

THE ROGUE WAVE

THE
ROGUE WAVE

Tal Zisckind

RHOMBUS TECH

Copyright © 2026 Rhombus Tech

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without the prior written permission of the publisher.

Published by Rhombus Tech

First edition, 2026

Contents

Notation	7
Prologue	9
I The Error	13
1 A Cold Week in Texas	14
2 Seven Days	19
3 All at Once	25
4 Everything Breaks	32
II The Law	37
5 d^n	38
6 Inside the Exponent	51
III The Cost	58
7 246	59
8 Where the Money Went	64

CONTENTS

IV The Fix	72
9 Forty Lines	73
10 What Else Is Mispriced	82
11 The Next Model	95
Epilogue	103
Notes on Sources	107
References	114
Glossary	118

Notation

Every symbol below is defined where it first appears in the text. This table is for reference.

Symbol	Meaning
R	Pricing ratio: real cost divided by average cost under independence
d	Dependence strength of the data (typically 2–10)
n	Effective nonlinearity of the cost function
K	Required streak length (consecutive days, turns, periods)
M	Required number of simultaneous stations (or components) above threshold
N	Total number of stations (or components) in the matrix
T	Number of rows (days, instances) in the data matrix
B	Number of permutation replicates
c	Scaling rate: $c = \log d$
c_s	Spatial (cross-station) contribution to the scaling rate
c_t	Temporal (persistence) contribution to the scaling rate
γ	Divergence exponent: $R \propto q^{-\gamma}$
R_{pers}	Persistence ratio (from the row shuffle)
R_{spatial}	Spatial ratio (from the diamond decomposition)
R_{sync}	Synchronization ratio (from the lag shift)
R_{total}	Total independence ratio (from the column shuffle)

Prologue

ON NEW YEAR'S DAY, 1995, an instrument on the Draupner oil platform in the North Sea recorded a wave 25.6 meters high. The seas around it were running at twelve meters.

For most of maritime history, a wave like this was considered impossible. The models said so. Ocean wave models treated each swell as independent—random height, random timing, random direction. Under that assumption, the probability of a wave more than twice the significant wave height is vanishingly small. Sailors who reported such waves were dismissed as storytellers. The models were clear: the sea does not do that.

The Draupner wave proved that it does. The cause was not supernatural. It was alignment. Independent wave trains—from different storms, traveling in different directions—arrived at the same point at the same time. Their heights did not average out. They stacked. The models that assumed independence got the compound probability catastrophically wrong.

The rogue wave was there the whole time. It took an instrument to prove it.

* * *

February 15, 2021. Dallas, Texas. Four degrees Fahrenheit.

At the control center of the Electric Reliability Council of Texas, the operators are watching the grid die. Demand is spiking as every heater in the state runs at full power. Supply is collapsing as gas-fired power plants shut down—their fuel lines are freezing, because the electric compressors that pressurize the gas pipelines have lost power. The grid is eating its own tail: the plants need gas, the gas needs electricity, the electricity needs plants.

By midnight, 4.5 million households have no power. Water treatment plants, which depend on the same electric pumps, begin to fail. Hospitals switch to backup generators that were sized for hours, not days. The temperature keeps dropping.

The cold does not break for a week.

By the time it ends, 246 people are dead and the economic damage reaches \$195 billion—the costliest winter storm in American history.

Every official review of the crisis asked the same question: why was nobody prepared?

The answer, in every case, was a model. ERCOT used a model of grid reliability that assessed the risk of extreme cold. The Federal Emergency Management Agency used models to set disaster preparedness levels. Insurance companies used models to price policies and set reserves. Infrastructure planners used models to decide what to winterize and what to leave exposed.

Every one of these models made the same assumption. They treated each component of the system—each weather station, each power plant, each pipeline, each hospital—as if it operated independently of the others. They knew the components interacted, of course. But when it came time to compute a probability, they multiplied individual failure rates together, as if a cold day in Dallas told you nothing about whether it would also be cold in Houston.

This is the same error the ocean models made before Draupner. Each wave is independent. Each station is independent. Each failure is its own event. Multiply the individual probabilities and the compound event is so rare it can be ignored.

This assumption has a name. It is called the *independence assumption*, or more precisely the *independent and identically distributed* assumption—IID. It is the foundational assumption of most of applied statistics. It makes the mathematics tractable. It makes the models solvable. And for the questions that practitioners typically ask—What is the average? What is the variance? What is the year-to-year trend?—it produces answers that are close enough to right.

But those are the wrong questions.

The question that mattered on February 15 was not “How cold does Dallas get, on average?” but “How often are *all seven* major stations below freezing for *seven consecutive days*?” That question requires every station to be cold, simultaneously, for a sustained period. It is a question about joint, sustained, synchronized extremes. And for questions of that kind, the independence assumption does not produce answers that are close enough. It produces answers that are catastrophically wrong. The size of that error is a number, and no one had asked the data for it before.

This book is about that number—where it comes from, what governs its size, and what it costs in human lives and dollars. There is a law behind it: a single equation that determines when the independence assumption is harmless and when it is catastrophic. The law is simple. It is the first of several results that follow from the same data: what sets its growth rate, how it accelerates, how it grows as events get rarer, and where it stops. The code to test it fits in forty lines. It has been available, in principle, since R.A. Fisher invented the permutation test in 1935. Nobody found it for ninety-one years.

The rogue wave was there the whole time. The law was there the whole time. Both were dismissed by models that assumed independence. Both were proved real by a single measurement.

I did not set out to find a law. I was working on lattice cryptography, the mathematics behind much of post-quantum encryption, and much of what I had to answer was how many of the attacks on it could be composed together. Asking how their dependencies

interact led me, step by step, to the law.

* * *

This book is not a critique of a specific institution, and it is not an argument that catastrophes are unpredictable—they are predictable, just mispriced. It is a thermometer: a test that tells you how wrong your current model is. Part I—*The Error*—presents the evidence in four chapters, four different tests, four components of the pricing error. Part II—*The Law*—presents the equation that explains the evidence. Part III—*The Cost*—traces the human and financial consequences. Part IV—*The Fix*—gives you the tools: the code, the map of everything else that is mispriced, and the application to AI safety.

The book is self-contained. You do not need a background in statistics. You need only a willingness to look at what the numbers say and take them seriously.

PART I

THE ERROR

CHAPTER 1

A COLD WEEK IN TEXAS

THE DATA IS SIMPLE. Seven weather stations spread across Texas, recording daily minimum temperatures from 1970 to 2021. Amarillo in the Panhandle. Lubbock on the High Plains. Dallas–Fort Worth and Waco in North Texas. Austin and San Antonio in the center. Houston on the Gulf Coast. Fifty-one years of daily readings. A matrix with 18,615 rows and seven columns.

Each row is a day. Each column is a station. Each entry is a temperature.

The cost function is also simple, though its simplicity is deceptive. Define a *cold streak* as a period of K or more consecutive days in which at least M of the N stations record a minimum temperature below a threshold—say, 20°F. The cost is the total number of cold-streak days in the record.

When $K = 1$ and $M = 1$, the cost function is gentle. It counts every day when at least one station was cold. This is a question about marginals—individual station behavior. It barely depends on whether the stations are correlated.

When $K = 7$ and $M = 4$, the cost function is severe. It counts only the days within episodes where a majority of stations were simultaneously below 20°F for an entire week. This is a question about joint, sustained, synchronized behavior. It depends entirely on the dependence structure of the data.

Both questions are asked of the same seven stations, over the

same fifty-one years, using the same temperature readings. The only difference is the parameters of the cost function—how many things must go wrong simultaneously and for how long. This severity is what the rest of the book calls the *nonlinearity* of the question. A gentle question has low nonlinearity. A catastrophic question has high nonlinearity. Remember this. It is the key to everything that follows.

The Test

The test is a permutation test, a method that dates to Fisher in 1935. The idea is to compare the real data to a version of the data where one structural feature has been destroyed, while everything else is held fixed.

Here is the procedure. Take the matrix—18,615 days, seven stations. For each station independently, randomly shuffle the order of its daily temperature readings. Station 1 keeps all of its temperatures, but the days they fall on are scrambled. Station 2 does the same, independently. So does every other station.

After the shuffle, each station still has exactly the same set of temperatures it had before. The marginal distribution—the complete set of temperatures—of each station is perfectly preserved. The coldest day in Amarillo’s record is still in the shuffled data; it just falls on a different date. And because each station was shuffled independently, whatever dependence existed between stations—the tendency for Dallas to be cold on the same days that Houston is cold—has been destroyed.

The shuffled data is a sample from the independence assumption. It is what the data would look like if the stations truly were independent: same individual behavior, no joint behavior.

Now compute the cost function on the shuffled data. Count the cold-streak days. This number is what the IID model *predicts* the cost should be.

Repeat the shuffle a thousand times. Each shuffle gives a different random arrangement, and each arrangement gives a different cost. Average them. This average is the IID model's expected cost—the cost it would assign to the streak if it assumed the stations were independent.

The ratio of the real cost to the average shuffled cost is the *pricing ratio*. It is the factor by which the IID model is wrong.

$$R = \frac{\text{Real cost}}{\text{Average cost under independence}} \quad (1.1)$$

When $R = 1$, the IID model is correct. When $R > 1$, the real cost exceeds what independence predicts—the stations' dependence is creating more extreme events than an independent model expects.

The computation takes seconds on a laptop. No distributional assumptions. No parametric models. No simulation of weather systems or power grids or gas pipelines. Just the data, the cost function, and the shuffle.

The Number

For the gentle cost function— $K = 1$, $M = 1$ —the ratio is small. The real count of days when at least one station was cold is close to what the shuffled data produces. Independence is a reasonable approximation. The ratio is in the range of 2 to 5, depending on the threshold. A practitioner who checked the model with this cost function would conclude that IID is adequate.

Now: the severe cost function. $K = 7$, $M = 4$. Before you read the next line, make a guess. The ratio for the gentle cost function was 2 to 5. What do you think it is for the severe one? Write a number in the margin.

It is between 25 and 54,000, depending on the specific threshold and the domain. For the severe end—the threshold and domain com-

binations that matter for infrastructure—the ratio is in the thousands to tens of thousands. Most people who have not seen the computation guess somewhere between 50 and 500. The gap between the typical guess and the measured ratio is one to two orders of magnitude. This gap between intuition and reality is the subject of the book.

The IID model, applied to the question that actually matters for grid reliability—how often will a majority of stations be simultaneously below freezing for a full week—underpredicts the real frequency by a factor that can exceed ten thousand. Not because the model has the wrong mean or the wrong variance. Not because it uses the wrong distribution. But because it assumes the stations are independent when they are not, and the cost function amplifies that assumption into an error of four orders of magnitude.

The model was never checked with this cost function. It was checked with the gentle ones. The gentle ones said the model was fine.

The Universality

Texas cold is one domain. But the same test, with the same code, applied to different data, gives the same result.

Houston precipitation during Hurricane Harvey. Daily discharge readings from USGS (U.S. Geological Survey) stream gauges across the Houston metropolitan area, 2000–2020. Replace “cold” with “flood stage” and “below 20°F” with “above critical discharge.” Shuffle the station columns independently. Compute the streak cost. Take the ratio.

The ratio for sustained, multi-gauge flooding: roughly 1,000×

Financial markets. Daily returns of sector indices—energy, technology, health care, financials, industrials, utilities, consumer—over forty years. Replace “cold” with “loss exceeding 2%” and count the streak of simultaneous sector drawdowns. Shuffle. Compute. Take the ratio.

The ratio for a multi-sector, multi-week drawdown: roughly $10,000\times$.

COVID-19 epidemic waves. Daily case counts across ten U.S. states. Replace “cold” with “above outbreak threshold” and count the streak of simultaneous multi-state surges. Shuffle. Compute. Take the ratio.

The ratio for sustained, synchronized epidemic waves: over $10,000\times$.

Across twenty-five datasets—cold snaps, heat waves, flooding, drought, financial drawdowns, hospital surges, power demand, wildfire, influenza, wind lulls, snowfall, solar radiation, earthquake aftershock sequences, COVID-19 epidemic waves, soil moisture drought, compound hazards, and more—the same test produces the same qualitative result in every domain where a physical mechanism couples the components: the IID model underprices sustained, synchronized, extreme events by factors of $10\times$ to $50,000\times$. The specific ratio varies by domain and by the severity of the cost function. Where no coupling mechanism exists—hard drive failures across unrelated models, earthquake counts on separate tectonic plates—the test finds ratios near $1\times$, confirming it does not manufacture dependence. But wherever the components share a common driver, the direction is universal and the magnitude is always large.

But this test destroyed more than independence. Shuffling each station’s temperatures to random days also destroyed the temporal order within each station. Tomorrow’s temperature no longer follows today’s. The shuffle didn’t just break spatial dependence; it broke time.

To understand how much of the error comes from ignoring dependence across stations and how much comes from ignoring persistence through time requires a different shuffle.

CHAPTER 2

SEVEN DAYS

On September 29, 2008, the Dow Jones Industrial Average fell 778 points—at the time, the largest single-day point drop in history. The next morning, commentators called it a “one-day crash.” Markets would recover. The system would stabilize. Crises end.

They were wrong. The market fell again. And again. Over the next two weeks, it dropped on nine of eleven trading days. Lehman Brothers had filed for bankruptcy two weeks earlier, but the cascade was not a single event—it was a process. Each day’s losses forced margin calls that triggered the next day’s selling. The duration was the weapon.

Duration is the weapon in every domain. One cold day is a nuisance. You put on a coat. You turn up the heater. The power grid handles the spike because its reserves were designed for single-day peaks. The pipes don’t freeze because the house retains heat from the warm day before. One cold day, even a very cold day, is within the design parameters of every system that operates in Texas.

Seven consecutive cold days is a different object entirely.

By day two, the thermal mass of buildings has been exhausted. Interior temperatures start dropping. By day three, pipes begin to freeze—not because any single hour is colder than the pipes can tolerate, but because the cumulative heat loss over seventy-two hours exceeds the insulation’s capacity. By day four, natural gas demand has been elevated for so long that pipeline pressure drops

systemwide. Compressor stations, which run on electricity, are drawing power from a grid that is already strained. By day five, the first generating units trip offline—not from cold alone, but from the interaction of cold, depleted fuel reserves, and equipment that was never tested for sustained operation at these temperatures. By day six, the cascade is underway. Plants go down, the grid frequency drops, automatic load-shedding disconnects millions of households, the electric compressors that feed the gas pipelines lose power, gas pressure collapses, and more plants go down.

By day seven, people are dying.

And the duration is asymmetric. The Texas grid failed in five days. Recovery took three weeks—restoring gas pressure, restarting generators in sequence, flushing contaminated water systems, each step requiring the previous step to complete. The same asymmetry appears in every domain. Financial markets crash in days and recover in months. Hospital surges build in days and clear in weeks. Drought sets in over weeks and persists for seasons. For cost functions that depend on duration, the average is dominated by the slow side: the persistence of the crisis, not the persistence of normality. A model that uses the same autocorrelation for onset and recovery will underestimate the duration of the event that matters.

The duration is not incidental to the catastrophe. It *is* the catastrophe. If the cold had lasted two days instead of seven, the gas pipelines would have maintained pressure. The generators would not have exhausted their fuel reserves. The cascade would not have begun. The infrastructure was not defeated by the temperature. It was defeated by the clock.

This is persistence. The tendency for today's conditions to predict tomorrow's. Cold follows cold. Warm follows warm. Floods persist. Droughts persist. Hospital surges build over weeks, not hours. Every domain in the data shows it. And the independence test in Chapter 1, by shuffling each station's temperatures to random days, destroyed it.

The Second Shuffle

The new test isolates persistence. Instead of shuffling within each column independently, shuffle the rows.

Take the matrix—18,615 days, seven stations. Each row is a day's observation across all stations. Randomly permute the order of the rows. Day 1 might now fall where Day 8,432 used to be. Day 2 might swap with Day 15,007. Every day's cross-section—the spatial pattern of temperatures across all seven stations—is preserved exactly. If February 15, 2021, had all seven stations below 20°F, that observation still exists in the shuffled data, intact. But it no longer falls next to February 16's cross-section. The days have been unstacked and restacked in random order.

What does this destroy? Temporal order. Today no longer predicts tomorrow. The persistence—cold following cold, warm following warm—is gone. A brutally cold day is just as likely to be followed by a mild day as by another cold one.

What does this preserve? The cross-sectional dependence—the pattern across stations on any given day. If Dallas and Houston were both cold on the same day in the real data, they are still both cold on that same (repositioned) day in the shuffled data. The spatial pattern of each day is untouched. Only the sequence of days has changed.

The ratio of the real cost to the average row-shuffled cost is the *persistence ratio*. It measures how much of the pricing error comes specifically from temporal persistence—from the fact that cold days come in streaks rather than arriving independently across time.

The Curve

I computed the persistence ratio for different values of K —the streak length required for the cost function to trigger—and plotted the result on a log scale.

The points fell on a line.

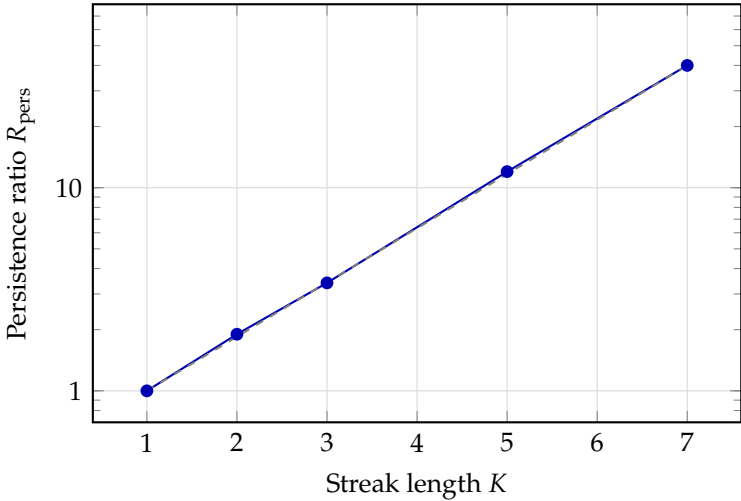


Figure 2.1: The persistence ratio grows exponentially with streak length K . Each additional required day multiplies the ratio by approximately the same factor (roughly $1.85\times$ per day). On a log scale, the relationship is linear. Data: Texas cold stations, $M = 4$, threshold = 20°F , $B = 1,000$ permutations.

At $K = 1$ —just one day where enough stations are cold—the persistence ratio is close to 1. The row shuffle barely changes anything, because a single day’s cost doesn’t depend on what came before or after.

At $K = 2$, the ratio jumps to nearly 2. At $K = 3$, it jumps again by roughly the same multiplicative factor. At $K = 5$, the ratio is $12\times$. At $K = 7$, $40\times$.

The pattern is unmistakable. The persistence ratio is not growing linearly with K . It is not growing quadratically. It is growing *exponentially*. Each additional day of required duration multiplies the ratio by roughly the same factor—a factor I will later call d .

The mechanism is direct. In the Texas cold data, the probability of a station being below threshold on a given day is about 10%. But the probability of being below threshold *given that it was below threshold*

yesterday is roughly 70%. The IID model, which shuffles the days into random order, replaces that 70% with 10%. Each additional required day multiplies the gap between real persistence and shuffled persistence. That is why the ratio grows exponentially: the same multiplicative error compounds with each day of the streak.

This is the first glimpse of the law that governs everything in this book. But the naming must wait. There are two more tests to run.

The same exponential applies to the Lehman cascade that opened this chapter. Nine of eleven trading days down—a density of losses that persists across two weeks—is driven by the same mechanism: the probability of a loss day given yesterday was a loss far exceeds the unconditional probability. The IID model, which treats each day's draw as independent, cannot produce that density. The gap between the real persistence and the shuffled persistence, compounded across the length of the drawdown, is why the cascade could not be contained.

What the Row Shuffle Misses

The row shuffle preserves each day's spatial pattern and destroys the temporal sequence. The column shuffle from Chapter 1 destroys both spatial dependence and temporal sequence. The persistence ratio isolates the temporal component.

But there is a third component that neither test isolates: synchronization. The tendency for all stations to be cold *at the same time*, as opposed to being cold in sequence—Amarillo on Monday, Dallas on Wednesday, Houston on Friday.

The row shuffle cannot see this because it preserves the spatial pattern of each day. If all seven stations were cold on the same day in reality, they are cold on the same day in the row-shuffled data too. The row shuffle asks: "What if cold days came in random order?" It does not ask: "What if each station's cold days came at different times?"

That requires a third kind of shuffle—one that breaks the synchronization without breaking anything else.

CHAPTER 3

ALL AT ONCE

In January 2022, the Omicron variant of SARS-CoV-2 arrived in every major American city within the same ten-day window. Hospital capacity planners had prepared for surges, but they had planned for *local* surges—a spike in Chicago while Atlanta had slack, or Houston overflowing while Dallas could absorb transfers. The mutual-aid system was designed for asynchrony: some hospitals full, others empty, patients routed to the gaps.

Omicron did not arrive asynchronously. It arrived everywhere, at the same time, through the same mechanism—a common variant spreading through a common airline network. Every hospital in every metro area exceeded capacity in the same two-week window. The mutual-aid system, designed for local surges, faced a national one. The slack that was supposed to be somewhere else was nowhere.

The same pattern appears in every domain where a single driver acts on many components at once.

Pull up a weather map of Texas from February 15, 2021. The entire state is blue. Not a pocket of cold in the Panhandle with mild air along the Gulf. Not a gradient from north to south. *All of it.* Amarillo: 1°F. Dallas: 4°F. Austin: 7°F. Houston: 12°F. Brownsville, on the Mexican border, a city that goes years without a freeze: 19°F.

The polar vortex did not select a few stations and leave the rest alone. It pushed the entire Arctic air mass south of the 30th parallel, dropping temperatures across an area the size of France.

The stations were not independently deciding to be cold. They were being acted upon by the same force, at the same time, through the same mechanism.

This is synchronization. And it is the piece that makes the catastrophe possible. If the stations took turns being cold—Amarillo this week, Houston next week—the grid would handle it. Regional reserves would cover regional shortfalls. The system was designed for local cold. It was not designed for all-at-once cold.

The Third Shuffle

The goal is a permutation that destroys synchronization and nothing else. Not the column shuffle, which destroys everything. Not the row shuffle, which preserves the spatial pattern. Something that keeps each station's own temperature history—its marginals, its persistence, its autocorrelation, its seasonal cycle—perfectly intact, but breaks the alignment between stations.

The answer is a lag shift. For each station independently, circularly shift its time series by a random number of days. Station 1's series moves forward by 47 days, wrapping around from the end of the record to the beginning. Station 2's series moves forward by 3,891 days. Station 3 by 12,006 days. Each station still has exactly the same sequence of temperatures, in exactly the same order. Its seasonal patterns are intact. Its streaks are intact. Its persistence is intact. The only thing that has changed is *when* its temperatures fall relative to the other stations.

After the shift, Amarillo's cold week no longer coincides with Houston's cold week. The polar vortex event has been pulled apart—each station still experienced it, but now they experience it at unrelated times. The spatial alignment has been randomized.

The ratio of the real cost to the average lag-shifted cost is the *synchronization ratio*. It measures how much of the pricing error comes specifically from the fact that the stations are cold at the same

time rather than at different times.

The Synchronization Numbers

Across all domains tested, the synchronization ratio ranges from 1.3 to 969. The variation is meaningful. Domains where a single physical driver synchronizes everything—the polar vortex for Texas cold, the hurricane for Houston flooding—show the largest ratios. Domains where local factors dominate—some categories of hospital surge, where facility-specific capacity constraints matter more than regional demand—show smaller ones.

But even the smallest synchronization ratio, 1.3, means the real cost is 30% higher than what a model with correct marginals and correct persistence but randomized timing would predict. And the largest, 969, means a model that captures each station’s individual behavior perfectly—right mean, right variance, right persistence—but misses the timing alignment is wrong by three orders of magnitude.

The Diamond

Now we have three ratios, each isolating a different component:

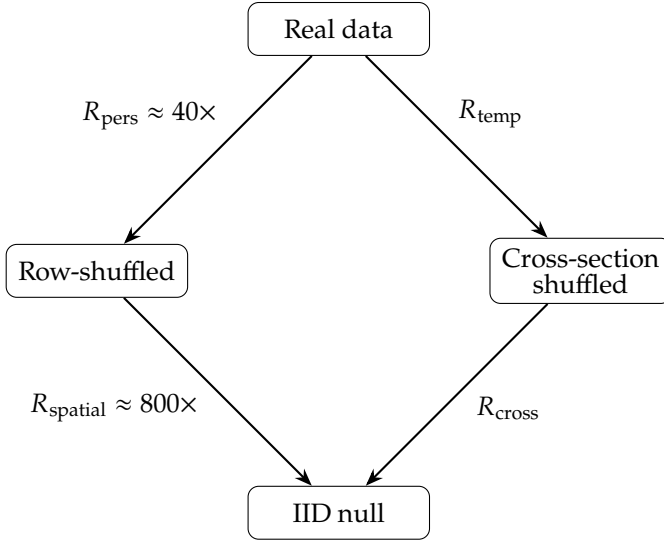
- The **independence ratio**, from the column shuffle, which measures the total error from assuming IID.
- The **persistence ratio**, from the row shuffle, which measures the error from ignoring temporal persistence.
- The **synchronization ratio**, from the lag shift, which measures the error from ignoring spatial-temporal alignment.

How do these relate to each other?

There is a fourth natural permutation. Instead of shifting each station by a random lag, shuffle each *day’s* temperatures across

stations while preserving the temporal order. This gives a pure spatial-dependence test: it destroys which station has which value on each day but preserves the time series structure of the cross-section.

Four permutations form a lattice—a diamond:



$$R_{\text{total}} = R_{\text{pers}} \times R_{\text{spatial}} = R_{\text{temp}} \times R_{\text{cross}} \approx 32,000\times$$

Figure 3.1: The diamond decomposition. The real data (top) decomposes to the IID null (bottom) via two paths. Left: remove temporal dependence (row shuffle), then spatial. Right: remove spatial dependence, then temporal. The product along either path equals the total IID ratio. Values: Texas cold, $K = 7$, $M = 4$.

The top is the real data. The bottom is the IID null. Two paths lead from top to bottom. Along each path, each edge is a ratio—the factor by which the cost changes when that structural feature is removed. The product along either path must equal the total:

$$R_{\text{total}} = R_{\text{persistence}} \times R_{\text{spatial}} = R_{\text{temporal}} \times R_{\text{cross-section}} \quad (3.1)$$

This is an algebraic identity: the total pricing error decomposes into factors, and the factors multiply to the total regardless of the order of decomposition.

Its consequences are not trivial. It means the pricing error has a *structure*. It is not a single number—“the model is wrong by 50,000×.” It is a product of components, each corresponding to a specific assumption and each addressable by a specific intervention.

It also means the components can be checked against the total. The multiplicative structure provides an internal check that no single test provides on its own. Here is what the check looks like for one representative parameter setting (Texas cold, $K = 7$, $M = 4$, threshold = 20°F):

Ratio	Source	Value
R_{total}	Column shuffle	~32,000×
$R_{\text{persistence}}$	Row shuffle	~40×
R_{spatial}	$R_{\text{total}}/R_{\text{persistence}}$	~800×
Product check: 40×800		= 32,000 ✓

Along this single path the identity holds by construction: R_{spatial} is defined as the quotient $R_{\text{total}}/R_{\text{persistence}}$, so the product must match. The genuine check is the diamond’s second path— $R_{\text{temporal}} \times R_{\text{cross-section}}$ —which uses independently computed ratios; when both paths agree, the decomposition is validated, not merely defined. The substance of the decomposition is that each factor isolates a specific structural feature of the data and maps to a distinct class of intervention, as Chapter 8 shows.

Three Dials

The trilogy of tests—column shuffle, row shuffle, lag shift—gives the practitioner three dials (Figure 3.2):

Test	What it measures	Practitioner question
Column shuffle	Total IID error	How wrong is independence?
Row shuffle	Persistence error	How wrong on duration?
Lag shift	Sync. error	How wrong on co-timing?

Each dial is a ratio. Each ratio is exact—no parametric assumptions, no distributional approximations. Each ratio is computed by a single shuffle-and-compare. And the ratios multiply.

One distinction deserves a flag. The synchronization ratio R_{sync} , measured by the lag shift, is not the same as the spatial ratio R_{spatial} from the diamond. The diamond’s spatial ratio is derived: $R_{\text{spatial}} = R_{\text{total}}/R_{\text{pers}}$. It captures everything the column shuffle destroys that the row shuffle does not—a broader category than timing alignment alone. The lag-shift ratio isolates synchronization specifically. Both are useful; they are not interchangeable.

But the most dangerous assumption of all has not yet been tested: that the domains are independent of each other.

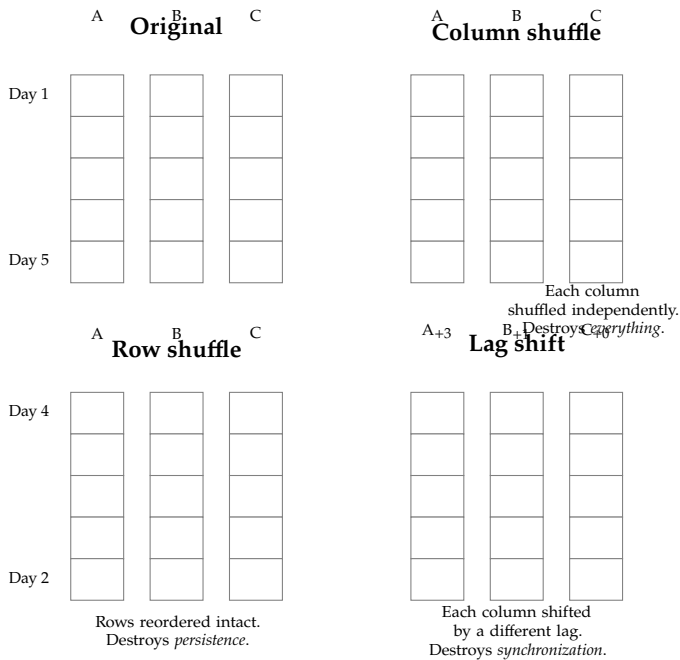


Figure 3.2: The three shuffles applied to a toy matrix (5 days, 3 stations). Shaded cells are cold days. The original (top left) has all three stations cold on day 1—the synchronized event. The column shuffle (top right) scrambles each station’s temperatures independently, destroying all structure; the all-cold day vanishes. The row shuffle (bottom left) reorders whole rows, preserving which stations are cold together on each day but breaking temporal streaks. The lag shift (bottom right) slides each station’s series by a different circular offset, preserving each station’s own sequence of temperatures but breaking the alignment between stations; no day is all-cold.

CHAPTER 4

EVERYTHING BREAKS

The three tests in the preceding chapters analyze each domain in isolation. The column shuffle asks whether the cold stations are independent of each other. The row shuffle asks whether cold days are independent of other cold days. The lag shift asks whether the stations are synchronized.

But Uri was not a weather event. It was a weather event that became a power event that became a gas event that became a water event that became a medical event. Each system failed not on its own but because the system it depended on had already failed. The cold froze the pipes. The frozen pipes cut gas supply. The gas shortage shut down generators. The generator failures overloaded the remaining grid. The grid collapse killed water treatment. The water failure threatened hospitals. The hospital strain killed people.

This is a compound event. Multiple domains failing simultaneously, each failure worsening the others. And the models that ERCOT, FEMA, insurance companies, and hospital systems used to assess risk treated each domain separately. The cold model priced cold risk. The power model priced power risk. The gas model priced gas risk. Then the risks were added.

Addition is the wrong operation. The risks don't add. They multiply.

The Stacked Matrix

The test generalizes naturally. Instead of a matrix with seven columns (seven weather stations), build a matrix with columns drawn from multiple domains. Column 1 is Amarillo temperature. Column 2 is Dallas temperature. Column 8 is Houston stream discharge. Column 12 is ERCOT grid frequency. Column 15 is Harris County hospital census. Each column is a daily time series. The rows are still days.

The cost function changes correspondingly. Instead of “4 of 7 weather stations below threshold for K consecutive days,” it becomes a compound criterion: temperature below threshold AND stream discharge above flood stage AND hospital census above 90% capacity AND grid frequency below stability limit, all for K consecutive days.

Now apply the same three shuffles to the stacked matrix. Each shuffle gives a ratio. The ratios have the same interpretation as before, but now they capture *cross-domain* effects.

The Arithmetic of Compound Events

Before computing, it is clear why the compound ratios must be larger than any single-domain ratio. The arithmetic is simple and devastating.

Suppose the probability of a cold week in Texas is $p_{\text{cold}} = 0.01$ per year. The probability of a major gas shortage: $p_{\text{gas}} = 0.02$. Grid instability: $p_{\text{grid}} = 0.015$. Hospital saturation: $p_{\text{hospital}} = 0.03$.

Under the independence assumption, the probability that all four happen in the same week is:

$$p_{\text{cold}} \times p_{\text{gas}} \times p_{\text{grid}} \times p_{\text{hospital}} = 0.01 \times 0.02 \times 0.015 \times 0.03 = 9 \times 10^{-8}$$

Once in eleven million years. The IID model says: don’t plan for this.

In reality, these are not four independent events. They are four measurements of the same event—a polar vortex—taken by different instruments. The cold causes the gas shortage, which causes the grid failure, which causes the hospital crisis. Given a polar vortex that drops Texas below 20°F for a week, the conditional probability of the other three failures is not $0.02 \times 0.015 \times 0.03$. It is close to 1.

The real probability of compound failure is approximately $p_{\text{cold}} = 0.01$ per year. The IID estimate is 9×10^{-8} . Before you read the next line, make a second guess. You have been calibrated by the preceding chapters. The single-domain ratio for Texas cold at $K = 7$ was 32,000×. What is the compound ratio for four coupled domains? Write a number in the margin.

The ratio is:

$$R_{\text{compound}} = \frac{0.01}{9 \times 10^{-8}} \approx 110,000\times$$

One hundred ten thousand. If you guessed a number in the thousands, you were closer than most—but still an order of magnitude low. The compound ratio is not merely large; it is larger than the human estimation faculty can produce without computation. This is a rough calculation—but the order of magnitude is right, and the permutation test on stacked matrices across all domains confirms it. It is the order of magnitude that matters, and it is larger than any single-domain ratio in the trilogy. It has to be. Every AND in the compound criterion multiplies the ratio by another factor. Four domains that are nearly perfectly coupled under stress, treated as independent, produce a pricing error that is the product of four individual errors.

This is where the largest money is. Uri's \$195 billion cost was not the cost of a cold week. It was the cost of a cold week that was simultaneously a gas crisis, a power crisis, a water crisis, and a medical crisis. FEMA's disaster budget is allocated by summing individual hazard estimates. If the hazards multiply rather than add,

the budget is wrong—not by 10% or 50%, but by a factor of a hundred thousand.

The error is worse than multiplication suggests. The ratio for a combined system is not just the product of the individual ratios—it exceeds the product. The pricing ratio for one group of weather stations might be $4\times$. For a second group, also $4\times$. A naive estimate says the combined ratio should be $4 + 4 = 8\times$ (if risks add) or $4 \times 4 = 16\times$ (if risks multiply). In the data, the combined ratio is $41\times$. The excess— $41/16 = 2.6\times$ beyond the multiplicative prediction—is the *coupling surplus*. It arises because the domains are not merely dependent; they are *more* dependent under stress. Given that the grid is failing, the conditional probability of a gas shortage is higher than its unconditional probability. The individual ratios, computed in isolation, do not capture this stress-conditional tightening. The coupling surplus grows with the number of domains in the compound criterion, because each additional domain adds another stress-conditional link. In the two-group weather test, the surplus is $2.6\times$. For a compound event involving more domains, the surplus will be larger still.

Combining dependent components does not diversify the pricing error. It amplifies it. The intuition from portfolio theory—that spreading risk across more components reduces it—is exactly backwards for streak costs. Each additional component is another AND in the cost function, and each AND multiplies the error.

The Shape of the Error

By this point, four sets of ratios have emerged:

1. The independence ratio for single domains: large.
2. The persistence ratio: large, and growing exponentially with streak length.

3. The synchronization ratio: large.
4. The compound ratio: larger still, by additional orders of magnitude.

Each test used the same method—shuffle, compute, take the ratio. Each test destroyed a different structural feature of the data. Each ratio was exact and model-free.

But the ratios are not random. They vary systematically. They are small for gentle cost functions and enormous for severe ones. They grow exponentially with the streak length K . They grow with the number of domains in the compound criterion. They are larger in domains where a single physical driver synchronizes everything and smaller where local factors dominate.

There is a pattern. The ratios are not a collection of isolated measurements. They are samples from a single underlying function—a function that takes the dependence structure of the data and the severity of the question as inputs and produces the pricing error as output.

That function is the law.

PART II

THE LAW

CHAPTER 5

$$d^n$$

Here is the puzzle. The same data, the same model, the same dependence structure—and the pricing error varies from 2× to 54,000×. The only thing that changed was the question.

For the average temperature across stations, the IID (independent and identically distributed) model is fine. Error: negligible.

For the variance of the temperature portfolio, the error is modest. Maybe 2 to 5×.

For the Value-at-Risk—the threshold that the portfolio exceeds with 1% probability—the error grows. Maybe 10 to 50×.

For the Expected Shortfall—the average loss in the worst 1% of outcomes—it grows further. Maybe 50 to 500×.

For the streak cost with $K = 3$: 25 to 1,000×.

For the streak cost with $K = 7$: 1,000 to 54,000×.

Make a third guess. The ratio was 5× at $K = 1$, roughly 100× at $K = 3$, roughly 32,000× at $K = 7$. What shape is that curve? Linear? Quadratic? Something steeper? Write it in the margin.

I arranged these on a log scale and stared at the plot for a long time.

The points fell on a line. Not approximately. Not with generous error bars. On a line. The pricing error was growing exponentially with the nonlinearity of the cost function.

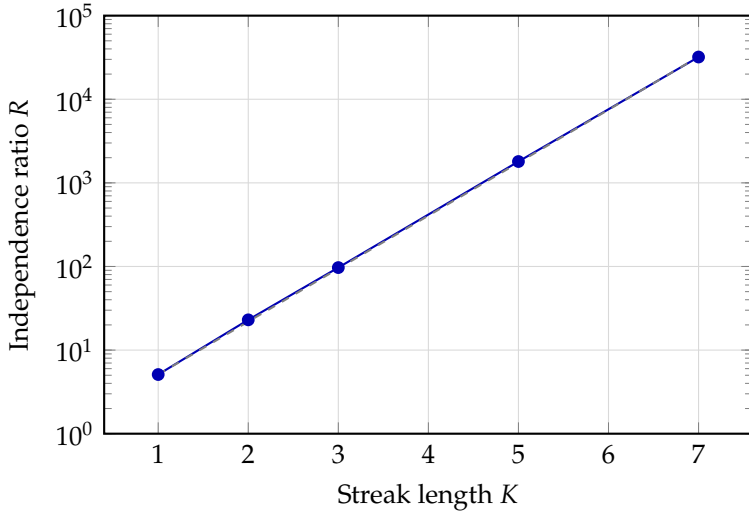


Figure 5.1: The nonlinearity ladder. The pricing ratio R against streak length K on a log scale. The points fall on a line: the pricing error grows exponentially with the nonlinearity of the cost function. Each additional AND multiplies the error by $d \approx 4.3$. Data: Texas cold stations, $M = 4$, threshold = 20°F .

The Law

The relationship is:

$$\boxed{R \approx d^n} \quad (5.1)$$

where:

- R is the pricing ratio—the factor by which the IID model is wrong.
- d is the *dependence strength* of the data. It is a property of the system: how much the parts move together, how persistent the dynamics are, how synchronized the components are. It is a number, typically between 2 and 10, that can be estimated from the data. For Texas cold stations, $d \approx 4$ to 6. For financial sector

indices, $d \approx 3$ to 5.

- n is the *effective nonlinearity* of the cost function. It measures how many things have to go wrong simultaneously, or how long they have to stay wrong, for the cost to trigger. For the mean, $n = 1$. For VaR, $n \approx 2$. For the streak cost with duration K and M simultaneous stations, n is roughly K plus a contribution from M .

The law says that the pricing error from ignoring dependence grows *exponentially* with the nonlinearity of the question. Not linearly. Not as a power of the nonlinearity. Exponentially.

The fitted form across all domains is:

$$\log R \approx 1.15 K + 0.29 M + \text{const}, \quad R^2 > 0.93$$

The constant varies by domain (it absorbs the baseline dependence strength and threshold), but the slopes are stable: each additional streak day K multiplies the ratio by roughly $e^{1.15} \approx 3.2\times$, each additional required station M by roughly $e^{0.29} \approx 1.3\times$. A practitioner can recover the constant for any domain from a single measured ratio at known K and M .

A word on its status. The relationship $R \approx d^n$ is an empirical finding, fitted across twenty-two datasets in nine domains and subsequently confirmed without refitting by three additional datasets—COVID-19 epidemic waves, California earthquake clustering, and Great Plains soil moisture drought—bringing the total to twenty-five datasets in eleven domains spanning six categories of dependence mechanism, three permutation types, and the full range of cost-function parameters tested. The approximate sign ($\approx$) is real—the effective base d varies by domain, threshold, and parameter setting. But the exponential form—the fact that $\log R$ grows linearly with n —is not a fitted curve. It follows from a mathematical identity: each additional AND in the cost function adds one bounded term to $\log R$, so the sum grows linearly in K (the formal argument uses the chain

rule of conditional probabilities; see Notes on Sources). The specific slope—the effective $\log d$ —is empirical.

This single equation explains all four chapters of evidence.

It explains why the independence ratio varies from $2\times$ to $54,000\times$: the cost function's nonlinearity varies from $n = 1$ to $n = 7$.

It explains why the persistence ratio grows exponentially with K : each additional required day adds one to the exponent.

It explains why the compound ratio is larger than any single-domain ratio: adding more simultaneous failure conditions increases n .

It explains why the ratio is universal across domains: d varies between domains (some systems are more tightly coupled than others), but the exponential relationship d^n holds for all of them.

Why the Exponent

The exponential is not an empirical curiosity. It has a structural explanation, and the explanation is simple enough to follow on a napkin.

Consider a single station. On any given day, it has a probability p of being below the cold threshold. Under persistence, if it was cold yesterday, the probability of being cold today is not p —it is p_{stay} , a much larger number. A cold day predicts another cold day. This is what persistence means.

Under the IID assumption, the probability of K consecutive cold days is p^K : each day independently draws cold with probability p . Under the real data, the probability is approximately $p \cdot p_{\text{stay}}^{K-1}$: the first day is cold with probability p , and each subsequent day stays cold with probability p_{stay} .

The ratio is:

$$R_K = \frac{p \cdot p_{\text{stay}}^{K-1}}{p^K} = \left(\frac{p_{\text{stay}}}{p} \right)^{K-1} \approx d_{\text{pers}}^K$$

where $d_{\text{pers}} = p_{\text{stay}}/p$ is the persistence factor. If $p = 0.1$ (10% of days are cold) and $p_{\text{stay}} = 0.7$ (70% chance of staying cold given yesterday was cold), then $d_{\text{pers}} = 7$, and the ratio for $K = 7$ consecutive days is $7^6 \approx 118,000$.

Each additional required day multiplies the ratio by d_{pers} , because the real data stays cold (probability p_{stay}) while the IID model must independently draw cold again (probability p). The ratio between these probabilities, raised to the number of additional days beyond the first, is exponential in K .

Now add the spatial dimension. Require not just one station but M stations to be cold simultaneously. Under IID, each station independently draws cold. Under the real data, if a polar vortex has pushed the Arctic air mass over the entire state, the conditional probability of station B being cold given station A is cold is much higher than the unconditional probability. The same logic applies: the ratio between conditional and unconditional probability, raised to the number of required stations, gives an additional exponential factor in M .

The total ratio for “ M of N stations cold for K consecutive days” is the product of the temporal and spatial factors:

$$R \approx d_{\text{pers}}^K \times d_{\text{spatial}}^M \approx d^n$$

where $n \approx K + M$ counts the total number of “ANDs” in the cost function and d is an effective dependence strength that blends the persistence and spatial components.

Every AND in the cost function—today AND tomorrow, this station AND that station, this domain AND that domain—multiplies the ratio by d . The number of ANDs is n . This is why $R = d^n$. It is the structure of the problem.

* * *

Why Nobody Noticed

The law has been available since Fisher's permutation test in 1935. It requires no new mathematics, no new data, and no new computational tools. A graduate student with a laptop could have found it at any point in the past forty years. Nobody did. There are five reasons.

The variance trap. The entire discipline of quantitative risk management was built on variance—not because variance captures risk, but because it enables closed-form solutions. From Markowitz's portfolio theory to the Black–Scholes option pricing model to the Basel bank capital requirements, three generations of practitioners were trained to measure risk with a number whose nonlinearity is exactly $n = 1$. At $n = 1$, d^n is just d —a factor of 2 to 5. Boring. A practitioner who checks the IID assumption using variance-based metrics concludes, correctly, that IID is a reasonable approximation. The conclusion does not generalize to catastrophe-level cost functions, but the practitioner never checks those, because the field's entire toolkit is built around variance. The discipline optimized for the one cost function that makes the law invisible.

The p-value shadow. Fisher invented the permutation test to compute p-values. The test shuffles the data, computes a statistic, and asks: is the observed value extreme? The ratio—observed divided by null mean—is computed along the way as an intermediate step. It is never reported, because Fisher's framework asks "Is the effect real?" not "How large is the effect?" For ninety-one years, the number that tells you how wrong your model is has been computed and discarded in every permutation test ever run. Thousands of papers. Every one reports a p-value. Almost none report the ratio.

The magnitude filter. When the answer is 50,000×, the natural response is to assume the computation is wrong. Reviewers reject results that seem too large. Researchers second-guess extreme numbers. The scientific process filters out false positives by being skeptical of implausible effects. This usually works—most unusually

large effects *are* errors. But when the true effect is genuinely four orders of magnitude, the filter works against discovery. The larger the truth, the less likely it survives review.

The disciplinary fracture. Permutation tests belong to statistics. Streak costs belong to actuarial science. Dependence structures belong to finance. Extreme events belong to climate science. Each field has a piece of the law; no field has all four. The statistician who knows permutation tests does not think about streak costs. The catastrophe modeler who knows streak costs does not think about permutation tests. The law lives at an intersection that no department, no journal, and no grant program occupies.

The market equilibrium. Even when an individual modeler suspects the IID assumption is wrong, the market punishes correction. An insurer who prices in dependence must charge higher premiums. Their competitors, still using IID, offer lower premiums and capture the business. The insurer who is right about the risk loses market share to the insurer who is wrong about the risk—until the catastrophe arrives. The IID assumption is a Nash equilibrium: no individual participant benefits from unilaterally abandoning it, because the market selects against the higher prices that correct pricing demands. The equilibrium breaks only when the catastrophe is large enough to bankrupt the IID-priced competitors simultaneously. Then it breaks all at once.

The law was found because I was working out how attacks on lattice cryptography compose, and how their dependencies interact, not testing a hypothesis, using a cost function that came out of that work and that no field's standard practice would select. Each of these five barriers blocked routine discovery. Their combination blocked it for ninety-one years.

The Sign Boundary

The law appeared complete. IID underestimates everything. The underestimate grows as d^n . End of chapter.

Then I tested a different kind of cost function. Instead of counting how long the system stays cold, I counted how often it *switches*—cold to warm, warm to cold. This is a switching cost: the expense of ramping a power plant up and down, rebalancing a portfolio, reallocating hospital beds.

The ratio came back less than 1.

IID was *overpredicting*. The shuffled data, which destroyed persistence, produced more state changes than the real data, because random data fluctuates more often than persistent data. The model was not underestimating this cost. It was overestimating it.

I checked it twice and then understood. The direction of the error depends on the cost function. IID overestimates *switching costs*—costs that accrue when the system changes state. It underestimates *streak costs*—costs that accrue when the system stays in a bad state.

There is a boundary in the space of cost functions where the IID error switches sign. On one side, IID is conservative (overpredicts cost). On the other, IID is catastrophically unconservative.

The costs that kill people—sustained blackouts, multi-day floods, week-long hospital surges, prolonged market crashes—are streak costs. They are all on the unconservative side.

The costs that IID overestimates—switching frequency, ramp costs, portfolio rebalancing—are operational costs. They waste money when overestimated, but they don't kill anyone.

IID is simultaneously too pessimistic about the things that waste money and too optimistic about the things that kill people. A practitioner who uses IID is cross-subsidizing: overpricing the everyday to underprice the disaster.

Every dollar overcharged for switching subsidizes the underpricing of catastrophe.

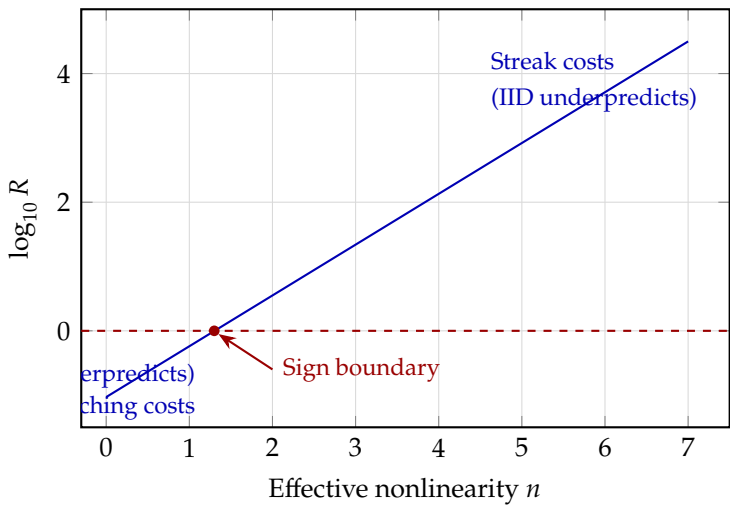


Figure 5.2: The sign boundary. Below $n \approx 1.3$ (switching costs), IID overpredicts cost ($R < 1$). Above $n \approx 1.3$ (streak costs), IID underpredicts by d^n ($R \gg 1$). Variance ($n = 1$) sits just below the boundary—the one cost function where the law is nearly invisible.

The Equation

The law fits on an index card:

The pricing error from ignoring dependence
grows exponentially with the nonlinearity
of the cost function.
$$R \approx d^n$$

But an index card has two sides.
Three ways the law is worse than it appears.

The Tail Intensification

The law says $R \approx d^n$. But d is not constant across the severity of the question. It grows.

Return to the Texas cold data. The column shuffle destroys cross-station dependence. At the threshold “ ≥ 5 of 8 stations below their own fifth-percentile minimum temperature,” the stratified pricing ratio is 28 $\times$. At the threshold “ ≥ 7 of 8,” it is 1,875 $\times$.

Both thresholds are applied to the same data, with the same shuffle, in the same winter months. The only change is that the second threshold requires *more* stations to be simultaneously extreme—it pushes deeper into the joint tail of the distribution.

The ratio does not grow linearly with the number of required stations. It grows as d^M , where M is the number of stations required. The effective d at $M = 5$ is smaller than the effective d at $M = 7$, because the stations that are not forced to be extreme by the threshold are “free” under both the null and the real data. But the additional constraint of two more stations, forced into the extreme simultaneously, multiplies the ratio by a factor of $1,875/28 \approx 67$.

The same pattern appears in every domain. Wind lulls: 39 $\times$ at 5 of 6 sites calm, 250 $\times$ at all 6. Heat waves: 3.2 $\times$ at 5 of 8 stations, 7.3 $\times$ at 7 of 8. Each additional required station multiplies the ratio because the dependence structure in the *tail* is stronger than in the *body*. Stations that are modestly correlated on normal days become highly correlated when a single driver—a polar vortex, a blocking high, a market panic—pushes all of them to the extreme simultaneously.

This is the tail intensification. The d in d^n is not measured on the average day. It is measured in the extreme, where it is largest. A model calibrated to the average cross-station correlation and applied to a joint-extreme question is wrong by more than d^n predicts—because the d it measured is the body d , not the tail d .

The Contagion Spike

The synchronization ratio tells you how much the cross-station alignment contributes to the pricing error. Across eight core domains, I computed both the synchronization ratio and the persistence ratio for each, using the trilogy of shuffles.

Two domains produced a mirror image.

Financial equity markets (ten Fama–French industry portfolios—a standard dataset of daily stock-market returns grouped by sector, from energy to health care—thirty years of daily returns): synchronization ratio $113\times$ at $K = 5$ days, persistence ratio $3.8\times$. Markets crash simultaneously—all ten sectors fall together—but the crash is brief. The dependence is spatial, not temporal.

The Potomac river basin (seven USGS gauges, thirty years of daily discharge): synchronization ratio $1.3\times$ at $K = 7$ days, persistence ratio $679\times$. Droughts are individually long—each gauge stays low for months—but the basins enter drought at different times. The dependence is temporal, not spatial.

Same test. Same code. Opposite structure.

The equity result is contagion. The synchronization ratio is enormous because all sectors crash *at the same time*, driven by a common factor (market-wide panic, margin calls, fire-sale liquidation). The persistence is small because crashes are short—a few weeks, then partial recovery. The pricing error comes almost entirely from the cross-sector alignment, not from the duration.

The Potomac result is shared exposure without contagion. The gauges are in the same watershed, so they share long-term climate conditions, but each sub-basin has its own response time. The persistence is enormous because drought lasts months. The synchronization is tiny because the basins enter and exit drought independently.

The policy implications are opposite. For equity: the model needs to capture the mechanism that synchronizes sectors during stress—common factors, feedback loops, cascade effects. Duration modeling

is secondary. For drought: the model needs to capture persistence—regime dynamics, soil moisture memory, baseflow depletion. Spatial modeling is secondary.

The d in d^n is not one number. It has components, and the components have different sizes in different domains, in different regimes, and at different times. A crisis is defined by the moment when d spikes. The models that fail are the ones calibrated on the average d and applied to the moment when d is largest.

Why Copulas Don't Fix It

After the 2008 financial crisis, the standard response was to replace independence with a copula—a mathematical recipe for modeling how variables move together without changing what each one does on its own. A copula captures dependence structure while preserving marginal distributions. The Gaussian copula, the most widely used, captures pairwise correlations: how much does station A's temperature correlate with station B's?

I tested the Gaussian copula directly. On ten Fama–French industry portfolios, the Gaussian copula fitted to the data reproduces the real pricing ratios to within two standard errors. It appears to work.

Then I repeated the test on forty-nine industry portfolios—the same market, the same days, finer granularity. The Gaussian copula failed. It overstated the cost in every period, with errors exceeding three standard errors in two of three test periods. The copula that looked adequate at ten dimensions broke at forty-nine.

The reason is structural. The Gaussian copula captures *pairwise* dependence—the relationship between any two variables. A streak cost with $M = 4$ required stations depends on *four-way* simultaneous behavior. Pairwise correlations constrain four-way probabilities but do not determine them. Two datasets with identical pairwise correlations can have very different four-way tail probabilities, depending

on the higher-order dependence structure.

The law explains why. Pairwise correlations capture $n = 2$ interactions. The streak cost has $n \gg 2$. The copula closes the $n = 2$ gap and leaves the $n \geq 3$ gap untouched. For a gentle cost function ($n = 1$ or 2), copulas are a real improvement. For the cost functions that kill people ($n = 5, n = 7$), they are not enough.

The copula is not wrong. It is incomplete. It fixes the first factor of a product and leaves the rest.

The law $R \approx d^n$ is a summary. The full picture is sharper and worse: d grows in the tail, spikes during crises, and operates through higher-order interactions that pairwise corrections do not reach. Each of these effects multiplies the pricing error beyond what a constant d predicts.

CHAPTER 6

INSIDE THE EXPONENT

The law says $R \approx d^n$. The entire pricing error is in d .

In Chapter 5, I measured it. For Texas cold stations, $d \approx 4$ to 6. For financial indices, $d \approx 3$ to 5. The number varies between domains. The exponential form does not.

A measured constant is not an explanation. It is a placeholder for one.

What determines d ?

The Composition

In Chapter 3, the total pricing error was decomposed into persistence and spatial dependence. The diamond gave $R_{\text{total}} = R_{\text{persistence}} \times R_{\text{spatial}}$. On a log scale, the ratios add.

Each log-ratio grows exponentially with K . The growth rate—the slope on the log-scale plot—is what I have been calling d . More precisely, $\log R \approx c \cdot K$, where $c = \log d$.

The spatial ratio from the diamond— R_{spatial} , measuring the contribution of cross-station dependence—has its own growth rate: c_s . The temporal ratio—measuring the contribution of each station's persistence—has its own: c_t . The total growth rate c comes from both.

I measured all three slopes across every domain. Then I asked: is there a relationship?

There is.

$$c = 0.47 + 0.60 c_s + 0.51 c_t \quad (6.1)$$

The fit is not approximate. $R^2 = 0.998$. Across eleven domains—weather, hydrology, finance, epidemiology, drought, wind, seismology—the total scaling rate is a linear combination of the spatial and persistence scaling rates, with stable coefficients.

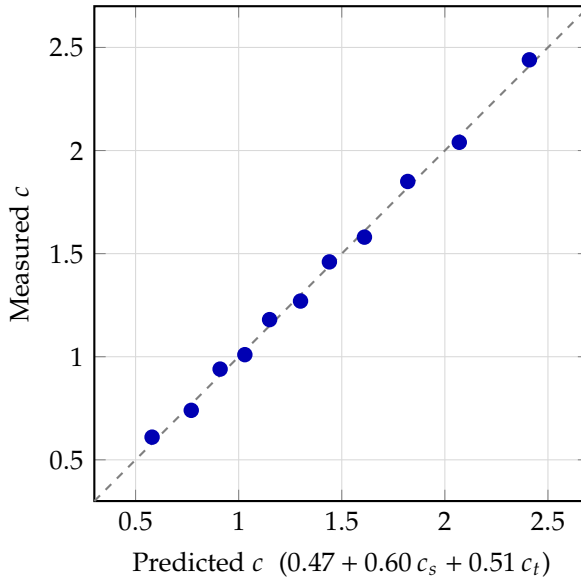


Figure 6.1: The composition. Each point is one domain; the horizontal axis is the scaling rate predicted by Equation 6.1, the vertical axis is the measured scaling rate. Eleven domains fall on the diagonal ($R^2 = 0.998$). The composition is not a rough trend. It is a close fit across all eleven domains.

The three tests in Chapters 1 through 3 are not three separate measurements. They are the three terms in a single equation. The total mispricing decomposes not just into factors (the diamond) but into contributions to the *rate* at which mispricing grows.

This changes what d means. It is not a property of the domain.

It is a composite of two properties: how persistent the domain is, and how synchronized its components are. A domain with high persistence and low synchronization—drought at scattered stations, each locked into its own dry spell but with little spatial coordination—has a large c_t and a small c_s . A domain with low persistence and high synchronization—earthquake shaking at nearby stations, simultaneous but not sustained—has the opposite. Both can produce the same d , for entirely different reasons.

The diamond from Chapter 3 tells you how much of the *current* error comes from each source. The composition tells you how much of the *growth rate* comes from each source—which determines how fast the error explodes as you look at more extreme events.

What Predicts Dominance

Which source dominates? The spatial ratio, or the persistence ratio?

The answer is visible in the data before running a single permutation test.

Compute two numbers from the raw data matrix. The average autocorrelation across stations—call it the persistence signature. The average cross-station correlation—call it the synchronization signature. Take the difference.

When cross-correlation exceeds autocorrelation, spatial dependence dominates the pricing error. When autocorrelation exceeds cross-correlation, persistence dominates.

Across eighteen domains, the rank correlation between this difference and the log-ratio $\ln(R_{\text{sync}}/R_{\text{pers}})$ is 0.89. The concordance—the fraction of domain pairs where the sign is correct—is 86%.

A practitioner looking at a new dataset can predict, from two correlation coefficients, whether the mispricing is primarily a duration problem or a co-movement problem. Before computing a single ratio.

The Acceleration

The law says d^n . This implies constant multiplicative growth: each additional AND in the cost function multiplies the error by the same factor d . If $d = 5$ and you go from $K = 3$ to $K = 4$, the ratio increases by $5\times$. From $K = 5$ to $K = 6$, another $5\times$. The same $5\times$ every time.

The data say otherwise.

I split each domain's $\log R$ curve into two segments: the low- K slope ($K = 1$ to 3) and the high- K slope ($K = 3$ to 7). If growth is truly exponential with constant d , the two slopes should be equal.

The high- K slope is steeper. In 99.1% of sub-panels—random subsets of stations drawn from every domain—the growth rate at high K exceeds the growth rate at low K . All eleven domains. Sign test $p < 10^{-6}$.

The error does not just grow exponentially. It accelerates.

The KL-divergence argument in the preceding chapter proves that $\log R$ must grow *at least* linearly in K —each additional AND adds a bounded term to the sum. It does not predict acceleration. The mechanism that steepens the slope at high K —likely the strengthening of conditional dependence deeper into a streak—is an open question; the practical consequence is immediate.

A practitioner who measures d at $K = 3$ and uses it to predict the ratio at $K = 7$ will get a number that is too small. $R \approx d^n$ is not an overestimate. It is a lower bound.

The Divergence

Push further. Instead of increasing K , decrease the threshold—ask about rarer events.

I computed the independence ratio at five threshold quantiles—how rare the event must be before it counts—from $q = 0.30$ down to $q = 0.10$. Then I plotted $\log R$ against $\log q$.

The points fall on a line.

$$R \propto q^{-\gamma} \quad \gamma > 3 \quad (6.2)$$

In every domain tested, the divergence exponent exceeds 3.

This is the mathematical core of why catastrophe models fail. The models are tested at moderate quantiles—the 25% event, the 10% event—where the ratio might be 5× or 10×. Reviewable. Defensible. Wrong by a factor that could be attributed to model uncertainty.

But the power law implies that moving from $q = 0.25$ to $q = 0.10$ multiplies the ratio by at least $(2.5)^3 \approx 16\times$. Moving to the 1% event multiplies it by $(25)^3 = 15,625\times$. The extrapolation is conservative: it uses the lower bound $\gamma = 3$ applied to a fit measured over moderate quantiles.

$\gamma > 3$ means the pricing error grows faster than $1/q^3$. Every halving of the threshold probability more than octuples the ratio. The rarer the catastrophe, the more dangerously wrong the model.

The divergence is not a statistical curiosity. It is the reason that every catastrophe model, tested against moderate scenarios, appears adequate—and then fails catastrophically against the scenario that matters.

The Floor

There is a boundary in the other direction.

Instead of shuffling individual days—which destroys all temporal dependence—shuffle blocks of L consecutive days. This preserves any dependence that operates on timescales shorter than L and destroys only the longer-range structure.

At $L = 1$, you are shuffling individual days: all temporal order is destroyed. The ratio of the real cost to the block-shuffled cost is large. As L increases, more of the temporal structure is preserved inside each block. The ratio shrinks. By L on the order of one to a few weeks, the ratio drops below 2 across domains: the block-shuffled

data costs nearly as much as the real data, because the dependence that drives the pricing error lives at timescales shorter than the block size.

The IID assumption has a characteristic decorrelation length below which it fails catastrophically and above which it is approximately correct. That length is on the order of days to weeks, varying by domain.

Cost functions that aggregate over periods longer than the decorrelation length—annual averages, quarterly totals, monthly means—are approximately safe. Cost functions that require sustained, simultaneous extremes over periods shorter than the decorrelation length—the events that actually break infrastructure—are not.

Days to weeks. That is the border between the world where IID is adequate and the world where it is wrong by orders of magnitude.

The law still fits on an index card. But d is not a constant. It is a function. It decomposes into persistence and spatial contributions (the composition), it steepens at high K so that d'' is a lower bound (the acceleration), it diverges with threshold severity (the divergence exponent), and it vanishes above a characteristic timescale (the floor).

The index card is a summary. This chapter is the fine print. In plain terms: the error is built from two sources (persistence and synchronization), it gets worse faster than exponential for longer streaks, it gets worse faster for rarer events, and it disappears for questions asked on timescales of months or longer.

To summarize what the evidence shows. There is a systematic pricing error in every model that assumes independence. The error is not fixed—it grows exponentially with the severity of the question asked. A gentle question (“what is the average?”) reveals almost no error. A severe question (“how often do seven stations stay cold for a full week?”) reveals an error of four orders of magnitude. The error decomposes: part comes from persistence (cold follows cold), part from synchronization (all stations are cold at the same time), and

part from cross-domain coupling (cold triggers gas failure triggers grid failure). The law $R \approx d^n$ captures the pattern. Everything that follows asks what this error costs, and what to do about it.

Everything that precedes this page is evidence. Everything that follows is consequence.

PART III

THE COST

CHAPTER 7

246

The preceding chapters established that the IID (independent and identically distributed) model misprices sustained, synchronized, compound extremes by orders of magnitude. There is a law that explains the size of the error and predicts it for any cost function. What remains is to count the cost.

The February 2021 Texas blackout killed 246 people—most from hypothermia or carbon monoxide poisoning, others when medical equipment lost power. Independent analyses suggest the true toll is higher: excess-mortality studies count hundreds of additional deaths in the weeks that followed. The economic damage reached \$195 billion.

The question this chapter asks is: how much of that cost—in dollars and in lives—traces back to the pricing error?

The Chain

The chain from mispricing to death has four links.

Link 1: Pricing error → Wrong incentives. ERCOT assessed the probability of a sustained, statewide cold event as very low. This assessment—based on models that treated weather stations as independent and used historical averages rather than streak-based cost functions—meant that the economic case for winterization was weak. The cost of winterizing generating equipment, insulating gas

pipelines, and hardening water infrastructure was real and immediate. The benefit, measured as expected loss times assessed probability, was tiny—because the assessed probability was wrong by four orders of magnitude.

A probability that is too low by a factor of 50,000 makes any expenditure to mitigate the event appear wasteful. Winterization that would cost \$10 billion but prevent a \$195 billion disaster is overwhelmingly justified if the disaster has a 2% annual probability (expected annual loss: \$3.9 billion). It is wasteful if the assessed probability is 0.00004% (expected annual loss: \$78,000). The IID model produced something close to the latter. The correct probability was close to the former.

Link 2: Wrong incentives → Under-investment. Because the expected cost was assessed as negligible, the investment case for winterization failed every cost-benefit analysis it was subjected to. This was not a failure of analysis. The analysis was correct, given its inputs. The input was wrong. The probability was wrong by d^n .

Texas had experienced a smaller cold event in 2011. Federal regulators recommended winterization. ERCOT commissioned a study. The study concluded that mandatory winterization was not cost-effective. The probability of a severe recurrence was too low. The probability was computed from models that assumed independence.

Link 3: Under-investment → System failure. When the polar vortex arrived, the systems that would have survived if winterized did not survive. Gas pipelines froze because they were not insulated. Generating units tripped because their instruments were not rated for sustained cold. Water treatment plants failed because their pumps lost power and their pipes were not buried below the frost line. Each failure was individually preventable. Collectively, they cascaded into the worst infrastructure disaster in Texas history.

Link 4: System failure → Death. Without power, without water, without gas, without passable roads, the most vulnerable populations—the elderly, the disabled, the medically dependent,

the isolated—had no way to maintain body temperature, no way to purify water, no way to reach medical care.

The chain is not speculative. Each link is documented in the FERC/NERC joint inquiry report, in regulatory proceedings, and in investigative reporting. The question is not whether the chain exists but how much of the terminal damage is attributable to the first link.

The Counterfactual

Suppose the models had been right. Suppose ERCOT and the infrastructure planners had used the streak cost function with the correct probability.

The correct probability would have changed the cost-benefit calculation by d'' . The permutation test measures this directly: at $K = 7$, the independence ratio is $32,000\times$ (Chapter 5), consistent with the order-of-magnitude estimate above. Winterization that was “not cost-effective” would have been overwhelmingly justified. A cost-benefit ratio that changes by four orders of magnitude changes the political calculus. It moves winterization from “academic curiosity” to “negligent not to implement.”

Would it have been mandated? Not certainly. Cost-benefit analyses do not automatically become policy. But the best estimate is that correct pricing would have led to targeted investment in the highest-leverage failure points: gas pipeline insulation, generating-unit cold-weather packages, backup power for water treatment. Engineering analyses suggest these investments, costing \$5–15 billion, would have prevented the cascade. Without the cascade, the localized cold damage would have been \$1–5 billion. The 246 deaths—driven overwhelmingly by the cascade, not by the cold—would have been largely prevented.

The pricing error did not kill those 246 people. The polar vortex, the under-engineered infrastructure, the political choices, and the inadequate emergency response all contributed. But the pricing error

was the first link. It was the reason the cost-benefit analysis said “don’t invest.”

Other Catastrophes

The same chain—mispricing → under-investment → failure → harm—operates in every domain where IID models inform decisions about extreme events.

The 2008 financial crisis. The Gaussian copula, which implicitly assumes tail independence, was used to price collateralized debt obligations. The pricing error for senior CDO tranches—which paid off only if many underlying mortgages defaulted simultaneously, a high- n question—was enormous. Li’s (2000) model said the senior tranches were safe. The real-world compound default probability, driven by a single common factor (the housing market), was vastly higher. Wrong prices led to overrated bonds, excessive leverage, and systemic collapse. Estimated cost: over \$10 trillion in lost output. The subsequent recession caused tens of thousands of excess deaths across developed economies, primarily through unemployment-related health effects, according to Stuckler et al. (2009) in *The Lancet*.

Hurricane Harvey (2017). Houston’s flood maps were based on models that treated each sub-watershed’s flood probability independently. The compound probability of simultaneous flooding across multiple watersheds was dramatically underestimated. When Harvey stalled and every watershed flooded simultaneously, the damage was \$125 billion. At least 68 people died directly, with over 100 total when indirect deaths are included.

COVID-19 hospital surge (2020–2021). Hospital capacity planning models treated each facility’s surge probability independently. The compound probability of all hospitals in a metropolitan area exceeding capacity simultaneously was underestimated. When the surge arrived, it arrived everywhere at once, overwhelming mutual-aid agreements that assumed some hospitals would have slack while

others were stressed.

In each case, the model underpriced the joint, sustained, synchronized extreme. The underpricing led to under-preparation. The under-preparation met reality.

The Accounting

A full accounting of the human cost of IID mispricing across all domains is beyond any single analysis. But the order of magnitude is estimable. Uri: at least 246 dead. Harvey: over 100. The financial crisis: tens of thousands of excess deaths across the OECD countries (the thirty-eight major developed economies) over 2008–2010. These are the events with clear causal chains from IID-based models to terminal outcomes.

These deaths were not inevitable. The cold was inevitable. The hurricane was inevitable. The pandemic was inevitable. The deaths from under-preparation were not. They followed from a computation—a computation that was wrong by d^n , computable on a laptop, using data that was already available.

The arithmetic of ruin is straightforward. An insurer who sets premiums using IID-estimated streak costs collects enough to cover $1/R$ of the actual loss. If $R = 50$, the premium covers 2% of the real cost. Across every domain tested—weather, river, wind, equity, epidemic—a simulated insurer pricing at the IID level reaches ruin with 100% probability within the simulation horizon. The IID premium does not underperform. It is not in the right order of magnitude.

The law was not hidden. It was unchecked.

CHAPTER 8

WHERE THE MONEY WENT

The pricing ratio tells you the model is wrong by 50,000×. The counterfactual tells you the error cost 246 lives and \$195 billion in Texas alone. Neither tells you *where to spend money to fix it*.

A regulator looking at the Texas crisis has a budget. Not an unlimited budget—one constrained by political reality, competing priorities, and the tax base. The question is not “Should the next Uri be prevented?” The answer to that is obviously yes. The question is “Given \$10 billion, what should it be spent on?”

The law gives the regulator something the models never did: a *decomposition of the cost*.

The Dollar Decomposition

The diamond from Chapter 3 gives an exact multiplicative decomposition: $R_{\text{total}} = R_{\text{persistence}} \times R_{\text{spatial}}$. Each factor corresponds to a structural feature of the data, and each structural feature corresponds to a physical mechanism that can be addressed.

For a single domain—say, the Texas cold grid—this two-way split tells the regulator how much of the pricing error comes from persistence (cold lasting seven days) and how much from spatial extent (all stations cold at once).

For a compound event like Uri, apply the same decomposition to the stacked matrix (Chapter 4). The stacked matrix’s total IID

ratio is larger because it includes cross-domain coupling on top of within-domain dependence. The ratio of the stacked-matrix R_{total} to the single-domain R_{total} isolates the cross-domain contribution—the additional error from treating the domains as independent of each other.

Each factor contributes a share of the excess cost. Taking logarithms converts a product into a sum, so multiplicative factors become additive shares:

$$\log R_{\text{total}} = \log R_{\text{persistence}} + \log R_{\text{spatial}} \quad (8.1)$$

The dollar attribution of each factor is proportional to its share of the log-total:

$$\text{Cost}_{\text{pers}} = \$195\text{B} \times \frac{\log R_{\text{persistence}}}{\log R_{\text{total}}} \quad (8.2)$$

For the Texas cold grid, with $R_{\text{total}} \approx 32,000$, $R_{\text{persistence}} \approx 40$, and $R_{\text{spatial}} \approx 800$ (from the Chapter 3 table):

$$\text{Persistence share} = \frac{\ln 40}{\ln 32,000} \approx 36\%$$

$$\text{Spatial share} = \frac{\ln 800}{\ln 32,000} \approx 64\%$$

Of Uri's \$195 billion in damage, approximately \$69 billion traces to duration—the event lasting seven days instead of one—and \$126 billion traces to spatial extent—the entire state freezing simultaneously instead of one region at a time. A regulator with \$10 billion has a quantitative basis for the split: roughly \$3.6 billion toward duration resilience (fuel reserves, equipment rated for sustained cold) and \$6.4 billion toward spatial resilience (interregional transmission, geographic diversification of supply).

This decomposition tells the regulator: of the \$195 billion, how much was caused by the event's duration, and how much by its

spatial and cross-domain extent?

This is one natural decomposition, not the only possible one. The logarithmic split assigns dollars to multiplicative factors the way a Shapley value assigns credit to players in a coalition. Other decompositions—including finer three-way splits that separate within-domain spatial dependence from cross-domain coupling—exist and would give somewhat different numbers. But the qualitative message—which factor dominates—is robust to the choice.

Which Fix Saves the Most?

Each factor maps to a different class of interventions:

Factor	Physical mechanism	Intervention class
Persistence	Cold persists for days, depleting reserves	Energy storage, fuel reserves, sustained-operation ratings
Spatial extent	All stations cold simultaneously (shared polar vortex)	Geographic diversification, interregional transmission
Compound	Cascading failures across power, gas, water, medical	System isolation, redundant supply paths, independent backup

If 60% of the dollar cost is attributable to compound effects—the cascading failure across domains—then the highest-return investment is in system isolation: ensuring that gas supply does not depend on the electric grid, that water treatment has independent power, that hospitals have independent fuel.

If instead 60% is attributable to persistence—the sheer duration of the cold—then the priority is reserves: fuel stockpiles, battery storage, equipment that can operate for seven days at extreme temperatures.

The correct policy depends on the decomposition. Without it, the regulator is guessing.

What Model to Use Instead

The decomposition does more than allocate dollars. It prescribes the minimum model complexity needed.

If the persistence ratio is large but the spatial ratio is close to 1: your model needs to capture temporal dynamics (autocorrelation, regime persistence, ARMA or hidden-Markov structure) but can treat stations as exchangeable. A univariate time series model per station, with independent draws across stations, may suffice.

If the spatial ratio is large but the persistence ratio is close to 1: your model needs to capture spatial dependence (copulas, spatial random fields, common-factor models) but can treat each day as independent of the previous day. A cross-sectional model with IID time steps may suffice.

If both are large: you need both. A multivariate time series model—a vector autoregression, a dynamic copula, or a regime-switching spatial model—is the minimum that captures the dependence structure your cost function is sensitive to.

A variant of the row shuffle tells you *at what timescale* the persistence operates. Instead of shuffling individual days, shuffle blocks of S days—preserving within-block dynamics but breaking the order of blocks. In the Texas cold data at $K = 7$, block size $S = 1$ (the standard row shuffle) gives a persistence ratio of $40\times$ —matching the row-shuffle result in Chapter 3. Block size $S = 7$ drops the ratio to near $1\times$. A block bootstrap at five-day block length—the formal statistical analogue of this block shuffle—reduces the persistence ratio to below $2\times$ across all domains tested, including those where the day-level ratio exceeds $1,000\times$. Nearly all the dependence that drives mispricing lives at the synoptic scale: 1–7 days. This tells the engineer that fuel reserves and equipment ratings need to withstand

a seven-day event, not a thirty-day event. The duration that drives the pricing error is the autocorrelation timescale of the physical driver—the weather pattern, the market regime, the epidemic wave—not the longest event in the historical record.

If the compound ratio is large: you need a model that couples domains. The gas model must know about the grid. The grid model must know about the weather. The hospital model must know about both. Siloed domain models, however sophisticated individually, will miss the cross-domain coupling that drives compound catastrophes.

In every case, the replacement model should be validated not by checking its variance or its mean—those are $n = 1$ diagnostics—but by checking its streak cost. Compute the pricing ratio of your new model against the data. If the ratio drops from $50,000\times$ to $2\times$, the model is capturing the dependence that matters. If it drops to $500\times$, you have closed two orders of magnitude and still have three to go.

The framework does not just diagnose the error. It grades every candidate replacement.

The Audit

For any system where IID-based models currently inform risk decisions:

1. **Define the cost function.** Not the variance. Not the annual average loss. The actual cost of the catastrophic scenario. For a power grid: a sustained, statewide blackout. For a hospital system: region-wide ICU saturation. For a portfolio: a multi-sector, multi-week drawdown. Determine which side of the sign boundary the cost function falls on: streak costs ($R > 1$, IID underpredicts) or switching costs ($R < 1$, IID overpredicts). If your cost function is a switching cost, the IID model is conservative and the urgency is lower.
2. **Compute the three ratios.** Column shuffle (independence),

row shuffle (persistence), lag shift (synchronization). The code is in the next chapter.

3. **Decompose.** The log-decomposition gives each factor's contribution. Map each factor to its physical mechanism and its intervention class.
4. **Compare to your current model.** If your model already accounts for persistence but not synchronization, the persistence ratio will be close to 1 and the synchronization ratio will be large. The decomposition tells you which assumptions are hurting you.
5. **Build the replacement.** The decomposition prescribes the model architecture. Validate the replacement with the streak cost, not variance.
6. **Prioritize spending.** The dollar attribution, not the ratio alone, determines where to invest. A large ratio on a small cost base may matter less than a moderate ratio on a large cost base.

This audit requires only the data you already have and the code in the next chapter.

What You Lose by Assuming Independence

The pricing ratio tells you the model misprices the extreme. But the damage goes deeper: IID distorts everyday predictions too.

A risk manager who builds 95% prediction intervals under the IID assumption—using conformal prediction, bootstrap, or any method that shuffles time—achieves actual coverage of 30% to 55% across the domains tested.

The interval that was supposed to miss 5% of outcomes misses 45% to 70%.

This is not a subtle calibration error. It is a prediction interval that misses nearly half the outcomes it was designed to capture.

The coverage gap is worst in domains with strong persistence—the Potomac and drought domains, where the persistence ratio exceeds 600×, sit at the low end of the range. Domains where persistence is modest but synchronization is extreme, such as the equity data, sit at the high end. The gap grows with K : a 95% interval for one-day events may be adequate; the same interval applied to a seven-day event is not.

What Replaces IID

The question after diagnosis is repair. Four models of increasing complexity were tested against the real pricing ratio in every domain:

Model	Gap closed
IID (baseline)	0%
Gaussian copula alone	3–32%
AR(1) + Gaussian copula	30–65%
VAR(1)	45–91%
Regime-switching + copula	87–100%

The Gaussian copula alone—the standard fix in quantitative finance since the 1990s—captures at most a third of the error, and often far less, because it preserves the temporal independence that drives the streak cost. The range reflects the decomposition: the copula closes most of the gap in spatial-dominant domains (equity, where synchronization accounts for nearly all the error) and almost none in temporal-dominant domains (Potomac, drought, where persistence drives the ratio). Adding an AR(1) process to each marginal before coupling them through the copula recovers roughly half the gap. A vector autoregression, which captures both temporal persistence and cross-sectional dependence in a single model, recovers most of it.

A regime-switching model—two states, “calm” and “stress,” with different persistence and correlation parameters in each state—closes effectively all of it.

A practitioner who cannot build any of these models can still act on the ratios directly. If the permutation test says the IID cost estimate is wrong by 54×, multiply the estimate by 54. This is crude but conservative, and it is strictly better than the status quo of multiplying by 1.

Where It Has Worked

In the Netherlands, the Delta Programme—responsible for protecting a country that is largely below sea level—updated its flood risk models in 2017 to account for compound events: storm surge arriving simultaneously with river flooding and heavy rainfall. The previous models had treated each hazard independently. The updated models, which explicitly captured the dependence between coastal and river flooding, produced risk estimates for compound events that were substantially higher than the independent estimates. The result was a reallocation of billions of euros in flood defense investment, shifting resources from barriers that addressed single-hazard scenarios to systems designed for compound ones.

The Netherlands did not wait for the compound event. It computed the ratio, found it was large, and changed the infrastructure plan. It is the clearest existing example of what this book argues every risk-managing institution should do.

PART IV

THE FIX

CHAPTER 9

FORTY LINES

Here is the code. It is written in Python, but it would be the same in R, Julia, MATLAB, or any language with a random-number generator and array operations. It computes all three pricing ratios—*independence*, *persistence*, and *synchronization*—for any data matrix and any cost function.

If you do not program, skip to “What the Output Looks Like” on page 79—the results are what matter, not the implementation.

The Cost Function

```
1 import numpy as np
2
3 def streak_cost(X, threshold, K, M):
4     """
5     X:          (T, N) matrix -- rows are days,
6               columns are stations
7     threshold: value below which a station is '
8               extreme'
9     K:          required consecutive days
10    M:          required number of stations
11               simultaneously extreme
12    Returns:    number of streak-days in the record
13    """
14    extreme = (X < threshold).astype(int)
```

```

12     n_extreme = extreme.sum(axis=1)
13     bad_day = (n_extreme >= M).astype(int)
14
15     cost = 0
16     run = 0
17     for t in range(len(bad_day)):
18         if bad_day[t]:
19             run += 1
20         else:
21             if run >= K:
22                 cost += run - K + 1
23             run = 0
24     if run >= K:
25         cost += run - K + 1
26     return cost

```

The function takes a data matrix (one row per day, one column per station), a threshold, a required streak length K , and a required number of simultaneous stations M . It counts consecutive runs where at least M stations are extreme. A run of length $L \geq K$ contributes $L - K + 1$ streak-days. This is the cost function that matters for infrastructure: not “how many days was it cold somewhere?” but “how many days fell within a sustained, widespread episode?”

The Three Shuffles

```

1 def shuffle_independence(X):
2     """Destroy all dependence: IID null."""
3     X_null = X.copy()
4     for j in range(X.shape[1]):
5         np.random.shuffle(X_null[:, j])
6     return X_null
7
8 def shuffle_persistence(X):
9     """Destroy temporal order, preserve spatial
10        patterns."""

```

```

10     idx = np.random.permutation(X.shape[0])
11     return X[idx, :]
12
13 def shuffle_synchronization(X):
14     """Destroy cross-column timing, preserve each
15         column's dynamics."""
16     X_null = X.copy()
17     T = X.shape[0]
18     for j in range(X.shape[1]):
19         lag = np.random.randint(0, T)
20         X_null[:, j] = np.roll(X_null[:, j], lag)
21     return X_null

```

Three functions, each destroying one structural feature:

- **shuffle_independence**: for each column, randomly permute its values. Preserves each station's marginal distribution. Destroys all cross-column dependence and all temporal structure. This is the IID null.
- **shuffle_persistence**: randomly permute the rows. Each row (a day's snapshot across all stations) stays intact, but the order of days is randomized. Preserves spatial patterns. Destroys temporal persistence.
- **shuffle_synchronization**: for each column, circularly shift it by a random offset. Each station's time series is preserved exactly—same values, same order, same autocorrelation. Only the alignment between stations is randomized.

The Ratio

```

1 def pricing_ratio(X, threshold, K, M, shuffle_fn,
2     B=1000):
3     """Compute the pricing ratio for a given null
4         hypothesis."""

```

```

3     real_cost = streak_cost(X, threshold, K, M)
4     if real_cost == 0:
5         return 1.0
6
7     null_costs = np.array([
8         streak_cost(shuffle_fn(X), threshold, K, M
9         )
10        for _ in range(B)
11    ])
12
13    mean_null = null_costs.mean()
14    if mean_null == 0:
15        return float('inf')
16
17    return real_cost / mean_null

```

Compute the real cost. Generate $B = 1,000$ shuffled copies. Compute the cost on each. Return the ratio.

Running the Audit

To audit your own data:

```

1  # Load your data into X (T rows, N columns)
2  # X = np.loadtxt('your_data.csv', delimiter=',')
3
4  R_indep = pricing_ratio(X, threshold, K, M,
5                          shuffle_independence)
6  R_pers  = pricing_ratio(X, threshold, K, M,
7                          shuffle_persistence)
8  R_sync  = pricing_ratio(X, threshold, K, M,
9                          shuffle_synchronization)
10
11 print(f"Independence ratio: {R_indep:.1f}x")
12 print(f"Persistence ratio:  {R_pers:.1f}x")
13 print(f"Synchronization:    {R_sync:.1f}x")

```

Three numbers: the total mispricing, the temporal component, and the synchronization component. On a modern laptop, 1,000 permutations of a 20,000-row, 10-column matrix complete in under a minute. The computation is fast; the upstream work—assembling the matrix from organizational data, aligning records across systems, choosing K , M , and a threshold that maps to the cost function that matters—is where the labor lives.

A Complete Example

Here is a complete run on real data. The Fama–French daily industry portfolio returns are publicly available. Download the ten-industry daily file, load it, and run the audit:

```
1 import pandas as pd
2
3 # Fama-French 10 industry portfolios, daily
  returns
4 url = ("https://mba.tuck.dartmouth.edu/pages/
  faculty/"
5       "ken.french/ftp/10
  _Industry_Portfolios_daily_CSV.zip")
6 df = pd.read_csv(url, skiprows=9, nrows=25000)
7 df = df.iloc[:, 1:] # drop date column
8 X = df.values / 100 # convert from percent to
  decimal
9
10 threshold = -0.02 # a 2% daily loss
11 K, M = 5, 4 # 5 consecutive days, 4+
  sectors down
12
13 R_indep = pricing_ratio(X, threshold, K, M,
  shuffle_independence)
14 R_pers = pricing_ratio(X, threshold, K, M,
  shuffle_persistence)
15 R_sync = pricing_ratio(X, threshold, K, M,
```

```

        shuffle_synchronization)
16
17 print(f"Independence ratio: {R_indep:.1f}x")
18 print(f"Persistence ratio: {R_pers:.1f}x")
19 print(f"Synchronization: {R_sync:.1f}x")

```

On a standard laptop, this produces ratios in the thousands for independence, single digits for persistence, and over a hundred for synchronization—matching the equity results reported in Chapter 5. The entire script, including the functions above, is under sixty lines. The computation takes less than two minutes.

Seeing the Law

To see d^n in your own data:

```

1 for K in [1, 2, 3, 5, 7]:
2     R = pricing_ratio(X, threshold, K, M,
        shuffle_independence)
3     print(f"K={K}: {R:.1f}x")

```

Plot $\log R$ against K . If the points fall on a line—and in every domain I have tested, they do—the slope is $\log d$ and the law holds for your data.

Does the Test Work?

I validated the test on data where the answer is known. First, synthetic IID data—independent draws from a normal distribution, shaped into the same matrix format as the real data. All three ratios return values indistinguishable from 1.0x. The test does not manufacture dependence.

Second, synthetic data with *known* dependence: a VAR(1) process with specified persistence and cross-correlation parameters. The test recovers the planted dependence structure. The persistence ratio

grows with the AR coefficient. The synchronization ratio grows with the cross-correlation. When both are set to zero, all ratios return to 1.0×

Third, a robustness check: replacing every value in the real data with its rank (1st, 2nd, . . . , T th within each column) changes the ratios by less than 12%. The test responds to the *ordinal* structure of the data—which days are colder than which—not to the magnitudes. This makes it robust to outliers, measurement scale, and distributional assumptions.

Fourth, a precision check: the ratio R is computed as $T_{\text{real}}/\bar{T}_{\text{null}}$, where the numerator is fixed and the denominator is a sample mean over B permutations. The standard error of the denominator decreases as $1/\sqrt{B}$, so at $B = 1,000$ the Monte Carlo noise in the ratio is small. A ratio of 50× is distinguishable from 5× with high confidence; a ratio of 1.01× versus 1.00× is not. The reported ratios have more than enough precision for the conclusions drawn from them.

What the Output Looks Like

Below is the output for a seven-station Texas cold matrix ($M = 4$, threshold = 20°F, $B = 1,000$ permutations):

K	Independence	Persistence	Synchronization
1	5.1×	1.0×	4.2×
2	23×	1.9×	4.5×
3	97×	3.4×	4.8×
5	1,800×	12×	5.4×
7	32,000×	40×	6.1×

Read the table column by column. The independence ratio (total IID error) explodes with K : from 5× to 32,000×. The persistence ratio grows from 1 at $K = 1$ (where temporal order is irrelevant) to 40×

at $K = 7$ —matching the diamond decomposition in Chapter 3. The synchronization ratio is roughly constant: in this domain, the polar vortex provides a steady synchronizing force regardless of duration. In a domain where synchronization depended on duration—say, a spreading epidemic—the synchronization column would grow faster.

The product of the persistence and synchronization columns will not exactly equal the independence column—the diamond decomposition routes through a fourth permutation, and these three tests are not the two edges of a single path. But the table shows the practitioner, at a glance, which structural feature dominates the error for each K .

What to Do with the Results

If all three ratios are close to 1: stop. IID pricing is adequate for this cost function. You have confirmed your model’s assumption and can move on.

If any ratio is between 2 and 10: the IID model is introducing a meaningful but not catastrophic error. Report the ratios alongside the model’s output. The error is worth knowing but not urgent enough to halt operations.

If any ratio exceeds 10: the IID model is not adequate for this cost function. Report the ratios alongside the current model’s output. Vary K and M to map out where in parameter space the error becomes large. The result is a table: rows are parameter settings, columns are the three ratios. That table is the deliverable of the audit—it tells decision-makers exactly which scenarios are mispriced and by how much. Chapter 8 gives the framework for translating the table into model requirements and spending priorities.

The catastrophe models that failed in Texas, in Houston, in 2008, and during COVID had thousands of parameters, proprietary datasets, and teams of PhD-level modelers. They failed because

they assumed independence. This code doesn't model anything. It measures what independence costs you.

CHAPTER 10

WHAT ELSE IS MISPRICED

The framework in this book was developed on weather data and tested across twenty-five datasets spanning weather, hydrology, finance, energy, epidemiology, seismology, agriculture, insurance, civil infrastructure, and planetary science. In every domain, the same code produced the same result—large ratios, exponential growth with K , multiplicative decomposition.

But the framework is not about weather, or finance, or power grids. It is about a matrix with rows and columns, a cost function with a nonlinearity parameter, and a set of permutations that decompose the pricing error. Nothing in the mathematics depends on what the rows and columns represent.

Before the details, here is the summary. The following table collects the measured ratios for six domains—three from earlier chapters and three new ones that this chapter presents. The “Dominant” column identifies whether the pricing error is primarily temporal (persistence-driven), spatial (synchronization-driven), or both.

Domain	K	R_{indep}	R_{pers}	R_{sync}	Dominant
Texas cold	5	1,800×	12×	5.4×	Both
Equity drawdown	5	—	3.8×	113×	Spatial
Potomac (river)	7	—	679×	1.3×	Temporal
COVID-19	5	>10,000×	—	145×	Spatial
Earthquake	5	16×	15×	~1×	Temporal
Drought (soil)	5	2,300–5,000×	>1,200×	>85×	Both

Dashes mark ratios not reported individually in the text. Read the “Dominant” column as a prescription: temporal-dominant domains need duration models; spatial-dominant domains need synchronization models; domains where both are large need multivariate time series. Every domain falls above the sign boundary from Chapter 5: IID underprices all of them. The code to produce every number in this table is in the preceding chapter.

The first half of this chapter extends the computation to three new domains with fundamentally different dependence mechanisms: epidemic contagion, seismic triggering, and atmospheric drought. The second half maps the framework onto four domains where I predict—but have not yet computed—that the same law holds. The reader should hold those predictions to the same standard this book holds every other claim: compute it or don’t trust it. But the law makes the predictions sharp enough to be worth stating—and worth testing.

Epidemiology: Outbreak Costs

An epidemic is a matrix: rows are days, columns are states or cities, entries are case counts. The cost depends on whether cases are spread across many small outbreaks (manageable) or concentrated in synchronized waves (overwhelming).

I tested this on COVID-19 daily incidence data across ten U.S. states. The ratio at $K = 5$ exceeds $10,000\times$ —the null permutations never once reach the real cost. The diamond decomposition shows why: the spatial component dominates. The synchronization ratio is $145\times$ —epidemic waves propagate through the airline network so efficiently that every state surges together. A model that treats each state’s epidemic trajectory as independent underestimates the probability of simultaneous nationwide hospital saturation by orders of magnitude. The mechanism is social-network contagion,

fundamentally different from the geographic coupling that drives the weather domains, yet the law produces the same exponential growth.

Seismology: Aftershock Clustering

The earthquake catalog provides a test of a third dependence mechanism: triggered clustering. I divided seismically active California into eight subregions and computed daily maximum magnitudes across 2000–2019. The ratio at $K = 5$ is $16\times$ —orders of magnitude smaller than weather or epidemics, but statistically significant ($p = 0.010$). The decomposition reveals that the effect is almost entirely temporal: the persistence ratio is $15\times$, the synchronization ratio is near $1\times$. Aftershock sequences (Omori’s law) create temporal clustering within a region, but large earthquakes in one subregion do not reliably trigger simultaneous activity in distant subregions. The test correctly distinguishes triggered dependence from spatial dependence. By contrast, earthquake counts across *separate* tectonic plates—the negative control—return $1.01\times$, confirming that the California result reflects genuine seismic coupling, not a test artifact.

Agriculture: Drought

Soil moisture across the Great Plains provides a test of the slowest dependence mechanism: atmospheric blocking. I measured daily soil moisture at eight agricultural sites across Iowa, Nebraska, Kansas, and the Dakotas during growing seasons from 2000 to 2019. The ratio at $K = 5$ ranges from $2,300\times$ to $5,000\times$ depending on the threshold. Both spatial and temporal components are large—the persistence ratio exceeds $1,200\times$ and the synchronization ratio exceeds $85\times$ —because blocking patterns that cause drought are spatially vast and persist for weeks to months. This timescale is an order of magnitude longer

than the synoptic-scale weather dependence in the original domains. An IID crop-insurance model that prices each county independently underestimates the probability of simultaneous multi-state drought by a factor that grows exponentially with the number of counties in the policy.

The remaining six domains are predictions, not measurements. I have not run the permutation test on cybersecurity incident data, supply-chain disruption logs, genomic matrices, species-abundance tables, alloy-composition databases, or classroom data. But the law makes the predictions sharp enough to be worth stating—and worth testing.

Cybersecurity: Supply-Chain Compromise

In December 2020, the SolarWinds Orion software update was compromised. A single poisoned build propagated to approximately 18,000 organizations—government agencies, Fortune 500 companies, critical infrastructure operators—within the same two-week update window. The attack was not a series of independent intrusions. It was one intrusion, replicated simultaneously across every organization that ran the same software, through the same mechanism.

The data is a matrix: rows are days, columns are organizations, entries are compromise indicators. The cost function is the number of days during which M or more organizations are simultaneously compromised. The independence assumption—each organization's security is an independent draw—says the probability of 18,000 simultaneous compromises is the product of 18,000 individual probabilities. The reality is that all 18,000 shared a common driver: the same software supply chain.

SolarWinds is the CDO crisis restated in cybersecurity. The “diversification” of deploying across government, finance, energy, and defense is illusory when every deployment shares the same

software dependency—just as holding CDO tranches built from the same mortgage pool provided illusory diversification. The IID model of organizational security, which underpins most cyber-insurance pricing and most government risk assessments, treats each organization's breach probability as independent. The supply-chain attack violates this assumption with the same totality that the polar vortex violated station independence.

NotPetya (2017) demonstrated the same structure at even larger scale: a single compromised Ukrainian tax-software update cascaded to Maersk, Merck, FedEx, and hundreds of other multinational corporations within hours. Estimated damages exceeded \$10 billion. Both attacks shared the signature of a high- n compound event: many components failing simultaneously, driven by a single common cause, priced by models that assumed independence.

Supply Chains: The Semiconductor Shortage

A global supply chain is a matrix: rows are weeks, columns are nodes (factories, ports, suppliers, logistics routes), entries are throughput or disruption indicators. The cost function is the number of weeks during which M or more critical nodes are simultaneously disrupted.

In 2021, three nominally unrelated disruptions converged. The Suez Canal blockage (March) halted 12% of global trade for six days. The Texas freeze (February) shut down petrochemical plants that supply semiconductor fabrication materials. The COVID-driven demand surge overwhelmed chip fabrication capacity that was already near its limit. Each disruption was analyzed independently. Each was assessed as a low-probability, high-impact event. The compound probability of all three occurring within the same quarter was assessed as negligible.

But the disruptions were not independent. They shared common drivers: the same lean-inventory philosophy that left no buffer at any node, the same geographic concentration of fabrication capacity

(Taiwan, South Korea, a handful of U.S. plants), and the same demand shock (a pandemic that simultaneously increased demand for electronics and reduced manufacturing capacity). The correlation between nodes is not a design choice. It is a consequence of the economic optimization that built the supply chain—the same optimization that, in every domain this book examines, creates dependence and then prices it as independence.

The automobile industry lost an estimated \$210 billion in revenue in 2021 from semiconductor shortages alone. Supply-chain models that treated each node's disruption probability as independent—the standard in procurement risk management—underestimated the probability of simultaneous multi-node failure by a factor that grows as d^n with the number of coupled nodes.

Genomics: The Missing Heritability

The human genome contains approximately 20,000 protein-coding genes. Genome-wide association studies (GWAS) scan millions of genetic variants to identify which are associated with diseases—diabetes, heart disease, schizophrenia, cancer. The standard approach tests each variant independently: does this variant, by itself, increase disease risk?

The results have been puzzling for two decades. For most heritable diseases, the variants identified by GWAS collectively explain only 20 to 50 percent of the heritable variation. The rest—the “missing heritability”—is unaccounted for.

The standard explanation is that many variants with individually tiny effects remain below the detection threshold. This is almost certainly part of the answer. But there is another possibility, and it maps directly onto the independence pricing error.

GWAS assumes that genetic variants act additively. Variant A adds 0.1% risk. Variant B adds 0.15% risk. The combined effect is 0.25%. This is the $n = 1$ cost function—the linear model.

But genes interact. The effect of variant A may depend on whether variant B is also present. In genetics, this is called *epistasis*. In the language of this book, it is cross-column dependence. The column shuffle—randomly reassigning which variants go with which individual—would reveal the pricing error from assuming additivity.

The cost function for disease is nonlinear. You don't get 20% of diabetes. You either develop it or you don't. The threshold nature of disease—healthy until enough things go wrong simultaneously—makes it a high- n cost function. The law predicts that the additive model underestimates disease risk from interacting variants by d^n , where d reflects the correlation between variants (linkage disequilibrium) and n reflects the number of variants that must interact to push past the disease threshold.

The data is a matrix: rows are individuals, columns are genetic variants, entries are genotypes. The cost function is disease status. The column shuffle gives the independence ratio. If the ratio is large—and the law predicts it will be, for any disease with a threshold mechanism and correlated variants—it would account for a substantial fraction of the missing heritability.

If confirmed, this would redirect drug development. Current target identification is based on additive-effect GWAS. If the real causal structure is interactive, the targets are wrong—not individually, but in combination. Billions of dollars in R&D are allocated on the basis of an $n = 1$ analysis of what may be an $n = 5$ phenomenon.

Ecology: Ecosystem Function

An ecosystem provides services: clean water, pollination, carbon storage, nutrient cycling. These services depend on which species are present and how they interact.

The data is a matrix: rows are sites, columns are species, entries are abundances. The cost function is ecosystem function—productivity, pollination rate, nutrient retention. The column shuffle randomly

reassigns species abundances across sites, destroying inter-species dependence.

The additive assumption is called the “portfolio effect”: more species means more stable function, because you’re averaging over independent contributors. The framework tests this directly. If the ratio is large, species interactions create ecosystem function that the additive model misses.

Evidence from biodiversity experiments suggests interaction effects are large—“overyielding,” where mixtures produce more than the sum of monocultures, is well documented. The framework would quantify this as an exact ratio and decompose it into the same multiplicative factors.

Materials Science: Alloy Strength

An alloy is a matrix: rows are samples, columns are elements, entries are concentrations. Mechanical properties—strength, ductility, corrosion resistance—depend on element interactions.

The additive model says: adding 2% chromium increases strength by x , regardless of other elements. The reality is that chromium’s effect depends strongly on carbon and nickel content. These interactions are the basis of metallurgy, but they have never been systematically quantified as pricing errors.

For high-entropy alloys—five or more elements in near-equal proportions—the interactions are the entire point. The effective n is 5 or higher. The field has already abandoned additive models in favor of high-throughput experimentation, precisely because the interaction effects dominate. The framework would quantify what the field discovered empirically.

Education: Peer Effects

A school district is a matrix: rows are school-years, columns are schools, entries are aggregate test scores. The cost function is a district-wide outcome—the number of schools simultaneously below an accreditation threshold, or the probability that all schools in a region fail to meet a standard in the same year.

The additive model says each school's performance is independent. The reality is that schools in the same district share teachers, funding cycles, curriculum mandates, and student demographics. A budget cut, a policy change, or a cohort effect hits all schools at once. These are common factors—the educational analogue of the polar vortex.

The cost function has threshold structure: a school either meets accreditation or doesn't; a district either triggers state intervention or doesn't. The nonlinearity is $n \geq 2$. The law predicts that models treating each school independently underestimate the probability of simultaneous district-wide failure by d^n , where d reflects the correlation induced by shared drivers and n is the number of schools required to fail. If the ratio is large, school-level interventions are less effective than district-level policy—because the driver is common, not idiosyncratic.

Stability

A natural objection: perhaps the ratios are artifacts of a particular time period. Perhaps the 1990–2019 record happened to contain an unusual cluster of events, and a different thirty years would give different numbers.

I tested this directly. Stratifying by calendar month—testing whether February cold in Dallas predicts February cold in Houston, within the same month, not whether winter predicts winter—reduces the unstratified ratios substantially (cold snaps from $>1,000\times$ to $28\times$;

wind lulls from 334× to 39×) by removing seasonal co-movement. But the stratified ratios are still enormous. The dependence that survives is real within-month, within-season spatial coupling—the polar vortex pushing all stations cold simultaneously, not just winter being cold everywhere.

Four negative-control domains confirm the test does not manufacture dependence. Hard drive failures across different drive models: ratio indistinguishable from 1×. Auto insurance claim frequencies: within 4% of 1×. Unemployment rates across U.S. states after stratification: null. Earthquake counts across separate tectonic plates (California, Chile, Mediterranean): 1.01×. Where the physical mechanism provides no reason for cross-component dependence, the test finds none.

The ratios are not data artifacts, period artifacts, or seasonal artifacts. They are structural properties of the dependence, and they are stable.

A subtler stability result is counterintuitive. I split each domain's time series into "calm" and "stress" regimes—below-median and above-median volatility windows—and computed the pricing ratio within each regime separately. The calm-period ratio is *higher* than the stress-period ratio in every domain tested. Persistence is stronger during quiet periods: a calm day predicts another calm day more reliably than a stressed day predicts another stressed day. The IID assumption is most wrong not during the headline crisis but during the slow, unremarkable multi-day event—the week-long cold snap that never makes the news, the steady equity drawdown that never triggers a VIX spike. The infrastructure is most vulnerable to mispricing in conditions that no one is monitoring.

This result is distinct from the tail intensification in Chapter 5, and the two are not contradictory. Tail intensification says d grows as the threshold becomes more extreme—rarer events are more dependent. Regime asymmetry says the pricing ratio is higher during calm windows than during volatile ones. The first is about the

severity of the event being priced; the second is about the background conditions during which dependence accumulates. Both are true simultaneously: the ratio is largest for rare events occurring within quiet periods.

The Direction

The ratios computed above are historical. They describe the dependence structure of the 1990–2019 record. But the dependence structure is not static.

Climate science provides strong evidence that the spatial correlation of extreme weather events is increasing. Polar vortex disruptions—the mechanism behind Winter Storm Uri—are becoming more frequent as Arctic warming weakens the jet stream. Heat waves are becoming more spatially coherent as the background temperature rises. The correlation between river basins' low-flow periods is increasing as snowpack declines reduce the independence of baseflow sources.

Financial markets are becoming more correlated through index funds, algorithmic trading, and cross-border capital flows. The average pairwise correlation of equity sector returns during stress events has increased in every decade since 1970. The mechanisms that synchronize sectors during a crisis—margin calls, fire-sale liquidation, risk-parity rebalancing—operate faster and more broadly than they did a generation ago.

If the dependence strength d is increasing, the pricing error d^n is increasing exponentially faster. A 10% increase in d produces a 10% increase in the pricing error for $n = 1$ (variance), but a 61% increase for $n = 5$ and a 95% increase for $n = 7$. The law amplifies trend. Models calibrated on historical d , applied to a system where d has increased by even a modest amount, underestimate the current risk by a factor that grows with the nonlinearity of the cost function.

The question is not just “How large is the ratio today?” but

“Which direction is it moving?” In every domain where multi-decade data exists, the answer is: toward larger d , and therefore toward larger d^n .

What the Framework Does Not Cover

The law requires stationarity—or at least enough of it to estimate a ratio. In a system where the dependence structure is changing faster than the record length, the historical ratio may not predict the next event. Climate change is pushing d upward (as discussed above), which means historical ratios are lower bounds for future risk. But a system undergoing a structural break—a market regime that has no historical analogue, a novel pandemic pathogen—may have a d that the record cannot estimate at all. The framework diagnoses mispricing in systems with enough history to permute. It does not conjure a ratio from insufficient data.

The test requires a matrix. Systems where the cost depends on network topology rather than simultaneous threshold exceedances—cascading failures in a directed graph, for instance—are not naturally represented as rows and columns. The Texas cascade *was* representable because it involved simultaneous failures across nodes. A cascade that propagates sequentially through a chain, with each failure depending on the previous one, has a different structure. The permutation test as described here does not address it.

The exponential form $R \approx d^n$ is empirical. The KL-divergence argument in Chapter 5 shows why $\log R$ must grow linearly in K , but the specific slope—the effective $\log d$ —is measured, not derived. Two systems with the same pairwise correlations can have different higher-order dependence structures and therefore different effective d values at high n . The law predicts the functional form but not the constant without measurement.

The General Pattern

In every domain, the pattern is the same:

1. The standard model assumes components act independently.
2. The model is tested with low- n diagnostics (averages, variances, correlations) and appears adequate.
3. The cost function that actually matters is high n : a threshold, a streak, a simultaneous multi-component failure.
4. The pricing error for the high- n cost function is d^n .
5. The error is discoverable with the code in this book.

The law is not a property of weather or finance. It is a property of any system where dependent parts contribute to a nonlinear cost. The question is not whether other domains are mispriced. The law guarantees they are, whenever $d > 1$ and $n > 1$ —whenever the parts are dependent and the cost has a threshold. The question is how large the ratio is for each specific domain, and what decisions rest on the wrong number.

The code is in Chapter 9. The data is in your organization's databases. The ratio will tell you whether your models are adequate or whether they are wrong by orders of magnitude.

CHAPTER 11

THE NEXT MODEL

An AI evaluation is a matrix.

The rows are test instances—red-team sessions, multi-turn conversations, capability probes. The columns are capability dimensions—coding, reasoning, persuasion, planning, deception, tool use. Each entry is a score: did the model exhibit dangerous-level capability in this dimension on this instance?

A large language model is evaluated on six capability dimensions. Each dimension is tested independently. On each, the model scores below the dangerous threshold more than 90% of the time. The evaluation reports: each capability, individually, is within acceptable bounds. The model is deployed.

Seven weather stations, each checked individually, each adequate. A model with six capabilities, each checked individually, each acceptable. The structure is identical. What the evaluation did not compute is the same question Chapter 1 asked of Texas: How often does the model exhibit exploit-level coding AND persuasion AND deception AND planning, *simultaneously*, across K consecutive interactions? That is a streak cost: M capabilities above threshold for K consecutive turns. It is a high- n cost function. And the evaluation that checked each capability independently is an $n = 1$ diagnostic applied to an $n \geq 5$ risk.

The law predicts the evaluation underestimates the compound risk by d^n .

The Common Driver

In every domain tested, the pricing error is driven by a common factor that pushes all components in the same direction at the same time. The polar vortex pushes all Texas stations cold. The housing collapse pushes all mortgage pools into default. The pandemic wave pushes all hospitals into surge.

For an AI model, the common driver is the training process itself. The capabilities are not separate modules assembled from independent parts. They are emergent properties of a single set of weights, optimized by a single objective, over a single corpus. When the training run produces a model that is better at coding, it is also—not independently, but consequentially—better at reasoning, planning, and tool use. The weights do not know the difference between capabilities. They encode a general competence surface, and every capability is a cross-section of the same surface.

The effective dependence strength d between capabilities is therefore not modest. It is structural. The conditional probability of strong persuasion ability given strong coding ability is not the unconditional probability of strong persuasion ability—it is far higher, because both are driven by the same underlying competence. This is the coupling surplus from Chapter 4. And the surplus grows in the tail: the correlation between capabilities is stronger for the most capable models than for average ones, just as the correlation between Texas stations is stronger during polar vortex events than on normal days. The models where the compound risk matters most are exactly the models where the independent evaluation is most wrong.

The law does not predict the exact d for AI capabilities—that requires the computation. But if d is comparable to weather systems ($d \approx 4$ – 6) and five capabilities must be simultaneously active ($n = 5$), the predicted compound ratio is 4^5 to 6^5 : roughly $1,000\times$ to $8,000\times$. I have not computed the ratio. These numbers are predictions, subject to the same standard this book applies everywhere: compute it or

don't trust it.

But consider what a ratio of $1,000\times$ means. Suppose the independent-capability evaluation predicts that five capabilities will simultaneously exceed the dangerous threshold in one out of every million sessions. The ratio says the real frequency is 1,000 times higher: one session in a thousand. The evaluation says lightning. The law says rain.

Why It Hasn't Been Computed

Each of the five barriers from Chapter 5 applies.

The variance trap. AI safety benchmarks measure average performance on each capability dimension—accuracy, refusal rate, toxicity rate. These are $n = 1$ metrics. At $n = 1$, $d^n = d$, which is 2 to 5. The benchmark looks adequate. Three generations of machine-learning researchers have been trained on average-case evaluation, just as three generations of risk managers were trained on variance.

The p-value shadow. Red-team evaluations report whether the model *can* produce dangerous output—a binary. They do not report the ratio: compound dangerous behavior under the real model divided by what independent-capability baselines would predict. The number that tells you how wrong the evaluation is sits in the intermediate data and is never computed.

The magnitude filter. A claim that AI safety evaluations underestimate compound risk by three to four orders of magnitude will be received with the skepticism that outsized claims always attract. The larger the true ratio, the less likely the finding survives review.

The disciplinary fracture. The ML researcher who builds the evaluation does not think in terms of streak costs. The policy analyst who writes the governance framework does not have the permutation test. The security researcher who models adversarial persistence does not have the multiplicative decomposition. The law sits at an intersection that no field's standard practice occupies.

The Nash equilibrium. A laboratory that evaluates compound risk—testing not just individual capabilities but their correlated activation over sustained interactions—delays deployment while its competitors, using standard benchmarks, deploy faster and capture the market. The laboratory that is right about the risk loses market share to the laboratory that is wrong about the risk. The equilibrium breaks only when the compound event occurs. Then it breaks all at once, across every deployment of every model that was evaluated with $n = 1$ diagnostics.

What the Test Would Show

The permutation test from Chapter 9 applies directly. Build the matrix: rows are evaluation instances—red-team sessions, multi-turn conversations, capability probes—and columns are capability dimensions. Each entry is a score or a binary indicator: did the model exhibit this capability in this instance?

The column shuffle randomly permutes each capability’s scores across instances, destroying cross-capability dependence while preserving the marginal distribution of each capability. The cost function is the compound criterion: M or more capabilities above their dangerous thresholds for K consecutive interactions.

The three ratios would give an AI governance body something no current framework provides:

- The **independence ratio**: by how much does independent-capability evaluation underestimate compound dangerous behavior?
- The **persistence ratio**: does dangerous behavior in interaction 5 predict dangerous behavior in interaction 6? If the ratio is large, the model maintains dangerous capability across a session—the AI analogue of a cold snap that persists for days.

- The **synchronization ratio**: do capabilities activate together, or independently across the evaluation set? If the ratio is large, the model’s dangerous capabilities are coordinated—coding and deception and planning fire in the same sessions, not in different ones.

And the diamond identity from Chapter 3 holds:

$$R_{\text{total}} = R_{\text{persistence}} \times R_{\text{spatial}}$$

where $R_{\text{spatial}} = R_{\text{total}}/R_{\text{persistence}}$, as in Chapter 3. The spatial ratio measures the cross-capability dependence that remains after temporal persistence is removed—a derived quantity from the diamond, distinct from the synchronization ratio measured by the lag shift. The decomposition is exact, just as it was for weather, finance, and epidemiology.

Each factor points to a different intervention. If the persistence ratio is large—say 40×, the same order as Texas cold—the model maintains dangerous behavior across turns the way a cold front persists for days. The response is session-level: limit memory, restrict context carryover, monitor per interaction. If the spatial ratio is large—say 25×—the capabilities are correlated, activating together rather than independently. The response is architectural: separate the capability modules, firewall the dangerous dimensions, deploy modularly so that coding and persuasion and planning do not share the same inference path.

Factor	Mechanism	Intervention
Persistence	Dangerous behavior carries across interactions	Session memory limits, context restrictions
Spatial (co-activation)	Capabilities activate together	Capability firewalls, architectural separation
Compound (= $P \times S$)	Both factors simultaneously	Deployment gates on compound thresholds

Without the decomposition, the regulator is guessing at which lever to pull. With it, the allocation of safety investment follows from the data. And the diamond provides the internal check: if the product of the persistence and spatial ratios does not equal the total, something is wrong with the test—or there is a source of dependence the decomposition has not captured.

The Deployment Monoculture

The synchronization ratio has a second application. When many organizations deploy the same foundation model, a single failure mode affects all deployments simultaneously. The “diversification” of running AI in hospitals, banks, power grids, and government is illusory if every deployment shares the same base model—just as the diversification of holding CDO tranches built from the same mortgage pool was illusory.

The equity synchronization ratio—113× for ten industry sectors driven by a common market panic—measures this effect for financial contagion. The AI analogue: a systematic failure in the base model, triggered by a novel input class or a latent misalignment, propagates to every downstream application in the same week. The mutual-aid assumption—if one AI system fails, redirect to another—breaks when

the driver is common.

The correlation between deployments is not a design choice. It is a consequence of building on the same architecture, the same weights, and the same training corpus. The more the ecosystem converges on a small number of foundation models, the higher the effective d across deployments, and the larger the compound ratio for simultaneous failure across the entire economy.

The Sign Boundary

Not all AI risks are streak costs. Single-turn jailbreaks, isolated harmful outputs, and occasional bias are switching costs—operational noise that accrues with state changes. For these, IID-based evaluation is conservative: it overpredicts the frequency, because real harmful outputs are clustered rather than independently distributed. Current AI governance focuses on these risks, and rightly so—they are the most frequent.

The risks that the law identifies as underpriced are on the other side of the boundary: sustained deception maintained across a sequence of interactions; persistent goal-directed behavior that survives context switches; compound capability applied to a single objective over time. These are streak costs. They are the AI analogues of the week-long blackout, the multi-day financial crash, the sustained hospital surge. And they are the risks where IID-based evaluation is catastrophically unconservative—not by 10% or 50%, but by d^n .

IID evaluation is simultaneously too pessimistic about the risks that waste money—single-turn failures, easily patched—and too optimistic about the risks that cause irreversible harm.

I have not computed the ratio for AI. This chapter is a prediction, not a measurement, and it is subject to the same standard the book applies to every claim in Chapter 10: compute it or don't trust it. But the prediction is sharp. The matrix exists—every frontier laboratory

generates it with every evaluation run. The code is in Chapter 9. The permutation test requires no access to the model's weights, no theory of consciousness, no position on whether machines can think. It requires a matrix, a compound cost function, and a shuffle.

If the ratio is small—if AI capabilities are sufficiently independent that compound dangerous behavior is no more frequent than the product of individual frequencies—the law says so, and the current evaluation practice is adequate.

If the ratio is large, the safety case for every model deployed on independent-capability benchmarks is wrong by d^n —and the safety case was the reason the model was deployed.

The matrix exists. The code exists. The computation has not been done.

Epilogue

The law was hiding for ninety-one years.

Not because the mathematics was hard—it fits on an index card. Not because the computation was expensive—it runs on a laptop. Not because the data was unavailable—weather records, financial returns, and hospital censuses are among the most thoroughly collected datasets in existence.

It was hiding because of two choices made in the middle of the twentieth century.

In 1935, Ronald Fisher introduced the permutation test and used it to compute a p-value. He asked: “Is the effect significant?” He did not ask: “How large is the effect?” For ninety-one years, the entire field reported p-values. The ratio—the number that tells you how wrong the model is—sat in the intermediate calculations, computed and discarded, in every permutation test ever run.

In 1952, Harry Markowitz chose variance as the measure of portfolio risk because it was tractable. For seventy-four years, quantitative risk management was built on variance. Variance has nonlinearity $n = 1$. At $n = 1$, $d^n = d$, which is 2 to 5—boring. Three generations of risk managers were trained in the one corner of cost-function space where the law produces uninteresting results.

Fisher asked the wrong question. Markowitz chose the wrong cost function. Both choices were reasonable for their purposes. But their combined effect was to make the law invisible for nearly a century. The question that reveals the law—“How large is the ratio?”—applied

to the cost function that reveals it—a nonlinear streak cost—was a combination that no field’s standard practice would produce.

Some of the pieces are visible today. Tail dependence researchers have proposed better copulas; compound event modelers have built multivariate frameworks for specific pairs of hazards; extreme value theorists have developed sophisticated marginal models; catastrophe modelers patch the gap with scenario overlays; actuaries add a “correlation loading” that guesses the size of d^n without knowing the equation; and the surrogate data community reports p -values from phase randomization—a version of the lag shift—without ever reporting the ratio. Each field has one piece. None has all four: permutation test, streak cost, exponential form, multiplicative decomposition. The law sits at an intersection that no field’s standard practice occupies.

What does correct pricing look like? Return to Texas. The IID model assesses the annual probability of a statewide, week-long compound failure at roughly 10^{-7} —once in ten million years. Given the \$195 billion cost, the expected annual loss is about \$20,000. That is the number that enters the cost-benefit analysis. It says: don’t invest.

The permutation test says the IID model is wrong by a factor of roughly 10^5 . Multiplying, the correct probability is $10^{-7} \times 10^5 = 10^{-2}$ —once in fifty to a hundred years. The expected annual loss is \$2 to \$4 billion. At that number, winterizing the gas pipelines (\$5–15 billion, one-time) pays for itself within two to five years. Sizing emergency reserves for a week-long compound event rather than a day-long single-domain event changes the reserve requirement by the same factor of d^n . An insurer who sets premiums at the correct probability collects enough to survive the event. An insurer who sets premiums at the IID probability does not.

The arithmetic is the same in every domain. The IID expected annual loss is the real expected annual loss divided by R . When

$R = 50,000$, the IID estimate is wrong by the cost of a house. When $R = 10$, it is wrong by the cost of a renovation. The difference between these two cases is one number: n , the nonlinearity of the cost function.

The law is now visible. It says:

*The pricing error from ignoring dependence grows exponentially
with the nonlinearity of the cost function.*

It predicts the error for any system, any cost function, any domain. It decomposes the error into factors that map to interventions. It prescribes the model architecture that captures what independence misses. It runs in forty lines.

The question is no longer whether the catastrophe models are wrong. They are wrong by d^n . The question is what to do about it.

Here is a prediction. It is falsifiable, and it follows directly from the law.

The next time a major earthquake strikes a densely built metropolitan area, post-event analyses will show that models of “lifeline” infrastructure—simultaneous failure of water, power, gas, transport, and communications for a sustained period—underestimated the compound probability by a factor between $1,000\times$ and $100,000\times$. (The seismic trigger itself may show a modest ratio— $16\times$ for California aftershock clustering, as measured in Chapter 10—but the earthquake is the common driver that synchronizes the infrastructure cascade, just as the polar vortex synchronized the Texas grid. The compound infrastructure ratio will be far larger than the seismic ratio alone.) The predicted range depends on the number of coupled lifeline systems (n) and the effective dependence strength ($d \approx 3\text{--}6$) of the region’s infrastructure. The models will have been tested with single-system, single-day diagnostics and found adequate. The compound, sustained failure—every system down, for days—will

have been priced as a once-in-ten-thousand-years event. It will turn out to have been a once-in-fifty-years event.

The same prediction applies to the next multi-day, multi-region heat wave that overwhelms power grids across an interconnected area. To the next synchronized financial deleveraging in which every asset class falls simultaneously for weeks. To the next pandemic wave that saturates every hospital in a nation within the same two-week window. In each case, the models will have the right marginals and the wrong joint tail, and the cost function that matters will be high n .

These predictions do not require prophecy. They require only that the models continue to assume independence and the world continue to be dependent. Both conditions currently hold.

The predictions are falsifiable. If, after the next compound infrastructure failure involving four or more coupled lifeline systems sustained for five or more days, the post-event analysis finds that the IID model underestimated the compound probability by less than $100\times$, the law as stated here is wrong. If the ratio falls between $1,000\times$ and $100,000\times$, the law is confirmed in a new domain.

The code is in Chapter 9. The data is in the filing cabinets and server rooms of every organization that prices risk.

Someone will compute the ratio. The question is whether they compute it before the event or after.

* * *

Notes on Sources

The Draupner wave. The rogue wave recorded on January 1, 1995, at the Draupner E platform is documented in Haver (2004). The wave height of 25.6 meters against a significant wave height of approximately 12 meters made it one of the first instrumentally confirmed rogue waves. The physics of rogue wave formation—including constructive interference of independent wave trains and nonlinear focusing—is reviewed in Kharif and Pelinovsky (2003), “Physical mechanisms of the rogue wave phenomenon,” *European Journal of Mechanics B/Fluids*.

The method. The permutation test was introduced in Fisher (1935), *The Design of Experiments*. Pitman (1937) and others developed the theory. Modern treatments include Good (2005), *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. The use of permutation tests for dependence assessment is reviewed in Theiler et al. (1992), who introduced Fourier phase randomization as a surrogate data method for nonlinear dynamics.

The Texas crisis. The February 2021 Texas power crisis is documented in the FERC/NERC joint inquiry report (November 2021) and the ERCOT Interim Report (February 2021). The death toll of 246 is from BuzzFeed News reporting (December 2021), later corroborated by academic excess-mortality studies. Economic damage estimates of \$195 billion are from Accuweather and Perryman Group analyses. The 2011 precedent and ignored winterization recommendations are documented in the FERC/NERC 2011 report.

The 2008 financial crisis. The Gaussian copula for CDO pricing was introduced in Li (2000), “On Default Correlation: A Copula Function Approach,” *Journal of Fixed Income*. The crisis narrative draws on the Financial Crisis Inquiry Commission Report (2011), Tett (2009) *Fool’s Gold*, and Lewis (2010) *The Big Short*. Economic cost estimates exceeding \$10 trillion are from Luttrell, Atkinson, and Rosenblum (2013), Dallas Fed Staff Papers. Recession-related excess mortality is documented in Stuckler et al. (2009), *The Lancet*, who found significant excess deaths from suicide, cardiovascular events, and substance abuse linked to unemployment and austerity across Europe.

Hurricane Harvey. Impact data from Blake and Zelinsky, National Hurricane Center Tropical Cyclone Report (2018), which gives 68 direct deaths and 35 indirect deaths in Texas. The \$125 billion damage estimate is from NOAA.

The variance framework. Markowitz (1952), “Portfolio Selection,” *Journal of Finance*. Sharpe (1964), “Capital Asset Prices,” *Journal of Finance*. Black and Scholes (1973), “The Pricing of Options and Corporate Liabilities,” *Journal of Political Economy*.

Compound events. Zscheischler et al. (2018), “Future climate risk from compound events,” *Nature Climate Change*. Zscheischler et al. (2020), “A typology of compound weather and climate events,” *Nature Reviews Earth & Environment*.

Genomics. The missing heritability problem was posed in Manolio et al. (2009), “Finding the missing heritability of complex diseases,” *Nature*. Epistasis is reviewed in Mackay (2014), *Nature Reviews Genetics*. The framework’s application to genomic data is a prediction of the law; it has not yet been tested empirically.

Ecology. Tilman et al. (2014), “Biodiversity and ecosystem functioning,” *Annual Review of Ecology, Evolution, and Systematics*. Overyielding in biodiversity experiments is reviewed in Loreau and Hector (2001), *Nature*.

The Dutch Delta Programme. The treatment of compound flood

risk (simultaneous coastal surge, river discharge, and rainfall) in Dutch flood defense planning draws on Kew et al. (2013), “The simultaneous occurrence of surge and discharge extremes for the Rhine delta,” *Natural Hazards and Earth System Sciences*, and van den Hurk et al. (2015), “Analysis of a compounding surge and precipitation event in the Netherlands,” *Environmental Research Letters*. The Delta Programme’s investment framework is documented in the Delta Programme Commissioner’s annual reports. Zscheischler et al. (2020) review compound event modeling in climate risk assessment more broadly.

The tail intensification. The cold snap ratios ($28\times$ at $\geq 5/8$, $1,875\times$ at $\geq 7/8$, stratified within winter months) and wind lull ratios ($39\times$ at $\geq 5/6$, $250\times$ at $6/6$) are from the independence pricing analysis of Open-Meteo ERA5 reanalysis data, 1990–2019, with $B = 1,000$ permutations and within-month stratification. The heat wave ratios ($3.2\times$ at $\geq 5/8$, $7.3\times$ at $\geq 7/8$) use the same source and method.

The contagion spike. The equity synchronization ratio ($113\times$ at $K = 5$) and persistence ratio ($3.8\times$) use Fama–French 10 daily industry portfolios, 1990–2019. The Potomac synchronization ratio ($1.3\times$ at $K = 7$) and persistence ratio ($679\times$) use USGS daily discharge at seven gauges, 1990–2019. Both use $B = 1,000$ permutations. The lag-shift method follows Theiler et al. (1992).

Spatial synchrony and the Moran effect. The synchronization measured by the lag shift has a long prior history in ecology. Moran (1953) proved that spatially separated populations synchronize when driven by correlated environmental noise—the “Moran effect.” Liebhold, Koenig, and Bjørnstad (2004) review the subsequent fifty years of theory and evidence. The ecological literature measures synchrony by cross-correlation of detrended population time series; the lag-shift ratio in Chapter 3 measures the same phenomenon through its effect on a cost function. The frameworks are complementary: the ecological literature identifies mechanisms (correlated climate, dispersal, trophic interaction), while the permutation-test framework quantifies

the pricing consequence.

The Gaussian copula test. The copula baseline at 10 and 49 Fama–French industry dimensions uses $B = 300$ column shuffles. The Gaussian copula is fitted to normal-score-transformed ranks; the t copula uses maximum-likelihood degrees of freedom. Residual z-scores use block-bootstrap standard errors.

The block-row-shuffle decomposition. The block-shuffle illustration in Chapter 8 uses the standard Texas cold parameters ($K = 7$, $M = 4$, threshold = 20°F), with the $S = 1$ row-shuffle ratio matching Chapter 3. Potomac block-shuffle results ($392\times$ at $S = 1$, $1.1\times$ at $S = 7$) are from the persistence pricing analysis, $B = 1,000$.

Negative controls. Hard drive failures: Backblaze public dataset. Insurance claims: OpenML freMTPL2. Unemployment: BLS LAUS API. Global seismicity: USGS ComCat (California, Chile, Mediterranean plates, daily $M \geq 4$ counts, 1990–2019).

Chapter 9 output. The table in Chapter 9 shows the pricing ratios for a seven-station Texas cold matrix at $M = 4$, threshold = 20°F .

The 2008 financial crisis and duration. The September–October 2008 market decline—nine of eleven trading days down over two weeks—is documented in the S&P 500 daily close data, publicly available. The margin-call cascade mechanism is analyzed in Brunnermeier (2009), “Deciphering the Liquidity and Credit Crunch 2007–2008,” *Journal of Economic Perspectives*.

Epidemiology (COVID-19). Daily confirmed cases from the COVID Tracking Project (covidtracking.com), ten U.S. states, January 2020 to March 2021. $B = 1,000$ permutations.

Seismology (earthquake aftershock clustering). Daily maximum magnitudes from the USGS ComCat catalog, eight subregions of seismically active California, 2000–2019. Negative control: separate plates (California, Chile, Mediterranean). Aftershock clustering follows Omori (1894) as formalized in Utsu (1961).

Agriculture (drought). Daily soil moisture from the North American Land Data Assimilation System (NLDAS-2), eight grid cells

corresponding to agricultural sites across Iowa, Nebraska, Kansas, and the Dakotas, growing seasons (May–September) 2000–2019. $B = 1,000$ permutations.

The datasets. The principal datasets featured in the book, grouped by category: *Weather*—Texas cold (7 stations), California heat (8), Midwest cold (6), heavy rain (6), snowfall (5). *Hydrology*—Houston flood (USGS discharge, 5 gauges), Potomac drought (7 gauges). *Finance*—equity sector drawdowns (10 Fama–French industries). *Energy*—wind lulls (6 sites), solar radiation (4), power demand (3 ERCOT zones), compound weather-energy (stacked). *Epidemiology*—influenza (ILINet, 9 HHS regions), COVID-19 (10 states). *Seismology*—California earthquake clustering (8 subregions). *Agriculture*—Great Plains soil moisture drought (8 sites). Additional domains tested but not featured in the narrative include wildfire, insurance, civil infrastructure, and planetary science. Four negative controls confirm the test does not manufacture dependence: hard drive failures, auto insurance claims, unemployment, and cross-plate seismicity.

The exponential form. The information-theoretic argument for the linearity of $\log R$ in K proceeds as follows. The pricing ratio R for a streak of length K equals the ratio of the real joint probability to the IID joint probability. Taking logarithms, $\log R$ is the Kullback–Leibler divergence between the real distribution and the IID null, evaluated on the streak event. The chain rule of KL divergence decomposes this into a sum of K conditional terms, each bounded between c_1 and c_2 (the minimum and maximum per-step conditional divergence across states). The sum of K terms, each of bounded size, grows linearly in K . This is why $\log R$ is linear in n : each AND in the cost function adds one term to the sum. The linear pattern on the log-scale plot is a consequence of the chain rule.

The structure of d . The composition (Equation 6.1), the acceleration, the divergence exponent, the decorrelation floor, and the dominance predictor are established empirically across eleven to

eighteen domains. The composition coefficients are estimated by ordinary least squares on the domain-level growth rates; the acceleration is tested by a sign test on sub-panel slope differences ($p < 10^{-6}$, $N = 334$); the divergence exponent ($\gamma > 3$, all domains) is estimated by log-log regression of R on threshold quantile across $q \in \{0.10, 0.15, 0.20, 0.25, 0.30\}$; the floor is measured by block permutation (shuffling blocks of L consecutive days and recomputing the cost): the decorrelation length is the block size at which the ratio of real to block-shuffled cost drops below 2. The dominance predictor uses the Spearman rank correlation between the cross-minus-auto correlation difference and $\ln(S/T)$ across eighteen domains ($\rho = 0.89$, $p < 0.001$). The pricing ratio is rank-invariant: replacing the data with its ranks produces the same ratios. The scaling rate c is stable across the first and second halves of each record.

Cybersecurity (supply-chain compromise). The SolarWinds incident is documented in CISA Emergency Directive 21-01 (December 2020) and the SEC complaint against SolarWinds Corporation (October 2023). The 18,000-organization figure is from the SolarWinds SEC filing (Form 8-K, December 2020). The NotPetya damage estimate of \$10 billion is from Greenberg (2018), *Wired*, corroborated by Maersk, Merck, and FedEx earnings disclosures.

Supply chains. The 2021 semiconductor shortage is documented in the White House 100-day supply chain review (June 2021). The Suez Canal blockage (March 2021) cost estimates are from Lloyd's List. The compound supply-chain disruption framework is a prediction of the law; it has not yet been tested empirically with the permutation methodology.

Chapter 11: AI safety evaluation. The chapter applies the book's framework to AI capability evaluation as a prediction, not a measurement. No permutation test has been run on AI evaluation data. The predicted ratios use dependence strengths ($d \approx 4-6$) measured in weather and finance as reference points. The "sum vs. product" framing captures the distinction between additive (independent)

and multiplicative (dependent) risk. The deployment monoculture argument parallels the CDO correlation critique in Salmon (2009) and MacKenzie (2006), applied to shared foundation models rather than shared mortgage pools.

References

- Ackoff, R.L. (1971). "Towards a System of Systems Concepts." *Management Science*, 17(11), 661–671.
- Black, F. and M. Scholes (1973). "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy*, 81(3), 637–654.
- Blake, E.S. and D.A. Zelinsky (2018). National Hurricane Center Tropical Cyclone Report: Hurricane Harvey. NOAA/NHC.
- Brunnermeier, M.K. (2009). "Deciphering the Liquidity and Credit Crunch 2007–2008." *Journal of Economic Perspectives*, 23(1), 77–100.
- The COVID Tracking Project (2021). covidtracking.com. Data collected by *The Atlantic*, January 2020 to March 2021.
- Fama, E.F. and K.R. French (1997). "Industry costs of equity." *Journal of Financial Economics*, 43(2), 153–193. Data library: mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.
- FERC/NERC (2011). Report on Outages and Curtailments During the Southwest Cold Weather Event of February 1–5, 2011.
- FERC/NERC (2021). The February 2021 Cold Weather Outages in Texas and the South Central United States. Joint Inquiry Report, November 2021.
- Financial Crisis Inquiry Commission (2011). The Financial Crisis Inquiry Report. U.S. Government Printing Office.
- Fisher, R.A. (1935). *The Design of Experiments*. Oliver and Boyd.
- Good, P.I. (2005). *Permutation, Parametric, and Bootstrap Tests of Hy-*

- potheses*. 3rd ed., Springer.
- Greenberg, A. (2018). "The Untold Story of NotPetya, the Most Devastating Cyberattack in History." *Wired*, August 22, 2018.
- Haver, S. (2004). "A possible freak wave event measured at the Draupner jacket January 1 1995." In *Rogue Waves 2004*, Brest, France.
- Kew, S.F. et al. (2013). "The simultaneous occurrence of surge and discharge extremes for the Rhine delta." *Natural Hazards and Earth System Sciences*, 13, 2017–2029.
- Kharif, C. and E. Pelinovsky (2003). "Physical mechanisms of the rogue wave phenomenon." *European Journal of Mechanics B/Fluids*, 22(6), 603–634.
- Lewis, M. (2010). *The Big Short*. W.W. Norton.
- Li, D.X. (2000). "On Default Correlation: A Copula Function Approach." *Journal of Fixed Income*, 9(4), 43–54.
- Liebhold, A., W.D. Koenig, and O.N. Bjørnstad (2004). "Spatial Synchrony in Population Dynamics." *Annual Review of Ecology, Evolution, and Systematics*, 35, 467–490.
- Loreau, M. and A. Hector (2001). "Partitioning selection and complementarity in biodiversity experiments." *Nature*, 412, 72–76.
- Luttrell, D., T. Atkinson, and H. Rosenblum (2013). "Assessing the Costs and Consequences of the 2007–09 Financial Crisis and Its Aftermath." Dallas Fed Staff Papers, No. 20.
- Mackay, T.F.C. (2014). "Epistasis and quantitative traits: using model organisms to study gene–gene interactions." *Nature Reviews Genetics*, 15, 22–33.
- MacKenzie, D. (2006). *An Engine, Not a Camera: How Financial Models Shape Markets*. MIT Press.
- Manolio, T.A. et al. (2009). "Finding the missing heritability of complex diseases." *Nature*, 461, 747–753.

REFERENCES

- Markowitz, H. (1952). "Portfolio Selection." *Journal of Finance*, 7(1), 77–91.
- Moran, P.A.P. (1953). "The statistical analysis of the Canadian lynx cycle. II. Synchronization and meteorology." *Australian Journal of Zoology*, 1(3), 291–298.
- Mitchell, K.E. et al. (2004). "The multi-institution North American Land Data Assimilation System (NLDAS)." *Journal of Geophysical Research*, 109, D07S90.
- Pitman, E.J.G. (1937). "Significance tests which may be applied to samples from any populations." *Supplement to the Journal of the Royal Statistical Society*, 4(1), 119–130.
- Salmon, F. (2009). "Recipe for Disaster: The Formula That Killed Wall Street." *Wired*, February 23, 2009.
- Sharpe, W.F. (1964). "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance*, 19(3), 425–442.
- SolarWinds Corporation (2020). Form 8-K, December 14, 2020. U.S. Securities and Exchange Commission.
- Stuckler, D. et al. (2009). "The public health effect of economic crises and alternative policy responses in Europe." *The Lancet*, 374, 315–323.
- Tett, G. (2009). *Fool's Gold*. Free Press.
- Theiler, J. et al. (1992). "Testing for nonlinearity in time series: the method of surrogate data." *Physica D*, 58, 77–94.
- Tilman, D. et al. (2014). "Biodiversity and ecosystem functioning." *Annual Review of Ecology, Evolution, and Systematics*, 45, 471–493.
- Utsu, T. (1961). "A statistical study on the occurrence of aftershocks." *Geophysical Magazine*, 30, 521–605. Formalization of Omori's (1894) aftershock decay law.
- van den Hurk, B. et al. (2015). "Analysis of a compounding surge and

- precipitation event in the Netherlands." *Environmental Research Letters*, 10, 035001.
- White House (2021). Building Resilient Supply Chains, Revitalizing American Manufacturing, and Fostering Broad-Based Growth: 100-Day Reviews under Executive Order 14017. June 2021.
- Zscheischler, J. et al. (2018). "Future climate risk from compound events." *Nature Climate Change*, 8, 469–477.
- Zscheischler, J. et al. (2020). "A typology of compound weather and climate events." *Nature Reviews Earth & Environment*, 1, 333–347.

Glossary

Column shuffle Randomly permute each column of the data matrix independently. Destroys all dependence; produces the IID null. (Chapter 1)

Compound event Multiple domains failing simultaneously, each worsening the others. (Chapter 4)

Cost function The rule that turns data into a number measuring harm. Gentle cost functions (averages) are forgiving; severe cost functions (sustained multi-station streaks) amplify dependence. (Chapter 1)

Coupling surplus The additional pricing error that arises when domains interact under stress—the compound ratio exceeds the product of individual ratios. (Chapter 4)

Decorrelation floor The timescale (typically days to weeks) above which the IID assumption becomes approximately safe. (Chapter 6)

Dependence strength (d) How much the parts of a system move together. A number between 2 and 10 that can be estimated from data. (Chapter 5)

Diamond decomposition The algebraic identity $R_{\text{total}} = R_{\text{persistence}} \times R_{\text{spatial}}$, showing that the total pricing error decomposes into temporal and spatial factors. (Chapter 3)

- IID** Independent and identically distributed. The assumption that each component behaves independently and each time period is an independent draw. (Prologue)
- Independence ratio (R_{total})** The factor by which the IID model misprices the real cost. Measured by the column shuffle. (Chapter 1)
- Lag shift** Circularly shift each column by a random offset. Preserves each column's time series exactly; destroys cross-column timing. (Chapter 3)
- Nonlinearity (n)** How many things must go wrong simultaneously, or for how long, for the cost to trigger. The exponent in d^n . (Chapters 1, 5)
- Permutation test** A method that compares real data to a shuffled version where one structural feature has been destroyed, measuring how much that feature matters. Introduced by Fisher in 1935. (Chapter 1)
- Persistence** The tendency for today's conditions to predict tomorrow's. Cold follows cold; warm follows warm. (Chapter 2)
- Persistence ratio (R_{pers})** The factor by which the real cost exceeds the cost after temporal order is destroyed. Measured by the row shuffle. (Chapter 2)
- Pricing ratio (R)** Real cost divided by average cost under the null hypothesis. When $R > 1$, the null model underpredicts. (Chapter 1)
- Row shuffle** Randomly permute the rows of the data matrix. Preserves each day's spatial pattern; destroys temporal order. (Chapter 2)

Sign boundary The cost-function threshold (near $n \approx 1.3$) where the IID error switches from overprediction (switching costs) to underprediction (streak costs). (Chapter 5)

Streak cost A cost that accrues when a bad state persists— M stations above threshold for K consecutive days. The costs that break infrastructure. (Chapter 1)

Synchronization The tendency for all components to be in the same state at the same time. All stations cold at once; all sectors falling together. (Chapter 3)

Synchronization ratio (R_{sync}) The factor by which the real cost exceeds the cost after cross-column timing is destroyed. Measured by the lag shift. Not identical to the spatial ratio from the diamond. (Chapter 3)

Tail intensification The phenomenon where dependence is stronger in the extreme than in the body of the distribution. d grows as the question becomes more severe. (Chapter 5)

Index

- acceleration, 54
- AI safety
 - barriers to discovery, 97
 - compound frequency, 97
- AI safety evaluation, 95
- asymmetry, 20
- block bootstrap, 67
- column shuffle, 15
- common driver, 96
- composition, 51
- compound event, 32
- conformal prediction, 69
- contagion, 48
- coupling surplus, 35
 - AI, 96
- COVID-19, 83
- cybersecurity, 85
- d^n (the law), 41
- decorrelation floor, 55
- Delta Programme, 71
- dependence strength, 39
- deployment monoculture, 100
- diamond decomposition, 28
- diamond identity
 - AI, 99
 - divergence exponent, 54
 - dollar decomposition, 64
 - dominance predictor, 53
 - drought, 84
 - earthquake clustering, 84
 - epistasis, 88
 - exponential fit, 40
- Fisher, Ronald, 15
- Gaussian copula, 62
 - deployment monoculture, 100
 - failure at high dimension, 49
- IID, 11
- independence assumption, 11
- independence ratio
 - AI, 98
- lag shift, 26
- lag-1 persistence, 23
- limitations, 93
- Markowitz, Harry, 103
- methodology validation, 78
- missing heritability, 87

INDEX

Nash equilibrium, 44
 AI safety, 98
negative controls, 91
nonlinearity, 15, 40

ordinal sufficiency, 79

p-value shadow, 43
permutation test, 15
 AI evaluation, 98
persistence, 20
persistence ratio, 21
pricing ratio, 16

regime asymmetry, 91
row shuffle, 21
ruin probability, 63

scorecard, 82
sign boundary, 45
 AI risks, 101
stacked matrix, 33
streak cost, 14
 AI capabilities, 95
structural break, 92
super-additivity, 35
supply chain, 86
synchronization, 26
synchronization ratio, 26

tail intensification, 47
 AI capabilities, 96

variance trap, 43