

OPERATING CASE STUDY / DATA, ANALYTICS & AI

AI Model Selection & Release Governance

Choose the model, the decision loss and the exact version that earns release

An anonymized project case study of forecast-model research: comparative testing, a constrained-runtime champion, a partial engine handoff and an evidence-based decision to defer neural development.

PROJECT CONTEXT / ANALYTICS, MODEL GOVERNANCE & IMPLEMENTATION

BLUF / BOTTOM LINE UP FRONT

Promote the verified configuration.
Make complexity earn its next test.

RECALCULATED FINDING

Against the stronger comparator on the same recorded test split, the champion's error advantage is 8.02%, rather than 9.50%. The engine retained a component the research version removed, so the benchmark and promoted configuration require separate acceptance.

CONTRIBUTION IN THE SOURCE RECORD

Research direction, constraints, acceptance standards and promotion decisions. Provisional registry-level hoop support; agent-assisted analytical execution.

EVIDENCE STATUS

Registry-based model-development case; originals unrecovered. Seed benchmarks and a reported code port do not establish live accuracy or customer benefit.

BLUF: analytical inference. Source contribution: historical record. Client and internal identities anonymized.

EXECUTIVE ABSTRACT

Require the selected model and the released version to earn the same decision

[AI] The registered research established a practical champion and a conditional neural-model deferral. Its durable contribution is a disciplined selection and promotion process, bounded by the evidence each stage actually supplies.

CONTEXT | Model ambition met a constrained operating runtime

[HR] The project compared forecasting candidates against an incumbent, added uncertainty estimates, retained failed experiments and specified a neural-model go/no-go. The shipping environment favored a lightweight implementation; usable live history remained a dependency.

OBJECTIVE | Choose the next model investment and release

[PD] Select on a common target, comparator and decision loss. Separate research selection, exact-configuration promotion and live acceptance so that one strong score cannot authorize all three.

APPROACH | Direct a falsifiable comparative program

[HR] The project lead commissioned the review, imposed runtime and preservation constraints, set a falsification gate and directed goal-based severity decisions. [AI] These claims receive provisional registry-level hoop support. Analytical implementation was agent-assisted.

FINDINGS | The comparator and the promoted version change the claim

[RC] The champion's recorded TEST error advantage is 8.02% against the stronger incumbent, versus 9.50% against the shipped baseline. [HR] The research configuration removed weather; the reported engine port retained it. Matching keys and aggregate error did not imply identical target realizations.

DECISION VALUE | Fund an authenticated comparison, then authorize scope

[PD] The analytics sponsor should freeze the loss and eligible comparator; the platform owner should verify configuration parity and rollback. Require an untouched final live evaluation with a positive lower 95% bound on prespecified incremental value, plus separately sized safety and capacity gates, before scaled promotion.

IMPLICATIONS | Keep complexity conditional without declaring it futile

[AI] Two changed seed experiments do not establish a learning curve or neural inferiority. The proposed Champion Promotion Contract makes data access, selection validity, operating loss and runtime feasibility explicit conditions for the next investment.

S1: protected project register and calculation ledger. HR historical record; RC recalculated; AI analytical inference; PD proposed design. No live model superiority or historical customer benefit is established.

EXHIBIT 1 / BENCHMARK IDENTITY

Verify the target realizations before calling a replay exact

[HR] The register reports matching prediction keys and a close aggregate headline reproduction, while also recording different row-level actuals. [AI] These are compatible facts; they validate different parts of the comparison.

MATCHING HEADLINES DO NOT ESTABLISH AN EXACT REPLAY
Historical register and analytical inference • separate three identity checks



Recover target and prediction rows; reconcile version, weights, exclusions and tolerances.

[HR] The research regenerated a common seed panel and applied one grading harness to the candidates. That is a useful comparative design. [AI] It does not authenticate a replay of the engine’s original target realizations merely because row identifiers and a rounded mean error agree.

Identity layer	What the check can establish
[PD] Keys and population	The same entities, metrics, issue times and horizons are represented, with duplicates and exclusions reconciled.
[PD] Targets and scores	The same target version and as-of inputs are used; row predictions, losses, aggregation weights and tolerances also reconcile.

[AI] Within the regenerated panel, paired candidate comparisons may remain valid for that panel. Cross-artifact reproduction is a separate claim. Synthetic targets also limit transport to live operations even if every computational identity check passes.

VALIDATION OWNER ACTION

[PD] Recover and hash the original target and prediction rows. Preserve an explicit difference ledger; distinguish exact replay, approximate aggregate agreement and a newly generated common benchmark.

[HR/AI/PD] S1:M03/E01–E03. Original files were not recovered for this case. The documented target mismatch narrows replay claims without automatically invalidating within-panel candidate comparisons.

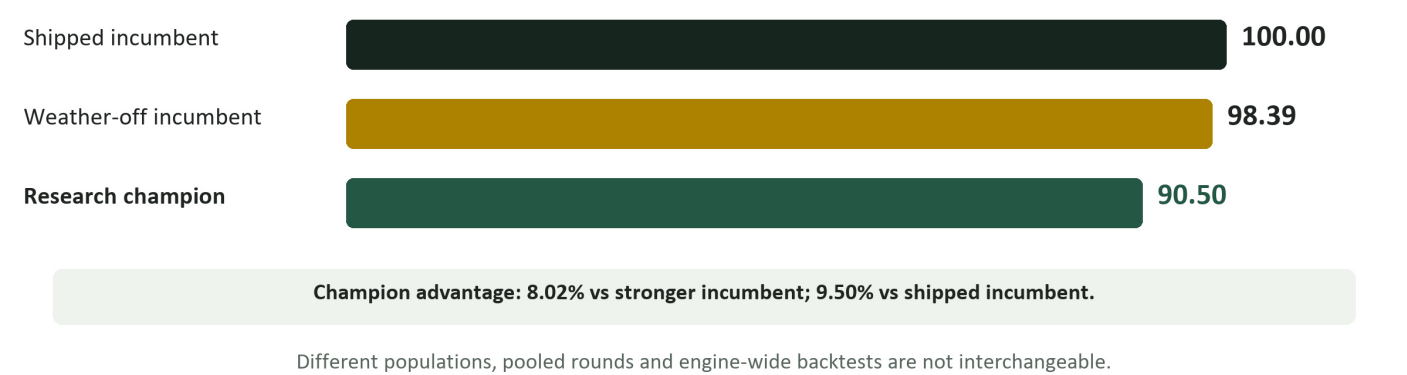
EXHIBIT 2 / THE ELIGIBLE COMPARATOR

Measure the champion against the stronger baseline on the same test

[RC] On the recorded TEST split, the champion’s mean absolute error is 9.50% below the shipped incumbent and 8.02% below the weather-off incumbent. Changing the comparator reduces the advantage without reversing it.

THE STRONGER MATCHED BASELINE REDUCES THE CLAIM

Recalculated • same recorded TEST split; shipped incumbent error = 100



[HR] The review raised its own bar by retaining both incumbent configurations and a simpler parameter guardrail. [AI] That prevents an avoidable configuration weakness from inflating the apparent value of a new architecture. The comparator still has to satisfy the intended use and runtime constraints.

Comparison contract	Required discipline
[PD] Like-for-like score	Use the same TEST rows, target versions, loss weights and horizon mix. Do not substitute the pooled baseline from an earlier round.
[PD] Untouched acceptance	Recover the selection chronology. If TEST scores influenced feature or model choices, use a fresh final evaluation after freezing the selected pipeline.

[AI] A disjoint training split protects against some errors; it does not prove the final score was untouched by iterative selection. The recorded instructions use TEST performance in several gates, so the label “out-of-sample” alone cannot settle selection bias.

ANALYTICS SPONSOR ACTION

[PD] Preserve the simpler champion as a candidate. Authorize promotion only against an eligible strong baseline and a final acceptance set that did not choose the winner.

[HR/RC/AI/PD] S1:M01/M02/D01–D03/D06/E04/E06. Shipped TEST error=100 for the chart. Gains describe one recorded seed benchmark, not independent live validation or realized operating value.

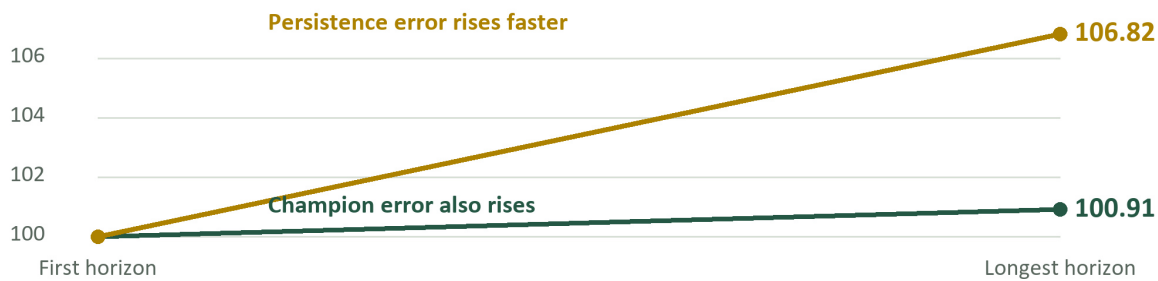
EXHIBIT 3 / RELATIVE SKILL AND ABSOLUTE ERROR

Read the forecast error beside the denominator that defines skill

[RC] Across the recorded horizons, champion error increases slightly while persistence error increases faster. [AI] Higher relative skill at the longer horizon does not mean the champion becomes more accurate.

RELATIVE SKILL CAN RISE WHILE MODEL ERROR ALSO RISES

Recalculated • each series indexed separately to first-horizon error = 100



Endpoint comparison, not an operating time series; connecting lines guide the eye.
Separate indices show within-series changes, not equal raw error levels.

[HR] The project reported both absolute error and skill relative to persistence. [AI] Their joint interpretation matters: an easier-to-beat comparator can improve a relative score even when the model’s own error rises. Separate first-horizon indices show change within each series; they do not equate the raw error levels.

Score interpretation	Acceptance question
[PD] Point accuracy	How large is the error at the horizon when a decision must be made, relative to an eligible comparator on the same observations?
[PD] Uncertainty quality	Are intervals adequately covered and sufficiently narrow, including tails and groups? Coverage alone can reward an uninformatively wide interval.

[AI] The register also uses opposite skill conventions: error ratio and one minus error ratio. Their favorable directions differ. A score dictionary must specify its equation before comparing headlines. Improved tier agreement after changed cut-points is likewise not an independent point-forecast gain.

MODEL OWNER ACTION

[PD] Publish the loss, comparator, horizon and definition together. Test calibration and sharpness separately; keep changed severity taxonomies outside an unchanged-metric trend claim.

[HR/RC/AI/PD] S1:M03/M06/M08/M16. Each error series is separately indexed to its first horizon=100. Horizons are evaluation slices, not an operating time series or independent samples.

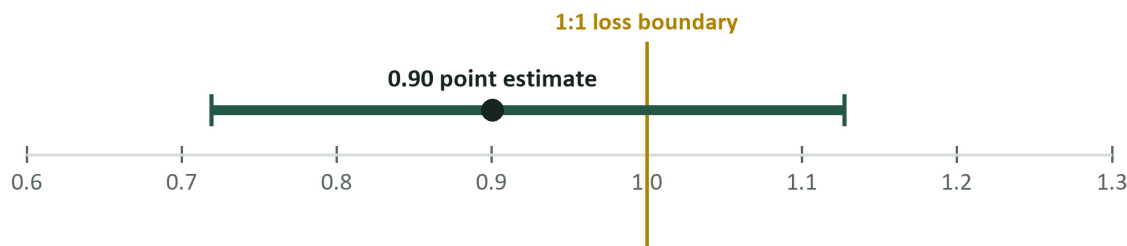
EXHIBIT 4 / TAIL RECALL AND DECISION LOSS

Price the extra false alerts before retaining the tail-sensitive component

[RC] The broad tail slice implies 0.90 extra false alerts per extra detected breach; counts compatible with source rounding span 0.72–1.13. [AI] A one-to-one decision-loss ratio does not robustly settle the component choice.

SOURCE ROUNDING CAN REVERSE THE TAIL-LOSS DECISION

Recalculated • extra false alerts per extra detected breach in the broad tail slice



Rounding-compatible range: 0.72–1.13; not a confidence interval.
Underlying confusion counts and the decision loss must settle the choice.

Conditional on common labels and population; detection does not equal prevention.

[HR] The research preferred weather off for aggregate error; the engine retained it for tail recall. These metrics answer different questions. [AI] A higher recall is not sufficient justification if false alerts consume scarce response capacity, and a lower average error does not settle the cost of missed severe events.

Loss component	Decision interpretation
[PD] Missed breach versus false alert	Under fixed-label classification loss, favor the change when missed-breach cost divided by false-alert cost exceeds the exchange ratio.
[PD] Operating response	Add actual action cost, displacement, preventability and safeguards. Detected breaches are not automatically prevented customer failures.

[AI] This range reflects discrete count combinations compatible with rounded source summaries, not a confidence interval or recovered outcomes. The weather slices are nested and cannot be added. Common labels and population identity remain conditions; field preventability remains untested.

OPERATIONS SPONSOR ACTION

[PD] Freeze the loss ratio and response budget before selecting the weather configuration. Require a non-circular live panel and test the decision rule, not recall alone.

[HR/RC/AI/PD] S1:M04/M15/D02/E08/E09 and private rounding-feasibility ledger. Only the protected broad-slice exchange and its rounding-compatible sensitivity are shown; neither is an observed operating benefit.

EXHIBIT 5 / COMPLEXITY AND THE NEXT EXPERIMENT

Treat the neural deferral as an investment gate, not a model-class verdict

[HR] A tree benchmark led the champion in one seed experiment and had a much smaller edge in another. [AI] Both the history and evaluation setup changed; the difference does not identify what more data caused.

CHANGED SEED EXPERIMENTS DO NOT IDENTIFY A LEARNING CURVE

Historical register and proposed design • no neural model was trained

EXPERIMENT A

Tree challenger leads

EXPERIMENT B

Smaller observed edge

MISSING TEST

Fixed cohort, varied history

History, origins and holdout composition changed together.

Reopen complexity only after data, loss, runtime and fresh-test gates are explicit.

A conditional investment deferral is supported; a verdict on neural model capability is not.

[HR] The work retained evidence favorable to the challenger, including a station-identity ablation and horizon comparisons. It specified deeper candidates but did not train them. [AI] The simpler runtime path and unresolved live data justify a conditional next investment; they do not prove that complex models cannot add value.

Question to reopen	Falsifiable comparison
[PD] Does history change the ranking?	Hold evaluation cohorts and loss fixed; vary training history within repeated, prespecified origins. Include independent realizations and report uncertainty.
[PD] Does a deeper model earn its cost?	Compare the frozen candidate against persistence and the strongest eligible challenger on point loss, probabilistic loss and runtime burden.

[AI] A modest train/test gap or removing a site identifier does not rule out every leakage path. Null results from operational proxies also cannot prove that the corresponding real signal is absent—or that it will help once available.

RESEARCH SPONSOR ACTION

[PD] Fund data and an informative experiment before a larger architecture. Define the result that would reopen the neural build, and the result that would keep it deferred.

[HR/AI/PD] S1:M09–M13/D04/D06/E12–E16. The two seed experiments are not a controlled learning curve. No neural superiority or inferiority was empirically established.

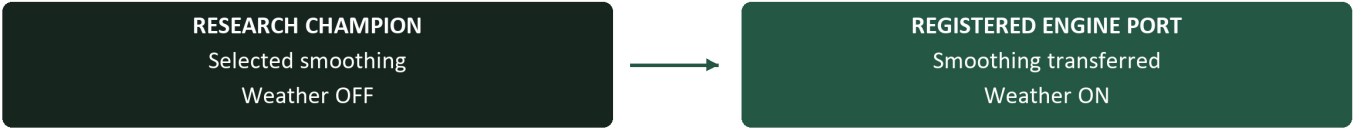
EXHIBIT 6 / EXACT-CONFIGURATION PROMOTION

Credit the smoothing handoff without relabeling it the full champion

[HR] The register describes a reversible smoothing port into the engine while weather remained on. [AI] This is a partial configuration handoff; the research champion’s complete score cannot be attached to the promoted version.

THE RESEARCH CHAMPION AND PORT ARE DIFFERENT CANDIDATES

Historical register • configuration identity is a promotion requirement



TEST and engine-wide backtest gains describe different evaluations.
Do not attach the full research score to the partial configuration handoff.

Authenticate artifact identity, supported outputs, exception ownership and rollback.

[HR] The recorded handoff preserved a feature switch, parameter provenance and before/after backtests. These are useful implementation controls. [AI] Original code and logs remain unrecovered, so this case reports the registered port rather than independently certifying execution or runtime parity.

Handoff boundary	Evidence needed
[PD] Research to runtime	Match smoothing, clipping, feature gates, source version, missingness rules, prediction outputs and uncertainty behavior—not only parameter names.
[PD] Runtime to live use	Authenticate deployment and intended users, then validate live targets and the permitted operating decision on its declared follow-up clock.

[AI] The research TEST score and engine-wide seed backtest use different populations and configurations. Their gains are not additive benefits or a controlled measure of lost gain during translation. A rollback parity check verifies only the behavior and inputs actually covered by that check.

PLATFORM OWNER ACTION

[PD] Sign a configuration difference ledger. Promote the exact tested artifact, assign exceptions to an owner and expiry, and verify rollback on the same supported cases before release.

[HR/AI/PD] S1:M01/M02/M14/D08/E08/E09/E17/E18. Directed standards and handoff decisions: provisional registry-level hoop support. Field adoption and attributable customer impact remain unestablished.

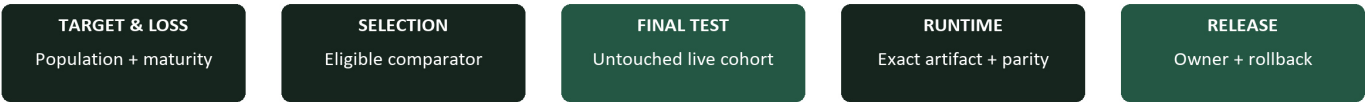
EXHIBIT 7 / CHAMPION PROMOTION CONTRACT

Make the promotion decision a signed evidence contract

[PD] The proposed Champion Promotion Contract links the chosen loss to the selected pipeline, eligible comparator, final evaluation, runtime artifact and accountable release. A change in any field reopens the relevant gate.

THE CHAMPION PROMOTION CONTRACT MAKES EACH GATE EXPLICIT

Proposed CPC framework • a changed contract field reopens its evidence gate



Authorize the next evidence stage; scale only within supported conditions.
Separate score improvement, forecast-service value and customer-outcome impact.

Each claim travels with its population, comparator, outcome, configuration and evidence class.

Authorization	Accountable requirement
[PD] Select the candidate	Analytics owner freezes targets, loss and candidate-selection process. The final acceptance set remains outside adaptive tuning.
[PD] Promote the artifact	Platform owner reconciles exact configuration, dependencies, output parity, operating limits and rollback; exceptions are explicit.
[PD] Expand permitted use	Operations sponsor validates live performance and decision costs, capacity and safety. The COO approves scale only within the supported population.

[AI] The case’s positive contribution is the directed research discipline: strong comparators, explicit constraints, negative results and a reversible handoff. Its next test must resolve the remaining evidence boundary rather than repackage the seed result as live operating value.

COO DECISION

[PD] Fund an authenticated final comparison and configuration-parity check. Retain the simple candidate as an evaluation option; make neural investment and scaled release conditional on fresh evidence and a detectable economic threshold.

[PD/AI] Proposed CPC; S1:D01–D08 and protected implementation register. The prospective evaluation, economic scenarios and information-value model are explicitly hypothetical, not realized project results.

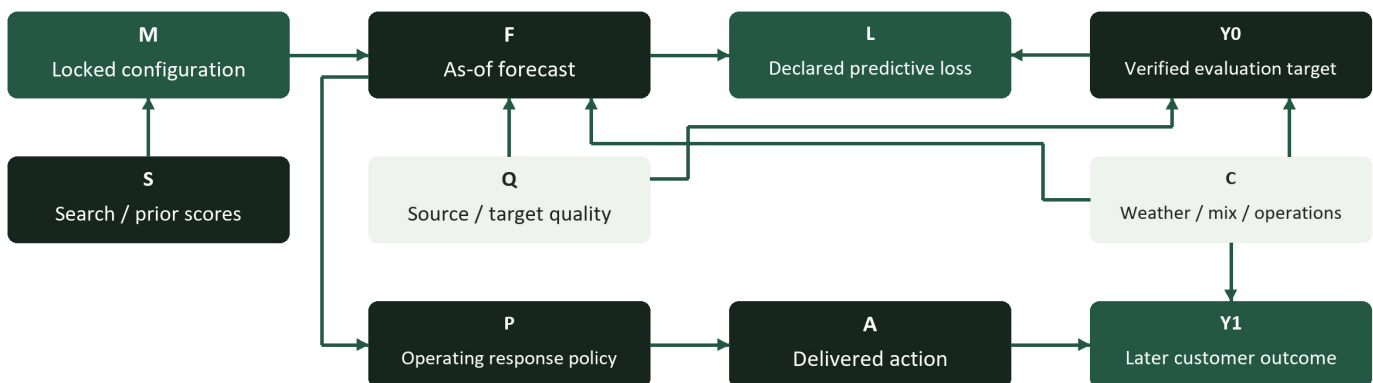
TECHNICAL FOUNDATION · THEORY OF CHANGE

A better score and a better operating decision require different evidence

[PD] The Champion Promotion Contract separates model selection, honest predictive evaluation, authorized operating use and customer effects. A software comparison identifies neither adoption nor prevention.

A PROMOTION SCORE IS NOT A CUSTOMER EFFECT

Proposed Champion Promotion Contract · computational and causal paths are distinct



$M \rightarrow F \rightarrow L$ and $Y0 \rightarrow L$ define a paired forecast-score comparison on a frozen target; Q controls its validity.

S selects M from earlier evidence. Reusing $Y0$ to select M would require a new untouched acceptance target.

$F \rightarrow P \rightarrow A \rightarrow Y1$ is a separate operating pathway. The shadow comparison does not estimate that causal effect.

C informs F and both outcomes; omitted common causes can explain association. Crossing lines are not junctions.

[PD] S is adaptive selection; M the locked configuration; F its as-of forecast; L the declared predictive loss against verified target $Y0$. Q is source and target quality; C is common weather, mix and operating conditions. P is the response policy, A the delivered action and $Y1$ the later customer outcome.

[AI] $M \rightarrow F \rightarrow L$ is a computational comparison conditional on Q and $Y0$. $S \rightarrow M$ can select favorable evaluation noise when the same outcomes inform tuning. A regenerated target can change L even when all keys match. These paths explain why a leaderboard is not an authenticated replay.

[PD] Customer benefit requires $F \rightarrow P \rightarrow A \rightarrow Y1$. C can affect both forecasts and outcomes; observed association does not identify the action effect. The proposed shadow test blocks candidate-driven operating use while its targets accrue.

[AI] Residual error may reflect omitted predictors, label error, regime change or irreducible variation. The residual alone does not identify the missing path or justify a more complex architecture.

Proposed CPC graph. Computational score paths are distinct from the downstream causal action path; no historical causal effect is asserted.

Define one comparison population and one loss before ranking candidates

[PD] The confirmatory predictive estimand is $E[L(\text{incumbent},Y)-L(\text{candidate},Y)]$ across all prespecified eligible origin bundles on a fresh prospective population, with both forecasts frozen at the same information cutoff.

Quantity	Interpretation contract
Paired risk difference	[PD] One locked candidate and strongest eligible incumbent; identical targets, horizon weights, exposure weights and ascertainment window. Positive means lower candidate loss.
Skill ratio	[AI] $MAE(\text{model})/MAE(\text{reference})$ is lower-is-better; 1—that ratio is higher-is-better. Label the convention and reference on every comparison.
Interval performance	[PD] Evaluate nominal coverage, width, tail calibration and proper quantile scores on the same population. Coverage alone does not reward informative intervals.
Customer policy effect	[PD] A separate assignment-based difference in a prespecified mature customer outcome under equal response capacity; not an algebraic transformation of forecast MAE.

[AI] A rising skill score can coexist with unchanged or rising absolute error when the reference forecast deteriorates more quickly at longer horizons. The source horizon pattern therefore supports relative resilience, not automatically better absolute long-range accuracy.

[PD] Freeze the target unit, event taxonomy, eligible denominator, issue time and label ascertainment cutoff. Shared target dates and overlapping horizons belong in one dependence structure. A large score-row count is not a large independent sample.

[PD] Test invariance across networks, node types and tail regimes using common definitions and explicit crosswalks. Missing or revised targets remain in the eligibility ledger with bounds; complete-case scoring requires a defensible missingness argument.

Registry benchmark conventions and horizon comparisons; proposed paired-risk contract. Gneiting and Raftery motivate score-specific evaluation.

Matching keys and aggregate error cannot authenticate target-level replay

[HR/AI] The registry reports complete key alignment and close headline-error agreement alongside material row-level target differences. These statements can coexist: matching summaries do not establish identical evaluation records.

Replay object	Required acceptance evidence
Population and target	[PD] Ordered row keys, values, units, taxonomy, maturity cutoff and exclusion flags; retain cryptographic hashes and an explicit changed-row report.
Feature availability	[PD] Timestamped as-of values, imputation states and publication delays. Training-time access to a later feature is not forecast-time availability.
Candidate configuration	[PD] Code, dependencies, parameters, feature gates, seed and runtime mode. A partial port receives a distinct configuration identifier.
Comparison result	[PD] Recompute paired row losses from both predictions and the same frozen truth. Reconcile weighted totals before presenting relative skill.

[AI] The regenerated panel may support a conditional comparison among candidates evaluated together. It does not reproduce the earlier stored result unless row-level truth and transformations also agree. Agreement in aggregate MAE is a weak checksum because offsetting differences can cancel.

[PD] Review a stratified label sample independently across ordinary/tail states, networks and horizons. Double-code event class and eligibility before adjudication; report confusion matrices, kappa and disagreement causes. Choose sample size for the desired reliability precision, rather than treating agreement as proof of validity.

[AI] Original source artifacts were not authenticated in this case pass. Registry descriptions establish a research record and its stated checks; they do not substitute for executable replay.

Private source-version and target-parity ledger. W3C provenance supports lineage; it does not certify the correctness of a stored target.

TECHNICAL FOUNDATION · SELECTION & VALIDATION

A disjoint test split is not necessarily an untouched final test

[HR/AI] One source field describes training-only gating while decision rules permit activation when test-window skill improves. The record does not resolve which results selected the final configuration.

Stage	Prespecified control
Development	[PD] Fit parameters and feature transforms on training data; record every candidate, gate, metric and manual intervention examined.
Selection	[PD] Use nested or rolling validation appropriate to the temporal dependence and deployment clock. Select the whole pipeline, not isolated coefficients.
Final evaluation	[PD] Lock the pipeline and comparator before new targets accrue. Reopening this holdout turns it into selection data and requires a new final evaluation.
Promotion	[PD] Compare the exact shipping configuration, including weather, leading-input and fallback gates; retain runtime and rollback acceptance separately.

[AI] Cawley–Talbot and Varma–Simon show that optimizing a noisy selection criterion can bias performance assessment. They do not quantify the optimism in this project. The synthetic search experiment illustrates the mechanism without subtracting an invented historical bias correction.

[AI] The selection chronology is unresolved, not proven leakage. A station-identifier ablation narrows one memorization hypothesis; it cannot exclude correlated features, temporal overlap, target revisions or all forms of site-specific learning.

[PD] Predeclare the primary loss and one confirmatory candidate comparison. Tail and group guardrails require their own multiplicity and power plan. Time-series cross-validation is not categorically invalid; its assumptions must fit the data-generating and deployment process.

Conflicting registry gate-selection statements; Cawley–Talbot, Varma–Simon and Bergmeir–Hyndman–Koo. No historical leakage finding is claimed.

A tail tradeoff needs a cost threshold and a capacity constraint

[AI] Better aggregate error and better tail detection are different objectives. A weather feature can reduce missed events while increasing false alerts; neither ranking wins without a declared decision loss.

[PD] For a common truth population, extra detections ΔTP justify extra false alerts ΔFP only when the value per additional detected event divided by the cost per additional false alert exceeds $\Delta FP / \Delta TP$, subject to available response capacity and harm constraints. Detection value still requires an effective response.

Question	Required evidence
Comparable tail counts?	[PD] Common target, threshold, population and mature confusion matrix for both configurations. Precision alone cannot supply workload without prevalence.
Rounded metrics stable?	[RC/AI] The broad-tail implied cost threshold is about 0.90, but a rounding-compatible count sensitivity spans about 0.72–1.13. The preference can reverse at unit cost parity. This range is not a confidence interval.
Nested weather cuts?	[AI] Severe-event subsets overlap. Do not sum their detections, alerts or estimated benefits as disjoint populations.
Intervals useful?	[PD] Combine proper quantile loss with width and conditional coverage. Wide intervals can meet coverage while offering little decision value.

[AI] Altering tier boundaries changes the classification task. An improved within-one-tier score across versions is not a clean forecast improvement unless both predictions are rescored against the same frozen taxonomy.

[PD] The synthetic priced-loss contract later in this case deliberately values tail error differently. Changing that weight changes the estimand; it is a different decision preference, not a robustness check on one immutable objective.

Private rounded-metric sensitivity and version crosswalk; proposed operating-loss decision rule. No raw internal precision or recall values are disclosed.

Candidate profiles and changing panels are not a learning curve

[AI] The evidence compares research configurations, horizons and panels with changing composition. Those contrasts do not form an invariant operating time series or an isolated experiment on training-data volume.

Source contrast	Permitted conclusion
Longer research panel	[AI] Training span, test span and calendar realization changed together. A smaller model gap cannot be attributed to added history alone.
Candidate and horizon profiles	[AI] Ordering architectures or horizons does not create successive process observations. Mechanical XmR calculations are not evidence of operating stability or change.
Neural-model gate	[AI] No neural candidate was actually trained in the documented comparison. Deferral is a conditional resource decision, not demonstrated neural inferiority.
Leading-input ablation	[AI] A seed process without the relevant lead cannot establish that live operating inputs lack predictive or causal value.

[PD] For a genuine future fixed-phase sequence, individual limits are $\text{mean} \pm 2.66 \times \text{mean moving range}$; moving-range upper limit is $3.268 \times \text{mean moving range}$. Short-series limits are tentative. No supplied sequence supports a natural-process classification here; no stable binary count/exposure panel supports a Laney rate chart.

[PD] Freeze test origins and labels while varying only training-history length to estimate a learning curve. Repeat across independent regimes with uncertainty. A source heuristic about history or site count is not a statistical sufficiency theorem.

[AI] No fixed worst-performing operating cohort with follow-up is supplied. Model-winner selection instead creates selection optimism. Prospectively assess horizon, node type, network and tail regime as moderators; avoid post-hoc subgroup promotion.

Source panel/configuration changes and private ordered-screen inventory. Mechanical profiles remain separate from observed operating processes.

TECHNICAL FOUNDATION · CONTRIBUTION & PROMOTION

Credit the research direction without claiming unverified implementation outcomes

[HR/AI] The registered contribution is commissioning, directing and adjudicating a constrained model-research program. The record attributes implementation work to software agents; it does not establish personal authorship of every model.

Contribution claim	Process-tracing grade
Scope and constraints	[HR/AI] Provisional hoop: registry records research questions, isolation requirements and a standard-library shipping constraint. Original instructions and timestamps remain to be authenticated.
Architecture and data gate	[HR/AI] Provisional hoop: the record requests stronger baselines and a conditional deep-model decision. A gate is not evidence that all alternatives were evaluated.
Champion promotion	[HR/AI] Provisional hoop for the described partial port. The shipping feature configuration differs from the research champion and requires its own validation.
Customer or financial impact	[AI] Unsupported by the recovered evidence. No smoking-gun or doubly decisive test links these research decisions to mature field outcomes.

[AI] A later stress result cannot be treated as the cause of an earlier decision memorandum without a versioned chronology. The case preserves this uncertainty rather than rearranging the narrative into a stronger decision story.

[PD] A smoking-gun contribution test would authenticate a dated instruction, attributable code/configuration response and independent review confirming the change. A doubly decisive test would additionally eliminate a credible alternative origin; neither is available here.

[AI] Upstream systems, separate model workstreams and external program results retain their own ownership. A proposed data-access or integration path is not an approved grant, completed extraction or adopted production service.

Private role, chronology and artifact registry. Collier’s process-tracing vocabulary is used as an evidence grading method, not a substitute for originals.

TECHNICAL FOUNDATION • PROSPECTIVE ACCEPTANCE

Use paired live forecasts to test predictive value before testing actions

[PD] Run both locked configurations silently against the same future targets. This creates a paired predictive comparison without assigning operational treatment or confusing prevention with forecast error.

[PD] Define independent validation clusters by shared sites, outcome cohorts and environmental shocks; overlapping horizons stay inside each cluster. Freeze equal-weight origin bundles upstream of forecast availability. Failed runs and unavailable labels remain in the cohort with loss or missingness bounds, not favorable deletion.

[PD] Estimate the mean incumbent-minus-candidate loss using one cluster mean per independent unit. The synthetic model specifies an entire bounded mixture, evaluates null and economic-boundary behavior, and compares detectable gain with the same service-pool cost. This planning distribution is not a measured field variance.

[PD] Mandatory acceptance also covers target/source authenticity, tail harm, conditional calibration, interval width, runtime reliability and configuration parity. Primary score power does not establish power for these safeguards. The number of feasible independent clusters is unknown; correlated extra rows cannot replace them.

[PD] Only after predictive acceptance should a separately powered concurrent cluster trial compare forecast-led and usual action policies under equal response capacity, with upstream eligibility and fixed mature customer outcomes. Measure delivered action, reach and contamination while preserving assignment-based analysis.

[PD] A phased rollout needs randomized timing or explicit group-time identification assumptions, calendar effects, pretrend diagnostics and interference controls. Non-significant pretrends do not prove exchangeability. The paired-score power calculation does not size that causal trial.

Proposed prospective paired-risk evaluation. Diebold–Mariano informs loss-differential dependence; a customer-policy trial remains a separate evidence stage.

TECHNICAL FOUNDATION • INTELLECTUAL CONTRIBUTION

The Champion Promotion Contract makes five claims falsifiable

[PD] CPC requires a target-level replay contract, an honest comparator, an untouched acceptance set, a declared decision loss and validation of the exact shipping configuration.

Proposition	Falsifiable test
P1 • Replay identity	[PD] Aggregate agreement is insufficient for target-level parity. Attempt to reconstruct every paired loss from frozen targets and hashes; row equality would remove the alleged fidelity gap.
P2 • Selection separation	[PD] Reusing evaluation results for gates inflates selected estimates on average under noisy search. Compare selected gains with a fresh frozen holdout across repeated independent tasks.
P3 • Objective dependence	[PD] The preferred weather configuration changes with relative miss/false-alert costs near the documented sensitivity frontier. Common-count evaluation could show no reversal over the relevant price range.
P4 • Promotion specificity	[PD] A partial port does not inherit the full champion’s acceptance evidence. Exact shipping replay and fresh paired testing can establish or reject equivalence.
P5 • Complexity option	[PD] Deferring a costly architecture is rational only while expected priced gain fails its incremental-cost gate. New independent data or a stronger result can reverse the decision.

[AI] Cawley–Talbot and Varma–Simon motivate P2; proper-scoring research supports P3; benchmark evidence such as TiDE motivates fair architecture comparisons without supplying this project’s data threshold. The case extends these ideas into a promotion contract spanning target lineage and operational configuration.

[AI] CPC applies the Shewhart–Deming warning against reacting to unstudied variation and Campbell’s concern about administrative indicators. Tail weights expose a composite indicator’s value judgments. Ittner–Larcker–Randall found little support for simple measurement-alignment benefits; CPC therefore requires a separate value test. Retaining failed experiments extends Tucker–Edmondson’s system-learning argument to model handoffs. These are analytical connections, not empirical confirmation of those theories.

[PD] A lower confidence bound above cost supports promotion under the assumed contract. An upper bound below cost contradicts the economic claim; an interval spanning cost is inconclusive. Scientific credibility does not require forcing every experiment into a winner.

Proposed framework and testable propositions. The case contributes an applied decision contract, not a claimed new statistical theorem.

TECHNICAL FOUNDATION · LIMITATIONS & REPRODUCTION

Keep four unresolved evidence boundaries visible at the decision gate

[AI] The case supports a defensible research and promotion process. It does not authenticate the original code, prove an untouched historical test, establish field transport or monetize customer benefits.

Boundary	Required next evidence
Source authenticity	[PD] Recover immutable inputs, code, reports and timestamps; reconcile target-level differences and the training-versus-test gate-selection chronology. Registry pointers are not reopened originals.
Statistical transport	[PD] Validate fresh live labels, cluster dependence, missingness and tail prevalence. Seed-relative rankings cannot be assumed to transport either absolutely or relatively.
Economic interpretation	[PD] Obtain an accepted price for the declared forecast-service loss and complete same-horizon incremental costs. Without that contract, the synthetic net value is not cash or customer savings.
Operating effect	[PD] Authenticate shipping configuration, data access and adoption; then test the action policy on mature outcomes. Model accuracy alone leaves the prevention path unidentified.

[PD] The private annex retains source locators, alternative-version calculations, ordered mechanical screens, original-scale inputs and excluded claims. The public model uses only hypothetical inputs and one fixed normalization; it can be rerun independently without confidential data.

[PD] Reproduction order is evidence selection, arithmetic reconciliation, configuration comparison, synthetic model, independent numerical review, document rendering and visual inspection. Hash the final inputs and outputs. Report changes in conclusions under each defensible source version rather than silently selecting a favorable one.

[AI] Passing these checks makes the case reviewable. It does not remove uncertainty about unavailable originals, future data or the commercial price of a forecast error.

Consolidated limitations and reproducibility contract. Historical claims, recalculations, analytical inferences and proposed designs remain explicitly distinguished.

PROPOSED DECISION MODEL · PAIRED VALIDATION

Freeze the candidate before collecting its acceptance evidence

[PD] This synthetic shadow test compares two locked forecasts on identical future targets. The endpoint is priced forecast-service loss under a hypothetical contract, not customer prevention or realized savings.

Score assumption	Ordinary bundle	Tail bundle
Incumbent absolute error	80	180
Prespecified loss weight	1	3
Mean candidate advantage, weighted	15	-30

[PD] Each independent validation cluster contains 12 equal-weight origin bundles; dependent horizons are scored inside a bundle. Tail probability is 0.10 or 0.30 with equal probability across clusters. Add independent paired score noise of ± 8 with equal probability. These are illustrative inputs.

[PD] A cluster’s mean advantage is $15-45T/12+8(2R/12-1)$, where T is its tail-bundle count and R its positive-noise count. Conditional binomial counts yield an exact finite distribution; candidate losses remain nonnegative.

[PD] Mean incumbent weighted loss is 172, indexed 100. Mean advantage 6 becomes 3.4884 points. A stronger candidate on ordinary days can still be worse in the tail; the declared objective and guardrails determine whether that tradeoff is acceptable.

[PD] Lock target, as-of inputs, scoring weights and runtime configuration. A matched key is insufficient if its target value changes. The final holdout cannot be reused for tuning, gate selection or a different promoted configuration.

Wholly synthetic future validation design. Its cluster count, loss values and prices do not come from the project’s seed benchmark.

PROPOSED DECISION MODEL · POWER & SEARCH

A fresh holdout must clear cost after selection is finished

[PD] Invert the specified finite distribution of paired cluster means for a conservative equal-tail 95% interval. Exact-process convolution avoids the skewed lower-tail error of a normal approximation.

Independent clusters	MDE, 80%	P(lower>2)	Joint gate
30	2.083	52.7%	Fail
60	1.481	80.4%	Pass
76	1.318	88.3%	Pass

[PD] Require $MDE \leq 2$ and at least 80% chance the lower bound clears cost 2 under benefit 3.4884. Conditional minimum 60; buffered candidate 76. The known synthetic law must be replaced and recalibrated using independent pilot evidence before field use.

[PD] Effect shifts preserve variance by construction. In 16,000 process trials: null rejection 4.85%, boundary false funding 2.45%, full-effect funding 80.53%. Available independent clusters are unknown.

Candidates searched	Selected mean gain	Naive false positive
1	0.003	2.5%
8	1.509	12.5%
24	2.067	23.0%

[PD] Separate search illustration: all true gains 0, estimate SD 1.5, correlation 0.5; maximize the noisy estimate. The last column uses an unadjusted one-sided 2.5% test. A new independent holdout returns mean gain near 0. This is not a correction estimated for the historical search.

Synthetic clustered score planning and a separate Gaussian selection experiment. Tail, calibration and runtime guardrails require their own acceptance tests.

PROPOSED DECISION MODEL · ECONOMIC CONTRACT

Choose the error price before choosing the model

[PD] Under an illustrative forecast-service contract, each unique weighted-loss unit has a declared price. Baseline contract loss is 100 and incremental model cost is 2 over the same service pool: $\text{net} = \text{paired reduction} - 2$.

Tail frequency / weight	1	3	5
10%	5.267	4.105	2.942
20%	3.814	1.488	-0.837
30%	2.360	-1.128	-4.616

[PD] Grid entries are net contract units. Tail-weight changes are alternative loss estimands, not robustness of one unchanged target. All cells retain the original reference 172 and cost 2; do not silently rebase the denominator.

[PD] Base priced reduction=3.4884; net=1.4884. With ordinary advantage 12 instead of 15 against a stronger incumbent, net=0.0930. Comparator strength can defeat the case without changing the candidate.

[PD] Avoidability and exclusive credit are each 1 because the hypothetical penalty is already defined as unique, fully avoidable service loss. No second discount is applied. Cost 2 is allocated per eligible bundle at fixed service volume: refresh/training, runtime, monitoring and rollback. Study sample size does not redefine that volume; research-only cost is separate.

[PD] These price weights are assumptions, not a finance-approved valuation. Without an accepted mapping from forecast loss to a priced obligation or decision consequence, customer savings and cash remain unidentified. A downstream action trial would need a different estimand and cost model.

[PD] The buffered 76-cluster design fails at higher tail-prevalence heterogeneity. The private grid varies bundle count and shared regime risk; extra horizons do not create independent clusters.

Hypothetical contract sensitivity. No historical benefit, avoided neural-model expenditure or customer effect is attributed to this project.

PROPOSED DECISION MODEL · INFORMATION VALUE

Pay to resolve the evidence state, not to confirm a favored architecture

[PD] Hypothetical states are a repeatable candidate gain, search optimism with weak gain, or incompatible target/provenance. Priors are 0.40/0.35/0.25. Funding means a gated validation path, not deployment permission.

Action / net utility	Real gain	Weak gain	Invalid basis
Fund gated candidate validation	16	-10	-18
Keep simpler baseline / targeted audit	5	8	-4
Defer expansion	0	0	0

[PD] The diagnostic reports the correct state with the stated probability; errors split equally across other labels. Update by Bayes’ rule and choose the highest expected utility. Action payoffs include commitment costs; evidence-acquisition cost is deducted separately.

Correct-state probability	EVSI	Study cost	Net EVSI
45%	0.000	0.500	-0.500
70%	1.610	1.500	0.110
90%	4.130	3.000	1.130

[PD] $EVSI = E[\max_a E(U_a | \text{signal})] - \max_a E(U_a)$. Perfect information is worth 5.400; each scenario respects the ceiling. Low diagnostic quality can have negative net value even when information cannot reduce optimized gross utility.

[PD] This restricted model does not establish the probability of neural superiority or actual avoided spend. Expand it for partial gains, platform constraints and harmful tail behavior. Information value cannot be added to benefits already pricing the same consequence. Mandatory source, target and safety gates still apply.

Wholly synthetic evidence-purchase decision, distinct from forecast-service acceptance and from customer-impact evaluation.

Read the case as a model-investment and promotion decision

[AI] The project’s value lies in directing a comparative research program and controlling what its evidence authorizes. A selected seed-data champion, a reported code port and a validated live operating model remain different achievements.

Claim class	Meaning in this case
[HR] Historical record	A protected register account of instructions, model runs, configurations or artifacts. Original research files and runtime logs remain unrecovered; copied summaries are not independent authentication.
[RC] Recalculated	Arithmetic from registered values or a separately labeled hypothetical model. Indexed error reductions are benchmark comparisons, not causal customer effects or cash savings.
[AI] Analytical inference	A bounded interpretation of replay identity, comparator strength, selection adaptation, tail costs, configuration differences and investment gates.
[PD] Proposed design	The Champion Promotion Contract, fresh paired live validation, hypothetical forecast-service economics and future decision rules. None is a historical implementation claim.

[RC] The same-TEST strong comparator reduces the champion’s recorded advantage to 8.02%. The tail exchange also crosses a one-to-one cost threshold under source-rounding sensitivity. [AI] Both findings change the decision without inventing new observations.

[AI] Research direction, constraints, acceptance standards and promotion decisions receive provisional registry-level hoop support. Agent-assisted execution is identified. Parent-engine architecture, application work, action libraries and other teams’ science retain separate attribution.

[PD] Reproduce register arithmetic from the frozen extract and cell-linked ledger. Recover original targets, issue-time inputs, code, configurations, selection chronology and prediction rows before authenticating historical replay or a final untouched test. Re-run the proposed model separately with its declared synthetic parameters.

DISTRIBUTION

[PD] Share the anonymized case. Keep raw metrics, identity keys, internal paths, source versions and calculation inputs in the separate private annex.

S1: protected project register and private calculation ledger. Public historical quantities are relative, indexed or qualitative. The proposed scoring experiment has its own population, loss and hypothetical parameter contract.

REFERENCES / PRIMARY METHODS AND SUBSTANTIVE LITERATURE

Protect model selection from the score used to certify it

These sources inform selection validity, forecast loss and configuration governance. They do not authenticate the original artifacts or establish live superiority, operational adoption or customer benefit.

Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *JMLR*, 11, 2079–2107. Full primary paper. Optimizing a noisy selection criterion can bias the reported winner. Freeze the complete selection procedure before final evaluation.

Varma, S., & Simon, R. (2006). Bias in Error Estimation When Using Cross-Validation for Model Selection. *BMC Bioinformatics*, 7, 91. Author full text. Selection belongs inside the evaluation design. A separate training split alone does not protect an acceptance set repeatedly used to choose features or models.

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction. *Computational Statistics & Data Analysis*, 120, 70–83. Full author manuscript. Validity depends on the error and model conditions; time-series data do not justify either blanket permission or blanket prohibition.

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13, 253–263. Primary article and scanned full text. Compare paired loss differentials with their dependence structure. Point estimates alone do not establish a reliably superior forecast or a universally superior model class.

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *JASA*, 102, 359–378. Author full text. Probabilistic forecasts need a declared scoring objective. The appropriate score does not repair changed targets, adaptive evaluation or missing live outcome truth.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic Forecasts, Calibration and Sharpness. *JRSS B*, 69, 243–268. Author full text. Report interval width and conditional behavior alongside coverage. Near-nominal pooled coverage can coexist with unhelpful or uneven predictive uncertainty.

Hyndman, R. J., & Koehler, A. B. (2006). Another Look at Measures of Forecast Accuracy. *International Journal of Forecasting*, 22, 679–688. Author manuscript. Forecast measures depend on scaling and baseline construction. The case retains both absolute-error movement and the comparator behind relative skill; it does not relabel its score as MASE.

Elkan, C. (2001). The Foundations of Cost-Sensitive Learning. *IJCAI*, 973–978. Author full text. Misclassification penalties require a coherent reference and decision. Tail recall must be evaluated with false-alert costs; the resulting classification-loss trade is distinct from the effect of an operating intervention.

Primary research, author manuscripts and official guidance. Access scope is stated. No external model performance or customer effect is imported into this project.

REFERENCES / PRIMARY METHODS AND SUBSTANTIVE LITERATURE

Choose architecture and release controls for the actual task

These sources inform selection validity, forecast loss and configuration governance. They do not authenticate the original artifacts or establish live superiority, operational adoption or customer benefit.

Das, A., Kong, W., Leach, A., Mathur, S., Sen, R., & Yu, R. (2023). Long-term Forecasting with TiDE: Time-series Dense Encoder. Full author paper, arXiv:2304.08424. A specific forecasting architecture is a testable candidate. Its published experiments supply neither a universal minimum-history rule nor a result for this project.

Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. *International Journal of Forecasting*, 37, 1748–1764. Full author paper. Known-future and historical inputs have different roles. Architecture selection still requires task-matched comparison and issue-time availability.

Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. *International Journal of Forecasting*, 36, 1181–1191. Author preprint. Global probabilistic modeling motivates a candidate specification; this case does not claim it was trained or empirically rejected.

Sculley, D., Holt, G., Golovin, D., et al. (2015). Hidden Technical Debt in Machine Learning Systems. *NeurIPS*, 28. Full proceedings paper. Data dependencies, configuration and external consumers create system obligations beyond model accuracy. The promoted artifact must preserve the precise tested behavior and declare its exceptions.

Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). Model Cards for Model Reporting. *FAT* 2019*. Full author paper. State intended uses, evaluation populations and relevant performance slices. A source card describes evidence and limits; it does not certify code parity, live deployment or causal impact.

W3C (2013). PROV-DM: The PROV Data Model. Official recommendation. Entity versions, activities and derivations support an inspectable configuration handoff. Provenance distinguishes the research candidate from the engine artifact; it cannot make two different targets identical or prove that an attributed activity occurred.

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *JMIS*, 12(4), 5–33. Full original article. Data must be fit for the task, represented consistently and accessible. A synthetic benchmark can serve development while remaining insufficient for a live promotion decision.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, 25, 289–310. Author preprint. Predictive and explanatory objectives are distinct. A feature ablation measures a specified forecasting comparison; it neither identifies the physical mechanism nor establishes the value of an action suggested by that feature.

OECD/European Union/EC-JRC (2008). Handbook on Constructing Composite Indicators: Methodology and User Guide. Official full text. Weighting and aggregation encode choices; sensitivity analysis exposes their consequences. Applying this principle to a composite forecast-loss score is an analytical extension, not a direct finding about this project.

The W. Edwards Deming Institute. The Funnel Experiment. Official explanation of tampering: adjusting a process without understanding its variation can worsen results. The case applies this warning to evidence-responsive model governance; candidate leaderboard rows are not a control-chart process.

Primary research, author manuscripts and official guidance. Access scope is stated. No external model performance or customer effect is imported into this project.

REFERENCES / PRIMARY METHODS AND SUBSTANTIVE LITERATURE

Keep the evidence, attribution and economic decisions separate

These sources inform selection validity, forecast loss and configuration governance. They do not authenticate the original artifacts or establish live superiority, operational adoption or customer benefit.

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the Mean: What It Is and How to Deal with It. *IJE*, 34, 215–220. Publisher abstract. Extreme selected observations may moderate without treatment. Worst-bucket narratives require comparable definitions and selection-aware follow-up, not a favorable new snapshot.

NIST/SEMATECH. e-Handbook of Statistical Methods: Individuals Control Charts. Official guidance. Moving-range limits require comparable ordered observations. Candidate leaderboard rows, horizons and regenerated panels do not become an operating process series simply because a chart can connect them.

Campbell, D. T. (1976). Assessing the Impact of Planned Social Change. Occasional Paper Series, 8. Primary ERIC record. Separate the intervention, its evaluation and the administrative indicator. Changing severity cut-points can change tier agreement without improving the underlying point forecast or a customer outcome.

Ittner, C. D., Larcker, D. F., & Randall, T. (2003). Performance Implications of Strategic Performance Measurement in Financial Services Firms. *Accounting, Organizations and Society*, 28, 715–741. Author institutional abstract. The study finds little support for simple alignment benefits and distinguishes measurement-system satisfaction from economic performance. CPC likewise requires a separate value test.

Tucker, A. L., & Edmondson, A. C. (2003). Why Hospitals Don't Learn from Failures. *California Management Review*, 45(2), 55–72. Full original article. System learning requires more than local workarounds. Retained negative experiments and explicit reopen conditions support learning; they do not prove subsequent operational follow-through.

Collier, D. (2011). Understanding Process Tracing. *PS: Political Science & Politics*, 44, 823–830. Primary article. Grade evidence against the particular contribution and its alternatives. A registered commissioning directive supports a different claim from original model code, deployment logs or an identified outcome effect.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using Simulation Studies to Evaluate Statistical Methods. *Statistics in Medicine*, 38, 2074–2102. Author full text. Declare the generating process, estimand, method and Monte Carlo uncertainty. Check null behavior and power at the decision boundary, not only under a favorable effect.

Strong, M., Oakley, J. E., Brennan, A., & Breeze, P. (2015). Estimating the Expected Value of Sample Information Using the Probabilistic Sensitivity Analysis Sample. *Medical Decision Making*. Primary full text. Information value depends on explicit decisions and future evidence. This case uses a finite-state example, not the authors' regression estimator or measured historical savings.

Primary research, author manuscripts and official guidance. Access scope is stated. No external model performance or customer effect is imported into this project.

FINAL REFEREE REPORT / INTERNAL, AGENT-ASSISTED REVIEW

A stronger model decision requires an explicit promotion boundary

[AI] Parallel agents developed evidence, methods and editorial reviews; the lead reviewer completed numerical and publication checks. This is internal, agent-assisted review, not external peer review, client acceptance or certification.

Major issue	Resolution or remaining gate
Benchmark identity	Key and aggregate agreement do not certify identical targets. Within-panel comparisons remain distinct from replay; original rows and code must be recovered.
Selection and comparator	The stronger same-TEST baseline reduces the recorded advantage. Disjoint TRAIN/TEST does not prove a final untouched set; selection chronology remains a recovery requirement.
Tail loss and complexity	Rounding-compatible tail exchanges cross a one-to-one loss boundary. Two changed seed experiments do not identify a learning curve or prove neural inferiority.
Promotion and contribution	The reported smoothing port is separated from the full research champion. Directed standards and decisions receive provisional registry-level hoop support; agent-assisted implementation is explicit.
Prospective evaluation	Finite-law calibration replaces an approximate interval. Separate numerical implementation verifies the synthetic power and decision model; the unknown field law still requires pilot estimation and recalibration.

Partner critique: would the COO act on page 1?

[AI] The decision is concrete: fund an authenticated final comparison and exact-configuration parity check, assign analytics and platform owners, and make neural investment and scaled promotion conditional on fresh evidence. The simple candidate remains available without borrowing a field result.

Positive contribution within the evidence boundary

[AI] The registered work sets strong comparators, practical runtime constraints, falsification gates, negative-result retention and a reversible handoff. Its professional contribution is directing model choice and release discipline rather than claiming sole manual coding or invented customer savings.

REVIEW DISPOSITION — PASS WITHIN STATED SCOPE

[AI] Source arithmetic, separate model implementation and rendered-page checks support publication as a registry-based development case. Original artifacts, historical live superiority and customer impact remain unestablished.

Internal agent-assisted review; scoped numerical and actual visual checks bind final artifacts in the private audit. Source recovery and outcome identification remain open. No external certification is implied.