

4th International Conference on Media, Economy, Communication and Intelligence Management (MECI 2026)

Article

Associated factors and spatiotemporal patterns of agricultural insurance risk control using K-means clustering and anomaly detection

Ruochen Ma^{1,*}

¹ Southwestern University of Finance and Economics, Chengdu, China

* Correspondence: Ruochen Ma, Southwestern University of Finance and Economics, Chengdu, China

Abstract: Agricultural insurance risk control requires comprehensive information about persistent exposure, geographically correlated losses, and exceptional portfolio conditions to ensure financial stability. This study combines United States Department of Agriculture insurance records with National Oceanic and Atmospheric Administration climate data for 48 contiguous states during 2010–2024. The 720 state-year observations are analyzed using standardized K-means clustering, Isolation Forest screening, and spatial randomization tests; fixed-effects models use 672 observations after constructing annual lags. Silhouette selection favors two distinct profiles. The elevated-loss profile contains 19.7% of observations and 15.9% of insured liabilities but accounts for 44.4% of total indemnities. Annual loss-cost ratios exhibit positive spatial association in 14 of 15 years after multiple-testing adjustment. Among 36 observations selected under a 5% review budget, 16 are also selected by Local Outlier Factor. Regression estimates associate greater dry-side moisture pressure and higher April–September temperature anomalies with higher insured losses, while wet-side and financial associations are more sensitive to specification. Cluster agreement remains high under alternative linear scaling but weakens substantially after logarithmic compression. These results support geographically coordinated monitoring and review of exceptional observations while showing why cluster labels, detector agreement, and aggregate financial ratios cannot independently establish management effectiveness or misconduct. Ultimately, the contribution is a reproducible framework linking portfolio classification to spatial diagnostics and explicitly qualified factor analysis, providing actionable insights for policymakers and insurance providers to mitigate climate-related financial risks effectively.

Keywords: agricultural insurance; risk control; clustering; anomaly detection; spatiotemporal analysis; climate risk

Received: 26 July 2026

Revised: 13 September 2026

Accepted: 26 September 2026

Published: 01 October 2026



Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Agricultural insurance helps distribute the financial impact of uncertain production and revenue outcomes, yet the risks covered are often geographically correlated [1]. Weather-related events can simultaneously affect numerous farms, reducing the effectiveness of diversification within a regional insurance portfolio. The evolution of federal crop insurance programs reflects an ongoing balance among participation rates, financial sustainability, and the provision of protection against severe adverse events.

Dollar-denominated indemnity amounts alone are insufficient for comprehensive risk assessment. A large agricultural state may report high total payments simply due to its greater insured liability, whereas a smaller state may suffer proportionally more severe losses relative to its exposure. Prior research shows that the geographic patterns of

insurance losses vary depending on the normalization method applied. Indemnities normalized by liability offer insight into downside risk exposure, and analyses integrating insurance data with climate variables reveal spatially varying associations with temperature trends and shifting distributions of weather-related payouts [1]. These insights support a monitoring framework that jointly incorporates standardized financial metrics with explicit spatial and temporal dimensions.

Three analytical objectives remain conceptually distinct: identifying state-year portfolios with comparable financial profiles; flagging observations whose indicator combinations deviate substantially from expected patterns; and estimating how observable climate and portfolio characteristics relate to changes in insured losses across states. Clustering methods address the first objective; anomaly detection supports the second; and panel regression modeling serves the third. Applying any single method to all three questions would conflate classification, screening, and causal interpretation [1].

This study applies these complementary methods to a consistent panel dataset covering the 48 contiguous U.S. states from 2010 through 2024. Its contribution lies in integrating a reproducible portfolio classification framework with tests for geographic dependence, analysis of temporal transitions, and a separately estimated model of factors associated with loss variation [2]. Special attention is paid to administrative measurement considerations: indemnified units do not correspond one-to-one with damaged farms; premium-based and liability-based loss metrics differ meaningfully; and aggregated commodity categories may not reflect actual on-farm crop diversification. These distinctions are essential for determining whether observed statistical patterns support actionable risk management interpretations.

Here, risk control refers to the systematic monitoring and prioritization of insured financial risk [3]. The available data do not capture insurers' internal control mechanisms, claims adjustment practices, reserve levels, reinsurance arrangements, or farm-level risk mitigation activities. Consequently, the empirical findings pertain to conditions relevant to risk control rather than providing a direct evaluation of managerial performance. The analysis is retrospective and does not constitute validated forecasting of future loss outcomes.

2. Literature Review

2.1. Insurance Incentives and the Measurement of Risk

The agricultural insurance literature identifies several mechanisms that can influence observed loss experience. Adverse selection can arise when rate setting and expected farm yields fail to reflect differences in underlying risk. Moral hazard concerns changes in behavior after insurance protection is obtained; evidence concerning insurance and chemical input decisions illustrates why this mechanism requires information about farm choices. Neither mechanism can be inferred simply from a high state-level loss ratio [4]. A large payment may follow an appropriately insured systemic event, and an unusual financial observation may reflect product design or the composition of the insured portfolio.

Research on insurance and climate adaptation further emphasizes heterogeneity. Some studies associate insurance participation with greater yield sensitivity to high-temperature stress. Others examine how participation relates to multiple dimensions of warming-induced yield risk. Farm-level evidence indicates that the relationship differs across crops and becomes more nuanced when endogeneity is addressed. These studies provide substantive reasons to investigate insurance characteristics, but they do not justify assigning a causal interpretation to aggregate subsidy or product-mix coefficients [3].

2.2. Climate Exposure and Geographically Differentiated Losses

Research on insurance-loss geography emphasizes normalization, insured exposure, and distinctions among weather-related hazards. Accordingly, this study measures liability-normalized financial loss and examines seasonal climate conditions, without attributing payments to specific hazards or anthropogenic climate change [1]. State-level

patterns extend this literature by enabling whole-portfolio comparisons within a common retrospective monitoring framework.

Actual risk-reducing practices require further empirical validation. Remote sensing evidence linking cover cropping to insurance losses reveals condition-dependent associations across different weather regimes. Such indicators reflect on-farm management behavior more directly than aggregate premiums or indemnity payouts. A state-level monitoring framework can help prioritize locations for deeper investigation; however, establishing causal effectiveness necessitates granular management data and a robust identification strategy [5].

2.3. Analytical Expectations

The literature supports examining both persistent regional patterns and exceptional years [6]. Dry-side moisture stress and warmer growing seasons are expected to be positively associated with insured losses, conditional on state and year effects. Wet-side moisture pressure may also correlate with higher losses, although a statewide moisture index cannot capture localized flooding events. Subsidy shares and insurance-product composition may reflect selection behavior, contract design features, or evolving exposure profiles; therefore, their coefficient signs are left unconstrained. Geographic association is assessed directly, as a cluster of individually high-risk observations does not, in itself, confirm a coherent regional pattern.

3. DATA and Methodology

3.1. Data Sources and Sample Construction

Insurance records are obtained from the USDA Risk Management Agency State/County/Crop/Coverage Level Summary of Business files for commodity years 2010–2024. The 15 annual files contain over two million source rows. Financial amounts and reported insurance-unit counts are aggregated to the state-year level after excluding Alaska, Hawaii, and U.S. territories. The resulting sample comprises 720 observations with positive liabilities and premiums. It represents the full scope of the agency’s crop-business portfolio, including pasture and forage, specialty products, and whole-farm coverage, while explicitly excluding livestock and dairy insurance plans reported separately. This dataset should not be interpreted as limited to field crops.

The files were retrieved in September 2026. The 2024 data align with the corresponding state report issued in early September 2026. For 2010 and 2024, all 48 contiguous states were validated against official state reports across six financial and operational metrics—liabilities, premiums, subsidies, indemnities, premium-bearing units, and indemnified units—yielding perfect agreement across all 576 comparisons. Record definitions follow the standard agency layout [7]. Inspection identified 43 rows with missing liability or premium values and zero premium-bearing policies, units, and indemnities; their financial fields were treated as zero during aggregation, and the resulting state-level totals satisfied all validation checks. No active-exposure financial values were imputed, and no state-year observations were excluded as anomalies in the primary analysis.

Monthly statewide climate data are drawn from NOAA’s nClimDiv dataset, released in September 2026. State names and postal abbreviations are used to harmonize climate and insurance records, as NOAA employs state identifiers distinct from federal FIPS codes. April–September dry and wet moisture pressures are derived from monthly Palmer Drought Severity Index values. Mean temperature over the same period is expressed as a Celsius anomaly relative to each state’s 1981–2010 climatological baseline. All 720 state-year climate–insurance matches are complete. Generalized state boundaries provided by NOAA support spatial mapping and neighbor construction [8].

3.2. Variables and Measurement Choices

Let indemnity, liability, total premium, indemnified units, and units with premium be denoted by I , L , P , U_i , and U_p . The three clustering indicators are the loss-cost ratio LCR

= I/L , the indemnified-unit share $IUS = U_i/U_p$, and the premium rate $PR = P/L$. The loss ratio $LR = I/P$ is reported separately. Because $LCR = LR \times PR$, including all three financial ratios as independent clustering dimensions would introduce an exact accounting relationship [9]. The selected feature set instead combines the severity of financial losses, the breadth of indemnified administrative units, and the effective premium charged relative to liability.

IUS is an administrative share, not a probability that a unique farmer or field experiences damage [10]. Units differ across products and may overlap through coverage arrangements. Likewise, the indemnities of index insurance need not coincide with an individual operation's realized production loss. The main product-mix covariate is the liability share of Rainfall Index insurance, identified consistently by plan code 13. Its contractual basis in selected grid and time-interval indices supports treating the variable as an insurance-design descriptor.

The panel model also uses the federal subsidy-to-total-premium ratio and the logarithm of insured liability. These financial variables are lagged by one year. All financial amounts and pooled dollar shares use nominal dollars; ratios remove common monetary scaling, and year effects absorb common annual changes in the regression. They do not fully adjust for heterogeneous changes in crop prices or insured asset composition. A commodity HHI is not used because public records pool some crops into an "All Other Crops" category; in New Hampshire and Rhode Island, this category accounts for all reported liability throughout the sample. A resulting HHI of one would describe the reporting category rather than indicate agricultural specialization. Coverage levels are also excluded because their bases differ across insurance plans.

For state s and year t , dry pressure averages $\max(-PDSI, 0)$, and wet pressure averages $\max(PDSI, 0)$, over April–September [11]. Averaging these portions separately prevents a dry month from simply canceling a wet month:

$$D_{st} = \frac{1}{6} \sum_{m=4}^9 \max(0, -PDSI_{stm}) \quad , \quad W_{st} = \frac{1}{6} \sum_{m=4}^9 \max(0, PDSI_{stm})(1)$$

Both measures are dimensionless [12]. Wet pressure indicates positive moisture balance and is not a measure of extreme precipitation events. April–September is a common seasonal window for this analysis, but it does not reproduce every crop's growing season or every index insurance contract interval. Table 1 reports definitions and pooled descriptive statistics.

Table 1. Variables and Pooled State-Year Descriptive Statistics, 2010–2024.

Indicator	Definition	Unit	Mean	SD
LCR	Indemnity / liability	%	10.04	8.43
IUS	Indemnified units / units with premium	%	29.09	15.41
Premium rate	Total premium / liability	%	11.04	4.38
Loss ratio	Indemnity / total premium	Ratio	0.89	0.59
Subsidy share	Federal subsidy / total premium	%	63.75	4.99
Rainfall Index share	Plan 13 liability / total liability	%	5.30	14.18
Liability	Insured liability	USD bn	2.55	3.42
Dry pressure	Apr–Sep mean $\max(0, -PDSI)$	Index	1.00	1.34
Wet pressure	Apr–Sep mean $\max(0, PDSI)$	Index	1.14	1.42
Temperature anomaly	Apr–Sep mean minus 1981–2010 baseline	°C	0.64	0.59

Note: $N = 720$. Financial variables are from RMA [13,14]; climate variables are derived from NOAA [15,16]. Statistics give each state-year equal weight. The regression uses one-year lags of subsidy share, Rainfall Index share, and log liability.

3.3. K-Means Portfolio Classification

Each of the three clustering indicators is centered and scaled using its pooled sample mean and population standard deviation [13]. Every state-year receives equal weight to ensure the classification reflects comparable administrative portfolios rather than

replicating the dollar distribution of the national program. K-means minimizes within-group squared Euclidean distances.

$$\min \sum_{k=1}^K \sum_{(s,t) \in C_k} \|z_{st} - \mu_k\|^2 \quad (2)$$

Here, z_{st} is the standardized feature vector, C_k is cluster k , and μ_k is its centroid. The implementation uses K-means++ initialization, 50 random starts, a maximum of 300 iterations per start, and random seed 2026. Candidate values of K ranging from two to six are compared using the average silhouette coefficient. The entire time period is fitted jointly to ensure consistent cluster label interpretation across years. Labels are assigned in ascending order of mean LCR after fitting; this naming convention does not imply causal interpretation.

Stability is evaluated using the adjusted Rand index. Initialization stability is assessed with twenty alternative random seeds. An additional 200 bootstrap samples draw states with replacement and retain all 15 years for each sampled state; each replicate refits both the scaler and K-means model before assigning cluster labels to all original observations. Median and interquartile range scaling, along with a $\log(1 + 100x)$ transformation of the three ratio features, test sensitivity to geometric assumptions. Clustering is also repeated after excluding primary anomaly flags, with agreement measured only on the remaining observations. These checks characterize dependence on sampling and measurement choices, not external classification accuracy.

3.4. Anomaly Detection and Review Thresholds

Isolation Forest is fitted to the same three standardized features using 500 trees, a subsample size of 256, and seed 2026. Higher reported scores indicate greater unusualness under the fitted forest. The primary review set contains the 36 highest-scoring observations, corresponding to a 5% review budget. This budget reflects an analyst-determined screening capacity, not an estimate of the incidence of irregularities, errors, or excessive loss. Continuous scores are retained, and thresholds of 2.5% and 10% are examined as sensitivity checks [7].

Local Outlier Factor (LOF), with 20 neighbors, provides an alternative ranking based on local density. Its top-ranked sets are compared with the corresponding Isolation Forest sets by their overlap and Jaccard index—the size of the intersection divided by the size of the union. No labeled anomaly ground truth is available; consequently, agreement cannot be interpreted as precision, recall, or detection accuracy. Both models are fitted retrospectively to the complete sample.

3.5. Spatial Dependence and Temporal Transitions

Annual global Moran statistics are calculated for the local clustering ratio. Two states are considered neighbors if their generalized boundaries share an edge or a point [14]. The resulting Queen adjacency graph contains 107 undirected links and no isolated states; weights are row standardized and diagonal elements are zero. One invalid polygon in the supplied Minnesota geometry is repaired before adjacency construction. With x_{st} denoting annual state local clustering ratio and $\bar{x}_t$ its cross-state mean, the statistic is

$$I_t = \frac{n}{S_0} \frac{\sum_s \sum_j w_{sj} (x_{st} - \bar{x}_t)(x_{jt} - \bar{x}_t)}{\sum_s (x_{st} - \bar{x}_t)^2} \quad (3)$$

where $n = 48$ and S_0 is the sum of all spatial weights. For each year, 9,999 permutations of the state values generate a one-sided test for positive geographic association. The permutation p-value is $(1 + \text{the number of permuted statistics at least as large as the observed statistic})/10,000$. Benjamini–Hochberg adjustment is applied across the 15 annual tests. These global tests do not locate individually significant local hotspots. Cluster maps and adjacent-year transitions provide complementary descriptive evidence.

3.6. Panel Model of Associated Factors

The dependent variable is $\log(1 + 100LCR)$, which moderates right skew and remains defined at zero. State and year fixed effects isolate within-state variation from persistent state-level differences and common annual shocks [15]. The adjusted specification is

$$\log(1 + 100LCR_{st}) = \alpha_s + \gamma_t + \beta_D D_{st} + \beta_W W_{st} + \beta_T T_{st} + X_{s,t-1}'\theta + \varepsilon_{st} \quad (4)$$

where D and W are dry and wet pressures, T is the temperature anomaly, and X contains lagged subsidy share, Rainfall Index liability share, and log liability. Share coefficients are scaled to a 10-percentage-point change. All regression specifications use 2011–2024 because lagging removes the initial year, leaving 672 observations in the main models. No climate or financial variable is selected based on statistical significance.

Standard errors are clustered by state, with finite-sample correction and t-based 95% confidence intervals using 47 degrees of freedom. A climate-only model provides a baseline. Sensitivity specifications exclude the 36 flagged observations, replace common year effects with separate year effects for four broad geographic regions, substitute full-calendar PDSI pressures while retaining April–September temperature anomalies, and omit temperature. The regional specification absorbs shared shocks within the Northeast, Midwest, South, and West, but it is not a complete spatial-error correction. Financial controls may remain endogenous despite lagging; the coefficients describe associations and cannot identify the causal effectiveness of subsidies or insurance design.

4. Results

4.1. Aggregate Loss Experience and Spatial Association

Pooled state-year LCR averages 10.04%, with a standard deviation of 8.43 percentage points and a range of 0.50–53.35%. Its mean exceeds its median of 7.42%, indicating substantial right skew. National aggregates are calculated as ratios of summed financial amounts, rather than as averages of state ratios. Figure 1 shows that aggregate LCR peaks at 14.91% in 2012, falls to 3.89% in 2016, and reaches 11.20% in 2022. The corresponding national loss ratios are 1.570, 0.419, and 1.056. Thus, the year of the greatest liability-normalized loss intensity is not determined solely by the scale of the insured program [16].

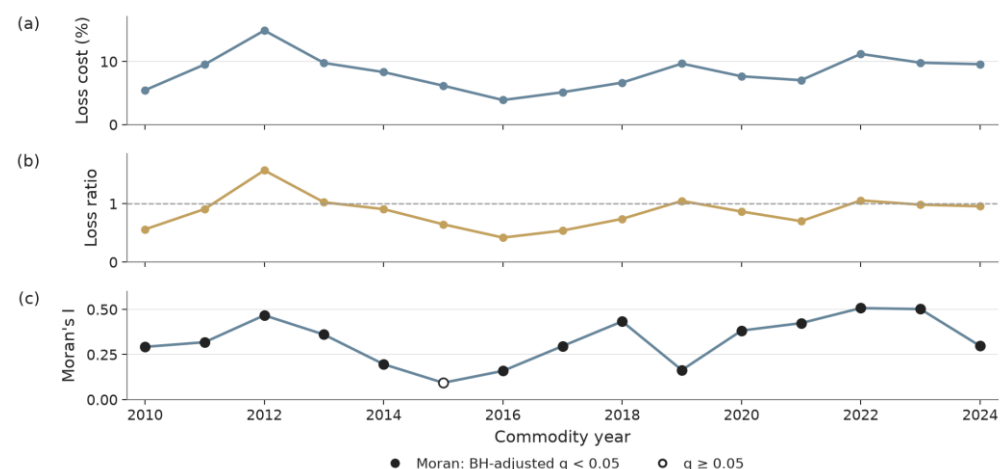


Figure 1. Aggregate Financial Loss Experience and Annual Spatial Dependence, 2010–2024. National Ratios Use Sums over the 48 Contiguous States. Moran I Uses Row-Standardized Queen Weights; Filled Markers Indicate Benjamini–Hochberg $Q < 0.05$ for One-Sided Permutation Tests. Source: Author Calculations from RMA and NOAA Data [13,17].

Positive spatial association occurs throughout the period. Moran I ranges from 0.093 in 2015 to 0.507 in 2022; 14 of the 15 annual tests yield adjusted q-values below 0.05. The 2015 value has $q = 0.0897$. The strong spatial pattern in 2022 is followed by $I = 0.501$ in 2023. These results indicate that neighboring states often exhibit similar insured loss intensity, consistent with correlated regional exposure. They do not distinguish whether the association arises from shared weather patterns, crop composition, insurance design, or other geographically structured factors.

4.2. Portfolio Profiles and Their Geographic Persistence

The silhouette coefficient is 0.527 for $K = 2$, compared with 0.375, 0.375, 0.347, and 0.313 for $K = 3, 4, 5$, and 6, respectively. The two-cluster partition therefore offers the

clearest average separation among the specified candidates. The lower-loss profile contains 578 observations, while the elevated-loss profile comprises 142. As Table 2 shows, elevated-loss observations combine a higher LCR, a greater indemnified-unit share, and a higher premium rate. These dimensions describe a financial profile; a high premium rate can reflect the pricing of risky coverage and does not constitute evidence that charging more causes losses [17].

Table 2. Financial Profiles of the Selected Two-Cluster Partition.

Measure	Cluster 1 Lower loss	Cluster 2 Elevated loss
Observations	578	142
Distinct states represented	48	33
Mean LCR (%)	6.77	23.35
Mean IUS (%)	23.69	51.10
Mean premium rate (%)	9.68	16.54
Mean loss ratio	0.736	1.508
Share of pooled liabilities (%)	84.07	15.93
Share of pooled indemnities (%)	55.55	44.45

Note: Means are unweighted across state-years. Dollar shares use pooled financial totals. A state can appear in both clusters in different years. Classification uses LCR, IUS, and premium rate; loss ratio is descriptive only.

Although the elevated-loss profile accounts for only 19.7% of state-year observations, it encompasses 15.9% of pooled liabilities and 44.4% of pooled indemnities. Its mean LCR is 23.35%, compared with 6.77% for the lower-loss profile [18]. These cluster averages assign equal weight to each observation, whereas the liability and indemnity shares utilize dollar totals. This distinction is crucial when transitioning from a profile classification to an assessment of national financial exposure.

Figure 2 illustrates that elevated-loss membership is persistent in certain states and episodic in others. Texas enters this profile in 14 of 15 years, Oklahoma in 13, New Mexico in 12, and Nevada in 11. In contrast, 15 states never enter it during the study period. The number of elevated-loss states increases from two in 2010 to 16 in 2024, with intermittent fluctuations [19]. This represents a descriptive shift in the pooled classification rather than an estimate of deterioration in management quality. Product diversification and changing exposure can contribute to this pattern.

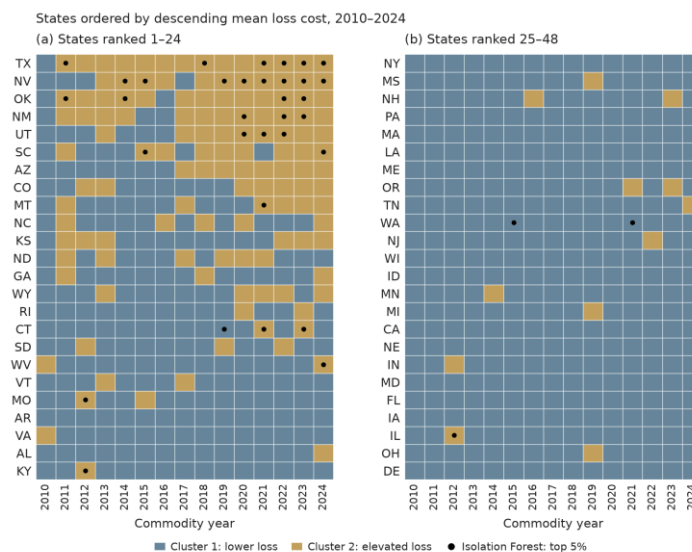


Figure 2. State-year Portfolio Profiles and Anomaly Flags. States Are Ordered by Their Mean LCR; the Order Continues from the Left Panel to the Right. Colors Show the Common Full-Period K-

means Classification. Black Dots Identify the Top 36 Isolation Forest Scores. A Dot Represents Review Priority, Not a Confirmed Data Error or Improper Claim. Source: Author Calculations from RMA Data [13].

Among 672 adjacent-year state transitions, 484 remain in the lower-loss profile, 78 remain in the elevated-loss profile, 62 transition from lower to elevated loss, and 48 move in the opposite direction. Conditional persistence is therefore 88.6% for the lower-loss profile and 61.9% for the elevated-loss profile. These transition counts document historical persistence and do not represent estimated probabilities for future forecasting. Figure 3 contextualizes this persistence alongside the geography of anomaly screening.

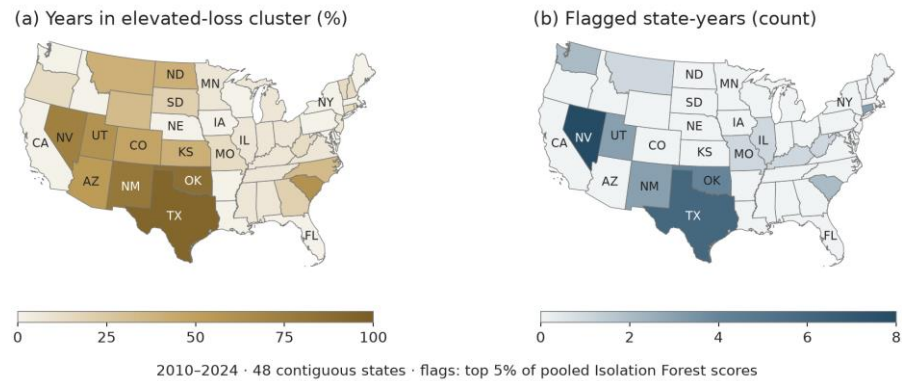


Figure 3. Geographic Persistence and Screening Frequency. The Left Map Shows the Fraction of 2010–2024 Observations Assigned to the Elevated-Loss Profile; the Right Shows the Number of Top-5% Isolation Forest Flags. Generalized Boundaries Support State Comparisons and Do Not Denote Statistically Significant Local Hotspots. Source: Author Calculations from RMA Records and NOAA Boundaries [13,17].

4.3. Exceptional Observations and Detector Agreement

Isolation Forest identifies 33 observations from the elevated-loss profile and three from the lower-loss profile. Nevada accounts for eight flagged cases, Texas for six, and Oklahoma for four. Thus, anomaly detection correlates with the higher-loss classification but does not replicate it. An observation may be flagged as unusual due to its distinctive combination of indicators, even when its loss cost ratio remains comparatively low.

Table 3 presents the eight observations with the highest Isolation Forest scores. Texas in 2022 exhibits the largest loss cost ratio in the sample—53.35%—alongside an indemnified-unit share of 73.79% and a premium rate of 24.92%. Nevada in 2020 and 2022 also displays unusually high indemnified-unit shares paired with elevated premium rates. The dry-pressure measure is 5.27 for Texas in 2022, 3.00 for Nevada in 2020, and 3.63 for Nevada in 2022. These independently observed moisture conditions align with drought-related conditions; however, the analysis does not attribute payments to specific perils or quantify the proportion attributable to drought.

Table 3. Eight Highest-Ranked Isolation Forest Observations.

State-year	LCR (%)	IUS (%)	Premium rate (%)	Loss ratio	IF score	LOF rank
TX 2022	53.35	73.79	24.92	2.141	0.744	7
NV 2020	47.43	79.96	24.84	1.909	0.742	29
NV 2022	44.03	78.40	26.42	1.667	0.737	34
OK 2022	51.58	70.61	23.04	2.239	0.713	22
NV 2015	41.51	80.72	14.96	2.774	0.702	10
UT 2020	44.04	77.70	20.01	2.201	0.690	62
TX 2011	47.25	70.48	20.05	2.357	0.673	57

NV 2024	30.41	66.31	26.25	1.158	0.658	26
---------	-------	-------	-------	-------	-------	----

Note: IF = Isolation Forest; larger scores indicate greater unusualness. LOF rank 1 is most unusual under Local Outlier Factor; ranks 1–36 enter its 5% review set. All scores are retrospective and are not calibrated probabilities.

At the 5% review threshold, Isolation Forest and Local Outlier Factor jointly flag 16 of 36 observations, yielding a Jaccard index of 0.286. Overlap occurs for three of 18 observations at the 2.5% threshold and 34 of 72 at the 10% threshold [20]. This partial agreement is substantively meaningful: global isolation and local density capture distinct manifestations of unusualness. Selecting only jointly flagged cases would reduce the primary review set from 36 to 16, yet the available data do not support a conclusion that such filtering improves detection accuracy. A practical review system should maintain both rankings separately and record the rationale for each selection.

4.4. Factors Associated with Insured Financial Losses

Table 4 separates the climate-only model from the adjusted model and the regional-year sensitivity specification. In the adjusted model, dry pressure has a coefficient of 0.160 with a 95% confidence interval of [0.107, 0.213]. The temperature anomaly coefficient is 0.200 [0.087, 0.314], and wet pressure is 0.061 [0.014, 0.108]. These positive associations remain after controlling for persistent state characteristics, common year shocks, and lagged financial descriptors [21]. Their units are changes in $\log(1 + 100\text{LCR})$, so the coefficients should not be interpreted as direct percentage-point changes in LCR.

Table 4. Fixed-effects Associations with $\log(1 + 100\text{LCR})$.

Associated factor	Climate only	Adjusted	Region-year effects
Dry pressure, 1 index unit	0.148 [0.096, 0.200]	0.160 [0.107, 0.213]	0.135 [0.087, 0.183]
Wet pressure, 1 index unit	0.072 [0.022, 0.121]	0.061 [0.014, 0.108]	0.051 [0.001, 0.102]
Temperature anomaly, 1°C	0.236 [0.114, 0.358]	0.200 [0.087, 0.314]	0.247 [0.133, 0.361]
Lagged subsidy share, 10 pp	-	-0.398 [-0.693, -0.104]	-0.286 [-0.581, 0.008]
Lagged Rainfall Index share, 10 pp	-	-0.032 [-0.104, 0.039]	-0.033 [-0.089, 0.023]
Lagged log liability, 1 log unit	-	0.306 [0.058, 0.554]	0.106 [-0.179, 0.392]
Observations	672	672	672
State fixed effects	Yes	Yes	Yes
Common year effects	Yes	Yes	Replaced
Overall R ²	0.580	0.602	0.687

Note: Brackets contain state-clustered 95% confidence intervals using $t(47)$. No significance stars are used. The regional specification replaces common year effects with region-specific year effects. Financial regressors are lagged one year; pp = percentage points. Overall R² includes fixed effects.

The lagged subsidy-share coefficient is -0.398 [-0.693, -0.104] per 10-percentage-point difference in the main adjusted model [5, 15]. The lagged log-liability coefficient is positive, 0.306 [0.058, 0.554], while the Rainfall Index share coefficient is -0.032 [-0.104, 0.039]. These estimates do not imply that subsidies reduce risk or that expanding liability increases it. Subsidy shares reflect coverage levels and product composition, and liability may vary with prices, participation rates, and the value of insured production. Unobserved within-state changes may jointly influence financial covariates and subsequent loss experience.

Regional-year effects preserve the positive dry-pressure and temperature associations but weaken the evidence for the financial covariates. The subsidy interval becomes [-0.581, 0.008], and the liability interval becomes [-0.179, 0.392]. The wet-

pressure interval remains just above zero, [0.001, 0.102]. This comparison places greater weight on the climate associations than on management interpretations of the subsidy or scale coefficients. The adjusted model's overall R^2 is 0.602; the substantive regressors reduce residual variation by 20.5% relative to a model containing only the same state and year effects. Neither measure constitutes an out-of-sample prediction score.

4.5. Sensitivity and Limits of Classification Stability

Repeated initialization yields identical primary classifications across 20 random seeds. State-block bootstrap agreement has a median adjusted Rand index (ARI) of 0.896, with empirical 2.5th and 97.5th percentiles of 0.652 and 1.000. Median and interquartile range scaling also preserves much of the partition, yielding $ARI = 0.937$. However, logarithmic compression reduces agreement substantially ($ARI = 0.270$), and refitting after removal of the 36 flagged observations yields $ARI = 0.486$ on the retained sample. These results indicate that the primary partition depends strongly on the absolute separation of high-loss observations. Consequently, initialization stability is greater than invariance to the substantive definition of distance (As shown in Table 5).

Table 5. Cluster Stability and Anomaly-Screening Sensitivity.

Sensitivity check	Result
20 alternative initializations	Minimum ARI = 1.000
200 state-block bootstrap samples	Median ARI = 0.896; central 95% range 0.652–1.000
Median/IQR scaling	$ARI = 0.937$
Log(1 + 100x) feature transformation	$ARI = 0.270$
Clustering after excluding 36 IF flags	$ARI = 0.486$ on 684 retained observations
IF–LOF overlap at 2.5% review budget	3 / 18; Jaccard = 0.091
IF–LOF overlap at 5% review budget	16 / 36; Jaccard = 0.286
IF–LOF overlap at 10% review budget	34 / 72; Jaccard = 0.309

Note: ARI compares an alternative partition with the primary $K = 2$ partition. Bootstrap percentiles describe the replicate distribution and are not confidence bounds on classification accuracy. Detector overlap has no known ground-truth interpretation.

Regression sensitivity is more consistent for dry pressure than for wet pressure. Removing flagged observations leaves 636 regression observations and yields dry and temperature coefficients of 0.153 and 0.175, respectively, with confidence intervals entirely above zero [16]. Using full-calendar PDSI produces a dry coefficient of 0.168 [0.104, 0.232], whereas the wet interval includes zero. Omitting temperature yields a dry coefficient of 0.198 [0.151, 0.245]; the wet interval again narrowly includes zero. The dry-side association is the most consistent finding across all specified checks. Temperature remains positively associated wherever included, but its coefficient cannot represent the isolated effect of temperature alone because PDSI also reflects climate-driven moisture balance.

5. Discussion

5.1. Implications for Risk Monitoring and Control

The findings support differentiated responses to persistent patterns and exceptional observations. Repeated elevated-loss membership can inform assessment of exposure distribution, contract composition, and geographically correlated payment scenarios. Texas, Oklahoma, New Mexico, and Nevada illustrate why such monitoring should extend beyond a single adverse year. Positive spatial dependence in most years also supports joint examination of neighboring portfolios when constructing stress scenarios. The present estimates cannot determine appropriate reserve levels or reinsurance coverage due to the absence of relevant contractual and balance-sheet data.

Anomaly screening serves a narrower operational purpose. A high-ranked state-year warrants verification of whether its premium rate, reported-unit structure, and payment

intensity align with underlying products and conditions. The limited overlap between Isolation Forest and LOF underscores the need for transparent, detector-specific review rationales. Automatically interpreting a flag as evidence of poor management or misconduct would conflate financial extremity, administrative configuration, and behavioral indicators. State-level aggregates can identify areas requiring scrutiny, but county-, policy-, and claim-level detail is essential to resolve underlying causes.

The factor analysis points toward climate-informed monitoring while offering limited support for direct policy prescriptions derived from financial coefficients. Dry-side moisture pressure remains associated with losses across the specified models. This supports incorporating independently measured moisture conditions into portfolio reviews. It does not imply that insurance can prevent drought, nor that reducing indemnity payments necessarily enhances protection. A well-functioning contract may trigger substantial payments precisely when protection is most needed. Complementary practices such as cover cropping require direct measurement and evaluation using methodologies established in applied agricultural management research.

5.2. Measurement and Methodological Implications

A central finding is that classification quality depends on the definition of the features [11, 17]. Absolute standardized ratios emphasize large financial departures, whereas logarithmic transformation gives relatively more emphasis to proportional differences among lower values. The resulting disagreement is not resolved by designating one partition as definitive; rather, it reflects a methodological choice that must be aligned with the monitoring objective. The primary analysis uses absolute standardized separation because exceptional financial severity is relevant to the task, while the sensitivity results ensure the profile boundaries are not interpreted as fixed natural categories.

Administrative aggregation and variations in insurance-unit definitions necessitate a compact set of interpretable indicators. Data dictionaries and reconciliation against published totals are therefore integral to establishing empirical validity.

5.3. Scope and Limitations

The state serves as a practical reporting unit but masks substantial heterogeneity across counties, crop types, and insurance contracts. State-level climate metrics are area-weighted rather than weighted by the actual distribution of insured agricultural exposure. This discrepancy is especially consequential for large states, high-value specialty crops, off-season (e.g., winter) production systems, and index-based insurance products with narrowly defined temporal windows. Consequently, aggregate portfolio-level findings must not be directly applied to individual crops, farms, insurers, or specific insurance programs.

Several methodological constraints also persist. State-clustered standard errors account for within-state error correlation but cannot capture all forms of cross-state dependence; regional-year effects offer only limited robustness verification [13, 18]. Fixed effects and lagged variables do not fully address time-varying selection into insurance coverage. The anomaly screening threshold reflects an explicit operational budget constraint, and no independently verified ground-truth labels exist for anomalies. Pooled scaling and full-period model fitting render outputs inherently retrospective: prospective deployment would require freezing preprocessing parameters from the training period, conducting rigorous temporal validation, establishing recalibration protocols, and evaluating associated review resource requirements. Furthermore, continuous administrative updates to underlying data necessitate preservation of archived extractions and source checksums to ensure exact reproducibility of reported numerical results.

6. Conclusion

This study integrates K-means clustering, anomaly detection, spatial diagnostics, and panel regression for 720 state-year observations of United States agricultural insurance

during 2010–2024. Two financial profiles provide the strongest silhouette separation. The elevated-loss profile contains 19.7% of observations but accounts for 44.4% of indemnities; several western and southern states appear in this profile repeatedly. Annual loss-cost ratios display positive spatial association in 14 of 15 years after adjustment for multiple testing.

Anomaly detection identifies a review set that overlaps only partially with an alternative local-density ranking. Associations with dry-pressure and temperature remain consistent across specified regression checks, whereas wet-side and financial findings show greater sensitivity. The principal implication is to combine regional exposure review, transparent exception screening, and climate information when monitoring insured risk. The results do not establish causal policy effects, validated forecasts, or managerial performance rankings. Reporting measurement limitations and sensitivity alongside primary estimates is necessary for responsible application of these methods in agricultural insurance risk control.

Data and Code Availability: The analysis uses publicly accessible RMA and NOAA records. The accompanying reproducibility files contain the analysis-ready panel, aggregate source tables, download manifests with SHA-256 checksums, spatial weights, and scripts that regenerate the reported statistics and figures. The extraction date is 11 September 2026. Python 3.12.14, pandas 2.2.3, scikit-learn 1.9.1, and statsmodels 0.15.0 were used. Data availability does not imply that later revisions of the public files will reproduce an earlier extraction byte for byte.

References

1. J. J. Reyes and E. Elias, "Spatio-temporal variation of crop loss in the United States from 2001 to 2016," *Environmental Research Letters*, vol. 14, no. 7, p. 074017, 2019.
2. M. J. Miranda and J. W. Glauber, "Systemic risk, reinsurance, and the failure of crop insurance markets," *American Journal of Agricultural Economics*, vol. 79, no. 1, pp. 206–215, 1997.
3. J. W. Glauber, "The growth of the federal crop insurance program, 1990–2011," *American Journal of Agricultural Economics*, vol. 95, no. 2, pp. 482–488, 2013.
4. E. D. Perry, J. Yu, and J. Tack, "Using insurance data to quantify the multidimensional impacts of warming temperatures on yield risk," *Nature Communications*, vol. 11, no. 1, p. 4542, 2020.
5. N. S. Diffenbaugh, F. V. Davenport, and M. Burke, "Historical warming has increased US crop insurance losses," *Environmental Research Letters*, vol. 16, no. 8, p. 084025, 2021.
6. L. R. Bundy, V. A. Gensini, and M. S. Russo, "Insured corn losses in the United States from weather and climate perils," *Journal of Applied Meteorology and Climatology*, vol. 61, no. 8, pp. 969–988, 2022.
7. V. H. Smith and B. K. Goodwin, "Crop insurance, moral hazard, and agricultural chemical use," *American Journal of Agricultural Economics*, vol. 78, no. 2, pp. 428–438, 1996.
8. F. Annan and W. Schlenker, "Federal crop insurance and the disincentive to adapt to extreme heat," *American Economic Review*, vol. 105, no. 5, pp. 262–266, 2015.
9. J. R. Skees and M. R. Reed, "Rate making for farm-level crop insurance: Implications for adverse selection," *American Journal of Agricultural Economics*, vol. 68, no. 3, pp. 653–659, 1986.
10. R. Wang, R. M. Rejesus, and S. Aglasan, "Warming temperatures, yield risk and crop insurance participation," *European Review of Agricultural Economics*, vol. 48, no. 5, pp. 1109–1131, 2021.
11. S. Aglasan, R. M. Rejesus, S. Hagen, and W. Salas, "Cover crops, crop insurance losses, and resilience to extreme weather events," *American Journal of Agricultural Economics*, vol. 106, no. 4, pp. 1410–1434, 2024.
12. M. Regmi, J. Tack, and A. M. Featherstone, "Does crop insurance influence crop yield impacts of warming temperatures? A farm-level analysis," *Journal of the Agricultural and Applied Economics Association*, vol. 2, no. 4, pp. 808–822, 2023.
13. D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, vol. 7, pp. 1027–1035, Jan. 2007.
14. S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.
15. L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.
16. P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.
17. F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, Dec. 2008.
18. M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104, May 2000.

19. P. A. Moran, "Notes on continuous stochastic phenomena," *Biometrika*, vol. 37, no. 1/2, pp. 17–23, 1950.
20. A. Colin Cameron and D. L. Miller, "A practitioner's guide to cluster-robust inference," *Journal of Human Resources*, vol. 50, no. 2, pp. 317–372, 2015.
21. Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

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