

Calibration Before Complexity: Forecasting Compound-Hazard Exposure for U.S. Freight Systems from Public Multi-Source Data

SUNDAY MICHAEL OYEBIYI

Information Technology, Baltimore, MD, United States

ORCID: [0009-0000-1747-4710](https://orcid.org/0009-0000-1747-4710)

Abstract- Freight systems face hazards that seldom arrive alone, yet predictive studies typically model weather, disaster, trade and financial shocks in isolation. We assemble a leakage-controlled state-day panel for the United States that aligns disaster signals (NOAA Storm Events and FEMA declarations), freight exposure (Commodity Flow Survey and Freight Analysis Framework version 5) and macro-geopolitical context (imports, the CBOE Volatility Index and the Geopolitical Risk index), and forecast whether a composite hazard shock index, a proxy for conditions under which freight disruption becomes likely, will exceed its training-period 90th percentile on the following day. Four learners were trained to December 2022 and evaluated once from January 2024 (22,617 and 11,296 state-days), with 2023 withheld. Discrimination was moderate and nearly indistinguishable across three models (ROC-AUC 0.725–0.733 for gradient boosting, logistic regression and histogram-based gradient boosting; 0.706 for random forest), but probabilistic accuracy diverged sharply. Only gradient boosting outperformed a constant base-rate forecast (Brier skill score ≈ 0.09); logistic regression and histogram-based boosting over-predicted risk at every level (skill ≈ -0.8 to -0.9). Event prevalence rose from 10% to 17.2% between windows. Ablations showed that hazard persistence carries most of the signal, freight exposure adds little, and a wider macroeconomic block reduces out-of-sample discrimination. At a seven-day horizon, logistic regression was the most stable learner. A newsvendor illustration shows how uncalibrated probabilities near a decision threshold translate into systematic over-stocking. The study provides a transparent public-data benchmark and shows that, for logistics risk screening, calibration and temporal validation matter more than model complexity.

Keywords: Compound hazards; Freight resilience; Probability calibration; Early warning; Machine learning; Distribution shift; Decision-focused evaluation

I. INTRODUCTION

Freight movement depends on physical, institutional and financial conditions holding simultaneously. Carriers need passable roads and functioning ports; shippers need predictable demand and credit; both need a trading environment that does not shift abruptly. A single failure is usually absorbed. When several coincide, disruption can propagate across regions and modes in ways no single indicator anticipates. The COVID-19 pandemic made this plain: factory closures, labour shortages, port congestion and demand swings combined into prolonged supply chain stress (Ivanov, 2020; Ivanov & Dolgui, 2020), and composite pressure indices built afterwards showed how many separate measures had moved together (Benigno et al., 2022). Extreme weather adds a growing physical dimension to this problem. Most of the world's ports are exposed to more than one natural hazard (Verschuur et al., 2023), and weather-related disruption of transport systems is now documented at both local and system-wide scales (Gössling et al., 2023).

Climate science describes hazards whose joint impact exceeds the sum of their parts as *compound events*: they coincide, follow in quick succession, or strike systems already weakened by an earlier shock (Zscheischler et al., 2018). The concept transfers naturally to logistics. A severe convective storm in a state already under a federal disaster declaration, during a month of elevated financial volatility and geopolitical tension, creates a different operating environment from the same storm in calm conditions. Freight planners, emergency managers and state transportation agencies would benefit from an early, even coarse, signal that such conditions are building.

Public data in the United States make such a signal feasible in principle. The National Oceanic and Atmospheric Administration (NOAA) publishes event-level severe-weather records (National Centers for Environmental Information, n.d.); the Federal Emergency Management Agency (FEMA) releases disaster declarations (Federal Emergency Management Agency, n.d.); the Commodity Flow Survey (CFS) and the Freight Analysis Framework (FAF) describe the volume and value of goods moving within and between states (U.S. Bureau of Transportation Statistics, n.d.; U.S. Census Bureau, n.d.); and macro-financial and news-based risk indices are openly available (Baker et al., 2016; Caldara & Iacoviello, 2022; Federal Reserve Bank of St. Louis, n.d.). Each source is widely used on its own. Few studies align them on a common spatial and temporal grain and ask whether their combination anticipates periods of stress.

Three methodological obstacles explain much of this gap. The first is *alignment*: storm reports are daily, declarations follow administrative timelines, freight estimates are annual, and macroeconomic series are monthly or quarterly. Combining them requires explicit rules that carry slow-moving values forward without leaking future information into the past. The second is *evaluation*: random train–test splits, still common in applied prediction, allow models to learn from observations dated after those they predict and so overstate performance on time-ordered data (Bergmeir & Benítez, 2012; Kaufman et al., 2012). The third is *decision relevance*: a model that ranks days well is not automatically useful, because operational choices such as how much inventory to hold depend on the probability level, not merely its rank (Elmachtoub & Grigas, 2022; Van Calster et al., 2019).

This study addresses these obstacles through four research questions:

- RQ1. How well can openly available multi-source data anticipate next-day compound-hazard stress at the state level under strict chronological validation?
- RQ2. Do standard learners that are indistinguishable in discrimination also agree in calibration and probabilistic accuracy?

- RQ3. What incremental signal do freight-exposure and macro-geopolitical information add beyond hazard persistence, and is it stable at a longer horizon?
- RQ4. How do calibration errors propagate into operational inventory decisions?

We construct a state–day panel from seven public sources, define a forward-looking stress label from a composite shock index, and compare logistic regression, random forest, gradient boosting and histogram-based gradient boosting (HGB) under identical preprocessing. Training uses data to 31 December 2022; the whole of 2023 is withheld; evaluation begins on 1 January 2024. We deliberately use standard learners: for moderately sized tabular data with heterogeneous features, tree ensembles remain strong and are easier to audit than neural networks (Grinsztajn et al., 2022), and logistic regression provides a transparent reference. If a simple, reproducible baseline already captures most of the achievable signal, that is worth knowing before investing in complex architectures.

The study makes four contributions. *First*, it specifies a reproducible, leakage-controlled compound-hazard stress label for U.S. states and documents every construction step, including temporal alignment rules for exogenous series. *Second*, it provides a public-data benchmark of standard learners for this task on a buffered chronological holdout, with bootstrap intervals and paired tests, reported in line with TRIPOD+AI principles (Collins et al., 2024). *Third*, it shows empirically that models with near-identical discrimination can differ by an order of magnitude in probabilistic skill, and traces the cause to the training objective rather than to distribution shift. *Fourth*, it uses block ablations, a second horizon and a stylised newsvendor model to show that adding conceptually relevant variables can harm out-of-sample performance and that miscalibration, rather than ranking ability, determines decision cost. Ablations are interpreted as associative stress tests of fitted models, not causal estimates.

The remainder of the paper is organised as follows. Section 2 reviews related work and states the research gap. Section 3 describes the data, panel construction, outcome, models and evaluation. Section 4 reports the

results. Section 5 discusses theoretical and practical implications and limitations, and Section 6 concludes.

II. RELATED WORK

2.1. Compound hazards and freight-system disruption

Hazard research has long held that losses depend on the interaction between a physical event and the systems it strikes. Social vulnerability research showed that identical hazards produce very different losses depending on the characteristics of the places affected (Cutter et al., 2003), and the same logic applies to infrastructure: a road network or port operating near capacity has less slack to absorb disruption, and a region coping with one shock is less able to cope with a second. The compound-event framework formalises this reasoning by distinguishing simultaneous, successive and preconditioned hazards, and warns that single-driver risk assessments systematically underestimate losses (Zscheischler et al., 2018). In freight systems, physical hazards interact further with market conditions, labour availability and policy. Asset-level analysis shows that roughly 86% of the world's ports are exposed to multiple natural hazards and that tens of billions of dollars of maritime trade are at risk annually (Verschuur et al., 2023), while a broad review of weather and transport documents disruptive effects across all modes at micro and macro scales (Gössling et al., 2023).

Empirical work on transport disruption also shows that systems adapt. Rose and Wei (2013) estimated that rerouting, inventory drawdown and input substitution could substantially reduce the losses from a port shutdown, and Rose (2004) distinguished inherent from adaptive resilience. At the same time, firm-level evidence from the 2011 Great East Japan Earthquake shows that disruptions propagate along supply chains to firms far from the directly affected area (Carvalho et al., 2021), which is why regional hazard exposure is relevant well beyond the places where hazards strike. For forecasting, the implication is that a stress indicator measures *exposure*, not realised loss; whether exposure becomes loss depends on how firms and agencies respond. Hazard data also carry their own measurement problems. Loss databases, including the National Weather Service Storm Events records, suffer from inconsistent reporting thresholds, under-reporting of small events and changes in collection

practice over time (Gall et al., 2009). A label built from such records inherits their error, so model performance should be read as performance against a proxy.

2.2. Supply chain risk, resilience and the estimation problem

Supply chain management developed its own vocabulary for disruption: risk typologies matched to mitigation strategies (Chopra & Sodhi, 2004), resilience through redundancy and flexibility (Sheffi & Rice, 2005), structural drivers of disruption severity such as network density and node criticality (Craighead et al., 2007), and the balance of vulnerabilities and capabilities (Pettit et al., 2010). Reviews note persistent inconsistency in how resilience is defined and measured (Tukamuhabwa et al., 2015), a caution that applies to any attempt to summarise resilience in a single index. The pandemic prompted a systems view in which disruptions interact across sectors and regions and resilience is understood as the viability of intertwined networks (Ivanov, 2020; Ivanov & Dolgui, 2020).

Operations research contributes a large body of models for planning under disruption (Hosseini et al., 2019; Snyder et al., 2016; Tang, 2006). A recurring feature of these models is that disruption probabilities are treated as known parameters. In practice they must be estimated, and the quality of the estimate governs the quality of the decision. This is the link that motivates the present study: even a modestly accurate predictive model is valuable if its probabilities are trustworthy, and harmful if they are not.

2.3. Macro-financial and geopolitical signals

Freight activity responds to economic uncertainty. The CBOE Volatility Index (VIX) captures expected equity-market volatility (Whaley, 2000), and uncertainty shocks cause firms to pause investment and hiring (Bloom, 2009); inventory and transport commitments are forms of investment and are likely to respond similarly. News-based indices extend this approach to economic policy uncertainty (EPU; Baker et al., 2016) and geopolitical risk (GPR; Caldara & Iacoviello, 2022), both associated with lower investment. Trade volumes such as U.S. imports of goods and services provide a more direct link to port and inland freight, and the Global Supply Chain

Pressure Index demonstrated that no single indicator captures the full extent of disruption (Benigno et al., 2022). These series are, however, national and monthly or quarterly. Carried forward onto a state–day panel, they vary over time but not across states, so they can help a model distinguish calm from turbulent periods while also exposing it to regime change. Our ablations examine this trade-off directly.

2.4. Machine learning for transport and supply chain risk

Machine learning has a long history in traffic forecasting, where reviews note that performance under normal conditions often degrades during disruptions, precisely the periods of greatest operational interest (Ermagun & Levinson, 2018; Vlahogianni et al., 2014). In supply chain risk management, most applications concern supplier selection, demand forecasting or risk identification within a single firm (Baryannis et al., 2019). Subsequent studies predict supplier delivery delays from purchase-order data (Brintrup et al., 2020), pool supplier-level risk information across firms through federated learning (Zheng et al., 2023), and use causal machine learning to support intervention planning (Wyrembek et al., 2025). All of these rely on proprietary transactional records and target firm-level outcomes; none uses public, system-level hazard and exposure signals, and calibration of predicted probabilities is rarely examined. Interpretability remains a recurring concern, and there is a strong argument that interpretable models should be preferred for high-stakes decisions when they perform comparably (Rudin, 2019).

Tree ensembles occupy a middle ground between transparency and flexibility. Random forests average decorrelated trees (Breiman, 2001); gradient boosting fits trees sequentially to the residual gradient (Friedman, 2001); histogram-based implementations bin features to accelerate split search and handle missing values natively (Chen & Guestrin, 2016; Ke et al., 2017). On heterogeneous tabular data, these ensembles generally outperform deep learning (Grinsztajn et al., 2022). Feature-attribution methods such as SHAP (Lundberg & Lee, 2017) explain individual predictions; block ablations, used here, answer the different question of whether a group of

variables improves out-of-sample performance at all. Both are associative.

2.5. Evaluation of time-ordered prediction

Out-of-sample evaluation should mimic deployment, with the test period following the training period (Hyndman & Athanasopoulos, 2021; Tashman, 2000); random cross-validation can be optimistic under temporal dependence (Bergmeir & Benítez, 2012). Leakage, the use of information unavailable at prediction time, can arise through back-filling, imputation on the full dataset, or preprocessing statistics computed with test observations (Kaufman et al., 2012). A chronological split also exposes a model to shifts in covariates, in outcome prevalence and in the relation between the two (Moreno-Torres et al., 2012).

Metrics must match the intended use. ROC-AUC measures ranking and is the probability that a random positive case outranks a random negative case (Fawcett, 2006; Hanley & McNeil, 1982), but it is insensitive to calibration and can flatter models on imbalanced data (Saito & Rehmsmeier, 2015). The Brier score is a strictly proper scoring rule that rewards both discrimination and calibration (Brier, 1950; Gneiting & Raftery, 2007), and reporting both kinds of measure is recommended practice (Steyerberg et al., 2010). Poor calibration has been described as the Achilles heel of predictive analytics (Van Calster et al., 2019), and flexible modern learners are often miscalibrated even when accurate (Guo et al., 2017). When the outcome prevalence changes between development and deployment, posterior probabilities can be adjusted to the new prior (Lipton et al., 2018; Saerens et al., 2002). It is closely tied to class imbalance: reweighting or resampling the minority class improves recall at a fixed threshold but inflates predicted probabilities (He & Garcia, 2009), requiring prior correction (King & Zeng, 2001) or post hoc recalibration (Niculescu-Mizil & Caruana, 2005; Zadrozny & Elkan, 2002). For comparing classifiers on one test set, McNemar’s test compares decisions at a fixed threshold (Dietterich, 1998), whereas DeLong’s test compares correlated ROC curves (DeLong et al., 1988); the bootstrap quantifies metric uncertainty (Efron & Tibshirani, 1993). Reporting standards for prediction models, notably TRIPOD and

its machine-learning extension, codify these practices (Collins et al., 2015, 2024).

2.6. From prediction to decision

The newsvendor problem, in which an order quantity is chosen before uncertain demand is realised, is the canonical inventory decision; its solution depends on the critical fractile and the demand distribution (Khouja, 1999). If that distribution is estimated by a predictive model, probability errors translate directly into sub-optimal orders. Prescriptive analytics (Bertsimas & Kallus, 2020) and Smart Predict-then-Optimize (Elmachtoub & Grigas, 2022) show that the most accurate predictor statistically need not yield the best decisions.

2.7. Research gap

Three gaps emerge. First, freight disruption arises from interacting physical, economic and political

conditions, but predictive studies rarely combine these sources at a common spatial and temporal resolution, and even more rarely using only public data. Second, applied machine learning in transport and supply chains often relies on random splits and reports discrimination without calibration, limiting its relevance for decisions. Third, the link between prediction quality and decision quality is well established in theory but seldom demonstrated with models fitted to real data. This study addresses all three within a single, reproducible design.

III. DATA AND METHODS

Fig. 1 summarises the study design, from source alignment to the decision analysis.

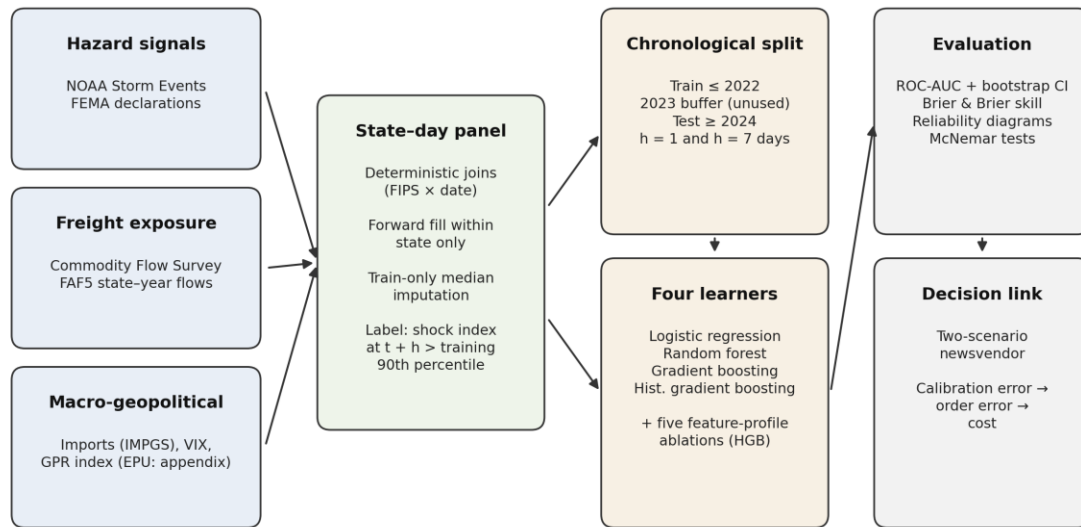


Fig. 1. Study framework: public data blocks are aligned on a leakage-controlled state-day panel, labelled with a forward-looking compound-hazard stress indicator, split chronologically with a one-year buffer, and used to compare four learners on discrimination, calibration and decision cost.

3.1. Study design and unit of analysis

We use a prognostic design in which information available on day t predicts a condition on day $t + h$. The unit of analysis is the state-day. States are the

level at which FEMA issues declarations, at which FAF reports origin-destination flows, and at which transportation agencies plan much of their emergency response; daily resolution matches storm reporting and supports short-horizon warning. The primary horizon is $h = 1$ day; $h = 7$ days is examined as a robustness check.

Evaluation uses a fixed-origin chronological split. Observations dated on or before 31 December 2022 form the training set; observations on or after 1

January 2024 form the test set; calendar year 2023 is excluded from both. The buffer prevents slow-moving exogenous variables, carried forward within state, from linking late-training and early-test observations, and it makes evaluation more demanding by widening the gap between learning and deployment. The cost is one year of unused data, which we later identify as a natural recalibration set (Section 5.2).

3.2. Data sources

Data were drawn from seven public sources grouped into three blocks (Table 1). All downloads are scripted so that the panel can be rebuilt from original files.

Table 1. Data sources, native resolution and representative variables.

Block	Source	Native resolution	Representative variables
Disaster and hazard signals	NOAA Storm Events Database	Event-level, daily	noaa_event_count;
	FEMA Disaster Declarations Summaries	Declaration-level, daily	noaa_reported_dmg_k fema_declaration_count
	Derived composite	State-day	shock_idx and lags at 1, 2 and 7 days
Freight exposure	Commodity Flow Survey (CFS)	Shipment records aggregated to state-day	cfs_val_musd_state_day; cfs_ton_k_state_day
	Freight Analysis Framework v5 (FAF5)	State-year	Domestic outbound/inbound, import and export tonnage and value
Macro-geopolitical context	Federal Reserve Economic Data (FRED)	Quarterly (IMPGS); daily (VIXCLS)	IMPGS; VIXCLS
	Geopolitical Risk (GPR) index	Monthly, national	gpr_index
	Economic Policy Uncertainty (EPU) index	Monthly, national	Appendix profile only

Note. Variable names follow the analysis code. IMPGS = U.S. imports of goods and services; VIXCLS = CBOE Volatility Index close.

Hazard signals. The NOAA Storm Events Database records severe weather reported by National Weather Service offices, including event type, location, date and estimated property and crop damage (National Centers for Environmental Information, n.d.). Events were aggregated to the state-day, counting events and summing reported damage in thousands of U.S. dollars. The FEMA Disaster Declarations Summaries list every major disaster, emergency and fire management assistance declaration by state and date (Federal Emergency Management Agency, n.d.); declarations were counted per state-day.

Freight exposure. The CFS, conducted jointly by the U.S. Census Bureau and the Bureau of Transportation

Statistics, samples shipments from mining, manufacturing, wholesale and selected retail and service establishments (U.S. Census Bureau, n.d.). Shipment value (million USD) and weight (thousand tons) were aggregated by origin state and shipment date. FAF5 integrates the CFS with other sources to estimate tonnage and value by origin, destination, commodity and mode (U.S. Bureau of Transportation Statistics, n.d.); flows were aggregated to the state-year separately for domestic outbound, domestic inbound, import and export movements and assigned to all days of the year. These variables describe how much freight a state handles and in which direction, conditioning its exposure when hazards strike.

Macro-geopolitical context. IMPGS and VIXCLS were obtained from FRED (Federal Reserve Bank of St. Louis, n.d.) and the GPR index from its authors' public release (Caldara & Iacoviello, 2022). They capture, respectively, trade volume entering the country, short-term financial uncertainty and geopolitical tension. EPU (Baker et al., 2016) and additional FRED series enter only a wider appendix profile.

3.3. Panel construction and temporal alignment

Construction proceeded in three stages. (i) Each source was cleaned and aggregated to its natural resolution: state–day for NOAA, FEMA and CFS; state–year for FAF5; national daily, monthly or quarterly for macro-geopolitical series. (ii) Sources were merged onto a complete grid of states and dates by deterministic joins on state FIPS code and date; national series were joined on date alone and so take a common value across states on a given day. (iii) Gaps arising from differing frequencies were filled *forward only* within each state after sorting by FIPS code and date, so that a day without a new observation inherits the most recent earlier value. No backward fill was applied; no day receives a value dated after it. Days preceding the first observation of a series remain missing and are imputed (Section 3.6).

Exogenous series were aligned by reference date rather than by publication date. Because monthly and quarterly statistics are released weeks after their reference period, this convention permits a small amount of look-ahead early in each period. The effect is confined to slow-moving national series and is revisited in Section 5.3. The build scripts produce a commodity-by-mode panel, the state–day modelling table and a table of FAF5 state–year exposures; the state–day table is the analytic dataset for all results.

3.4. Outcome definition

For state s and day t , the composite shock index sums three hazard components on their native scales:

$$S_{s,t} = N^{\text{NOAA}}_{s,t} + N^{\text{FEMA}}_{s,t} + D^{\text{NOAA}}_{s,t}, \quad (1)$$

where N^{NOAA} is the storm-event count, N^{FEMA} the declaration count and D^{NOAA} reported damage in thousands of dollars. The binary outcome at horizon h is

$$y^{(h)}_{s,t} = 1[S_{s,t+h} > q_h], \text{ with } q_h = Q_{0.90} \{ S_{s,t+h} : t \leq 31 \text{ Dec } 2022 \}, \quad (2)$$

that is, whether the forward shock exceeds the 90th percentile of its training-period distribution. The 90th percentile was chosen *a priori* to target the upper tail that is operationally relevant for contingency planning while retaining enough positive cases (about 2,260 in training) for stable estimation; fixing it on training data only ensures that the label definition uses no test-period information. For $h = 1$, $q_1 = 15,882.83$ index units. For $h = 7$, the forward shock is the index exactly seven days ahead (not the weekly maximum), with its own training threshold. Rows whose forward value is undefined were dropped, which is why the seven-day test set is slightly smaller (10,990 versus 11,296 rows). Two properties of this definition shape interpretation. First, because damage is measured in thousands of dollars while the other components are counts, the damage term dominates the upper tail; a positive label therefore identifies, above all, state–days followed by large reported storm damage. Standardising components before summation would give storm frequency and declarations more weight and is a natural extension. Damage is recorded in nominal dollars and was not deflated; because the threshold is fixed in nominal terms, price growth mechanically raises the share of later state–days that exceed it, a point we return to in Section 4.1. Second, because q_h is fixed using training data only, prevalence is approximately 10% in training by construction but is free to change in the test window, which makes any shift in the frequency of severe shocks directly observable.

We emphasise that the outcome is a *proxy*. It measures recorded hazard intensity and assistance declarations, not delays, cancellations or cost increases. A positive label indicates conditions under which freight disruption is plausible; whether disruption occurs depends on factors the index does not observe.

3.5. Predictors

The primary feature set (*paper primary* profile) contains 22 variables. The hazard block comprises the current shock index; its lags at one, two and seven days; the NOAA event count; the FEMA declaration count; and NOAA reported damage. The freight block comprises CFS state–day value and tonnage and FAF5 domestic outbound, domestic inbound, import and export tonnage and value. The macro-geopolitical block comprises IMPGS, VIXCLS and the GPR index.

Day of week and day of year capture weekly reporting patterns and seasonal hazard cycles.

Hazard activity is persistent: storm systems last several days, declarations cluster, and damage reports accumulate. Current and lagged shock values therefore carry much of the next-day signal, and the model partly learns persistence. This is appropriate for early warning, and the ablation design (Section 3.9) retains the hazard block in every profile so that it measures the value *added* by freight and macro information on top of persistence.

3.6. Missing data and preprocessing

After forward-filling, remaining missing values were imputed with each variable's training-set median. The imputer was fitted on training rows only and applied unchanged to test rows, and it was embedded with each model in a scikit-learn pipeline (Pedregosa et al., 2011), guaranteeing identical preprocessing at fitting and prediction and preventing test-period information from influencing imputed values.

3.7. Models

Four classifiers were fitted to the same features and outcome. *Logistic regression* with L2 regularisation provides a linear, interpretable baseline. A *random forest* averages trees grown on bootstrap samples with random feature subsets at each split (Breiman, 2001). *Gradient boosting* fits a sequence of shallow trees to the gradient of the loss at the current ensemble (Friedman, 2001). *HGB* follows the same principle but bins continuous features before split search and routes missing values within trees (Ke et al., 2017). Logistic regression was fitted with balanced class weights, which assign each class a weight inversely proportional to its frequency. The tree ensembles were fitted with the settings recorded in the replication code. HGB was designated *a priori* as the primary nonlinear model for its efficiency on mixed-type features and native missing-value handling; this designation was not revised after evaluation. Hyperparameters were fixed before evaluation, no tuning used the test window, and random seeds were fixed so that runs are exactly reproducible.

3.8. Evaluation

Each fitted model was applied once to the test window. Discrimination was summarised by ROC-AUC

(Fawcett, 2006). Probabilistic accuracy was summarised by the Brier score,

$$BS = (1/n) \sum_{i=1 \dots n} (\hat{p}_i - y_i)^2, \quad (3)$$

and by the Brier skill score relative to a constant forecast at the test-window prevalence π ,

$$BSS = 1 - BS / BS_{\text{ref}}, \text{ with } BS_{\text{ref}} = \pi(1 - \pi). \quad (4)$$

At $\pi = 0.172$, $BS_{\text{ref}} \approx 0.142$. A negative BSS means that a model's probabilities are, on average, less accurate than stating the base rate for every state-day, however well the model ranks. Because this reference uses the test prevalence, which a forecaster would not know in advance, it is a demanding benchmark. We therefore also report the skill relative to the operationally available alternative, a constant forecast equal to the training prevalence of 0.10, whose Brier score on the test window is $0.172 \times 0.9^2 + 0.828 \times 0.1^2 \approx 0.148$. The F1 score at a probability threshold of 0.5 summarises performance at a conventional operating point.

Calibration was examined with reliability diagrams that group test predictions into deciles and plot mean prediction against observed event rate (Niculescu-Mizil & Caruana, 2005). Uncertainty in ROC-AUC was quantified by nonparametric bootstrap of the test set, reporting the bootstrap mean and 2.5th and 97.5th percentiles (Efron & Tibshirani, 1993). Paired comparisons used McNemar's test at a threshold of 0.5 (Dietterich, 1998).

3.9. Feature-profile ablations and horizon sensitivity

To examine which information blocks contribute to discrimination, HGB was refitted under five profiles: *hazard + CFS* (hazard block, calendar variables and CFS flows); *hazard + FAF* (CFS replaced by FAF5 exposures); *paper primary* (all 22 variables); *appendix macro* (primary set plus additional FRED series and EPU); and *full features* (every constructed variable). All profiles used the same training and test rows. The full pipeline, including all models and profiles, was repeated at $h = 7$ with its own threshold. Results are read as associative stress tests of incremental signal, not causal effects.

3.10. Link to inventory decisions

To illustrate how predicted probabilities drive operational choices, we use a two-scenario newsvendor problem (Khouja, 1999). Demand is 80

units on a normal day and 120 units on a severe-shock day, which occurs with probability p . Overage cost is $c_o = 1.0$ and underage cost $c_u = 4.0$ per unit, so the critical fractile is $c_u/(c_u + c_o) = 0.80$. The optimal order is the smallest Q whose cumulative demand probability reaches the fractile. Since $Pr(D \leq 80) = 1 - p$,

$$Q^*(p) = 80 \text{ if } p \leq 0.20; Q^*(p) = 120 \text{ if } p > 0.20. \quad (5)$$

The example has no lead times, capacity limits or multi-period dynamics, and its costs are illustrative. Its purpose is to show that the decision depends on the *level* of the predicted probability relative to a threshold, so a model that systematically over- or under-predicts risk will make systematically wrong decisions even if its ranking is accurate.

3.11. Software and reproducibility

All analyses were conducted in Python with scikit-learn (Pedregosa et al., 2011). A single driver script runs the pipeline from data download to figure generation and writes all metrics to machine-readable files from which the tables were generated. Software versions are recorded with the replication materials. Reporting follows the TRIPOD and TRIPOD+AI principles (Collins et al., 2015, 2024); Appendix A maps the relevant items to manuscript sections.

IV. RESULTS

4.1. Sample and outcome distribution

Table 2 summarises the analytic sample. The training set contains 22,617 state-days and the test set 11,296. The most notable feature of the sample is the change in outcome prevalence: roughly one training day in ten is positive by construction, but in the test window the proportion rises to 17.2%, so state-days exceeding the training threshold were about 70% more frequent after 2023 than before. Plausible, non-exclusive explanations include higher severe-weather activity in 2024, changes in NOAA reporting practice, and

growth of the nominal damage component with inflation and asset values (Section 3.4). Distinguishing among them is beyond our scope, but the shift matters: any model trained on the earlier period enters the test window expecting a lower base rate than it will encounter.

Table 2. Study design and sample characteristics (one-day horizon).

Characteristic	Value
Unit of analysis	State-day
Outcome	Forward shock index at $t + 1$ above the training 90th percentile
Outcome threshold (index units)	15,882.83
Training window	On or before 31 December 2022
Buffer (excluded)	Calendar year 2023
Test window	On or after 1 January 2024
Training rows / test rows	22,617 / 11,296
Positive prevalence: training / test	≈10% (by construction) / 17.2%
Primary feature profile	22 variables (hazard, freight, macro-geopolitical, calendar)

4.2. Discrimination (RQI)

Table 3 reports test-window performance at $h = 1$, and Fig. 2 shows the ROC curve for HGB. Three models achieved very similar discrimination: gradient boosting (ROC-AUC 0.733; bootstrap 95% interval 0.721–0.744), logistic regression (0.732; 0.720–0.742) and HGB (0.725; 0.714–0.737). Their intervals overlap substantially, and differences between point estimates are smaller than the width of any interval. The random forest was weaker (0.706; 0.693–0.717), with only marginal overlap.

Table 3. Test-window performance at the one-day horizon.

Model	ROC-AUC	Bootstrap AUC, mean (95% interval)	Brier score	BSS, test rate	BSS, train rate	F1 at 0.5
Logistic regression	0.7319	0.732 (0.720–0.742)	0.2698	−0.89	−0.83	0.3847
Random forest	0.7059	0.706 (0.693–0.717)	0.1572	−0.10	−0.07	0.2731
Gradient boosting	0.7330	0.733 (0.721–0.744)	0.1297	+0.09	+0.12	0.0162

Model	ROC-AUC	Bootstrap AUC, mean (95% interval)	Brier score	BSS, test rate	BSS, train rate	F1 at 0.5
Histogram-based gradient boosting	0.7248	0.725 (0.714–0.737)	0.2565	−0.80	−0.74	0.3827

Note. BSS = Brier skill score = $1 - \text{BS}/\text{BS}_{\text{ref}}$. “Test rate”: $\text{BS}_{\text{ref}} = \pi(1 - \pi) \approx 0.142$ for a constant forecast at the test prevalence $\pi = 0.172$. “Train rate”: $\text{BS}_{\text{ref}} \approx 0.148$ for a constant forecast of 0.10 evaluated on the test window. Positive values indicate improvement over the constant forecast; values are approximate because π is rounded. Best value per column in bold (excluding F1; see Section 4.3).

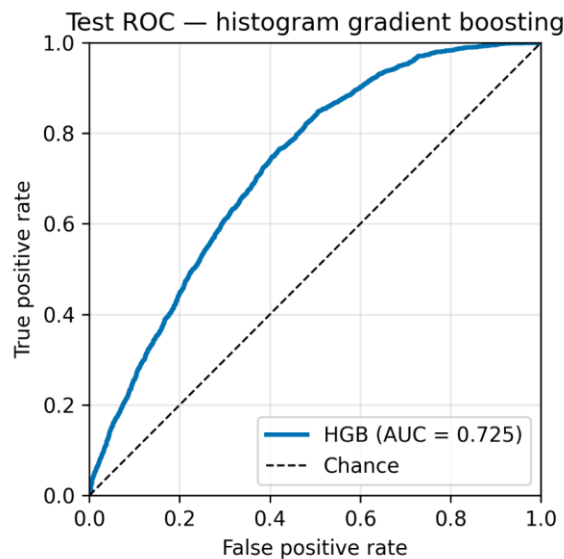


Fig. 2. Receiver operating characteristic curve for the histogram-based gradient boosting model on the test window (one-day horizon).

Two points follow. First, the model designated in advance as the primary nonlinear learner was not the best discriminator, and neither boosting implementation offered a meaningful advantage over regularised logistic regression; we report this as observed rather than revising the designation. Second, the parity between logistic regression and gradient boosting indicates that most available signal is close to additive and monotone, consistent with the dominance of current and lagged shock values whose relation to the next-day shock is roughly monotone. Nonlinear learners gain an advantage chiefly when interactions or thresholds carry substantial information, and the results give little sign that they do here. Because the

current and lagged shock values enter every model, all four learners are benchmarked against persistence implicitly, and the ablations (Section 4.5) quantify what non-hazard information adds on top of it.

An ROC-AUC of about 0.73 means that a randomly chosen positive state–day receives a higher score than a randomly chosen negative one about 73% of the time (Hanley & McNeil, 1982). This is clearly better than chance but far from the level needed to trigger costly interventions automatically; it suits *screening*, narrowing attention to states and days for closer review, rather than direct operational control.

4.3. Calibration and probabilistic accuracy (RQ2)

Brier scores differ far more across models than ROC-AUC does (Table 3). Gradient boosting achieved the lowest score (0.130), followed by random forest (0.157), HGB (0.257) and logistic regression (0.270). Only gradient boosting improved on the constant base-rate forecast ($\text{BSS} \approx +0.09$ against the test rate and $+0.12$ against the operationally available training rate); the ranking of models is identical under both references. Random forest was slightly worse than the constant forecast, and logistic regression and HGB were substantially worse ($\text{BSS} \approx -0.8$ to -0.9): their probabilities were less accurate, on average, than a forecast that ignored every predictor.

Fig. 3 locates the error. The HGB reliability curve is monotone, consistent with its reasonable discrimination, but lies well below the diagonal throughout. In the lowest decile the mean prediction is about 0.07 against an observed rate near 0.01; in the highest it is about 0.76 against about 0.36. At every level, predicted probability overstates observed frequency, often by a factor of two or more.

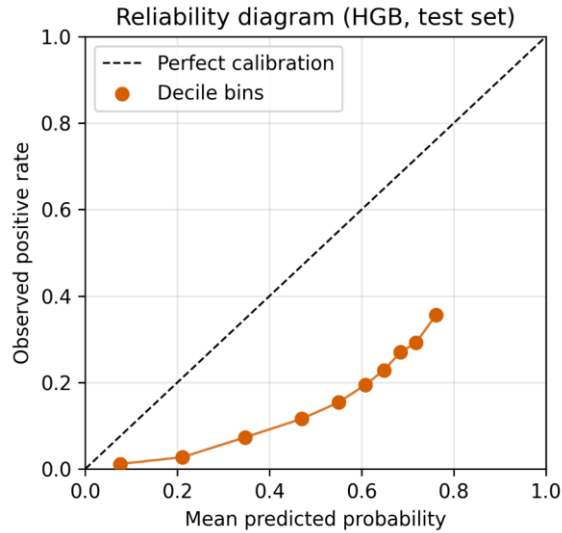


Fig. 3. Reliability diagram for the histogram-based gradient boosting model on the test window. Points show mean predicted probability and observed positive rate by decile of predicted probability; the dashed line indicates perfect calibration.

The *direction* of this error is diagnostic. Test prevalence (17.2%) exceeds training prevalence ($\approx 10\%$), so a model calibrated to training data would be expected to *under*-predict in the test window, placing its curve above the diagonal. The observed pattern is the opposite, which identifies the training objective rather than distribution shift as the dominant source of miscalibration. Class reweighting inflates the influence of positive cases during fitting and raises predicted probabilities across the board (He & Garcia, 2009; King & Zeng, 2001). Logistic regression was explicitly class-weighted, and the near-identical Brier score of HGB indicates that its probabilities were inflated in a comparable way during fitting. The upward prevalence shift partly offsets the inflation but not nearly enough to restore calibration. Conversely, gradient boosting, which shows no sign of reweighting-induced inflation, is exposed to the prior shift in the expected direction; its positive skill despite a 70% rise in prevalence indicates that it remains usable, although a prior-shift correction that reweights posterior odds by the ratio of new to old class priors (Saerens et al., 2002), with the new prior estimated from recent unlabelled data (Lipton et al., 2018), would be expected to improve it further.

The implication is that the raw outputs of logistic regression and HGB are *scores*, not probabilities. They

rank state-days about as well as gradient boosting, but an output of 0.6 does not mean a 60% chance of a severe shock. Where probabilities are needed, as in Section 4.7, outputs must first be recalibrated, for example by prior correction of the logistic intercept (King & Zeng, 2001) or by isotonic or logistic recalibration on a held-out period (Guo et al., 2017; Niculescu-Mizil & Caruana, 2005; Zadrozny & Elkan, 2002).

Gradient boosting presents the mirror-image trade-off. It was best on both discrimination and probabilistic accuracy, yet its F1 at 0.5 was only 0.016 because it rarely assigned any day a probability above 0.5. This reflects better calibration, not worse performance: with a 17% event rate and moderate discrimination, a well-calibrated model should seldom be more than 50% confident that a given day will be severe. F1 at 0.5 thus reveals where each model places its probabilities rather than how well it separates classes, and any operational threshold must be set for the application, not by convention.

4.4. Paired comparisons

McNemar tests comparing the random forest with logistic regression and with HGB at a threshold of 0.5 were both highly significant (Table 4). The discordant counts explain why: 5,384 test days were classified differently in one direction and only 5 in the other for logistic regression, and 5,398 against 3 for HGB. Nearly half of all test days were classified differently, almost all in the same direction.

Table 4. McNemar tests on test-window predictions at a threshold of 0.5.

Contrast	Discordant b	Discordant c	χ^2 statistic	p-value
Logistic regression vs. random forest	5,384	5	5,367.02	< .001
HGB vs. random forest	5,398	3	5,387.01	< .001

Note. HGB = histogram-based gradient boosting. Discordant counts are test days on which one model was correct and the other incorrect, in each direction. Given ROC-AUC differences of at most 0.027, these results show that the models operate at very different points on their ROC curves, not that one ranks days much better than another: logistic regression and HGB push many days above 0.5 and the random forest does not. McNemar's test is valid for comparing decisions at a fixed threshold (Dietterich, 1998), but here it chiefly detects the calibration difference already visible in Table 3. Comparisons of discrimination per se are better served by DeLong's test (DeLong et al., 1988).

4.5. Feature-profile ablations (RQ3)

Table 5 and Fig. 4 (right) report HGB ROC-AUC under each profile. At $h = 1$, hazard + FAF achieved the highest value (0.729), marginally above paper primary (0.725) and hazard + CFS (0.723). The two

wider profiles both reached 0.705, about 0.02 below the primary profile.

Table 5. Histogram-based gradient boosting ROC-AUC by feature profile.

Feature profile	ROC-AUC ($h = 1$)	ROC-AUC ($h = 7$)
Hazard + CFS	0.7229	0.7045
Hazard + FAF	0.7291	0.7067
Paper primary	0.7248	0.6815
Appendix macro	0.7050	0.6846
Full features	0.7050	0.6846

Note. Each profile was fitted on the same training rows and evaluated on the same test rows for its horizon. The appendix-macro and full-features profiles yield identical results to four decimal places at both horizons. CFS = Commodity Flow Survey; FAF = Freight Analysis Framework.

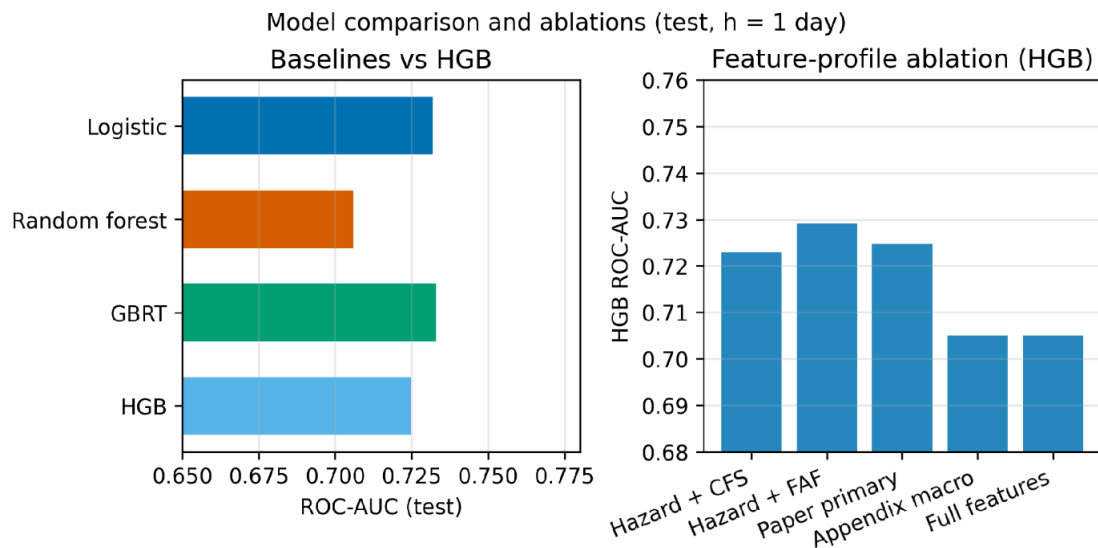


Fig. 4. Left: test-window ROC-AUC by model (one-day horizon). Right: histogram-based gradient boosting ROC-AUC under alternative feature profiles. Three patterns stand out. First, differences among the three compact profiles are at most 0.006, well within bootstrap uncertainty; the data do not favour CFS over FAF5 as the freight source, nor adding the macro block, when judged by discrimination. Second, the wider macroeconomic block *reduced* discrimination by about 0.02. The most plausible mechanism is distribution shift: national series vary only over time,

so a tree model can use them to learn period-specific patterns, including those tied to the atypical conditions of 2020–2022, that do not carry forward to 2024. Additional series give the model more ways to fit the training period without adding information that generalises. Third, the two wider profiles are indistinguishable to four decimal places at both horizons: beyond the appendix-macro set, the remaining constructed variables contribute nothing to the fitted model's discrimination.

For the question that motivated the study, whether combining sources helps, the answer is qualified. Hazard persistence carries most of the signal; freight exposure adds little on top of it at one day; and broad macroeconomic information can hurt. This does not imply that freight and macroeconomic conditions are unimportant for logistics stress. It reflects the fact that the outcome is defined from hazard records alone, for which hazard variables are the most direct predictors. Freight and macro variables would be expected to matter more for outcomes that measure logistics performance directly, such as delays or freight rates.

4.6. Horizon sensitivity

Discrimination declined for every model at $h = 7$ (Table 6), as expected given that hazard persistence weakens over a week. The decline was smallest for logistic regression ($0.732 \rightarrow 0.717$) and largest for HGB ($0.725 \rightarrow 0.682$). At seven days, logistic regression ranked first, followed by gradient boosting (0.710), random forest (0.688) and HGB. The model ranking thus changed with the horizon: HGB, close to the best at one day, was the weakest at seven. Brier scores rose slightly for three models and were unchanged for gradient boosting (0.1297 vs. 0.1295).

Table 6. Test-window performance at one-day and seven-day horizons.

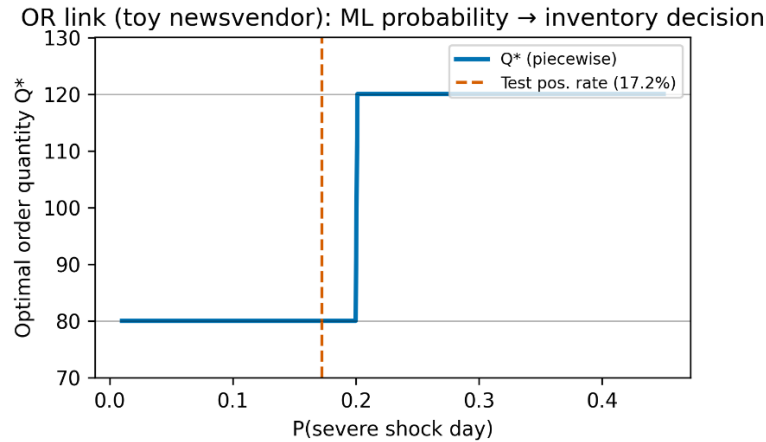
	ROC		ROC		
	-	-	-	-	-
	AUC	AUC		Brier	Brier
	($h =$	($h =$	ΔAU	($h =$	($h =$
Model	1)	7)	C	1)	7)
Logistic	0.731	0.717	-0.01	0.269	0.285
regressio	9	4	5	8	4
n					

	ROC		ROC		
	-	-	-	-	-
	AUC	AUC		Brier	Brier
	($h =$	($h =$	ΔAU	($h =$	($h =$
Model	1)	7)	C	1)	7)
Random	0.705	0.687	-0.01	0.157	0.165
forest	9	9	8	2	6
Gradient	0.733	0.709	-0.02	0.129	0.129
boosting	0	5	4	7	5
Histogra	0.724	0.681	-0.04	0.256	0.267
m-based	8	5	3	5	4
gradient					
boosting					

Note. Test rows: 11,296 at $h = 1$ and 10,990 at $h = 7$. Each horizon uses its own training-period threshold. The seven-day ablations (Table 5) sharpen the RQ3 finding. The paper primary profile had the *lowest* ROC-AUC of all five (0.682), below both compact hazard-plus-freight profiles (0.705 and 0.707) and slightly below the wider profiles (0.685). At the longer horizon, adding macro-geopolitical variables to hazard and freight features is the least effective choice for HGB. The greater stability of logistic regression is consistent with the view that simpler models generalise better when predictor–outcome relations weaken or shift: HGB fits more complex training-period patterns and therefore has more to lose when those patterns are unstable.

4.7. From probabilities to inventory decisions (RQ4)

Fig. 5 shows the optimal order quantity of Eq. (5) as a function of the probability of a severe-shock day. The test-window prevalence of 17.2% lies just below the switching point of 0.20, so a decision-maker who knew only the base rate would order 80 units every day.



Costs $c_{\text{hold}}=1.0$, $c_{\text{short}}=4.0$ → critical fractile 0.80; demand scenarios [80, 120] (see `or_newsvendor_toy.py`).

Fig. 5. Optimal order quantity in a two-scenario newsvendor problem as a function of the probability of a severe-shock day (holding cost 1.0, shortage cost 4.0, critical fractile 0.80, demand of 80 or 120 units). The dashed line marks the test-window prevalence (17.2%).

Because the switching point lies close to the base rate, the decision is highly sensitive to small probability errors. A well-calibrated model would move some days above 0.20 and leave most below, ordering extra stock only when evidence warrants. The HGB reliability curve (Fig. 3), however, assigns even the lowest-risk decile a mean prediction of about 0.07 and every decile from the third upward a mean prediction above 0.30. Used without recalibration, HGB would trigger the high order on most test days, including many on which the observed severe-shock rate was around 10% or lower. The resulting excess inventory would cost far more than the model's modest discrimination gain is worth. Two models with almost identical ROC-AUC would therefore produce very different stocking policies, and the one with the lower Brier score would produce better ones.

V. DISCUSSION

5.1. Theoretical implications

The results speak to three strands of literature. For *supply chain risk analytics*, they show that a transparent model built entirely from public data yields a moderate but genuine early signal of hazard-driven stress, and that simple learners capture most of it. This supports the argument that interpretable models should be preferred when their performance is comparable (Rudin, 2019), and it extends the tabular-learning evidence of Grinsztajn et al. (2022) with a

qualification: when the signal is dominated by monotone persistence, flexible ensembles offer little discrimination gain and can be *less* robust under temporal shift, as the horizon results show.

For *prediction-model evaluation*, the study provides a clear operational instance of the discrimination–calibration gap (Steyerberg et al., 2010; Van Calster et al., 2019). Four models within 0.027 in ROC-AUC differed by nearly a factor of two in Brier score, and the sign of the skill score split them into one model that beats the base rate and three that do not. The direction of miscalibration, opposite to what distribution shift alone would produce, points to the training objective, and class reweighting in particular, as its dominant cause. Because reweighting remains a default response to imbalance in applied work, this is a caution of general relevance: reweighted models should be reported and used as rankers unless explicitly recalibrated.

For *compound-hazard and multi-source forecasting*, the ablations show that conceptual relevance is not predictive value. Adding national macro-geopolitical series, all plausibly related to freight stress, reduced out-of-sample discrimination at both horizons, most likely because time-only variables let models memorise training-period regimes. Multi-source designs therefore need out-of-time evidence for each block rather than inclusion on theoretical grounds

alone. Finally, the newsvendor analysis makes concrete the prescriptive-analytics argument that prediction quality should be judged by decision quality (Bertsimas & Kallus, 2020; Elmachoub & Grigas, 2022): when a decision threshold lies near the base rate, calibration errors of a few percentage points, not differences in ranking, determine cost.

5.2. Managerial and policy implications

Choose the model by the use. For screening, where the goal is to rank states and days for attention, any of the three stronger models will serve. For decisions that depend on probability levels, such as pre-positioning inventory, staffing or routing, only gradient boosting produced usable probabilities without adjustment; logistic regression and HGB require recalibration. For binary alerts, thresholds should reflect the relative costs of false alarms and missed events, never a default of 0.5.

Recalibrate and monitor. The 70% increase in severe-shock frequency between windows would undermine any fixed threshold or calibration map. Operational deployments should recalibrate periodically on recent data and monitor observed against predicted event rates. In this design the unused 2023 buffer is a natural recalibration set, because it lies between training and deployment and was not used for fitting. Climate trends and changes in reporting practice make this especially important for hazard-driven outcomes.

Justify every data block out of time. More variables did not yield better performance. Agencies assembling multi-source dashboards should retain a block only if it improves performance on later held-out periods.

5.3. Limitations and future research

The most fundamental limitation concerns the outcome. The composite index measures recorded hazard activity and assistance declarations, not logistics performance; the model predicts when a state is likely to face intense hazard conditions, a plausible precursor of freight disruption but not the same thing. Its damage-dominated construction further narrows what a positive label represents. Validating the approach against direct disruption measures, such as truck travel times, port dwell times or freight rates, and testing standardised index constructions are the most important next steps.

The data impose further constraints. NOAA damage estimates are subject to inconsistent reporting and revision (Gall et al., 2009). CFS and FAF5 values are survey-based and temporally coarse, and assigning annual FAF values to every day removes within-year variation. National macro-geopolitical series cannot distinguish states on the same day, and aligning them by reference rather than release date may allow slight look-ahead early in each period; a vintage-aware alignment would remove this.

Findings at the state-day level cannot be transferred to firms, facilities or routes (Robinson, 1950). States differ greatly in size and freight activity, and observations within a state are serially correlated; the i.i.d. bootstrap intervals therefore likely understate uncertainty, and block or state-clustered resampling would be more conservative. The study uses a single fixed-origin split, so results describe one test period; rolling-origin evaluation across several years would test stability. Hyperparameters were not tuned, which is conservative but may understate achievable performance. Paired discrimination comparisons would be strengthened by DeLong's test, and an explicit persistence-only rule (for example, predicting from the current shock alone) would make the value added by learning more transparent. Sensitivity of the conclusions to the percentile used for the label and to deflating reported damage should also be examined. The ablations describe fitted-model behaviour, not causal effects. The inventory example is stylised. Sub-state geographies and network-aware models that represent highway and port connections could capture spatial spillovers that the present state-level design ignores, and decision-focused training and evaluation would bring assessment closer to operational use.

VI. CONCLUSION

This study asked whether openly available data on disasters, freight exposure and macro-geopolitical conditions can be combined into a useful early signal of compound-hazard stress for U.S. states. Using a leakage-controlled state-day panel from seven public sources and a chronological holdout separated from training by a full year, we find a moderate signal (ROC-AUC ≈ 0.73 at one day) that a regularised logistic regression captures as well as the best tree ensemble. The model designated in advance as

primary, histogram-based gradient boosting, did not outperform simpler alternatives and was the weakest at seven days. Calibration varied far more than discrimination: only gradient boosting beat a constant base-rate forecast, while logistic regression and HGB over-predicted risk at every level. Hazard persistence carried most of the signal, freight exposure added little, and a wider macroeconomic block reduced out-of-sample performance. Severe-shock frequency rose by about 70% between training and test periods.

The broader lesson extends beyond this application. Models with nearly identical ROC-AUC can support very different decisions, because decisions depend on the level of the predicted probability, not only its rank. Studies that report discrimination alone, as much applied work in transport and supply chain prediction still does, omit the information decision-makers most need. Public data will not replace operational telemetry for real-time freight management, but a transparent, reproducible public-data baseline can provide a moderate early signal of hazard-driven stress, provided its outputs are calibrated, its validity is tested out of time, and its thresholds are chosen with the intended decision in mind.

Declarations

CRediT authorship contribution statement. Sunday Michael Oyebiyi: Conceptualization, Methodology, Software, Data curation, Formal analysis, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest. The author declares no known competing financial interests or personal

relationships that could have appeared to influence the work reported in this paper.

Funding. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data availability. All data used in this study are publicly available from the federal and research sources listed in Table 1 and cited in the reference list (Baker et al., 2016; Caldara & Iacoviello, 2022; Federal Emergency Management Agency, n.d.; Federal Reserve Bank of St. Louis, n.d.; National Centers for Environmental Information, n.d.; U.S. Bureau of Transportation Statistics, n.d.; U.S. Census Bureau, n.d.).

Code availability. The panel construction, outcome definition, models and evaluation procedures are described in Section 3 in sufficient detail to allow replication with the cited public data.

Declaration of generative AI and AI-assisted technologies. During the preparation of this work the author used large language model-based assistants to support software implementation, code refactoring, and drafting and editing of the manuscript text. After using these tools, the author reviewed, verified and edited all outputs and takes full responsibility for the content of the publication.

Acknowledgements. The author thanks the U.S. federal agencies and research groups that maintain the public datasets and indices used in this study.

Appendix A. Reporting checklist (TRIPOD+AI, selected items)

Item	Where reported
Source of data and study dates	Sections 3.1–3.2; Tables 1–2
Unit of analysis and setting	Section 3.1
Outcome definition, timing and threshold	Section 3.4; Eqs. (1)–(2)
Predictors and timing of measurement	Sections 3.3, 3.5
Handling of missing data	Sections 3.3, 3.6
Leakage controls (forward fill, train-only preprocessing, buffer)	Sections 3.1, 3.3, 3.6
Model specification and class weighting	Section 3.7

Item	Where reported
Hyperparameter selection	Section 3.7 (fixed a priori; no test-window tuning)
Discrimination measures with uncertainty	Section 4.2; Table 3
Calibration measures and plots	Section 4.3; Table 3; Fig. 3
Distribution shift between development and evaluation	Section 4.1
Model comparison tests	Section 4.4; Table 4
Limitations and intended use	Sections 5.2–5.3
Code and data availability	Declarations

REFERENCES

- [1] Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, 131(4), 1593–1636. [Oxford Academic](#)
- [2] Baryannis, G., Validi, S., Dani, S., & Antoniou, G. (2019). Supply chain risk management and artificial intelligence: State of the art and future research directions. *International Journal of Production Research*, 57(7), 2179–2202. [Taylor & Francis](#)
- [3] Benigno, G., di Giovanni, J., Groen, J. J. J., & Noble, A. I. (2022). The GSCPI: A new barometer of global supply chain pressures (Staff Report No. 1017). Federal Reserve Bank of New York. [Federal Reserve Bank of New York](#)
- [4] Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192–213. [ScienceDirect](#)
- [5] Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3), 1025–1044. [INFORMS](#)
- [6] Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica*, 77(3), 623–685. [Wiley](#)
- [7] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. [Springer](#)
- [8] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3. [American Meteorological Society](#)
- [9] Brintrup, A., Pak, J., Ratiney, D., Pearce, T., Wichmann, P., Woodall, P., & McFarlane, D. (2020). Supply chain data analytics for predicting supplier disruptions: A case study in complex asset manufacturing. *International Journal of Production Research*, 58(11), 3330–3341. [Taylor & Francis](#)
- [10] Caldara, D., & Iacoviello, M. (2022). Measuring geopolitical risk. *American Economic Review*, 112(4), 1194–1225. [American Economic Association](#)
- [11] Carvalho, V. M., Nirei, M., Saito, Y. U., & Tahbaz-Salehi, A. (2021). Supply chain disruptions: Evidence from the Great East Japan Earthquake. *The Quarterly Journal of Economics*, 136(2), 1255–1321. [Oxford Academic](#)
- [12] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. [ACM](#)
- [13] Chopra, S., & Sodhi, M. S. (2004). Managing risk to avoid supply-chain breakdown. *MIT Sloan Management Review*, 46(1), 53–61.
- [14] Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378. [BMJ](#)
- [15] Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ*, 350, g7594. [BMJ](#)

- [16] Craighead, C. W., Blackhurst, J., Rungtusanatham, M. J., & Handfield, R. B. (2007). The severity of supply chain disruptions: Design characteristics and mitigation capabilities. *Decision Sciences*, 38(1), 131–156. [Wiley](#)
- [17] Cutter, S. L., Boruff, B. J., & Shirley, W. L. (2003). Social vulnerability to environmental hazards. *Social Science Quarterly*, 84(2), 242–261. [Wiley](#)
- [18] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3), 837–845. [Crossref](#)
- [19] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7), 1895–1923. [MIT Press](#)
- [20] Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.
- [21] Elmachoub, A. N., & Grigas, P. (2022). Smart “predict, then optimize.” *Management Science*, 68(1), 9–26. [INFORMS](#)
- [22] Ermagun, A., & Levinson, D. (2018). Spatiotemporal traffic forecasting: Review and proposed directions. *Transport Reviews*, 38(6), 786–814. [Taylor & Francis](#)
- [23] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. [ScienceDirect](#)
- [24] Federal Emergency Management Agency. (n.d.). *OpenFEMA dataset: Disaster declarations summaries – v2* [Data set]. Retrieved April 20, 2026, from [FEMA](#)
- [25] Federal Reserve Bank of St. Louis. (n.d.). *Federal Reserve Economic Data (FRED)* [Data set]. Retrieved April 20, 2026, from [Federal Reserve Bank of St. Louis](#)
- [26] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. [Project Euclid](#)
- [27] Gall, M., Borden, K. A., & Cutter, S. L. (2009). When do losses count? Six fallacies of natural hazards loss data. *Bulletin of the American Meteorological Society*, 90(6), 799–809. [American Meteorological Society](#)
- [28] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378. [Taylor & Francis](#)
- [29] Gössling, S., Neger, C., Steiger, R., & Bell, R. (2023). Weather, climate change, and transport: A review. *Natural Hazards*, 118(2), 1341–1360. [Springer](#)
- [30] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 507–520).
- [31] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (PMLR Vol. 70, pp. 1321–1330).
- [32] Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36. [Radiology](#)
- [33] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. [IEEE](#)
- [34] Hosseini, S., Ivanov, D., & Dolgui, A. (2019). Review of quantitative methods for supply chain resilience analysis. *Transportation Research Part E: Logistics and Transportation Review*, 125, 285–307. [ScienceDirect](#)
- [35] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. [OTexts](#)
- [36] Ivanov, D. (2020). Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. *Transportation Research Part E: Logistics and Transportation Review*, 136, 101922. [ScienceDirect](#)
- [37] Ivanov, D., & Dolgui, A. (2020). Viability of intertwined supply networks: Extending the

- supply chain resilience angles towards survivability. A position paper motivated by COVID-19 outbreak. *International Journal of Production Research*, 58(10), 2904–2915. [Taylor & Francis](#)
- [38] Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), Article 15. [ACM](#)
- [39] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 3146–3154).
- [40] Khouja, M. (1999). The single-period (news-vendor) problem: Literature review and suggestions for future research. *Omega*, 27(5), 537–553. [ScienceDirect](#)
- [41] King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2), 137–163. [Oxford Academic](#)
- [42] Lipton, Z. C., Wang, Y.-X., & Smola, A. J. (2018). Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning* (PMLR Vol. 80, pp. 3122–3130).
- [43] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774).
- [44] Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1), 521–530. [ScienceDirect](#)
- [45] National Centers for Environmental Information. (n.d.). *Storm events database* [Data set]. National Oceanic and Atmospheric Administration. Retrieved April 20, 2026, from [NOAA](#)
- [46] Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). ACM. [ACM](#)
- [47] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [48] Pettit, T. J., Fiksel, J., & Croxton, K. L. (2010). Ensuring supply chain resilience: Development of a conceptual framework. *Journal of Business Logistics*, 31(1), 1–21. [Wiley](#)
- [49] Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357. [Crossref](#)
- [50] Rose, A. (2004). Defining and measuring economic resilience to disasters. *Disaster Prevention and Management*, 13(4), 307–314. [Emerald](#)
- [51] Rose, A., & Wei, D. (2013). Estimating the economic consequences of a port shutdown: The special role of resilience. *Economic Systems Research*, 25(2), 212–232. [Taylor & Francis](#)
- [52] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. [Nature](#)
- [53] Saelens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1), 21–41. [MIT Press](#)
- [54] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. [PLOS](#)
- [55] Sheffi, Y., & Rice, J. B., Jr. (2005). A supply chain view of the resilient enterprise. *MIT Sloan Management Review*, 47(1), 41–48.
- [56] Snyder, L. V., Atan, Z., Peng, P., Rong, Y., Schmitt, A. J., & Sinsoysal, B. (2016). OR/MS models for supply chain disruptions: A review. *IIE Transactions*, 48(2), 89–109. [Taylor & Francis](#)

- [57] Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. *Epidemiology*, 21(1), 128–138. [PubMed](#)
- [58] Tang, C. S. (2006). Perspectives in supply chain risk management. *International Journal of Production Economics*, 103(2), 451–488. [ScienceDirect](#)
- [59] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4), 437–450. [ScienceDirect](#)
- [60] Tukamuhabwa, B. R., Stevenson, M., Busby, J., & Zorzini, M. (2015). Supply chain resilience: Definition, review and theoretical foundations for further study. *International Journal of Production Research*, 53(18), 5592–5623. [Taylor & Francis](#)
- [61] U.S. Bureau of Transportation Statistics. (n.d.). *Freight Analysis Framework (FAF)* [Data set]. Retrieved April 20, 2026, from [U.S. Bureau of Transportation Statistics](#)
- [62] U.S. Census Bureau. (n.d.). *Commodity Flow Survey (CFS)* [Data set]. Retrieved April 20, 2026, from [U.S. Census Bureau](#)
- [63] Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine*, 17, Article 230. [BMC](#)
- [64] Verschuur, J., Koks, E. E., Li, S., & Hall, J. W. (2023). Multi-hazard risk to global port infrastructure and resulting trade and logistics losses. *Communications Earth & Environment*, 4, Article 5. [Nature](#)
- [65] Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2014). Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies*, 43, 3–19. [ScienceDirect](#)
- [66] Whaley, R. E. (2000). The investor fear gauge. *The Journal of Portfolio Management*, 26(3), 12–17. [The Journal of Portfolio Management](#)
- [67] Wyrembek, M., Baryannis, G., & Brintrup, A. (2025). Causal machine learning for supply chain risk prediction and intervention planning. *International Journal of Production Research*, 63(15), 5629–5648. [Taylor & Francis](#)
- [68] Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 694–699). ACM. [ACM](#)
- [69] Zheng, G., Kong, L., & Brintrup, A. (2023). Federated machine learning for privacy preserving, collective supply chain risk prediction. *International Journal of Production Research*, 61(23), 8115–8132. [Taylor & Francis](#)
- [70] Zscheischler, J., Westra, S., van den Hurk, B. J. J. M., Seneviratne, S. I., Ward, P. J., Pitman, A., AghaKouchak, A., Bresch, D. N., Leonard, M., Wahl, T., & Zhang, X. (2018). Future climate risk from compound events. *Nature Climate Change*, 8(6), 469–477. [Nature](#)