risks

MDPI

Article

Risk Scoring for Crop Insurance at Enrollment: Evidence and Limits

Constantin Colonescu [1,*], Subhadip Ghosh [2] and Shahidul Islam [1]

[1] Department of Economics, MacEwan University, Edmonton, AB T5J 4S2, Canada; islams@macewan.ca
[2] Treffo School of Business, MacEwan University, Edmonton, AB T5J 4S2, Canada; ghoshs3@macewan.ca
* Correspondence: colonescuc@macewan.ca

Abstract

Crop insurance has to be priced and screened before the season's main losses are known. This paper asks how far an insurer can get using only the information already available when a contract is enrolled: where the farm is, what crop and practice are insured, the chosen coverage level, and the local hail rate history. Using administrative records from 2006 to 2024, we build a practical underwriting score that separates the chance of a claim from the likely size of a loss, then test it on recent years and compare it with simpler rules and alternative models. The score ranks contracts better than regional averages, hail rate rules, or premium-based sorting, and the highest-ranked fifth of contracts contains about one third of the realized losses. Still, it misses much of the most severe loss risk. Coverage choice helps with prediction but also reflects farmer decisions, and thus the score should be read as a contract risk measure rather than a causal measure of hazard. These results suggest an enrollment-time information constraint for the specifications tested here, while leaving room for richer administrative and hazard-based extensions. Better tail prediction may require weather, farm history, and spatial information not used in the strict score.

Keywords: crop insurance; underwriting; risk classification; calibration; administrative data; two-part models

1. Introduction

Crop insurance underwriting has to happen before the most important seasonal shocks are observed. When a contract is written, the insurer can see where the farm is located, what crop is insured, how it is produced, and what contract terms the producer selects. It cannot yet see the weather path, final yields, or adjustment information that helps explain large claims later. That timing problem makes ex ante underwriting intrinsically difficult.

The issue matters for both administration and policy. Public crop insurance programs carry large financial exposure, while underwriting staff have a limited capacity to examine every file in depth. Even a moderate improvement in early risk ranking can therefore help direct scarce attention toward the contracts that matter most. At the same time, an enrollment-time score must remain reproducible, stable, and operationally explainable. A system that works only after harvest may be useful for monitoring, but it is not an underwriting tool in the narrow sense.

This paper asks a practical question: how much can be learned from the routine administrative information that is actually available when the contract is written? To answer this question, this paper builds a strict underwriting-time benchmark from five

Academic Editor: Alexandru Valentin Asimit

Received: 4 April 2026
Revised: 30 April 2026
Accepted: 6 May 2026
Published: 13 May 2026

inputs observed at enrollment—region, commodity, production practice, coverage, and hail rate—and evaluates whether that narrow information set can support useful risk ranking. This paper then studies four follow-up questions. First, can holdout recalibration or simple annual updating repair the upper tail of the score? Second, does a nonlinear machine learning model using the same inputs materially change the answer? Third, do same-input tail models such as log-normal severity or Tweedie expected loss models close the severity gap? Fourth, how should the score be interpreted when coverage is observed by the insurer but is also partly chosen by the producer?

The empirical design is organized around five testable expectations. (1) The strict enrollment-time score should rank losses better than simple administrative rules. (2) Because large crop losses may depend on hazards not yet observed at enrollment, the same score should be better at ranking than at upper-tail level prediction. (3) Recalibration and level updating should improve calibration more than ranking. (4) More flexible models using the same static inputs may still leave a tail gap; such evidence would identify a limitation of the tested information set rather than a universal limit on crop insurance prediction. (5) Coverage should be treated as an informative contract choice and not as a causal agronomic hazard variable. These expectations are evaluated directly in the holdout tables rather than treated as informal narrative claims.

This paper connects four related sets of works. The first is the economics of insurance classification, where poor risk classification can intensify adverse selection, cross-subsidization, and pricing distortions (Chiappori and Salanié 2000; Cohen and Siegelman 2010; Rothschild and Stiglitz 1976). Crop insurance is a demanding application because losses are correlated across farms, and because many of the shocks that matter most arrive after enrollment (Goodwin and Mahul 2004; Miranda and Glauber 1997). The second sets of works studied agricultural insurance design and risk measurement. Classic works emphasize the difficulty of insuring systemic agricultural risk, while newer works ask whether richer data can improve insurance design, zone construction, or hazard measurement (Babcock et al. 2004; Barnett 2000; Goodwin and Smith 1995; Kirchner et al. 2026; Nguyen et al. 2025; Nourali and Shortridge 2026; Wijesena and Pradhan 2025). The present paper addresses a prior question: what can be done with routine enrollment-time administrative data alone? The third set of works concerns forecast evaluation and recalibration, especially the distinction between ranking, reliability, and resolution (Brier 1950; Gneiting and Raftery 2007; Murphy 1973; Murphy and Winkler 1987). Related works show that recalibration can improve reliability while preserving much of the original ordering (Kull et al. 2017; Silva Filho et al. 2023). The fourth set of works concerns limited dependent variables and heavy-tailed losses, including two-part models, selection diagnostics, Tweedie models, and extreme value approaches (Coles 2001; Cragg 1971; Heckman 1979; Jørgensen 1997; Mullahy 1986). This paper brings these ideas into an actuarial setting where the target is the expected loss per acre at enrollment.

The main findings are straightforward. The enrollment-time baseline is useful; it ranks risky contracts better than simple administrative rules and therefore has value for file triage and portfolio review. At the same time, it materially understates severe losses, especially in the upper tail and especially after the loss environment shifts. Holdout recalibration, previous-year level updating, and heavier-tailed same-input specifications can narrow some level gaps, but they do not jointly improve the ranking, calibration, or mean absolute error. The evidence supports the ranking expectation, confirms the upper-tail limitation, and shows that coverage is a selected but operationally informative contract term. This paper therefore identifies the value of the static administrative benchmark and a boundary for the specifications tested here.

The remainder of this paper is organized as follows. Section 2 describes the data, underwriting framework, and empirical design. Section 3 presents the main results. Section 4 discusses the implications and limitations. Finally, Section 5 concludes the paper.

2. Materials and Methods

2.1. Administrative Setting and Data

The underlying dataset was a large annual contract-year panel assembled from administrative crop insurance records. Each observation contained contract characteristics available to the program, including the location, crop, production practice, coverage, pricing information, insured acreage, and realized loss outcomes.

The administrative records were supplied by Agriculture Financial Services Corporation (AFSC, Lacombe, AB, Canada). All empirical calculations were implemented in Python 3.14.3 (Python Software Foundation, Beaverton, OR, USA), using open-source scientific-computing and machine-learning libraries, including scikit-learn and XGBoost where applicable.

The institutional setting was a publicly supported crop insurance program with standardized contract records, program-administered premiums, and realized indemnity outcomes. This context is similar to other public or public–private crop insurance systems in the key respect that coverage and exposure are recorded before the growing season's outcome is known. It differs from purely private crop-hail markets and from index-insurance designs because indemnities, claim adjustment, premium support, and producer participation rules are institutionally mediated. External validity is therefore context-dependent. Program rules, premium subsidies, claim adjustment procedures, reinsurance arrangements, and the regional crop mix can all affect both producer selection and realized losses. The estimates should be read as evidence of what a strict enrollment-time administrative information set can do in this program context and not as a universal pricing formula.

This paper draws a sharp timing distinction between information available at underwriting and information realized later. The headline score uses only variables that are cleanly observable when the contract is written: agricultural region, commodity, production practice, coverage level, and township hail base rate. The seeded day of the year and season length appeared in the raw files, but they were reserved for robustness checks because they were not part of a strict enrollment-time underwriting rule.

Table 1 summarizes the portfolio. The outcome was heavily zero-inflated and strongly right-skewed, making a two-part strategy natural. The table also shows that the evaluation period was riskier than the training period. That shift matters throughout this paper because a static score can preserve useful ranking while still missing the level of losses when the environment changes.

Table 1. Summary statistics for the master panel.

Variable	Mean	SD	P10	Med.	P90	Train Mean	Test Mean
Insured acres	290.863	323.379	65	182	617	279.796	321.145
Coverage level	0.739	0.076	0.600	0.800	0.800	0.730	0.765
Hail rate	0.062	0.025	0.040	0.060	0.090	0.062	0.062
Loss ratio	1.734	4.219	0	0	5.492	1.209	3.174
Loss frequency	0.363	0.481	0	0	1	0.325	0.469
Loss per acre	40.076	87.349	0	0	139.049	26.148	78.189
Premium per acre	25.850	14.990	10.994	22.190	45.450	24.065	30.732
Seeded day of year	135.562	15.598	122	135	148	135.919	134.585
Season length (days)	141.013	230.621	114	136	165	144.372	131.801

Notes: Summary statistics were computed at the contract-year level. Train means used 2006–2019, and test means used 2020–2024. A mean loss ratio above one reflects indemnities relative to premium in a panel with several high-loss years; it is descriptive of this portfolio, and it was not used as a pricing target in the scoring model.

2.2. Underwriting Target and Interpretation

This paper does not seek a structural model of farm production or insurance demand. It seeks a prospective underwriting rule. Let L_{it} denote the realized loss per acre, and let X_{it} collect the contract characteristics observed at enrollment. The central predictive object is

$$m_t(X_{it}) = \mathbb{E}_t[L_{it} \mid X_{it}],$$

with the year-t expected loss conditional on the underwriting-time information set.

For a risk-neutral insurer or regulator deciding which contracts require extra review, $m_t(X)$ is the natural target. If only a fraction q of files can be reviewed, then the optimal triage rule ranks contracts by the expected loss and sends the highest-ranked q share to deeper review. This does not require a structural welfare model; it requires only that review resources be scarce and that the higher expected loss be more underwriting-relevant than the lower expected loss.

The framework also clarifies what the score is not . We write $X_{it} = (W_{it}, C_{it})$, where W_{it} contains the region, commodity, practice, and hail rate and C_{it} is the chosen coverage level. A score based on $m_t(W, C)$ is a contract risk score. It uses the contract terms actually observed by the insurer at enrollment. It is not a pure agronomic hazard score, and it is not a causal estimate of how changing coverage would change the losses. This distinction is important because coverage can reveal a producer's private information. A positive association between coverage and the later loss is therefore informative for underwriting triage but also evidence of selection, and thus coverage coefficients are not interpreted structurally. This paper therefore does not estimate an instrumental variable model for coverage; the administrative file does not contain a credible enrollment-time exclusion restriction that shifts the coverage choice while affecting the realized losses only through coverage. In practice, an instrument would need to change the chosen coverage level while being unrelated to the producer risk, location, insured crop, practice, and expected loss; the observed administrative variables either enter the underwriting risk directly or plausibly select a proxy. Consequently, estimates involving coverage are used for predictive screening and sensitivity analysis and not for causal policy claims about coverage levels. Section 3 reports the specifications without coverage, within coverage bands, coverage choice dynamics, and a residual correlation diagnostic to make this interpretation explicit.

Finally, the omitted information problem can be written directly. Let Z_{it} denote later-season hazards and other variables unavailable at the underwriting time. Then, we have

$$\mathbb{E}_t[L_{it} \mid X_{it}] = \mathbb{E}_t[\mathbb{E}_t[L_{it} \mid X_{it}, Z_{it}] \mid X_{it}].$$

Even with a well-estimated baseline, a missing Z_{it} value compresses the right tail if large losses are triggered by hazard realizations that are weakly proxied by X_{it}. If the environment also shifts over time so that $m_t(X) \neq m_{t'}(X)$, then a well-ranked score from the training period may remain miscalibrated in later years. This simple decomposition motivates both the recalibration analysis and the year-by-year evidence reported below.

2.3. Information Sets and Preprocessing

Table 2 summarizes this paper's information sets. The strict underwriting-time baseline is the only central specification. Later timing models are retained only as robustness checks.

Several preprocessing choices matter. A missing agricultural region is coded as a separate "Missing" category, because those records lack a usable region crosswalk at the time the score is constructed. Retaining them keeps the rule prospective and avoids dropping files that an operational system would still have to handle. The resulting region

coefficient is therefore an operational missingness signal and not a substantive regional effect. The rare raw practice label "Dry, Fert" is merged into Dryland. Continuous predictors are standardized using training-sample means and standard deviations. Missing numeric predictors are imputed using training-sample means. This affects only a rather small number of hail rate records in the strict baseline and a larger number of season length records in the later timing robustness check. All headline estimates use insured acre weights. Acreage is the relevant exposure measure in this setting because the target is the losses per acre, while the operational cost of misclassification scales with the acres exposed. Liability weighting would partly reweight observations by price and chosen coverage, which are themselves predictors in the contract risk score. Appendix A Table A7 also reports the contract-weighted AUC to show that the main ranking result is not an artifact of acre weighting.

Table 2. Information taxonomy.

Information Set	Variables	Operational Use	Role in Paper
Strict underwriting-time baseline	Region, commodity, practice, coverage, hail rate	Enrollment-time underwriting score	Headline specification
Early-season extension	Strict baseline + seeded day of year	Post-planting monitoring	Timing robustness
Season-complete administrative monitor	Early-season extension + season length	End-of-season monitoring	Timing robustness
Without coverage level	Region, commodity, practice, hail rate	Separates selected coverage from other signals	Coverage robustness
Separate models within coverage bands	Baseline covariates estimated within each coverage band	Checks ranking within contract menus	Coverage robustness

2.4. Two-Part Estimation

The loss process is modeled as a standard two-part system (Cragg 1971; Mullahy 1986). Let $D_{it} = \mathbf{1}\{L_{it} > 0\}$. The frequency block is

$$\Pr(D_{it} = 1 \mid X_{it}) = \Lambda(X_{it}'\beta),$$

as estimated by insured acre-weighted logistic regression with an L_2 penalty parameter $C = 1$. The choice of $C = 1$ was a shrinkage default chosen in advance for numerical stability and ease of interpretation, rather than being tuned to maximize the holdout fit. Appendix A Table A9 reports a training window validation grid over C and the severity penalty. In that grid, the chosen C had a negligible numerical effect on the first-stage probabilities, whereas α affected the severity-level metrics but not the frequency AUC or Brier score. The severity block is

$$\mathbb{E}[L_{it} \mid D_{it} = 1, X_{it}] = \exp(X_{it}'\gamma),$$

as estimated on positive loss contracts using an insured acre-weighted Gamma generalized linear model with log link, a standard choice for positive and strongly right-skewed outcomes (McCullagh and Nelder 1989), implemented through `GammaRegressor` with the regularization parameter $\alpha = 1$. Again, $\alpha = 1$ is a stabilizing default. The contract score is the predicted expected loss per acre:

$$\widehat{S}_{it} = \widehat{p}_{it}\widehat{\mu}_{it}.$$

In plain language, the model first predicts whether any loss occurs and then predicts how large the loss would be if a loss occurred. Multiplying those two pieces produces an expected loss per acre for each contract. Each category is represented by its own indicator variable rather than being folded into a single omitted reference category. With regularization, this makes coefficient magnitudes easier to interpret as symmetric shrinkage around the intercept. The two-part system is a predictive factorization of the conditional mean and not a joint structural model of the latent weather or farm shocks. It therefore does not require the latent shocks driving loss occurrence and loss severity to be independent. In crop insurance, the same event can determine both margins, and thus dependence between the frequency and severity pieces is expected. The severity-power diagnostic, the positive-loss frequency–severity diagnostic, the same-input tail alternatives, and the coverage residual diagnostic below are therefore interpreted as specification checks rather than as structural tests of producer behavior.

As an additional formal check on the residual dependence between occurrence and positive-loss severity, the Appendix reports a Heckman-style inverse Mills diagnostic for the two-part specification. Within each sample, a probit loss-occurrence equation is estimated with the strict enrollment-time controls. The inverse Mills ratio from that first stage is then added to an insured acre-weighted log severity regression on the same controls among the positive-loss contracts. A significant Mills coefficient would indicate residual selection in severity, conditional on the observed underwriting controls. Because the same variables identify both stages, and no exclusion restriction is available, this is a diagnostic test of the conditional-independence concern rather than a structural sample selection correction.

2.5. Benchmarks and Evaluation Metrics

The strict underwriting-time baseline was compared with three simple operational benchmarks: (1) a hail-rate-only two-part model; (2) region–crop–practice means estimated on the training window; and (3) a premium-per-acre two-part model.

The headline evaluation used the 2020–2024 holdout. This paper reports the insured acre-weighted AUC and Brier score for the loss frequency, insured acre-weighted mean absolute error (MAE) for the expected loss per acre, and a contract-based triage measure: the share of realized total losses captured by the top 10%, 20%, and 30% of contracts ranked by the score. The acre-weighted AUC was used because a misranked 2000-acre file is more important operationally than a misranked 20-acre file when the outcome is in loss per acre. The contract-weighted AUC is reported as a diagnostic in Appendix A Table A7.

These metrics answer different practical questions. The AUC asks whether the score ranked higher-risk contracts above lower-risk contracts. The Brier score asks whether the predicted loss probabilities were numerically close to the realized frequencies. Appendix A Table A7 decomposes the Brier score into binned reliability, resolution, and uncertainty components, following (Murphy 1973). The expected loss MAE asks whether the score got the dollar magnitudes right on average. The top capture statistics are the operational triage metrics; if underwriting staff can review only a fixed fraction of the files, then how much realized loss falls inside that flagged set?

2.6. Holdout Recalibration

The recalibration design used contract-level holdout data throughout. The calibration window was 2020–2022, and the evaluation window was 2023–2024.

First, this paper applies frequency recalibration by fitting

$$\Pr(D_i = 1) = \Lambda(a_0 + a_1 \operatorname{logit}(\widehat{p}_i))$$

on the calibration window and evaluating the transformed probabilities in 2023–2024. This is a standard logistic recalibration of the frequency block. In plain language, recalibration does not add new underwriting variables; it rescales the model's fitted values so that the predicted levels line up better with what is observed in a later window.

Second, this paper estimates a power-based severity correction on positive-loss observations:

$$\log L_i = b_0 + b_1 \log \widehat{\mu}_i + \varepsilon_i,$$

which, again, is on the calibration window with the insured acre weights. The resulting map

$$\widetilde{\mu}_i = \exp(b_0)\widehat{\mu}_i^{b_1}$$

is a contract-level power recalibration of the severity block. When $b_1 > 1$, the severity predictions were too compressed in the upper tail. The exact slope estimates and their confidence intervals are reported later in Section 3 and in Appendix A Table A3. Importantly, all of these corrections were evaluated on fresh holdout periods rather than on grouped deciles used both for fitting and display.

2.7. Uncertainty and Comparative Inference

To compare the baseline with the benchmarks, this paper used year-block bootstrap resampling on the 2020–2024 holdout. Entire crop years were resampled with replacement. This preserved the within-year dependence structure created by the shared weather shocks. Because the evaluation window contained only five post-2019 years, these intervals should be read as rough evidence about the stability rather than as a highly precise large-sample inference. This paper therefore complemented the bootstrap with leave-one-year-out stability checks, reported in Appendix A Table A2. The coefficient tables are descriptive; this paper did not base its conclusions on contract-level coefficient standard errors. Spatial dependence across townships remained part of the interpretation of the holdout results rather than something fully solved by asymptotic clustering.

3. Results

3.1. Headline Out-of-Sample Performance

Table 3 reports the main benchmark comparison on the 2020–2024 holdout. The strict underwriting-time baseline was the strongest enrollment-time model overall. Its insured acre-weighted AUC was 0.646, compared with 0.628 for the region–crop–practice means, 0.580 for the hail rate alone, and 0.549 for the premium per acre. Its contract-weighted AUC was 0.640 (Appendix A Table A7), so the ranking result was not driven by acre weighting alone. Its Brier score was 0.250, which again was better than the simple benchmarks, and its expected loss MAE was USD 81.247 per acre, slightly below the comparable figures for the alternatives.

Table 3. Out -of-sample performance on the 2020–2024 holdout.

Model	AUC	Brier	MAE EL/ac	Top-Decile Gap
Strict underwriting-time baseline	0.646	0.250	81.247	106.581
Season-complete administrative monitor	0.638	0.254	81.279	107.337
Hail rate-only benchmark	0.580	0.271	82.546	87.240
Segment means (region × crop × practice)	0.628	0.264	81.431	70.067
Premium-based benchmark	0.549	0.266	82.713	61.256

Notes: The top-decile gap is the insured acre-weighted difference between the observed and predicted loss per acre in the highest score decile.

The season-complete administrative monitor was included only as a later-timing robustness check. It used information that was not available when the contract was written,

yet it still posted weaker headline metrics than the strict baseline: an AUC of 0.638, a Brier score of 0.254, and an MAE of 81.279. The positive result here is important but also limited; the baseline clearly improved the ranking and probability fit, while the gain in the average dollar error was real but modest. Appendix A Table A7 also shows why the Brier score worsened in the holdout; the test-period reliability penalty was 0.017 rather than essentially zero in the training window, while the resolution remained modest.

3.2. Comparative Inference Against Simple Benchmarks

Table 4 shows where the baseline's advantage was statistically most convincing. Relative to the region–crop–practice means, the strict baseline improved the AUC by 0.019 with a 95% interval of [0.006, 0.032], improved the Brier score by 0.014, and raised the top 20% loss capture by 0.060. Its MAE advantage over the segment means was much smaller at USD 0.184 per acre in absolute value, and the interval included zero. Thus, the ranking and probability gains over the segment means are reliable, while the improvement in the average dollar fit was less decisive.

Relative to the hail rate alone, the baseline improved the AUC by 0.066, improved the Brier score by 0.021, lowered the MAE by USD 1.299 per acre, and raised the top 20% capture by 0.082. Relative to the premium per acre, the baseline improved the AUC by 0.098 and lowered the MAE by USD 1.466 per acre. The top 20% capture gain over premium was positive as well, though less precisely estimated. Appendix A Table A2 shows that these signs were highly stable when one evaluation year was omitted at a time.

This is an important qualification. The baseline did not win by the same margin for every metric, but it did deliver a clearer enrollment-time ranking signal and better probability fit than the operational benchmarks. The hard part that remains is getting the level of expected losses right in the upper tail.

Table 4. Year-block bootstrap differences: strict baseline minus benchmark.

Comparator	Metric	Point Diff.	95% Interval
Segment means	AUC	0.019	[0.006, 0.032]
	Brier	−0.014	[−0.024, −0.006]
	MAE EL/ac	−0.184	[−1.468, 0.781]
	Top 20% capture	0.060	[0.020, 0.078]
Hail-rate-only	AUC	0.066	[0.036, 0.103]
	Brier	−0.021	[−0.031, −0.014]
	MAE EL/ac	−1.299	[−2.325, −0.562]
	Top 20% capture	0.082	[0.052, 0.124]
Premium-based	AUC	0.098	[0.015, 0.178]
	Brier	−0.016	[−0.033, −0.006]
	MAE EL/ac	−1.466	[−2.947, −0.206]
	Top 20% capture	0.071	[−0.017, 0.131]

Notes: Positive AUC and top 20% capture differences favored the strict baseline. Negative Brier and MAE differences favored the strict baseline.

3.3. Operational Triage Value

Table 5 translates the benchmark comparison into underwriting language. The strict baseline captured 19.2% of the realized losses in the highest-risk 10% of contracts, 34.4% in the highest-risk 20%, and 48.5% in the highest-risk 30%. The comparable figures were 12.9%, 28.4%, and 39.9% for the region–crop–practice means; 12.5%, 26.1%, and 37.9% for the hail rate alone; and 15.5%, 27.2%, and 37.5% for the premium per acre, respectively.

In plain language, if the underwriting staff can review only one fifth of the files, then the strict baseline directs them toward about one third of the realized losses, while the simpler rules usually point them toward only about one quarter. The season-complete

monitor captured slightly more losses than the strict baseline, but it used information that was not available at enrollment and was therefore not the relevant headline comparator.

Table 5. Loss capture among the highest-ranked contracts on the 2020–2024 holdout.

Ranking Rule	Top 10% of Contracts	Top 20% of Contracts	Top 30% of Contracts
Strict underwriting-time baseline	0.192	0.344	0.485
Season-complete administrative monitor	0.195	0.354	0.488
Segment means (region × crop × practice)	0.129	0.284	0.399
Hail rate-only benchmark	0.125	0.261	0.379
Premium-based benchmark	0.155	0.272	0.375
Random selection	0.100	0.200	0.300

Notes: Contracts ranked by the model score. Reported values are the shares of realized total losses captured by the highest-ranked contracts.

3.4. The Main Empirical Limit: Upper-Tail Underprediction

The central negative result remained upper-tail underprediction. Table 6 and Figure 1 show that higher score deciles were associated with higher realized losses, but the level fit deteriorated sharply at the top of the distribution. In the highest score decile, the predicted loss frequency was 0.530, while the observed loss frequency was 0.695. The predicted positive-loss severity was USD 84.750 per acre, while the observed severity was USD 217.917. Put together, the predicted expected loss was USD 44.896 per acre, and the observed loss was USD 151.478.

Table 6. Strict baseline deciles on the 2020–2024 holdout.

Decile	Acres	Pred. p(Loss)	Obs. Freq.	Pred. Sev.	Obs. Sev.	Pred. EL/ac	Obs. Loss/ac
1	6,629,663	0.206	0.256	60.488	94.849	12.555	24.257
2	8,002,234	0.279	0.356	67.283	109.251	18.754	38.942
3	7,131,421	0.300	0.399	69.376	121.470	20.811	48.504
4	6,933,416	0.336	0.476	72.838	147.675	24.463	70.321
5	9,077,808	0.373	0.485	72.193	147.679	26.891	71.553
6	7,741,804	0.386	0.498	74.152	152.856	28.628	76.181
7	7,112,897	0.404	0.553	75.717	163.747	30.556	90.618
8	8,022,963	0.431	0.602	78.237	185.017	33.711	111.301
9	7,316,355	0.473	0.638	80.112	197.460	37.778	125.970
10	7,820,650	0.530	0.695	84.750	217.917	44.896	151.478

Notes: Contracts sorted by the strict underwriting-time score and divided into equal-count deciles.

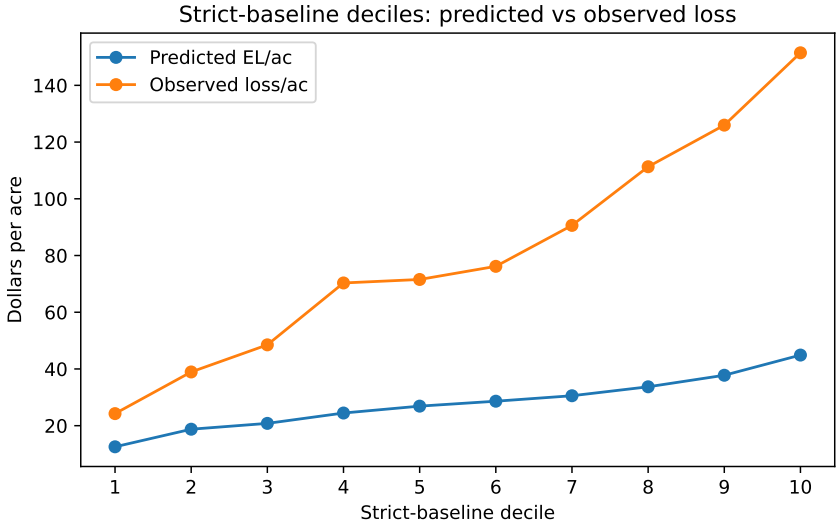

Figure 1. Strict-baseline deciles: predicted versus observed loss per acre. In the top decile, the predicted expected loss is USD 44.896 per acre, while the observed loss is USD 151.478 per acre.

This is not a small miss. The score clearly ordered the deciles in the right direction, which is why it remains useful for ranking, but it was much too low where the most

important losses occurred. The same pattern was already visible one decile below the top; in decile 9, the predicted expected loss was USD 37.778 per acre, and the observed loss was USD 125.970. The paper's main negative finding is therefore not poor ranking; it is severe underprediction in the upper tail.

3.5. A Nonlinear Benchmark with the Same Information Set

A natural concern is that the tail gap reflects the linear two-part specification rather than the limited underwriting-time information set. To probe that distinction, this paper estimated a two-part model based on gradient-boosted trees (Chen and Guestrin 2016; Friedman 2001) using the exact same five strict-baseline inputs: region, commodity, practice, coverage, and hail rate. This method can pick up nonlinear relationships and interactions without requiring the researcher to write them down one by one. The goal was not to replace the baseline as the headline score but to test whether allowing a more complicated relationship between the same variables and losses changed the substantive conclusion.

Table 7 shows a mixed pattern. The boosted-tree model lowered the MAE from 81.247 to 80.874 USD per acre and narrowed the top-decile gap from 106.581 to 87.694. However, it also lowered the AUC from 0.646 to 0.640 and reduced the top 20% loss capture from 0.344 to 0.313. The Brier score was essentially unchanged at 0.250.

For the underwriting task studied here, that trade-off matters. The score is mainly used to rank files for review. A small gain in the average dollar error is not enough to outweigh a weaker ranking signal. The remaining limitation is therefore consistent with a restricted underwriting-time information set, although this exercise tested only one same-input nonlinear extension and should not be read as proving that all useful within-set interactions or transformations were exhausted.

Table 7. Nonlinear benchmark with the same information set on the 2020–2024 holdout.

Model	AUC	Brier	MAE EL/ac	Top 20% Capture	Top-Decile Gap
Strict underwriting-time two-part baseline	0.646	0.250	81.247	0.344	106.581
Gradient-boosted trees	0.640	0.250	80.874	0.313	87.694

Notes: Both rows use the same five underwriting-time inputs. The results were mixed across metrics, so the exercise was interpreted as a model-form diagnostic rather than as a replacement headline score.

The practical takeaway is straightforward: the boosted-tree benchmark trimmed the average dollar error a little, but it weakened the ranking signal used for underwriting triage. That is why the simpler two-part model remained the headline score.

3.6. Tail-Specific Alternatives Using the Same Enrollment Inputs

The severity-power slope reported below was well above one, and thus it is natural to ask whether a heavier-tailed severity specification could close the gap without adding new information. Table 8 compares the headline Gamma severity block with two same-input alternatives: a log-normal severity model fitted on positive losses and back-transformed with a smearing factor and a one-part Tweedie expected-loss GLM. These are diagnostic alternatives and not a search for a new preferred production model.

The log-normal severity model modestly improved the positive-loss fit; the log RMSE fell from 1.189 to 1.180, and the Gamma deviance fell from 1.668 to 1.613. It also narrowed the top-decile gap from USD 106.576 to USD 76.723 per acre. However, the top 20% loss capture fell from 0.344 to 0.303. The one-part Tweedie GLM did not improve the MAE and still left a large top-decile gap. Thus, the tail-aware same-input specifications confirm that the Gamma block was compressed in the upper tail, but they did not remove the

main underwriting trade-off. The power slope should therefore be read as a substantive misspecification signal and not as a cosmetic calibration adjustment.

Table 8. Same-input tail alternatives on the 2020–2024 holdout.

Tail Model	Log RMSE	Gamma dev.	MAE EL/ac	Pred. EL/ac	Top 20% Capture	Top-Decile Gap
Gamma severity GLM (headline)	1.189	1.668	81.249	28.146	0.344	106.576
Log-normal severity GLM	1.180	1.613	81.038	29.980	0.303	76.723
One-part Tweedie GLM			81.522	28.176	0.316	97.823

Notes: Log RMSE and Gamma deviance computed on positive-loss holdout observations. The Tweedie row is a one-part expected-loss model, so the positive-loss severity diagnostics are not reported. All rows used only the strict enrollment-time inputs.

Table 9 adds a simple frequency–severity dependence diagnostic to the positive-loss holdout observations. A regression in log-realized severity on the log baseline severity fit gave a power slope of 2.230, close to the split-specific recalibration slopes in Appendix A Table A3. The residual correlation between the log severity error and the first-stage frequency logit was 0.159. This is consistent with the interpretation that common loss shocks affected both margins and that the baseline severity block was too compressed. It does not invalidate the expected loss factorization, but it reinforces the need to interpret the two parts as predictive components rather than independent structural equations.

Table 9. Positive-loss frequency–severity dependence diagnostic on the 2020–2024 holdout.

Sample	Positive Contracts	Positive Acres (m)	Power Slope	Residual Corr.
Holdout 2020–2024	110,621	37.820	2.230	0.159

Notes: The power slope came from a weighted regression of the log realized positive loss on the log predicted severity. The residual correlation was between the log severity error $\log L - \log \widehat{\mu}$ and the first-stage frequency logit among positive-loss contracts. Both quantities are descriptive diagnostics.

Appendix A Table A8 reports the complementary inverse Mills diagnostic. The Mills coefficient was positive but not statistically distinguishable from zero in either the training sample (0.724, $p = 0.201$) or the holdout sample (0.369, $p = 0.226$). Equivalently, the zero-coefficient null was not rejected in either sample. This does not prove structural independence because the diagnostic had no exclusion restriction and used the same administrative controls in both stages. For readers concerned that the two-part set-up imposes an independence assumption, this formal residual inclusion check does not add evidence against the predictive factorization while also underscoring that the result is diagnostic rather than a structural proof.

3.7. Holdout Recalibration Improved the Tail but Not the MAE

Table 10 reports the contract-level holdout recalibration results. On the 2023–2024 evaluation window, the unrecalibrated baseline had an AUC of 0.666, Brier score of 0.242, and MAE of 69.541. Its mean predicted expected loss was USD 28.319 per acre, far below the realized mean of USD 69.160.

Frequency recalibration left the AUC unchanged at 0.666, improved the Brier score to 0.230, and raised the mean predicted expected loss to USD 38.833 per acre. However, the MAE worsened to 70.931. Adding the severity power correction left the AUC and Brier score unchanged at 0.666 and 0.230, respectively, raised the mean predicted expected loss to USD 59.483, and cut the top-decile gap from 97.313 to 27.045. Yet, the MAE worsened further to 74.144.

The recalibration evidence is stronger than a grouped-decile exercise because the maps were estimated on contract-level data and evaluated on fresh holdout periods. Appendix A Table A3 shows that the key severity-power slope was stable across the calibration splits: 2.319 with a 95% interval of [2.179, 2.458] in the main split and 2.301 with [2.144, 2.458] in the alternative split. The substantive message is therefore not that recalibration fails to move the tail. It does move the tail. The message is that recalibration is mostly a level-correction device. Because the recalibration maps preserved the original ordering of the fitted values, they largely preserved the ranking while shifting the predicted levels.

Table 10. Holdout recalibration of the strict baseline, evaluated on 2023–2024.

Specification	AUC	Brier	MAE EL/ac	Obs. Freq.	Pred. Freq.	Obs. EL/ac	Pred. EL/ac
Baseline two-part	0.666	0.242	69.541	0.481	0.376	69.160	28.319
Frequency recalibration	0.666	0.230	70.931	0.481	0.517	69.160	38.833
Freq. + severity power	0.666	0.230	74.144	0.481	0.517	69.160	59.483

Notes: The calibration sample used 2020–2022 (139,248 contracts; 43,735,542 insured acres). The evaluation sample used 2023–2024 (96,749 contracts; 32,053,669 insured acres).

Table 10 should therefore be read as a level-adjustment exercise. Recalibration moved the predicted means toward the realized means in the upper deciles, but it did not create a better ranking model or a lower-MAE expected loss predictor.

Figure 2 illustrates this level-adjustment interpretation at the decile level: severity-power recalibration narrows the upper-tail gap, especially in the top decile, but does not overturn the higher-MAE result reported in Table 10.

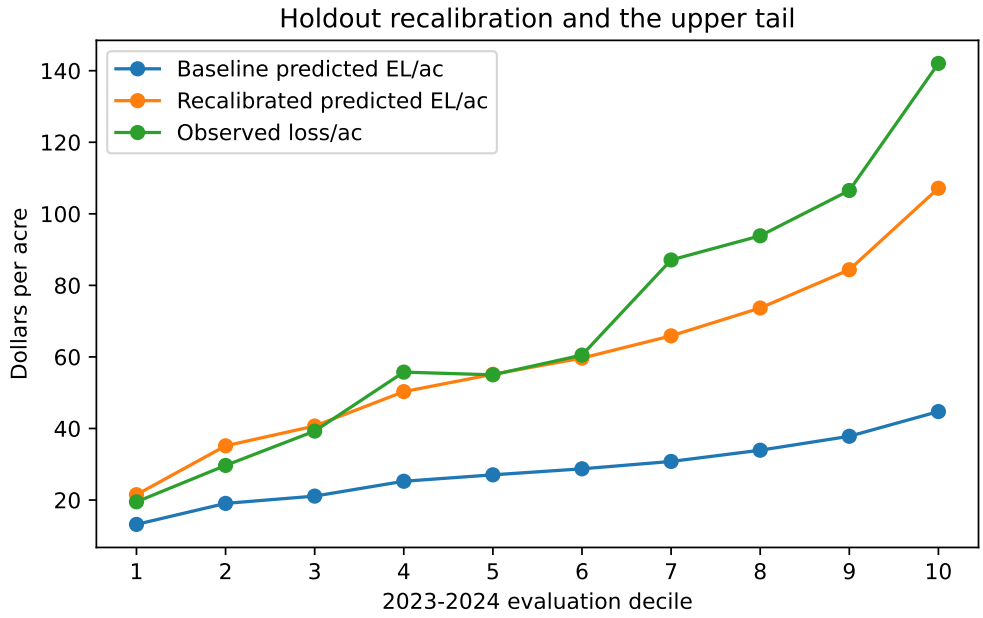

Figure 2. Holdout recalibration on 2023–2024 evaluation deciles. Frequency plus severity-power recalibration cut the top decile gap from USD 97.313 to USD 27.045 per acre, while the overall MAE worsened.

3.8. Distribution Shift Was Central and Not Peripheral

The tail problem is inseparable from the distribution shift. Table 11 shows that the evaluation period was far riskier than the training period. The loss frequency rose from 0.342 in 2006–2019 to 0.499 in 2020–2024. The mean loss per acre rose from USD 25.252 to USD 81.661. The 95th percentile of the loss per acre rose from USD 148.805 to USD 343.310,

and the mean severity conditional on loss rose from USD 73.757 to USD 163.643. This is not a small drift.

Table 12 brings that shift to the center of the paper. The predicted yearly mean was nearly flat, ranging only from USD 27.354 to USD 28.482 per acre, for a simple reason: the baseline was estimated once in the earlier years of the sample and contained no year effect or dynamic hazard updates. Thus, the yearly predicted means could move only through changes in contract composition across the evaluation years. The observed yearly mean loss, by contrast, ranged from USD 45.641 in 2020 to USD 162.297 in 2021.

The yearly table also clarifies the paper's main interpretation. In 2021, the AUC was still 0.687, but the MAE jumped to 146.926, and the top decile gap reached 234.412. In 2020, the top decile gap was only 2.371. Thus, the model could continue to rank contracts reasonably well while missing loss levels badly when the environment changed. The ordering across contracts was more stable than the year-to-year level fit.

Table 11. Distribution shift from the 2006–2019 training period to the 2020–2024 evaluation period.

Period	Loss Freq.	Mean Loss/ac	P90 Loss/ac	P95 Loss/ac	Mean Sev. If Loss
Train 2006–2019	0.342	25.252	91.166	148.805	73.757
Test 2020–2024	0.499	81.661	249.328	343.310	163.643

Table 12. Strict baseline performance by evaluation year, 2020–2024.

Year	AUC	Brier	MAE EL/ac	Pred. EL/ac	Obs. EL/ac	Top-Decile Gap
2020	0.578	0.241	51.880	27.354	45.641	2.371
2021	0.687	0.311	146.926	28.400	162.297	234.412
2022	0.676	0.216	68.911	28.188	62.638	80.875
2023	0.692	0.238	80.518	28.152	80.838	120.148
2024	0.639	0.246	58.719	28.482	57.648	73.970

Notes: Predicted EL/ac and observed EL/ac are insured acre-weighted means within each year.

3.9. Can Simple Dynamic Updating Repair the Shift?

Table 13 evaluates two simple updating devices on the same 2020–2024 window. The first kept the frozen 2006–2019 ranking model but rescaled each year's expected loss level using the previous year's observed-to-predicted mean loss ratio. This update narrowed the pooled top decile gap from USD 106.576 to USD 30.350 per acre and raised the predicted mean closer to the realized mean. It also worsened the MAE from USD 81.249 to USD 98.216 per acre because a one-year level factor overreacted to the highly variable annual loss environment.

Table 13. Simple dynamic updating diagnostics on the 2020–2024 holdout.

Updating Rule	AUC	Brier	MAE EL/ac	Pred. EL/ac	Top 20% Capture	Top-Decile Gap
Frozen 2006–2019 score	0.646	0.250	81.249	28.146	0.344	106.576
Frozen score + previous-year level update	0.646	0.250	98.216	77.705	0.344	30.350
Expanding annual refit	0.625	0.247	82.724	37.506	0.278	62.100

Notes: The level update row preserved the frozen model's ranking and rescaled the expected loss in year t by the observed-to-predicted mean loss ratio from year $t-1$; 2020 was left unscaled. The expanding refit row trained on all years before each evaluation year.

The second device refitted the strict two-part model each year using all prior years and then predicted the next year. This expanding annual refit lowered the Brier score slightly

from 0.250 to 0.247 and narrowed the top decile gap to USD 62.100 per acre. It also lowered the AUC to 0.625, lowered the top 20% capture to 0.278, and raised the MAE to USD 82.724 per acre. Simple annual updating is therefore useful as a calibration monitor, but it is not a substitute for hazard information or in-season updating when the environment changes sharply. Operationally, the evidence favors at least annual post-season recalibration checks, with stronger updating reserved for settings where current-season hazard measures can be observed before or during the underwriting review window.

3.10. Coverage Choice and Robustness

Coverage is observed at the underwriting time, but it is also chosen by producers. That dual role matters for interpretation. Table 14 shows that higher-coverage contracts within the same region–commodity–practice cells tended to have higher realized loss frequencies. For example, contracts with 0.80 coverage had higher loss frequencies than contracts with 0.70 coverage in 92.5% of comparable cells, with a mean frequency difference of 0.095. Against 0.60 coverage, the share positive was 94.9%, and the mean difference was 0.153. Coverage is therefore informative, but some of that information clearly reflects selection rather than a causal effect of coverage.

Table 14. Coverage choice descriptive evidence within region–crop–practice cells, 2006–2019 training sample.

Comparison	Cells Compared	Mean Freq. Diff.	Median Diff.	Share Positive
0.80 vs. 0.50	37	0.276	0.305	0.946
0.80 vs. 0.60	39	0.153	0.135	0.949
0.80 vs. 0.70	40	0.095	0.093	0.925

Notes: Each row compares the higher-coverage group with a lower-coverage group within the same training sample region–commodity–practice cell. The table is descriptive, not causal.

Table 15 adds a time dimension. The insured acre-weighted share of acres at 0.80 coverage rose from 0.423 in 2006–2010 to 0.771 in 2020–2024, while the mean loss per acre rose from USD 19.690 to USD 81.661. Looking across the years, the share at 0.80 coverage was 0.662 in the years following a high-loss year and 0.440 in the years following lower-loss years, where "high loss" means that the previous year's mean loss per acre exceeded the 2006–2019 median. This pattern does not identify adverse selection by itself, but it reinforces the contract risk interpretation: coverage carries information about exposure choices and producer sorting.

Table 15. Coverage dynamics and lagged-loss comparisons.

Window or Comparison	Mean Coverage	Share 0.80	Share $\geq$0.70	Loss/ac	Lag Loss/ac
2006–2010	0.720	0.423	0.798	19.690	
2011–2015	0.727	0.476	0.809	24.478	
2016–2019	0.745	0.601	0.865	32.091	
2020–2024	0.770	0.771	0.935	81.661	
Following high-loss year	0.754	0.662	0.890		51.856
Following lower-loss year	0.721	0.440	0.794		18.370

Notes: All entries were insured acre-weighted. The lagged-loss comparison classified year t by whether the mean loss per acre of year $t-1$ was above the 2006–2019 median. It is descriptive and should not be read as a causal response estimate.

Table 16 addresses the interpretive concern directly. Removing coverage lowered the AUC from 0.646 to 0.625, worsened the Brier score from 0.250 to 0.263, and raised the MAE from 81.247 to 81.609. By contrast, estimating separate models within coverage

bands yielded an AUC 0.645 and MAE of 81.196, both quite close to the full baseline. Appendix A Table A10 reports a Chiappori–Salanié-style residual correlation diagnostic. After noncoverage controls for the region, commodity, practice, and hail rate, the residual correlation between high coverage and loss remained about 0.116 in both the training and holdout samples. The operational ranking value of the strict score therefore does not come only from using coverage as a contract choice variable, but coverage remains an endogenous signal and is not interpreted causally. A deeper causal design would require an excluded determinant of coverage choice that is not also a proxy for producer risk, location, contract terms, or expected loss; the present administrative panel does not provide such an instrument, so the coverage results are deliberately descriptive and predictive.

Table 16. Coverage robustness checks on the 2020–2024 holdout.

Specification	AUC	Brier	MAE EL/ac	Top-Decile Gap
Strict underwriting-time baseline	0.646	0.250	81.247	106.581
Without coverage level	0.625	0.263	81.609	104.762
Separate models within coverage bands	0.645	0.250	81.196	105.339

For completeness, Table 17 reports the main timing and specification robustness checks. The early-season seeded-date extension posted an AUC of 0.645, Brier score of 0.250, and MAE of 81.259, essentially the same as the strict baseline. The season-complete administrative monitor posted an AUC of 0.638, Brier score of 0.254, and MAE of 81.279. These checks do not overturn the paper's central conclusion. The strict underwriting-time baseline remained the cleanest ex ante score and the right headline specification for this paper.

Table 17. Selected timing robustness checks on the 2020–2024 holdout.

Specification	AUC	Brier	MAE EL/ac
Strict underwriting-time baseline	0.646	0.250	81.247
Early-season seeded-date extension	0.645	0.250	81.259
Season-complete administrative monitor	0.638	0.254	81.279
Without coverage level	0.625	0.263	81.609
Separate models within coverage bands	0.645	0.250	81.196

4. Discussion

The results have a clear positive side. A score built only from routine enrollment-time records improved the underwriting-relevant ranking relative to simple administrative rules. That is meaningful because underwriting resources are scarce. In an operational environment where only a subset of files can be reviewed more closely, the relevant question is not whether a model perfectly predicts realized losses but whether it directs attention toward contracts that are more likely to matter. By that standard, the strict administrative baseline is useful.

The results also identify a boundary for the strict specifications tested here. The static enrollment-time variables and same-input extensions examined in the paper did not recover severe losses once the loss environment changed. The sharp increase in realized losses after 2019, combined with the near-flat yearly predicted means, shows why the model can preserve some ranking while losing level accuracy. This is precisely the kind of setting in which the distinction between ranking, reliability, resolution, and expected-loss calibration matters most. The score can still sort contracts meaningfully even when its predicted levels are too low.

The additional diagnostics refine the interpretation. Coverage is an informative contract term, but it is also a selected choice. The positive within-cell coverage loss pattern and the positive residual correlation between high coverage and loss support the contract risk interpretation and make a causal reading of the coverage coefficient inappropriate. The score is useful for triage because it uses information available to the underwriter, including the contract menu selected by the producer. It is not a structural adverse selection estimate, and it is not a basis for claiming that coverage itself causes higher agronomic risk.

The model form checks also narrow the conclusion. Gradient-boosted trees, log-normal severity, and a one-part Tweedie GLM all use the same strict enrollment-time inputs. They can trim one metric or another, especially the top decile level gap, but they do not jointly improve the ranking, calibration, and average expected loss error. The annual updating exercise leads to the same conclusion from a different direction. A previous-year level factor and an expanding annual refit are operationally simple and informative as monitoring devices, but neither is enough to stabilize performance in a sharply shifting environment. More frequent updating should therefore focus on observables that actually move within the season, such as weather, crop condition, remotely sensed vegetation, or near-real-time regional loss signals and not only on re-estimating the same static fields.

That interpretation should be qualified rather than overstated. The evidence is consistent with a bottleneck for the strict enrollment-time information set and for the same-input alternatives tested here but not a proof that all possible models built from related administrative data would fail. It does not rule out gains from producer history, farm fixed effects for repeat participants, spatial correlation models, township-level random effects, richer interaction structures, or dynamic hazard data. Some of those extensions may be useful for monitoring or pricing review, although they can be harder to apply to new producers and newly configured contracts. The study therefore identifies the boundary of a reproducible enrollment-time benchmark and not the boundary of all possible crop insurance prediction systems.

Several limitations should be kept in mind. This study used confidential administrative data from one program context, and thus replication requires licensed access, and external validity depends on the program rules, claim adjustment, crop mix, and reinsurance structure. The analysis was intentionally prospective and operational rather than structural; it did not estimate the welfare effects, optimal contract design, or a causal model of coverage demand. The same-input robustness checks were also finite; they examined representative nonlinear, tail, recalibration, and updating alternatives rather than all possible transformations of the administrative fields. The evaluation period contained only five post-2019 years. Thus, the year-block intervals were rough stability checks rather than precise asymptotic inference. Finally, this paper focused on variables available at enrollment, which is the right restriction for underwriting but also the reason why later-season hazards remained out of reach by design.

5. Conclusions

This paper evaluated a strict underwriting-time score for crop insurance using only information observed when contracts are written. The score is simple, reproducible, and operationally relevant. It is best understood as an administrative underwriting benchmark and triage device rather than as a causal farm risk model or a full dynamic claims forecasting system.

The findings have two sides, and both matter. On the positive side, routine enrollment-time variables contain enough information to improve risk ranking and support better triage than simple administrative rules in this portfolio. On the limiting side, the same variables, in the same-input alternatives tested here, did not recover severe losses in the

upper tail once the loss environment shifted. Holdout recalibration, heavier-tailed same-input severity models, and annual refitting can move predicted levels, but they do not jointly deliver a better ranking, calibration, or mean absolute error.

For practice, the implication is straightforward. A strict underwriting-time administrative score is worth using as a benchmark for review and monitoring. It should not be mistaken for a full solution to crop insurance tail risk. A practical deployment would pair the enrollment score with annual post-season calibration checks and, where available, in-season hazard updates while treating coverage as an informative but endogenous contract choice. Future research should test whether producer histories, spatial dependence, remote-sensing indicators, weather nowcasts, and township-level dynamic signals can improve the tail without sacrificing the transparency that makes the enrollment-time benchmark useful.

Author Contributions: Conceptualization, C.C., S.G. and S.I.; Software, C.C.; Validation, S.G. and S.I.; Formal analysis, C.C.; Resources, S.G. and S.I.; Data curation, C.C., S.G. and S.I.; Writing—original draft, C.C.; Writing—review and editing, S.G. and S.I.; Supervision, S.I.; Project administration, S.G. and S.I.; Funding acquisition, S.G. and S.I. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge a generous grant from the Agriculture Financial Services Corporation (AFSC) in Alberta, Canada, Grant Number: RES0001936.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The underlying administrative microdata are confidential and are not publicly available. A replication package containing the estimation scripts, table-generation scripts, and figure-generation scripts is available from the authors to credentialed researchers or other parties who are authorized to use the data, subject to the data provider's access conditions.

Acknowledgments: The authors are responsible for any errors.

Conflicts of Interest: We declare no known conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area under the receiver operating characteristic curve
Brier	Brier score for loss frequency
EL/ac	Expected loss per acre
MAE	Mean absolute error

Appendix A. Additional Appendix Tables

Appendix Table A1 reports bootstrap intervals; Appendix Table A4 reports regional performance; and Appendix Tables A5 and A6 report coefficient summaries.

Table A1. Bootstrap intervals for the strict baseline and the season-complete monitor.

Model	Metric	Point	95% Interval
Strict baseline	AUC	0.646	[0.618, 0.673]
	Brier	0.250	[0.227, 0.284]
	MAE EL/ac	81.247	[59.601, 114.433]
	Top 20% capture	0.344	[0.320, 0.367]
Season-complete monitor	AUC	0.638	[0.605, 0.665]
	Brier	0.254	[0.229, 0.289]
	MAE EL/ac	81.279	[59.629, 114.585]
	Top 20% capture	0.354	[0.326, 0.371]

Notes: These intervals were based on year-block bootstrap resampling over the five evaluation years. Because the number of year clusters was small, the table is best read together with the leave-one-year-out stability check in Table A2.

Table A2. Leave-one-year-out stability of strict-baseline benchmark comparisons.

Comparator	Metric	Favorable Omissions	Min. Diff.	Max. Diff.
Segment means	AUC	5/5	0.012	0.024
	Brier	5/5	−0.016	−0.010
	MAE EL/ac	4/5	−0.521	0.352
	Top 20% capture	5/5	0.026	0.063
Hail rate only	AUC	5/5	0.052	0.078
	Brier	5/5	−0.024	−0.017
	MAE EL/ac	5/5	−1.518	−0.848
	Top 20% capture	5/5	0.068	0.098
Premium-based	AUC	5/5	0.042	0.137
	Brier	5/5	−0.019	−0.008
	MAE EL/ac	5/5	−2.073	−0.901
	Top 20% capture	4/5	−0.001	0.102

Notes: For AUC and top 20% capture, positive differences favor the strict baseline. For the Brier score and MAE, negative differences favor the strict baseline. "Favorable omissions" counts how many of the five leave-one-year-out samples preserved the favorable sign.

Table A3. Severity-power recalibration uncertainty and split sensitivity.

Calibration and Evaluation Split	Severity Power Slope	95% Interval	Base Gap	Recal. Gap	Base MAE	Recal. MAE
2020–2022 → 2023–2024	2.319	[2.179, 2.458]	97.313	27.045	69.541	74.144
2020–2021 → 2022–2024	2.301	[2.144, 2.458]	92.249	6.776	69.335	80.337

Notes: The table focuses on the severity power slope because that is the key tail shape parameter. In both splits, the slope was well above one, the top decile gap narrowed sharply after power recalibration, and the overall MAE worsened.

Table A4. Strict-baseline performance by region.

Region	Acres Test	AUC	Brier	MAE EL/ac
Central	17,245,241	0.572	0.280	101.479
Missing	490,555	0.637	0.276	70.301
Northeast	17,958,501	0.628	0.231	51.053
Northwest	8,478,287	0.633	0.244	75.837
Peace	10,998,187	0.596	0.235	57.926
South	20,618,440	0.636	0.251	105.547

Table A5. Largest strict-baseline frequency coefficients.

Variable	Coef.
Practice: irrigated	−0.412
Region: missing	−0.382
Commodity: wheat	−0.380
Coverage level	0.294
Commodity: field peas	0.263
Region: northeast	−0.261
Commodity: barley	−0.257
Region: south	0.230
Region: peace	−0.216
Region: northwest	0.176
Commodity: oats	−0.165
Hail rate	0.101

Table A6. Largest strict-baseline severity coefficients.

Variable	Coef.
Coverage level	0.079
Hail rate	0.063
Region: northeast	−0.029
Region: south	0.021
Commodity: wheat	−0.020
Practice: irrigated	0.020
Practice: dryland	−0.020
Region: northwest	0.019
Commodity: canola	0.017
Commodity: field peas	0.017
Commodity: barley	−0.012
Region: peace	−0.011

Table A7. Additional frequency diagnostics and Brier score decomposition.

Sample	AUC acre-wtd.	AUC Contract	Brier	Reliability	Resolution	Uncertainty
Training 2006–2019	0.621	0.613	0.216	0.000	0.009	0.225
Holdout 2020–2024	0.646	0.640	0.250	0.017	0.017	0.250

Notes: The Brier score decomposition used 10 bins of predicted probabilities. Reliability is the calibration penalty, resolution is the sharpness term, and uncertainty is $\bar{y}(1-\bar{y})$.

Table A8. Heckman-style inverse Mills diagnostic for the two-part specification.

Sample	Positive Contracts	Positive Acres (m)	First-Stage AUC	Mills Coef.	z	p Value
Training 2006–2019	209,627	61.864	0.621	0.724	1.280	0.201
Holdout 2020–2024	110,621	37.820	0.669	0.369	1.210	0.226

Notes: The first stage is a probit model for loss occurrence using the strict enrollment-time controls. The second stage regresses the log positive loss on the same controls plus the inverse Mills ratio, using the insured acre weights and heteroskedasticity-robust standard errors for the reported z statistics. The test was diagnostic only because no exclusion restriction was used.

Table A9. Regularization sensitivity on a 2018–2019 validation window.

C	α	AUC	Brier	MAE EL/ac	Pred. EL/ac	Top 20% Capture
0.25	0.25	0.627	0.239	41.599	25.782	0.318
0.25	1.00	0.627	0.239	41.498	25.230	0.314
0.25	4.00	0.627	0.239	41.462	24.874	0.334
1.00	0.25	0.627	0.239	41.599	25.782	0.318
1.00	1.00	0.627	0.239	41.498	25.230	0.314
1.00	4.00	0.627	0.239	41.462	24.874	0.334
4.00	0.25	0.627	0.239	41.599	25.782	0.318
4.00	1.00	0.627	0.239	41.498	25.230	0.314
4.00	4.00	0.627	0.239	41.462	24.874	0.334

Notes: Models were trained on 2006–2017 and evaluated on 2018–2019 using the strict enrollment-time inputs. C is the logistic L_2-inverse penalty, and α is the Gamma severity penalty. The AUC and Brier score evaluated only the first-stage loss-occurrence model and therefore did not vary with α. In this validation grid, changing C altered the fitted occurrence probabilities only at numerical precision; the largest absolute change in fitted probability between $C = 0.25$ and $C = 4.00$ was approximately 1.8×10^{-6}, leaving the AUC and Brier score unchanged at the displayed precision.

Table A10. Residual coverage-selection diagnostic.

Sample	Share 0.80	Loss Freq.	Raw Corr.	Residual Corr.	Loss-Model AUC
2006–2019	0.497	0.342	0.122	0.116	0.594
2020–2024	0.771	0.499	0.129	0.116	0.654

Notes: High coverage is defined as 0.80 coverage. Residual correlations were computed after separate logistic models for high coverage and loss occurrence using the region, commodity, practice, and hail rate as controls. The diagnostic is descriptive and does not use an exclusion restriction.

References

Babcock, Bruce A., Chad E. Hart, and Dermot J. Hayes. 2004. Actuarial Fairness of Crop Insurance Rates with Constant Rate Relativities. *American Journal of Agricultural Economics* 86: 563–75. [CrossRef]

Barnett, Barry J. 2000. The U.S. Federal Crop Insurance Program. *Canadian Journal of Agricultural Economics* 48: 539–51. [CrossRef]

Brier, Glenn W. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* 78: 1–3. [CrossRef]

Chen, Tianqi, and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*. New York: Association for Computing Machinery, pp. 785–94. [CrossRef]

Chiappori, Pierre-André, and Bernard Salanié. 2000. Testing for Asymmetric Information in Insurance Markets. *Journal of Political Economy* 108: 56–78. [CrossRef]

Cohen, Alma, and Peter Siegelman. 2010. Testing for Adverse Selection in Insurance Markets. *Journal of Risk and Insurance* 77: 39–84. [CrossRef]

Coles, Stuart. 2001. *An Introduction to Statistical Modeling of Extreme Values*. London: Springer. [CrossRef]

Cragg, John G. 1971. Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica* 39: 829–44. [CrossRef]

Friedman, Jerome H. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29: 1189–232. [CrossRef]

Gneiting, Tilmann, and Adrian E. Raftery. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 102: 359–78. [CrossRef]

Goodwin, Barry K., and Olivier Mahul. 2004. *Risk Modeling Concepts Relating to the Design and Rating of Agricultural Insurance Contracts*. Policy Research Working Paper 3392. Washington, DC: The World Bank. [CrossRef]

Goodwin, Barry K., and Vincent H. Smith. 1995. *The Economics of Crop Insurance and Disaster Aid*. Washington, DC: AEI Press, ISBN 978-0-8447-3908-3.

Heckman, James J. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47: 153–61. [CrossRef]

Jørgensen, Bent. 1997. *The Theory of Dispersion Models*. London: Chapman & Hall, ISBN 978-0-412-99711-2.

Kirchner, Ella, Elinor Benami, Andrew Hobbs, Michael R. Carter, and Zhenong Jin. 2026. Get in the Zone: The Risk-Adjusted Welfare Effects of Data-Driven vs. Administrative Borders for Index Insurance Zones. *Journal of Development Economics* 180: 103658. [CrossRef]

Kull, Meelis, Telmo de Menezes e Silva Filho, and Peter Flach. 2017. Beta Calibration: A Well-Founded and Easily Implemented Improvement on Logistic Calibration for Binary Classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017*. Proceedings of Machine Learning Research. Brookline: PMLR, vol. 54, pp. 623–31. Available online: https://proceedings.mlr.press/v54/kull17a.html (accessed on 5 May 2026).

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models*, 2nd ed. London: Chapman & Hall. [CrossRef]

Miranda, Mario J., and Joseph W. Glauber. 1997. Systemic Risk, Reinsurance, and the Failure of Crop Insurance Markets. *American Journal of Agricultural Economics* 79: 206–15. [CrossRef]

Mullahy, John. 1986. Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics* 33: 341–65. [CrossRef]

Murphy, Allan H. 1973. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology* 12: 595–600. [CrossRef]

Murphy, Allan H., and Robert L. Winkler. 1987. A General Framework for Forecast Verification. *Monthly Weather Review* 115: 1330–38. [CrossRef]

Nguyen, Thuy T., Shahbaz Mushtaq, Jarrod Kath, Thong Nguyen-Huy, and Louis Reymondin. 2025. Satellite-Based Data for Agricultural Index Insurance: A Systematic Quantitative Literature Review. *Natural Hazards and Earth System Sciences* 25: 913–27. [CrossRef]

Nourali, Zahra, and Julie E. Shortridge. 2026. Predicting the Impact of Extreme Weather on Agricultural Losses in the Delmarva Peninsula Using Multi-Step Machine Learning and Financial Crop Loss Data. *Stochastic Environmental Research and Risk Assessment* 40: 13. [CrossRef]

Rothschild, Michael, and Joseph E. Stiglitz. 1976. Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information. *The Quarterly Journal of Economics* 90: 629–49. [CrossRef]

Silva Filho, Telmo de Menezes e, Hao Song, Miquel Perelló Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter A. Flach. 2023. Classifier Calibration: A Survey on How to Assess and Improve Predicted Class Probabilities. *Machine Learning* 112: 3211–60. [CrossRef]

Wijesena, Sachini, and Biswajeet Pradhan. 2025. Advancements in Weather Index Insurance: A Review of Data-Driven Approaches to Design, Pricing and Risk Management. *Earth Systems and Environment* 9: 2355–79. [CrossRef]