check, rolling-origin folds evaluate $e \in \{2021, 2022, 2023, 2024\}$, each time fitting on all years strictly earlier than e and predicting the 19 county observations in year e. The rolling-origin design is summarized by

$$\mathcal{D}_{\text{train}}(e) = \{(c,t) : t < e\}, \qquad \mathcal{D}_{\text{test}}(e) = \{(c,e)\}, \qquad e = 2021, 2022, 2023, 2024. \tag{26}$$

This design is temporally disciplined but still limited: each fold contains only 19 test rows, and counties are repeatedly observed through time. The resulting fold-level evidence is therefore descriptive rather than formal proof of superiority.

For a test set $\mathcal{D}_{\text{test}}(e)$ of size n_e, predictive accuracy is summarized by

$$\text{MAE}_e = \frac{1}{n_e} \sum_{(c,e)\in\mathcal{D}_{\text{test}}(e)} \left| y_{c,e} - \widehat{y}_{c,e} \right|, \tag{27}$$

which preserves the dollar-per-policy scale, and by

$$\mathrm{RMSLE}_e = \left\{ \frac{1}{n_e} \sum_{(c,e)\in\mathcal{D}_{\mathrm{test}}(e)} \left[\log(1+y_{c,e}) - \log(1+\widehat{y}^{+}_{c,e})\right]^2 \right\}^{1/2}, \tag{28}$$
$$\widehat{y}^{+}_{c,e} = \max\{0, \widehat{y}_{c,e}\}.$$

MAE emphasizes absolute dollar errors and is sensitive to dominant tail counties. RMSLE compresses the extreme right tail and evaluates multiplicative error on the nonnegative pure-premium scale. Reporting both metrics is useful because a model can reduce large-dollar stress errors while still being poorly calibrated for lower-loss counties, or vice versa.

4.5 Implemented exploratory constants and interpretation discipline

Table 5 reports the implemented constants used by the exploratory benchmark. These constants are included for reproducibility and transparency, not as statistically validated optimal parameters. In particular, the selective tail architecture was refined during project-level development after inspection of extreme-row behavior. The 2024 result should therefore be interpreted as a stress-year development check and hypothesis-generating benchmark evidence rather than as evidence from a pristine untouched holdout.

The comparison discipline is intentionally conservative. Occurrence metrics are not used as headline evidence because the 2024 stress-year test set contains only positive claim-occurrence observations. Naive row-bootstrap intervals are not emphasized because the panel is small and county-years are likely dependent within states, storm systems, and repeated county histories. The reported MAE and RMSLE should therefore be read as transparent benchmark-comparison diagnostics. They address a narrow question—whether the fused public-data stress benchmark can improve realized county-year pure-premium estimation in this selected Florida–Texas panel—rather than a broad claim of national NFIP pricing validity, climate attribution, or deployable actuarial superiority.

5 Results and Interpretation

This section reports the empirical evidence from the selected Florida–Texas stress panel. The results are interpreted in three layers. First, the 2024 holdout year is treated as the main severe stress-year evaluation. Second, rolling-origin folds are used to assess whether the improvement is stable across earlier evaluation years. Third, the diagnostic figure and loss-concentration table identify where the benchmark is informative and where the evidence remains fragile. Throughout, the results should be read as descriptive benchmark evidence for event-conditioned realized pure premium, not as a locked actuarial tariff, a national NFIP adequacy estimate, or a climate-attribution result.

Table 5 Implemented exploratory calibration settings

Component	Setting	Value
Insurance ensemble	Learner weights	Equal across three learners
Anchor release	Candidate minimum insurance weights	0.15, 0.25, 0.35, 0.50
Anchor release	Selected minimum insurance weight in the 2024 main split	0.15
Residual gate	Gate target	Large absolute residual with concurrent stress signal
Residual correction	Correction cap	90th percentile of training absolute residual
Stress index	Shock and cascade combination	0.55 and 0.45
Shock score	Storm-damage and flood-event weights	0.7 and 0.3
Shock score	Scaling constants	1,000 and 10
Cascade score	Premium and claims weights	0.5 and 0.5
Cascade score	Scaling constants	20,000 and 60,000
Depth selector	Minimum prior yearly histories for tail-expert activation	4
Tail expert	Exceedance threshold τ	1,000
Tail expert	Quantile level	0.9
Regime labeling	Score threshold	0.45
Structural tail heads	Cascade and shock activation thresholds	Same as Section 2; used only in deep-history mode

5.1 Stress-year performance in 2024

Table 6 reports the main 2024 stress-year comparison. The response variable is county-year pure premium measured in U.S. dollars per policy-year. Hence, the MAE values have a direct adequacy interpretation: they measure the average absolute dollar error in realized paid loss per policy-year across the 19 retained counties. The RMSLE values complement MAE by measuring relative and multiplicative error; this is important because flood losses are strongly right-skewed and a few severe county-years can dominate dollar-scale accuracy.

The fused benchmark gives the best value for both metrics in this split. Relative to the insurance-only ensemble, MAE falls from 5,829.04 to 1,887.97, a reduction of approximately 67.6%. RMSLE falls from 2.208 to 1.213, a reduction of approximately

Table 6 Exploratory 2024 stress-year pure-premium results

Feature set	Model	MAE	RMSLE
Insurance-only	Ensemble mean	5,829.04	2.208
Storm-event-fused	Depth-aware selective ensemble	1,887.97	1.213
Benchmark	Train mean	6,033.50	2.572
Benchmark	Lagged pure premium	5,174.61	3.858
Benchmark	Premium scaled	6,148.28	4.182

45.1%. The improvement is also large relative to the three simple actuarial benchmarks. Compared with the training-sample mean, the fused benchmark reduces MAE by about 68.7% and RMSLE by about 52.8%. Compared with lagged pure premium, it reduces MAE by about 63.5% and RMSLE by about 68.6%. Compared with the premium-scaled rule, it reduces MAE by about 69.3% and RMSLE by about 71.0%.

The numerical message is clear but bounded. In the severe 2024 stress split, the insurance-only view under-reacts to the realized event-year loss environment, whereas the fused benchmark moves predictions materially closer to the realized tail. This improvement is consistent with the proposed architecture: the anchored residual regime keeps short-history behavior conservative, while the deep-history selector permits the structural tail expert to govern only after sufficient recurrence information is available. However, Table 6 does not isolate the marginal effect of NOAA Storm Events alone. The fused benchmark differs from the insurance-only benchmark in several respects; consequently, the appropriate interpretation is that the combined storm-event-plus-history selective architecture performs better in this 2024 split, not that any individual feature block has been causally validated.

The 2024 result should also not be treated as confirmatory validation. The test set contains only 19 county-years, the largest losses are concentrated in a small number of Florida counties, and the selective ensemble architecture was refined during project-level model development. The result is therefore best described as hypothesis-generating stress-year evidence. The fused benchmark appears useful for retrospective event-conditioned adequacy measurement under severe loss pressure, but locked external validation is required before any claim of general superiority can be made.

5.2 Rolling-origin stability

Table 7 reports the fold-level rolling-origin comparison. Each test year is predicted using only earlier years for model fitting and historical feature construction. This design prevents the 2024 result from being the sole source of evidence and shows whether the benchmark behaves consistently across different loss regimes.

The rolling-origin evidence reinforces the 2024 result. Relative to the insurance-only ensemble, the fused benchmark improves both MAE and RMSLE in every fold. In 2021, when only one historical year is available for fitting, the gain is modest in absolute dollars but directionally favorable. In 2022, the MAE improvement is small, yet RMSLE also improves, suggesting that the anchored residual regime stabilizes

Table 7 Fold-level rolling-origin pure-premium comparison

Test year	Train window	Test n	Ins. MAE	Fused MAE	Ins. RMSLE	Fused RMSLE
2021	2020–2020	19	177.19	144.51	2.792	2.446
2022	2020–2021	19	2295.65	2262.98	3.521	3.324
2023	2020–2022	19	1872.65	1462.38	3.481	2.928
2024	2020–2023	19	5829.04	1887.97	2.208	1.213

multiplicative scaling without inducing excessive over-reaction. In 2023, the fused benchmark is better on both metrics. In 2024, the deep-history selector activates the structural tail expert and produces the largest gain in both metrics.

Averaging across the four rolling-origin folds, the fused benchmark has a lower mean MAE than the insurance-only ensemble, 1,439.46 versus 2,543.63, and a lower mean RMSLE, 2.478 versus 3.001. These averages correspond to approximate reductions of 43.4% in MAE and 17.4% in RMSLE. At the same time, the result should still be interpreted carefully. The improvement is uniform against the insurance-only ensemble, but not against every simple benchmark in every calm-year fold; for example, the premium-scaled rule remains competitive in the low-loss 2021 and 2023 comparisons. Thus, the rolling evidence supports a narrow but credible conclusion: the fused benchmark is a stronger and more stable event-conditioned county-year predictor than the lower-information insurance-only ensemble in this selected panel, but it is still not a validated universal pricing rule.

5.3 Diagnostic interpretation of the fitted benchmark

Figure 1 provides a compact diagnostic view of the fused benchmark. The figure is descriptive rather than inferential because it is based on a small stress-test sample. It is nevertheless useful for explaining why the numerical results should be interpreted as evidence for a tail-stress benchmark rather than as validation of a general pricing model.

Panel (a) shows that the leading random-forest variables combine portfolio composition, premium intensity, coverage structure, and storm-event damage. Variables such as the share of post-FIRM policies, the share of higher-risk-zone policies, total policy cost, premium per policy, total premium, coverage ratio, storm property damage, and storm damage per policy appear among the dominant predictors. This pattern is substantively plausible. Realized pure premium is not governed solely by the presence of a storm event; it is also shaped by the exposed policy base, risk-zone composition, coverage levels, and the county's historical loss memory. The feature-importance plot therefore supports the interpretation of the fused view as a joint exposure–hazard–history representation. It should not be interpreted as causal evidence that any individual variable independently drives losses.

Panel (b) compares predicted and realized pure premium in the 2024 stress year. Most low- and moderate-loss counties lie closer to the diagonal, whereas the largest

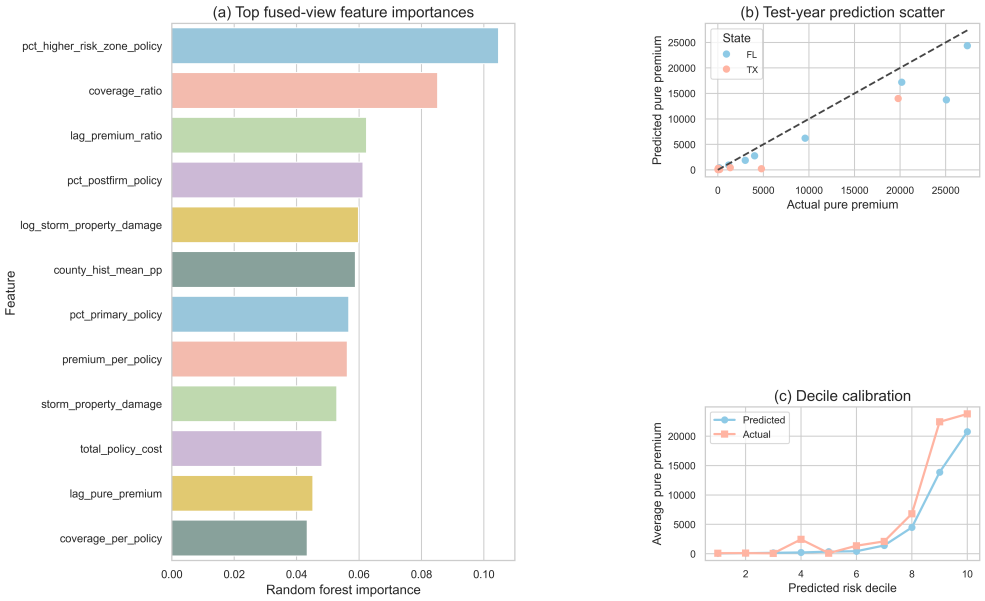


Fig. 1 Compact diagnostics for the fused pure-premium benchmark on the Florida–Texas stress design. Panel (a) reports leading random-forest feature importances for one submodel, panel (b) compares predicted and realized pure premium in the 2024 stress year, and panel (c) summarizes predicted and realized pure premium by predicted decile. The figure is descriptive and should not be read as a stand-alone validation result.

deviations occur among the extreme-loss counties. This is an expected error structure for a public county-year flood benchmark: the model can rank and scale many observations reasonably, but the most severe counties remain difficult because realized paid losses depend on local exposure concentration, claims settlement, event path, and extreme tail severity. The depth-aware selector reduces the most severe under-reaction of the insurance-only benchmark by allowing the structural tail expert to govern only in the deep-history regime, but it does not eliminate tail error. This point is important because an apparently strong aggregate MAE can still conceal residual fragility in the financially dominant counties.

Panel (c) summarizes predicted and realized pure premium by predicted decile. The upward movement in the highest predicted deciles indicates that the fused benchmark has some ability to separate high-stress observations from the rest of the sample. At the same time, the decile plot should not be read as stable calibration evidence. With only 19 observations in the 2024 split, each decile contains very few county-years, and the top deciles are highly sensitive to one or two extreme counties. The diagnostic value of the plot is therefore qualitative: it suggests that the model identifies a high-risk tail region, but it does not establish that predicted dollar magnitudes are calibrated throughout the distribution.

5.4 Loss concentration and scope of inference

Table 8 explains why the benchmark must be presented as a selected stress portfolio. The retained sample is not financially balanced across counties. The five highest-loss counties are all in Florida, and they account for approximately 9.28 billion dollars more paid loss than all retained Texas counties combined. Within the selected Florida–Texas panel, the five counties in Table 8 account for about 88.8% of total paid loss, while Lee and Pinellas alone account for about 69.3%. This concentration is not a secondary detail; it is the main empirical challenge.

Table 8 Highest-loss counties in the selected benchmark panel

State	County (FIPS)	Paid loss	Written premium	Loss ratio
Florida	Lee (12071)	3,810,104,029.83	673,784,116	5.655
Florida	Pinellas (12103)	3,548,917,657.43	634,885,903	5.590
Florida	Collier (12021)	1,051,915,748.45	389,062,863	2.704
Florida	Pasco (12101)	555,599,662.85	110,268,996	5.039
Florida	Volusia (12127)	466,045,424.24	99,527,202	4.683

The loss ratios in Table 8 are also informative. Lee, Pinellas, Pasco, and Volusia each have loss ratios above 4.6 in the retained sample, meaning that paid losses are several times larger than written premium over the study window. Collier has a lower but still severe loss ratio of 2.704. These values show that the selected panel contains genuine adequacy stress rather than ordinary annual fluctuation. They also explain why the premium-scaled benchmark performs poorly in the 2024 stress split: premium levels alone cannot absorb the realized tail when loss severity is concentrated in a small set of counties.

This concentration limits inference in two ways. First, equal-county MAE and RMSLE treat each retained county as one unit, even though dollar exposure and paid-loss contribution differ sharply across counties. Exposure-weighted metrics could therefore produce a different assessment. Second, the apparent success or failure of the fused benchmark is sensitive to a few Florida tail counties. A future confirmatory design should include leave-county-out, leave-state-out, or leave-event-out validation and should report uncertainty measures that respect spatial, temporal, and event-level dependence.

5.5 Overall interpretation

The full evidence supports a disciplined position. The 2024 stress-year result shows that the storm-event-fused benchmark can materially improve retrospective realized pure-premium estimation in a severe event-conditioned setting. The rolling-origin results show consistent improvement against the insurance-only ensemble in every evaluation year, with the largest gain occurring when the evaluation year contains severe tail pressure and sufficient pre-period history to justify the structural tail expert. The

diagnostics indicate that the model uses a plausible combination of policy composition, premium intensity, coverage structure, storm damage, and county history, but the largest errors remain concentrated in the extreme tail. The loss-concentration analysis further shows that a small group of Florida counties governs much of the empirical difficulty.

Accordingly, the contribution of the results is not a claim that the proposed architecture is a validated NFIP pricing model. Rather, the contribution is a transparent stress-test benchmark showing how realized pure-premium evaluation depends on provenance control, temporal feature discipline, and explicit treatment of tail counties. The current evidence is sufficient to motivate broader validation, stronger actuarial baselines, exposure-weighted scoring, and dependence-aware uncertainty analysis. It is not sufficient to claim national representativeness, prospective pricing validity, climate attribution, or universal dominance over every simple benchmark in every regime.

6 Conclusion

This paper develops a provenance-controlled stress-test benchmark for public NFIP pure-premium analysis in a selected Florida–Texas county-year panel. Its central contribution is not a claim that a final actuarial tariff has been obtained. Instead, the contribution is a disciplined empirical workflow showing how FEMA claims files, FEMA policy files, NOAA Storm Events summaries, and ACS county controls can be joined only after explicit sample-governance checks. By enforcing filename validation, state–county consistency, FIPS-prefix verification, and a complete-history rule, the benchmark prevents incomplete public retrievals from being silently converted into zero-loss observations.

The empirical findings support four conclusions. First, provenance control is not a secondary data-cleaning step; it determines the estimand itself. After the governance rules are imposed, the usable benchmark becomes a narrow high-risk Florida–Texas panel rather than a broad national NFIP sample. Second, county-history variables are essential for stress-year adequacy measurement because large realized losses are path-dependent and highly concentrated. Lagged pure premium, historical claim severity, recurrence pressure, and premium-scaled anchoring provide the memory needed to stabilize predictions under tail exposure. Third, storm-event information is useful when interpreted as event-conditioned evidence rather than as pre-season pricing information. The depth-aware selective ensemble improves the severe 2024 stress-year comparison and also improves both MAE and RMSLE against the insurance-only ensemble in every rolling-origin fold. Fourth, the largest errors and losses remain concentrated in a small number of Florida tail counties. This concentration explains why equal-county error metrics can be informative for geographic stress testing while still being insufficient for full actuarial adequacy assessment.

The scope of inference is therefore deliberately limited. The benchmark estimates realized county-year pure premium after the relevant policy portfolio and storm-event summaries are observed. It does not estimate national NFIP pricing adequacy, reproduce FEMA's property-level rating system, provide a prospective underwriting rule, or identify climate attribution. The depth-aware selective architecture should also be

interpreted cautiously because it was refined during model development and evaluated on a small, spatially dependent panel. These features make the results useful for benchmark construction and hypothesis generation, but not yet sufficient for confirmatory model validation.

Several extensions are needed before the framework can support stronger actuarial or policy claims. A next-stage study should lock the model specification before evaluation, expand the panel across additional states and years, and use leave-county-out, leave-state-out, or leave-event-out validation to test portability across spatial and disaster regimes. The scoring design should add exposure-weighted and premium-weighted metrics alongside equal-county errors because solvency-oriented adequacy and geographic stress testing answer different questions. Uncertainty assessment should account for dependence within states, counties, and storm systems rather than relying on naive row resampling. Finally, the benchmark should be compared with stronger actuarial baselines, including generalized linear models, frequency–severity models, hierarchical credibility models, and conformal or distributional prediction methods.

Appendix A Benchmark County List

Table A1 Counties retained in the final benchmark panel

State	County	County FIPS
Florida	Broward	12011
Florida	Charlotte	12015
Florida	Citrus	12017
Florida	Collier	12021
Florida	Lee	12071
Florida	Pasco	12101
Florida	Pinellas	12103
Florida	Volusia	12127
Texas	Aransas	48007
Texas	Brazoria	48039
Texas	Cameron	48061
Texas	Dallas	48113
Texas	Galveston	48167
Texas	Hidalgo	48215
Texas	Jefferson	48245
Texas	Liberty	48291
Texas	Montgomery	48339
Texas	Tarrant	48439
Texas	Walker	48471

Declarations

Funding

No external funding is declared for the present version of the manuscript.

Conflict of interest/Competing interests

The authors declare that they have no competing interests.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability

The reported benchmark is constructed from public FEMA, NOAA, and U.S. Census sources cited in the manuscript. The retained county list, sample-construction summary, and feature-availability audit are reported in the paper and appendix. Proprietary insurer-market overlays are excluded from the core benchmark and are not required to reproduce the reported figure and tables.

References

[1] Federal Emergency Management Agency: NFIP's Pricing Approach: Risk Rating 2.0. https://www.fema.gov/flood-insurance/risk-rating. Accessed 2026-06-05 (2026)

[2] U.S. Government Accountability Office: Flood Insurance: FEMA's New Rate-Setting Methodology Improves Actuarial Soundness but Highlights Need for Broader Program Reform. https://www.gao.gov/products/gao-23-105977. GAO-23-105977 (2023)

[3] Ortega, F., Petkov, I.: To improve is to change: The effects of Risk Rating 2.0 on flood insurance demand. Journal of Environmental Economics and Management **134**, 103228 (2025) https://doi.org/10.1016/j.jeem.2025.103228

[4] Gourevitch, J.D., Snyder, M., Kousky, C.: Effects of risk-based pricing reform on flood insurance uptake. Journal of Catastrophe Risk and Resilience **3** (2025) https://doi.org/10.63024/32za-vmy3

[5] Wing, O.E.J., Lehman, W., Bates, P.D., Sampson, C.C., Quinn, N., Smith, A.M., Neal, J.C., Porter, J.R., Kousky, C.: Inequitable patterns of US flood risk in the Anthropocene. Nature Climate Change **12**(2), 156–162 (2022) https://doi.org/10.1038/s41558-021-01265-6

[6] Poudel, S., Elliott, R., Anyah, R., Grabowski, Z., Knighton, J.: Differential flood insurance participation and housing market trajectories under future coastal flooding in the United States. Communications Earth & Environment **5**, 668 (2024) https://doi.org/10.1038/s43247-024-01848-z

[7] Nayak, A., Zhang, M., Gentine, P., Lall, U.: Catastrophic "hyperclustering" and recurrent losses: Diagnosing U.S. flood insurance insolvency triggers. npj Natural Hazards **2**, 83 (2025) https://doi.org/10.1038/s44304-025-00136-w

[8] Green, J., Haigh, I.D., Quinn, N., Neal, J., Wahl, T., Wood, M., Eilander, D., Ruiter, M., Ward, P., Camus, P.: Review article: A comprehensive review of compound flooding literature with a focus on coastal and estuarine regions. Natural Hazards and Earth System Sciences **25**(2), 747–816 (2025) https://doi.org/10.5194/nhess-25-747-2025

[9] Salas, J., Saha, A., Ravela, S.: Learning inter-annual flood loss risk models from historical flood insurance claims and extreme rainfall data. https://arxiv.org/abs/2212.08660 (2022)

[10] Bhattacharyya, A., Hastak, M.: Empirical causal analysis of flood risk factors on U.S. flood insurance payouts: Implications for solvency and risk reduction. Journal of Environmental Management **352**, 120075 (2024) https://doi.org/10.1016/j.jenvman.2024.120075

[11] Bowers, C., Serafin, K.A., Baker, J.W.: Uncovering drivers of atmospheric river flood damage using interpretable machine learning. Natural Hazards Review **25**(3), 04024022 (2024) https://doi.org/10.1061/NHREFO.NHENG-1995

[12] Liu, C., Mostafavi, A.: FloodGenome: Interpretable machine learning for decoding features shaping property flood risk predisposition in cities. https://arxiv.org/abs/2403.10625 (2024)

[13] Graziadei, H., Marques F., P.C., Melo, E.F.L., Targino, R.S.: Conformal prediction for frequency-severity modeling. https://arxiv.org/abs/2307.13124 (2025)

[14] Hong, L.: Conformal prediction of future insurance claims in the regression problem. https://arxiv.org/abs/2503.03659 (2025)

[15] Federal Emergency Management Agency: FIMA NFIP Redacted Claims v2. https://www.fema.gov/openfema-data-page/fima-nfip-redacted-claims-v2. OpenFEMA dataset page; accessed 2026-06-05 (2026)

[16] Federal Emergency Management Agency: FIMA NFIP Redacted Policies v2. https://www.fema.gov/openfema-data-page/fima-nfip-redacted-policies-v2. OpenFEMA dataset page; accessed 2026-06-05 (2026)

[17] NOAA National Centers for Environmental Information: Storm Events Database.

https://www.ncei.noaa.gov/stormevents/. Accessed 2026-06-05 (2026)

[18] U.S. Census Bureau: American Community Survey 5-Year Data. https://www.census.gov/data/developers/data-sets/acs-5year.html. Accessed 2026-06-05 (2026)

[19] Santos, R.P.: Some comments on the reliability of NOAA's Storm Events Database. https://arxiv.org/abs/1606.06973 (2016)