When the Average Doesn't Exist: Why Learned Insurance Payouts Fail Under Distribution Shift

Anonymous authors
Paper under double-blind review

Abstract

Parametric insurance is the leading hope for extending climate cover to people conventional insurance cannot reach: a satellite reading crosses a threshold and the money moves, with no inspector and no delay. Everything then rests on how closely the index tracks the loss, and machine learning is being recruited to close that gap. We show that at the granularity where 44% of real contracts settle, the quantity being predicted has no average to predict. On the complete US federal flood record (2.1 million claims) an individual claim is heavy-tailed but tractable (Hill tail index $\alpha = 2.43$). Aggregate to the county-month totals that area-index and sovereign products actually pay on, and α is 0.75. Below 1 there is no finite mean to estimate: the running average never settles over any range a claims history can observe, a single event holds 10% of all losses, and further pooling, across counties or states, does not recover it. Because this is a property of the target, it binds every model. Payment schemes defined through an expected loss, including the expectile-optimal designs proposed for this exact problem, are undefined here. For the quantile schemes that survive, we derive and verify a regret law capping how far a finite history lets a payout lean toward the policyholder before estimation error dominates. Can a model succeed regardless? We answer by elimination: six configurations, three model families, ten seeds, three regimes. In-distribution, gradient boosting genuinely works: it beats holding cash in 10 of 10 seeds and cuts basis risk by 19%. Under temporal or climate shift, nothing does, in any of 110 runs: not a budget constraint, a frequency–severity decomposition, quadrupled capacity, a seed ensemble, an oracle rescaling, or a real natural-capital dependency graph, which performs no better than a scrambled one. What makes the result actionable is its cause. The driver is not the hazard but the unevenness of the exposure being pooled: normalising by policies in force lifts $\hat{\alpha}$ from 0.71 to 1.00 on the matched subset. The binding constraint is therefore a contract parameter, not a fact about weather, i.e., the same losses, measured per unit of exposure, are estimable. The lever is therefore, we conclude, not the model; It is what a contract insures, and at what level it pools.

1 Introduction

In the 2015/16 season, Malawi suffered an El Niño drought that left some 6.5 million people needing food aid. The government held a parametric drought policy with African Risk Capacity, designed so that relief would not wait on a manual damage assessment. The policy did not trigger. Its crop model had been written around long-cycle maize, which largely survived the drought; most Malawian farmers had planted a shorter-cycle variety, far more sensitive at flowering, and the index could not see the mismatch. ARC members subsequently agreed an $8.1M payment — the amount Malawi would have received had short-cycle maize been the contractual base — which reached the country in January 2017, roughly a year after the harvest failed (ActionAid International, 2017). In a registry we compiled of 84 confirmed parametric and index insurance products spanning more than 100 countries and territories (Appendix A), 33 show a documented pattern of under-payment relative to independently verified loss, and a further 23 carry a structural under-payment risk that no specific event has yet tested.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The gap between what an index pays and what was actually lost is *basis risk*, and it is the principal obstacle to scaling parametric insurance to the populations that most need it (Jensen & Barrett, 2017). The field's response has been to build better indices: denser reference stations, more sensors, and lately supervised models over multi-modal Earth observation. With EO now providing global, continuous covariates, a single learned basis-risk model — trained on claims data, deployed wherever the satellite signal reaches — has become an explicit aim of the sustainable finance community.

We argue that this aim has never been posed as a learning problem. Before asking which architecture to use, three questions come first. Does the quantity a payment scheme is supposed to match exist as a population object? Is it estimable from the claims histories that exist? And does an estimator fitted in one regime transfer to another, which is the whole premise of deploying one model globally? This paper answers all three, and the answers are more restrictive than the field's current trajectory assumes.

The insured loss has no mean. On the complete NFIP record, individual claim severities have Hill tail index $\alpha = 2.43$ — heavy-tailed, but with finite mean and nearly finite variance. Aggregate them to the county-month totals on which area-index and sovereign products settle, and $\alpha = 0.75$: below the threshold at which the expectation ceases to exist (§3). Controls attribute this to aggregation over a heavy-tailed claim count rather than to event dependence, and further pooling — to state, to nation, or across randomized geography — does not restore it. Which side of $\alpha = 1$ a contract sits on is set by its risk-pooling level, a design choice; 44% of catalogued products sit on the wrong side.

Two consequences follow before any model is fit. Payment schemes defined through expectations of the loss, including the expectile-optimal schemes recently proposed for this exact problem (Maier & Scherer, 2026), are undefined at that granularity (§4). And for the quantile schemes that survive, we derive and verify a regret law showing that a finite claims history supports only a bounded degree of shortfall-aversion: past that bound, caring more about the policyholder pays them worse (§5).

We then ask what is learnable, and answer by elimination. Six configurations across three model families — a recurrent neural model and three principled repairs of it, a gradient-boosted ensemble, and a fixed-indemnity product that learns no magnitude at all — are evaluated across in-distribution, temporal-shift and climate-shift holdouts, ten seeds each (§6). In-distribution, boosting beats holding cash in 10 of 10 seeds. Under either shift, *no configuration beats cash in any run*. Nor does domain knowledge help: a real natural-capital dependency graph, coupled four ways, contributes nothing a randomly scrambled graph does not (§7).

Contributions.

1. The tail index of real catastrophe losses as a function of aggregation in time and space, on 2.1M federal claims, with controls separating claim-count heterogeneity, event dependence and geographic correlation (§3); and a registry placing 44% of 84 real products in the regime where the mean does not exist (Appendix A).
2. Two formal consequences: expectile-optimal schemes are undefined there (§4), and a derived, numerically verified regret law bounding shortfall-aversion (§5).
3. A learnability characterization across six configurations spanning three model families and three regimes (§6): in-distribution basis risk is learnable; under regime shift it is not, for any configuration we tested.
4. A well-powered null on neurosymbolic grounding, with a content-specificity control the literature rarely applies (§7).

2 Background and Related Work

A parametric contract specifies an observable index and a threshold. When the index crosses it, a pre-agreed sum is paid with no loss adjustment. This eliminates the settlement delay that makes indemnity insurance impractical in low-capacity markets, and removes moral hazard and claims fraud with it. Everything then rests on how faithfully the index tracks the loss it stands in for.

Index insurance and basis risk. The economics literature establishes basis risk as a quantified barrier to adoption and welfare gain (Chantarat et al., 2013; Jensen & Barrett, 2017), and shows that combining index scales reduces it without abandoning the automatic trigger (Elabed et al., 2013).

Maier & Scherer (2026) give the closest antecedent to our decision-theoretic material: they show expectiles are the basis-risk-optimal payment scheme under an asymmetric quadratic criterion, and note that shortfall-only criteria are degenerate because they reward over-payment. Our contribution is not to rediscover that but to observe that their optimal object requires a moment the data at contract granularity does not have, and to supply the estimation-side result their known-distribution setting cannot express. Constrained allocation under a budget is classical (Hadley & Whitin, 1963): a binding budget moves every product to a common critical fractile, so we treat total payout as a reported statistic rather than a free parameter.

Heavy tails and elicitability. Regularly varying loss distributions are standard in reinsurance (Embrechts et al., 1997; Mikosch, 2009), and the Hill estimator (Hill, 1975) is the standard tail-index descriptor, with well-documented sensitivity to the number of order statistics used (Danielsson et al., 2001; Clauset et al., 2009); we therefore report it alongside the moment estimator of Dekkers et al. (1989) and a block bootstrap (Beirlant et al., 2004; Resnick, 2007). The statistical-learning literature on heavy tails concerns estimating means robustly under infinite variance (Catoni, 2012; Lugosi & Mendelson, 2019); our setting is one step more severe, in that the mean itself is not a usable population object over the observable range, so the question becomes which functional to target rather than how to estimate a given one. Quantiles are elicited by the pinball loss (Koenker & Bassett, 1978) and expectiles by asymmetric least squares (Newey & Powell, 1987); Gneiting (2011) catalogs which functionals are elicitable by which losses. We use these tools rather than extend them.

Distribution shift and tabular learning. Transfer failure under covariate and label shift is a mature subject (Quiñonero-Candela et al., 2009; Koh et al., 2021); what is specific here is that the shift moves probability mass into a tail region the training sample cannot resolve at any density, so the failure is a property of the target rather than of the covariate map. That a gradient-boosted ensemble (Friedman, 2001; Ke et al., 2017) outperforms a recurrent network on the same features is consistent with tree ensembles remaining strong on heterogeneous tabular data (Grinsztajn et al., 2022), and we report it as such rather than as a claim about architectures in general.

Multi-modal EO learning and neurosymbolic AI. Deep learning on Earth observation is established (Reichstein et al., 2019), with gated fusion over heterogeneous streams standard practice (Baltrušaitis et al., 2019) and missing-modality robustness an active area (Wu et al., 2024). Grounding neural representations in expert symbolic structure is a recognized route to domain-consistent representations (d'Avila Garcez & Lamb, 2020); recent surveys categorize knowledge-graph coupling into logically-informed embeddings, embeddings with logical constraints, and rule learning (DeLong et al., 2024), and our mechanisms fall in the middle category. Methodologically we follow studies that test a popular direction rather than extending it (Lucic et al., 2018; Liu et al., 2019), with the seed discipline Henderson et al. (2018) showed to be necessary.

3 The insured loss behaves as if it has no mean

We work throughout with the complete NFIP loss-row record: 2,721,780 claim rows across 2,941 county-level entities, of which 2,114,319 carry a strictly positive indemnity. Each row is one property's claim, though its location is redacted to the county, which will matter later for what we can and cannot supervise. Tail indices are Hill estimates (Hill, 1975) at the top 5% of each series, reported with the asymptotic interval $\hat{\alpha} \pm 1.96\,\hat{\alpha}/\sqrt{k}$.

Before the numbers, a clarification we think is owed. Every NFIP claim is bounded by a policy limit, so the population moments do exist and "the loss has no mean" is, read literally, false. What we claim is the weaker and more useful statement that the aggregate *behaves as infinite-mean over the range a finite claims history can observe*: the running sample mean never settles, and the largest single observation keeps a stubbornly non-vanishing share of the total. That is measurable rather than rhetorical, and Figure 2B measures it — across 35,227 county-months the largest single event accounts for **10.2%** of the entire sum, against 1.0% per policy and 0.01% for individual claims. An estimator fitted on such a sample is, to a first approximation, fitting one hurricane.

With that said, Figure 1 reports the estimates, and Appendix B tabulates them with k, a moment estimator and block-bootstrap intervals. Individual claims sit at $\hat{\alpha} = 2.43$ (95% CI $[2.415, 2.444]$, $n = 2{,}114{,}319$) — heavy-tailed, but comfortably within the regime where ordinary statistical practice applies. Every aggregate falls below 1: county-month at $\hat{\alpha} = 0.747$ $[0.726, 0.768]$, county-day

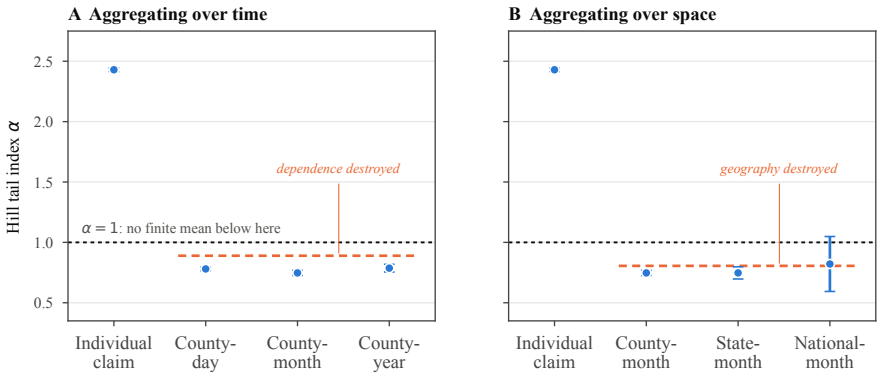

Figure 1: **Aggregation removes the mean, and no amount of further pooling brings it back.** Hill tail index with 95% intervals. **(A)** widening the time bucket with the pool fixed at one county; dashed line destroys within-event dependence while preserving claim counts. **(B)** widening the pool with time fixed at one month; dashed line reassigns counties to random pseudo-states, destroying geographic adjacency while holding pool size fixed. The transition happens once, at the first aggregation step, and neither more time nor more space recovers it.

at 0.780, county-year at 0.787. The gap between those two numbers is the whole paper, so it is worth asking what produces it.

Our first guess was event dependence. Claims within one county-month share a cause — one storm — so their severities are positively dependent, and summing dependent heavy-tailed variables fattens the tail. That turns out to be the smaller effect. The claim *count* is itself heavy-tailed: some county-months carry a single claim, others tens of thousands, and summing independent severities over a heavy-tailed count already produces a heavy-tailed total. Shuffling indemnity values across all claims while holding county-month group sizes fixed — destroying dependence, preserving the count distribution — recovers $\hat{\alpha} = 0.889$ (sd 0.005), most of the way back but not all. Aggregation over heterogeneous counts therefore carries the bulk of the drop from 2.43, with event dependence adding a further clear increment from 0.889 to 0.747. Both mechanisms are real; the first is the larger, which is not what we expected going in.

If aggregation is what removes the mean, a natural next question is whether more of it helps — pooling is, after all, the founding mechanism of insurance. It does not. Aggregation in *space* behaves the same way (Figure 1B): holding the time bucket at one month, $\hat{\alpha}$ is 0.747 at county, 0.747 at state (pooling 2,888 counties into 55), and 0.821 nationally — though that last rests on 581 months and its interval spans 1. A second control tests the obvious explanation: reassigning counties to *pseudo-states* at random, preserving pool count and size but destroying geographic adjacency, gives $\hat{\alpha} = 0.805$ (sd 0.006) — a real improvement over 0.747, and nowhere near 1.

This is what heavy-tail theory predicts rather than a surprise: for $\alpha < 1$, sums of independent copies preserve the tail index, growing like $n^{1/\alpha}$ rather than n, so averaging cannot recover a finite mean even across independent risks (Embrechts et al., 1997). The empirical content is that it holds on real claim records at three pooling scales, and the consequence is concrete: *spatial diversification does not move the statistical regime*. Sovereign and regional risk pools — ARC, CCRIF, PCRIC — exist partly on a diversification rationale; pooling delivers liquidity, speed and cross-subsidy, all real, but not a loss distribution with a well-defined mean.

Is it just exposure heterogeneity? The most serious alternative explanation is that the claim count is heavy-tailed because counties differ in size: pooling Harris County with a rural county produces a heavy-tailed count whatever the hazard does. We test it directly by renormalising, on the 35,227 county-months of 93,998 for which NFIP publishes exposure; the raw estimate on that matched subset is $\hat{\alpha} = 0.714$ rather than the full-sample 0.747, and every comparison below is within it. Per policy in force, $\hat{\alpha}$ rises from 0.714 to 1.001 (block-bootstrap $[0.952, 1.056]$); per dollar of coverage, to 0.979; and the claim count per policy to 1.338. *Exposure heterogeneity is therefore a substantial driver of the tail we measure*, and we report this prominently because it bounds the claim.

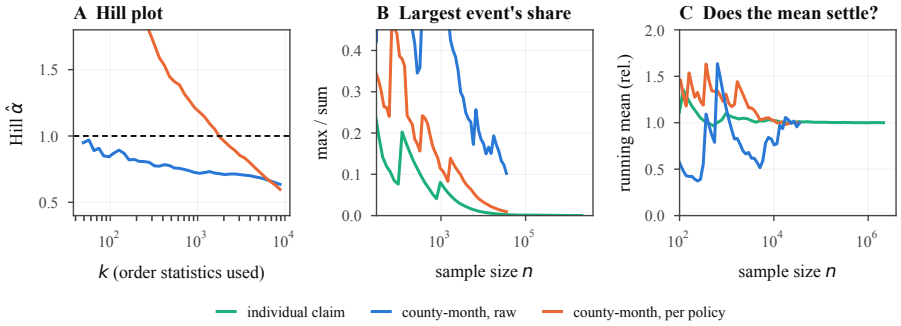

Figure 2: **Robustness of the tail estimate.** **(A)** Hill estimate against the number of order statistics k. The raw county-month aggregate stays below 1 everywhere; the exposure-normalised series is strongly k-dependent and crosses 1. **(B)** share of the total held by the single largest observation — the observable-range form of the infinite-mean claim. **(C)** running sample mean, normalised by its final value: individual claims settle, the aggregates do not.

Three things survive it. First, the raw aggregate is what aggregate-settled contracts actually pay on — a sovereign or area-index product indemnifies the region's total loss, not its loss per policy — so $\hat{\alpha} \approx 0.71$ remains the decision-relevant figure there even though the mechanism is partly compositional. Second, the raw estimate is robust where the normalised one is not: across every k from 50 to a quarter of the sample the raw Hill estimate stays below 1, while the per-policy estimate ranges from 1.8 to 0.6 and crosses 1 near $k \approx 2{,}000$ (Figure 2A); a moment estimator and a county-block bootstrap agree with Hill on the raw series. Third, and most usefully, it turns a limitation into a design instrument: if heterogeneous exposure is what makes the target unestimable, then normalising it — paying on loss per policy, or stratifying pools by exposure — is a contract-design lever that moves $\hat{\alpha}$ across 1. That is consistent with the pooling result above: widening a pool cannot fix heterogeneity by adding more of it.

Which side of $\alpha = 1$ a product sits on is a design choice. The transition happens once, at the first aggregation step. So the relevant question for a contract is not where losses are recorded but at what level risk is pooled. We classify the 84 confirmed parametric products in our registry by insured entity (Appendix A lists them all): 43 (51%) insure an individual policyholder — a farmer, a herder, a household, a firm — whose own loss sits in the well-behaved $\hat{\alpha} \approx 2.4$ regime; 34 (41%) insure a government or public entity against a national or state aggregate; 3 (4%) insure a community or trust; 4 are unclear. Roughly **44% of catalogued products settle on an aggregate**, and those are the ones this paper constrains. Malawi's ARC policy, with which we opened, is a sovereign aggregate product at the coarsest end of that range.

Caveats. The Hill estimator assumes iid draws and these are not — the dependence is the object of study — so we report $\hat{\alpha}$ as a tail descriptor, not a fitted generative model, with both controls built so the real-versus-shuffled comparison is meaningful even where the level is not. The pseudo-state control also randomizes geography within one country; pools spanning continents, as ARC does, may face genuinely lower hazard correlation.

4 Consequence I: Expectile schemes are undefined at that granularity

The first consequence is one of availability rather than performance, and it rules out a method proposed for this exact problem only months ago. The τ-expectile μ_τ of Y is defined by $\tau\,\mathbb{E}(Y-\mu_\tau)^+ = (1-\tau)\,\mathbb{E}(\mu_\tau - Y)^+$ (Newey & Powell, 1987), which requires $\mathbb{E}|Y| < \infty$ for both sides to be finite. Maier & Scherer (2026) show the expectile is the basis-risk-optimal parametric payment under an asymmetric quadratic criterion, a result we take as given.

Proposition 1. *For a loss variable with tail index $\alpha < 1$, $\mathbb{E}|Y| = \infty$ and the τ-expectile does not exist for any $\tau \in (0,1)$. The τ-quantile exists for every $\tau \in (0,1)$ regardless of α.*

We want to be careful about the scope here, because it is narrower than it might sound. Expectile-optimal schemes are well defined for individual-claim settlement, where $\hat{\alpha} = 2.43$. They are not well defined for the 44% of catalogued products that settle on an aggregate — and settling per claim would reintroduce the loss adjustment parametric insurance exists to avoid. The choice between expectile and quantile schemes is therefore not a matter of taste about which asymmetry to encode; at aggregate granularity only one of the two is available.

Layered contracts. Real sovereign covers are written with an attachment point a and an exhaustion point $a + L$, and $\mathbb{E}[\min((Y - a)^+, L)]$ exists for any Y, so a practitioner may object that Proposition 1 does not bite on real products. It does not bite on the *layer's expected value*, bounded by construction, but it bites on everything downstream: pricing the layer needs $\Pr(Y > a)$ and the conditional severity above a, both from the same tail, and the insured's basis risk is still measured against the unbounded Y. Layering caps the insurer's exposure, not the policyholder's residual risk.

5 Consequence II: shortfall-aversion is statistically bounded

Quantile schemes survive the previous section, but they do not survive freely, and the price they pay is the subject of this one. Let $\kappa \geq 1$ price a dollar under-paid relative to a dollar over-paid, and write the asymmetric criterion

$$L_\kappa(p, y) \;=\; \frac{\kappa \underbrace{\sum_i (y_i - p_i)^+}_{\text{under-paid}} \;+\; \underbrace{\sum_i (p_i - y_i)^+}_{\text{over-paid}}}{\underbrace{\sum_i y_i}_{\text{realized loss}}}. \tag{1}$$

Pointwise this is the pinball loss ρ_τ up to the factor $(1 + \kappa)$, with $\tau = \kappa/(1 + \kappa)$, so the population minimizer is the τ-quantile of the conditional loss (Koenker & Bassett, 1978). The aggregate normalizer is constant in p and does not move the minimizer. Two endpoints matter: $\kappa = 1$ recovers the symmetric two-sided basis-risk ratio, and $\kappa \to \infty$ recovers a shortfall-only measure whose minimizer is to pay the policy maximum always — which is why we report the symmetric case throughout and treat one-sided coverage measures as uninformative on their own.

Suppose we estimate that quantile from n samples. Writing $R(q) = \mathbb{E}\rho_\tau(Y, q)$, we have $R'(q_\tau) = F(q_\tau) - \tau = 0$ and $R''(q_\tau) = f(q_\tau)$, so the regret of a plug-in estimator $\hat{q}$ is $R(\hat{q}) - R(q_\tau) \approx \frac{1}{2} f(q_\tau)(\hat{q} - q_\tau)^2$. With the standard sample-quantile variance $\tau(1 - \tau)/(nf(q_\tau)^2)$, this gives $\mathbb{E}[\text{regret}] \approx \tau(1 - \tau)/(2nf(q_\tau))$. For a regularly varying tail with index α, $q_\tau = (1 - \tau)^{-1/\alpha}$ and $f(q_\tau) = \alpha(1 - \tau)^{(\alpha+1)/\alpha}$, yielding:

Proposition 2. *For losses with tail index α, the plug-in quantile regret at level τ satisfies*

$$\mathbb{E}[\text{regret}] \;\approx\; \frac{\tau\,(1 - \tau)^{-1/\alpha}}{2n\alpha}, \tag{2}$$

which diverges as $\tau \to 1$ at a rate governed by α.

Equation 2 is an estimation regret at level τ under τ's own loss, so it does not by itself compare criteria across τ. The reading we do claim is weaker and follows directly: the *estimation* component of basis risk grows without bound in τ at a rate set by α, so pushing a scheme toward the policyholder buys oracle improvement at a rapidly rising estimation cost, and beyond some sample-dependent level the second dominates. Establishing the crossover exactly needs the oracle risk too, which is infinite in the population for $\alpha < 1$; we do not attempt it here.

We define regret as $\mathbb{E}[\rho_\tau(Y, \hat{q}) - \rho_\tau(Y, q_\tau)]$, finite even where each term is not, and verify Equation 2 by simulation against Pareto populations at $\alpha \in \{0.75, 1.5, 2.5\}$ and $\tau \in \{0.5, 0.9, 0.99\}$ with $n = 500$: the ratio of empirical to predicted regret lies in $[0.86, 1.18]$ across all nine cells, and in $[0.96, 1.07]$ at $\alpha = 0.75$. At the $\hat{\alpha} \approx 0.75$ of §3 the divergence is as $(1 - \tau)^{-1.33}$: the shortfall-averse levels a policyholder advocate would want are precisely the levels a finite claims history cannot support. This is the result the known-distribution settings of both Maier & Scherer (2026) and Hadley & Whitin (1963) cannot express.

Table 1: **Two-sided basis risk and loss ratio, median over ten seeds** (interquartile range in brackets). *Lower basis risk is better;* 1.000 *is what an uninsured party achieves by holding cash. Loss ratio* is total payouts over total realized losses: a sellable product sits near 0.6–0.9, and a very low value means the method has retreated to paying almost nothing, so neither column is interpretable alone. In-distribution, boosting beats cash in every seed while paying a plausible 0.32; under either shift, no configuration beats cash in any run, and those that come closest do so by not paying.

	spatial (in-dist.)		temporal shift		climate shift	
configuration	basis risk	loss ratio	basis risk	loss ratio	basis risk	loss ratio
Recurrent neural, four variants						
direct	1.004 [1.00,1.03]	0.30	1.444 [1.00,1.58]	0.63	2.414 [1.00,2.74]	1.75
+ budget constraint	1.175 [1.17,1.18]	0.22	2.348 [2.26,2.64]	1.53	7.943 [7.64,8.36]	7.33
+ freq.–severity	1.018 [0.95,1.05]	0.23	1.344 [1.02,1.71]	0.41	2.782 [1.15,3.62]	1.95
+ capacity $\times 4$	1.021 [1.00,1.05]	0.26	1.481 [1.08,1.58]	0.66	2.535 [1.35,2.84]	1.93
Gradient-boosted ensemble						
boosting	**0.814** [0.79,0.85]	0.32	1.042 [1.04,1.05]	0.11	1.494 [1.46,1.52]	0.74
No learned magnitude						
fixed indemnity	1.000 [1.00,1.00]	0.00	1.001 [1.00,1.00]	0.00	1.014 [1.01,1.01]	0.03
SPI threshold baseline	1.176 [1.18,1.18]	0.21	3.006 [3.01,3.01]	2.21	8.428 [8.43,8.43]	7.80
Paying nothing	*1.000*	*0.00*	*1.000*	*0.00*	*1.000*	*0.00*

6 Is basis risk learnable?

The theory so far says what a payment scheme can and cannot be. Whether a model can find a good one is a separate question, and the rest of the paper is an attempt to answer it by elimination. Our testbed is a densified registry of 9,686 real United States assets spanning agriculture, power generation, mining and aquaculture, each crosswalked to its county and supervised by NFIP county-month indemnity. Five EO modalities: CHIRPS precipitation (Funk et al., 2015), ERA5-Land (Muñoz-Sabater et al., 2021), GRACE (Tapley et al., 2004), Sentinel-1 (Torres et al., 2012) and Landsat. Three holdout regimes, ten seeds each: *spatial* (whole counties held out, in-distribution), *temporal* (train pre-2018, test after) and *climate* (train on 189 temperate Köppen counties, test on 71 arid, continental and tropical ones). We report Equation 1 at $\kappa = 1$ alongside the loss ratio (total payouts over total losses), because either alone is uninformative: coverage on its own invites the degenerate solution of §5, and a loss ratio on its own says nothing about allocation. **Paying nothing scores exactly 1.000**, and that is the benchmark every method must clear to be worth writing.

We compare six configurations spanning three model families. (i) A recurrent neural model: per-modality encoders, gated fusion, a temporal state and a monotonic trigger, with the payout supervised by L_κ at $\kappa = 2$ and the loss weight selected on an inner-validation split of *training* counties. (ii) The same with a budget constraint making the zero payout infeasible by construction, the training-time analogue of the constrained-newsvendor optimum (Hadley & Whitin, 1963). (iii) The same with a frequency–severity decomposition: the county-month count has $\hat{\alpha} = 0.88$ but the mean severity per claim has $\hat{\alpha} = 2.03$, so we regress the two factors separately and quarantine the heavy tail in one. (iv) The same at four times the width. (v) A gradient-boosted ensemble (Friedman, 2001) — an occurrence classifier and a positive-cell magnitude regressor — on the same features and splits. (vi) A fixed-indemnity product learning no magnitude at all: the learned trigger paired with the constant minimising training L_κ, the strongest fixed product any trigger can support. Every payout is scaled by a factor fitted on training data only.

We began this section expecting to report that magnitude is not learnable at all, and the first several configurations we ran supported that. The gradient-boosted ensemble changed our minds. It scores 0.814, beating cash in 10 of 10 seeds and the recurrent model in 10 of 10 (p=0.0020, paired Wilcoxon), while paying a loss ratio of 0.32 — a plausible book rather than a retreat. It refutes the conclusion we ourselves first drew from the recurrent model alone: a learned payout can reduce two-sided basis risk by 19% against doing nothing, and by 31% against the calibrated rainfall threshold most deployed products use.

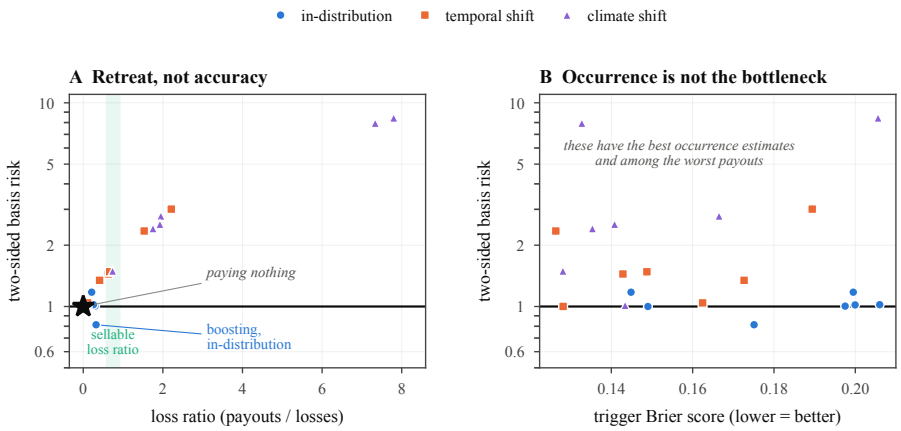

Figure 3: **Under shift, methods do not become accurate — they retreat.** **(A)** Loss ratio against two-sided basis risk, one point per configuration per regime; the star at $(0, 1)$ is selling no policy at all, the horizontal line is its basis risk, and the shaded band is the loss ratio a sellable product occupies. A genuinely useful product sits below the line *and* inside the band; only boosting in-distribution does. Configurations that approach the line under shift do so by sliding toward the star, paying almost nothing. **(B)** Trigger Brier score against basis risk: the two are essentially unrelated. Configurations with the best occurrence estimates (left) include some of the worst payouts, so occurrence skill is not what limits the product.

That result is confined to the in-distribution split, and the confinement is sharp. Across temporal and climate holdouts, *no run of any configuration* beat cash — 110 model–seed runs. Boosting remains best by a wide margin (1.042 and 1.494 against the threshold's 3.006 and 8.428), but its temporal loss ratio is 0.110: it learned to pay 11% of realized losses. The fixed-indemnity arm goes further, at 0.00–0.03 — the best constant any trigger can support is essentially zero. Both retreated, which is what §5 predicts a rational learner does when the tail it must price is not in the sample.

A reader may reasonably suspect something duller than a statistical limit — that the payouts are simply mis-scaled. NFIP indemnities are nominal, and median claim severity in our test window is $1.51\times$ the training window (Spearman $\rho = 0.82$, $p < 10^{-8}$), so a scale fitted on training data is fitted to cheaper dollars. That is a real confound and we test it directly. Granting each arm an *oracle* rescale — the constant minimising its score on the *test* set, which no deployable system could know — boosting improves from 1.042 to 0.999 under temporal shift and 1.494 to 1.008 under climate shift. The oracle-optimal factor is 0.05, not the >1 uncorrected inflation would require — the correction runs opposite to the inflation story — and even with test-set knowledge of the best scale, the product only ties cash under temporal shift and still loses under climate shift. The failure is in the allocation, not the level.

The recurrent model deserves separate comment, because its failure is of a different kind and we spent most of this project inside it. It converges to the degenerate policy — total payout 0.000, score exactly 1.000 — in **6 of 20 arms in every holdout** (an *arm* is one seed $\times$ one of the two model heads), 18 of 60 overall, with identical data, objective and hyperparameters. A budget constraint removes all collapses and makes allocation significantly worse everywhere (p=0.0020). The frequency–severity decomposition is indistinguishable from direct (p=0.68 spatial, p=0.14 temporal), and quadrupling capacity changes nothing (medians $1.021/1.481/2.535$ against $1.004/1.444/2.414$). A seed ensemble cannot collapse by construction and improves the score by 7–17% under shift, yet still pays 30–100% worse than nothing there, so the misallocation is systematic rather than seed noise. Since boosting on the same features succeeds in-distribution, this is a fact about recurrent optimisation on heavy-tailed targets, not about the target alone.

Set against all of this, one capability survives every regime we tested, and it is the most useful thing the paper offers a practitioner. Trigger-probability Brier score is proper and invariant to payout scale. Both learned models beat the threshold on it in every regime, and the margin *widens* under

shift: 0.135 (recurrent) and 0.128 (boosting) against the threshold's 0.206 under climate shift, versus 0.198/0.175 against 0.200 in-distribution. The models know when a claim event is happening, increasingly well as the regime moves away from training. What they cannot say is how much.

7 Domain knowledge does not close the gap

Faced with a model that will not transfer, the reflex in this community — and it was ours too, for most of the project — is to reach for domain knowledge: if the data does not generalise, perhaps structure will. From ENCORE, a published natural-capital database linking ISIC activities to 25 ecosystem services, we build a modality-relevance matrix $\Omega \in [0,1]^{N\times 5}$ and couple it to the recurrent model four ways: passive FiLM conditioning on the asset's graph embedding; *hard causal masking*, zeroing modality k for asset i when $\Omega_{i,k} < 0.3$; *causal divergence regularization*, adding $\sum_k \alpha_k(1-\Omega_k)$ to the loss; and the combination. Across 490 runs the graph shows no significant benefit under any holdout — spatial, temporal, anomaly-year or climate-zone, p=0.08–0.65 throughout — and no enforcement mechanism improves on passive conditioning. The control that settles it: a *randomly permuted* matrix, giving each asset another real asset's dependency profile, performs indistinguishably from the real one (p=1.00 spatial, p=0.16 temporal). Whatever the graph pathway contributes, it is not the semantic content ENCORE encodes — though it is coupled only to the recurrent model, so this is a null on a base model that itself fails.

8 Discussion and limitations

It is worth stepping back from the individual results to what they jointly imply. The field's trajectory is to fit one basis-risk model on the claims histories that exist and deploy it wherever the satellite signal reaches. Our results say this works only where the deployment regime resembles the training regime — the opposite of the intended use — and the failure is not architectural: it survives every repair we tried across three model families, including removing magnitude learning entirely. Two things follow constructively. Occurrence transfers best exactly where the baseline fails worst, so a learned trigger paired with a contractually fixed indemnity is deployable today on evidence; what it cannot claim is that the amount is right. And exposure heterogeneity being a substantial driver makes normalising it a contract-design lever that moves $\hat{\alpha}$ across 1.

What would change our minds. This is one country and one peril; whether $\alpha < 1$ at contract granularity generalizes to other perils or jurisdictions is untested, and is the most important extension. Equation 2 is asymptotic and derived for plug-in sample quantiles, so its qualitative divergence carries to our estimators but its constant does not. Because NFIP redacts claim location to the county, our labels are county-month aggregates and our experiments sit in the aggregate regime whatever a contract does; the individual-settled regime is the natural next study, though its $\hat{\alpha} = 2.43$ partly reflects truncation at policy limits. Indemnities are nominal and undeflated, which §6's oracle-rescale diagnostic indicates is not what drives the shift failure, though a deflated replication would settle it. The registry classification (Appendix A) is keyword-based. Finally, our graph null bounds the construction tested, not cross-attention or message passing.

9 Conclusion

We set out to build a better index. What we ended up with is an argument about what the index is being asked to predict. Machine learning for parametric insurance is framed as a modeling problem — better representations, better indices — and we report a property of the target that sits underneath every such choice. At the granularity where 44% of real products settle, catastrophe losses behave over every observable range as though they had no mean, and no further pooling undoes it. Expectation-based schemes are undefined there; quantile schemes face a sample-dependent ceiling on the protection they can express; a learned payout beats holding cash in-distribution but none of our configurations does so under shift. Basis risk at this granularity is therefore not primarily a modeling problem, and its driver is substantially the heterogeneity of the exposure being pooled — a contract parameter rather than a fact about weather. That is the more useful reading of a negative result: the same losses, measured per unit of exposure, are estimable. The lever is what a contract insures, and at what level it pools.

Reproducibility Statement

Every number reported is a real experimental result. Tail-index estimates, their intervals, and both shuffling controls are produced by a single script over the public NFIP loss record; the regret verification is a self-contained simulation. Per-run logs for all runs across six method families, the compilation scripts, the derived anomaly-year and Köppen splits, and the materiality matrix construction are included in the supplementary material. Seeds are fixed and enumerated; the scramble permutation is seed-derived and reproducible.

Ethics Statement

This work uses aggregated county-level federal insurance records containing no individual policyholder information. The central result is a limit on how much protection a learned payment scheme can deliver from a finite claims history, and we note the distributional consequence directly: the tail levels hardest to estimate are the catastrophic events that matter most to the least-insured populations. We are explicit that a negative transfer result should not be read as an argument against extending insurance to those populations, but as an argument about which contract designs evidence can support. A scheme performing well on pooled statistics can still fail individual claimants, and deploying any automatic trigger in a humanitarian context carries a duty to monitor per-claimant outcomes rather than pooled averages.

Use of Large Language Models

This paper was drafted with the assistance of an LLM-based coding assistant, which was also used throughout the project for implementation, experiment orchestration and analysis scripting, in an interactive loop with the authors. The scientific work was directed manually: every experimental design, the choice of holdout splits and seed protocol, the decision to run each control and ablation, and every interpretation reported here were authored and checked by us, and no result was accepted without our inspection of the underlying logs. Two substantive course corrections during the work — the discovery that an earlier headline metric was measuring an unsupervised payout, and the ablation that overturned an earlier and stronger negative claim — arose from that loop and are reported in the paper rather than elided. All numbers were regenerated from logged experimental output prior to submission.

References

ActionAid International. The wrong model for resilience: How G7-backed drought insurance failed Malawi, and what we must learn from it. Technical report, ActionAid International, May 2017.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019.

Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef Teugels. *Statistics of Extremes: Theory and Applications*. Wiley, 2004.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 48(4):1148–1185, 2012.

Sommarat Chantarat, Andrew G. Mude, Christopher B. Barrett, and Michael R. Carter. Designing index-based livestock insurance for managing asset risk in northern Kenya. *Journal of Risk and Insurance*, 80(1):205–237, 2013.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

Jon Danielsson, Laurens de Haan, Liang Peng, and Casper G. de Vries. Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis*, 76(2): 226–248, 2001.

Artur d'Avila Garcez and Luis C. Lamb. Neurosymbolic AI: The 3rd wave. *arXiv preprint arXiv:2012.05876*, 2020.

Arnold L. M. Dekkers, John H. J. Einmahl, and Laurens de Haan. A moment estimator for the index of an extreme-value distribution. *The Annals of Statistics*, 17(4):1833–1855, 1989.

Lauren Nicole DeLong, Ramon Fernández Mir, and Jacques D. Fleuriot. Neurosymbolic AI for reasoning over knowledge graphs: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 36(5):7822–7842, 2024.

Ghada Elabed, Marc F. Bellemare, Michael R. Carter, and Catherine Guirkinger. Managing basis risk with multiscale index insurance. *Agricultural Economics*, 44(4–5):419–431, 2013. doi: 10.1111/agec.12025.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer, 1997.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

Chris Funk, Pete Peterson, Martin Landsfeld, Diego Pedreros, James Verdin, Shraddhanand Shukla, Gregory Husak, James Rowland, Laura Harrison, Andrew Hoell, and Joel Michaelsen. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific Data*, 2:150066, 2015.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2022.

G. Hadley and T. M. Whitin. *Analysis of Inventory Systems*. Prentice-Hall, 1963.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018.

Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3(5):1163–1174, 1975.

Nathaniel D. Jensen and Christopher B. Barrett. Agricultural index insurance for development. *Applied Economic Perspectives and Policy*, 39(2):199–219, 2017.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, et al. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations (ICLR)*, 2019.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? a large-scale study. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 698–707, 2018.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

Markus Johannes Maier and Matthias Scherer. Expectiles as basis risk-optimal payment schemes in parametric insurance. *European Actuarial Journal*, 2026. doi: 10.1007/s13385-026-00447-w. Preprint arXiv:2505.02607, 2025.

Thomas Mikosch. *Non-Life Insurance Mathematics: An Introduction with the Poisson Process*. Springer, 2nd edition, 2009.

Joaquín Muñoz-Sabater, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, Margarita Choulga, Shaun Harrigan, Hans Hersbach, et al. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9):4349–4383, 2021.

Whitney K. Newey and James L. Powell. Asymmetric least squares estimation and testing. *Econometrica*, 55(4):819–847, 1987.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. MIT Press, 2009.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204, 2019.

Sidney I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, 2007.

Byron D. Tapley, Srinivas Bettadpur, Michael Watkins, and Christoph Reigber. GRACE measurements of mass variability in the Earth system. *Science*, 305(5683):503–505, 2004.

Ramon Torres, Paul Snoeij, Dirk Geudtner, David Bibby, Malcolm Davidson, Evert Attema, Pierre Potin, Betlem Bérgada, Berthyl Duesmann, Betlem Rosich, et al. GMES Sentinel-1 mission. *Remote Sensing of Environment*, 120:9–24, 2012.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

A THE 84-PRODUCT REGISTRY

Table 2 lists every confirmed parametric or index insurance product in the registry underlying §1 and §3. Inclusion requires a confirmed weather-, satellite- or event-index trigger with an automatic payout mechanism; conventional state-supported agricultural programmes without an index trigger, pre-launch market activity, and countries searched with no product found are recorded in the source registry but excluded here. *Pooling* is the level at which the insured entity holds the contract, which is what determines the aggregation regime of §3: INDIV. an individual policyholder, SOV. a government or public entity, GRP. a community, cooperative or trust. *Outcome* records the documented failure direction: **U** a documented under-payment relative to independently verified loss, u an under-payment risk that is structural but not documented as a specific event, O a documented over-payment, — no dispute documented in the sources found. Counts: 33 documented under-payments, 23 structural under-payment risks, 3 over-payments, 24 with no documented dispute, and one design-innovation entry. The pooling split is 43 individual (51%), 34 sovereign (41%), 3 group (4%), 4 unclear.

Classification is by keyword over the registry's free-text insured-entity and failure-direction fields and is approximate at the margins; the source registry with full provider, mechanism and citation detail is included in the supplementary material.

Table 2: Confirmed parametric and index insurance products ($n = 84$).

Programme	Region	Peril	Pooling	Outcome
AAIB bundled weather-index insurance pilot	Sri Lanka	Drought…	Indiv.	u
ACRE Africa	Kenya, Rwanda, Tanzania…	Drought, excess…	Indiv.	u

continued on next page

Table 2, continued

Programme	Region	Peril	Pooling	Outcome
AgriCat Fund	Italy (nationwide)	Drought, frost, flood …	—	**U**
Agroasemex / CADENA drought	Mexico (26 states)	Drought	Sov.	—
Agroseguro drought coverage	Spain (nationwide)	Drought (cereals and …	Indiv.	—
ARC 2015 Sahel drought payout	Mauritania	Drought	—	u
ARC Madagascar – cyclone and flood	Madagascar	Tropical cyclone …	Sov.	—
ARC Malawi	Malawi	Drought	Sov.	**U**
ARC Mozambique / Cyclone Idai	Mozambique	Tropical cyclone …	Sov.	**U**
ARC Senegal	Senegal	Drought	Sov.	—
ARC West Africa (Niger, Mali, Burkina Faso …	Sahel	Drought	Sov.	**U**
ARC Zambia	Zambia	Drought	Sov.	u
ARC Zimbabwe – updated with 2024 El Niño …	Zimbabwe	Drought	Sov.	**U**
Argentina parametric agricultural insurance	Argentina	Drought, water …	Indiv.	—
Assurance Récolte Sahel (ARS)	Senegal, Mali, Burkina …	Drought/flood …	Indiv.	—
Brazil parametric ag insurance (PROAGRO/PSR)	Brazil	Rainfall …	Indiv.	—
Cambodia hybrid weather-index insurance	Cambodia	Flood (70% of …	Indiv.	—
CCRIF Antigua and Barbuda – country breakdown	Antigua and Barbuda	Hurricane	Sov.	u
CCRIF Bahamas – country breakdown	Bahamas	Hurricane, excess …	Sov.	**U**
CCRIF Barbados – country breakdown	Barbados	Hurricane, excess …	Sov.	**U**
CCRIF Belize – country breakdown	Belize	Hurricane, excess …	Sov.	**U**
CCRIF Dominica – country breakdown	Dominica	Hurricane, excess …	Sov.	**U**
CCRIF Haiti earthquake	Haiti	Earthquake	Sov.	**U**
CCRIF Nicaragua – country breakdown	Nicaragua	Hurricane, earthquake	Sov.	**U**
CCRIF Saint Kitts and Nevis – country …	Saint Kitts and Nevis	Hurricane	Sov.	u
CCRIF Saint Lucia – country breakdown	Saint Lucia	Hurricane, earthquake	Sov.	**U**
CCRIF SPC	Caribbean and Central …	Hurricane …	Sov.	**U**
CCRIF Trinidad and Tobago – country breakdown	Trinidad and Tobago	Excess rainfall	Sov.	u
China catastrophe index insurance …	China – 74 cities across …	Typhoon, rainfall …	Sov.	O
Cirebon rice climate-index insurance	Indonesia (Cirebon …	Drought/flood (rice)	Indiv.	—
Croatia Osiguranje / Dunav soil moisture …	Croatia	Drought (soil …	Indiv.	u
Estonia satellite soil-moisture drought …	Estonia (nationwide)	Drought (also covers …	Indiv.	—
FarmIndex Dürre	Germany (+ parallel …	Drought	Indiv.	—
FloodFlash	United Kingdom	Flood	Indiv.	—
FONDEN catastrophe bond	Mexico (nationwide)	Earthquake, hurricane	Sov.	**U**
GAIP (Ghana Agricultural Insurance Pool)	Ghana	Drought (maize, soya …	Indiv.	u
Hungary drought/yield insurance	Hungary	Drought (wheat and …	Indiv.	**U**
IBLI Kenya	Northern Kenya	Livestock/drought	Indiv.	**U**
IBLI Mongolia	Mongolia – 330 soums …	Livestock/weather	Indiv.	**U**
India/Bangladesh IBFI (Index-Based Flood …	Bihar (India), Bangladesh	Flood …	Indiv.	u
INS first parametric coffee insurance	Costa Rica	Excess/deficit …	Grp.	u
Japan parametric earthquake/typhoon (private …	Japan	Earthquake (JMA …	Sov.	—
Jordan Climate Disaster Index (CDI) insurance	Jordan (incl. Tafileh …	Drought (rainfed …	Indiv.	u
Kazakhstan soil-moisture index insurance	Kazakhstan (all regions)	Drought	Indiv.	**U**
Korea parametric typhoon/earthquake (private …	South Korea	Typhoon (central …	Indiv.	—
Kyrgyzstan forecast-based/weather-index …	Kyrgyzstan	Drought, harsh …	Indiv.	u
Laos SEADRIF anticipatory drought insurance	Laos	Drought (extension of …	Sov.	—
Latin America regional parametric ag insurance	Ecuador, Colombia, Peru …	Drought, intense …	Sov.	O
Latvia SPI drought insurance	Latvia (nationwide)	Drought, excess …	Indiv.	—
Lithuania SPI-2 drought insurance	Lithuania (nationwide …	Drought	Indiv.	**U**
Moldova subsidized drought insurance …	Moldova	Drought	Indiv.	u
Myanmar coastal rice flood-index insurance …	Myanmar (coastal region)	Flood (rice)	Indiv.	—
NAIC index insurance (incl. Sharia-compliant …	Nigeria (nationwide)	Drought, flood	Indiv.	u
Nepal multi-peril index insurance	Nepal (incl. Jumla for …	Frost, excess rain …	Indiv.	**U**
Pacific Cat Risk Pilot (PCRIC predecessor)	Vanuatu, Tonga, Marshall …	Cyclone	Sov.	**U**
Pakistan satellite soil-moisture insurance …	Pakistan (Sindh/southern …	Drought (soil …	Indiv.	—
Papua New Guinea PCRAFI	Papua New Guinea	Drought, earthquake …	Sov.	—

continued on next page

Table 2, continued

Programme	Region	Peril	Pooling	Outcome
PCRIC	Pacific (Cook Islands …	Cyclone, earthquake …	Sov.	**U**
PCRIC Cook Islands – country breakdown	Cook Islands	Cyclone, earthquake …	Sov.	**U**
PCRIC Kiribati – country breakdown	Kiribati	Cyclone, excess …	Sov.	**U**
Philippines cat bond/parametric program	Philippines	Typhoon, earthquake	Sov.	O
PICAP Fiji micro-parametric	Fiji	Cyclone/weather …	Grp.	—
PMFBY farmer critique	India (nationwide)	Rainfall/weather …	Indiv.	**U**
Previsico × Descartes Underwriting parametric …	Ireland	Flood	Indiv.	u
Quintana Roo Coral Reef Insurance	Mexico (Yucatán/Quintana …	Hurricane wind	Sov.	**U**
R4 Rural Resilience Initiative	Ethiopia (Tigray, Amhara)	Drought/agriculture	Indiv.	**U**
Redicova parametric cyclone insurance	Northern Australia	Tropical cyclone	Indiv.	—
Romania "Index de Secetă"	Romania	Drought (maize …	Indiv.	—
Russia agricultural yield-index insurance	Russia (national)	Drought (region-wide)	Indiv.	**U**
Schweizer Hagel cross-border coverage	Liechtenstein	Hail/frost/drought	Indiv.	u
Schweizer Hagel KLIMA+ weather-index insurance	Switzerland	Hail/frost/drought	Indiv.	u
Seguro Agrario Universal Pachamama / PIRWA + …	Bolivia	Frost, flood …	Indiv.	**U**
Seguro Agropecuario (MAGA/MiCRO) + CCRIF …	Guatemala	Excess rainfall …	Sov.	u
Solomon Islands PCRAFI (documented double …	Solomon Islands	Earthquake (2013) …	Sov.	u
South Sudan WFP drought anticipatory action	South Sudan (Budi …	Drought	—	u
Swiss Re Insur8 typhoon warning cover	Hong Kong	Typhoon (business …	Indiv.	**U**
Tahoe Donner wildfire cover	California, USA	Wildfire	Grp.	–
Taiwan parametric earthquake and typhoon …	Taiwan	Earthquake, typhoon	Indiv.	—
TARSIM drought/yield insurance	Turkey (village-level)	Drought (wheat …	Indiv.	**U**
Togo ARC sovereign drought cover + FarmCover …	Togo	Drought	Sov.	u
Ukraine Copernicus soil-moisture drought …	Ukraine (5 regions)	Drought (with a …	Indiv.	—
Vietnam Storm Index Insurance	Vietnam	Typhoon/storm …	Indiv.	**U**
WBCIS (Weather Based Crop Insurance Scheme)	India	Rainfall …	Indiv.	**U**
WFP/Allianz/Humanity Insured drought policy	Syria	Drought	—	u

B TAIL-INDEX ESTIMATES IN FULL

Table 3 accompanies §3 and Figure 1. Hill and its asymptotic interval assume iid draws; the block bootstrap resamples whole counties with replacement and so preserves both the spatial and the within-county temporal dependence that assumption violates. The moment estimator of Dekkers et al. (1989) does not assume the Pareto form Hill does, and k-sensitivity is a known hazard of Hill in practice (Danielsson et al., 2001). All estimates use the top 5% of each series ($k = \lceil 0.05n \rceil$, floor 50); Figure 2A shows the dependence on k directly. Rows below the rule are the two controls and the exposure renormalisations, computed on the 35,227 county-months for which NFIP publishes exposure.

Individual claims are truncated at policy limits, so $\hat{\alpha} = 2.43$ there partly reflects truncation rather than the underlying severity law. The argument in §3 uses the *comparison* across aggregation levels, which truncation affects in the same direction at every level, rather than the level of any single estimate.

Table 3: Tail index by aggregation level and target, with three estimators.

target	n	Hill	Hill 95% CI	moment	block bootstrap 95%
Individual claim	2,114,319	2.429	[2.415, 2.444]	3.459	—
County-day	197,420	0.780	[0.765, 0.796]	—	—
County-month	93,998	0.747	[0.726, 0.768]	—	—
County-year	47,727	0.787	[0.755, 0.818]	—	—
State-month	16,978	0.747	[0.697, 0.797]	—	—
National-month	581	0.821	[0.593, 1.049]	—	—
County-month (exposure subset)	35,227	0.714	[0.681, 0.748]	0.745	[0.679, 0.761]
per policy in force	35,227	1.001	[0.954, 1.047]	1.455	[0.952, 1.056]
per dollar of coverage	35,227	0.979	[0.933, 1.024]	1.393	[0.930, 1.031]
claim count	35,227	0.821	[0.783, 0.860]	0.856	[0.777, 0.876]
County-month, dependence destroyed	93,998	0.889	sd 0.005	—	—
State-month, geography destroyed	16,978	0.805	sd 0.006	—	—